INTERPRETABILITY DRIVEN EVOLUTIONARY APPROACH FOR THE DESIGN OF BIOLOGICAL SEQUENCES

Akash Pandey* **& Wei Chen**
Department of Mechanical Engineering
Northwestern University
Evanston, IL 60208, USA
{akashpandey2026@u., 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

Designing biological sequences such as proteins and DNA for desired properties is challenging due to vast search spaces and limited wet lab evaluation budgets. Current evolutionary approaches ignore sequential dependencies and rely on random mutations, which scale poorly for long sequences. In contrast, reinforcement learning (RL) and generative models that explicitly model sequence structure, require large datasets to guide generation toward the target properties. These limitations suggest the need for a method that combines the sample efficiency of evolutionary approaches with the ability to exploit sequential structure. In this work, we propose a novel evolutionary approach, IDEAS, in which mutations are guided by an explainable model. The model identifies critical motifs in high-fitness sequences and uses them to mutate non-critical positions. Across six continuous-property datasets, seven baselines, and three evaluation budgets, IDEAS achieves a **19%** acceleration in design while maintaining a favorable position on the Pareto curve balancing acceleration, diversity, and novelty.

1 INTRODUCTION

Biological sequences such as proteins and DNA play a central role in therapeutics and biotechnology, making the ability to design sequences with desired properties highly valuable (Zimmer, 2002; Lorenz et al., 2011; Barrera et al., 2016a; Ogden et al., 2019). However, the combinatorially large search space of biological sequences renders this design problem inherently challenging. Prior work has approached this problem using evolutionary methods (Sinai et al., 2020; Hansen, 2006; Arnold, 1998), reinforcement learning (RL) (Angermueller et al., 2019; Jain et al., 2022), and generative model–based (Brookes et al., 2019; Brookes & Listgarten, 2018; Gupta & Zou, 2019a) approaches.

Evolutionary approaches (Sinai et al., 2020; Hansen, 2006; Bloom & Arnold, 2009) iteratively improve candidate biological sequences through mutation and oracle-based selection, either experimentally or via simulations. Their main advantage is requiring no training data or oracle constraints, but random mutations ignore sequential dependencies, leading to poor scalability and high oracle demands in large search spaces.

Reinforcement learning (RL)–based methods, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Generative Flow Networks (GFlowNets) (Bengio et al., 2021), have been applied to biological sequence design. On-policy methods like DynaPPO (Angermueller et al., 2019)

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

train only on samples from the current policy, limiting exploration, whereas off-policy approaches like GFN-AL (Jain et al., 2022) leverage past sequences to generate high-performing, diverse, and novel designs. This increased novelty, however, can produce out-of-distribution sequences, leading to unreliable reward estimates and potentially degrading policy learning (Trabucco et al., 2021; Yu et al., 2021).

Generative model-based design methods, including generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma & Welling, 2019), iteratively shift the generative distribution toward desirable sequences. Design by Adaptive Sampling (Dbas) (Brookes & Listgarten, 2018) weights samples according to oracle-evaluated properties, while Conditioning by Adaptive Sampling (Cbas) (Brookes et al., 2019) introduces a penalty for divergence from the initial distribution. However, these approaches often require thousands of oracle evaluations, limiting their applicability under tight evaluation budgets (Brookes & Listgarten, 2018; Brookes et al., 2019; Gupta & Zou, 2019a).

Existing biological sequence design methods either fail to scale to long sequences due to random mutations that ignore sequential dependencies, or require a large number of oracle evaluations (often exceeding 10^2) to learn distributions over high-performing sequences. In practical experimental and computational settings, only a limited number of oracle evaluations can be performed per design cycle, making such requirements prohibitive (Graham et al., 2025; Tokareva et al., 2014). Consequently, there remains a critical gap for methods that combine the sample efficiency of evolutionary approaches with the ability of RL- and generative models to capture sequential dependencies. Bridging this gap is essential for developing design frameworks that are both sample-efficient and effective under tightly constrained oracle budgets.

Our Contribution In this work, we introduce `IDEAS` (**I**nterpretable **E**volutionary **A**pproach for biological **S**equences), which combines explainable models (XAI) that capture sequential dependencies with evolutionary design strategies. By leveraging attribution scores from XAI, `IDEAS` identifies the critical contiguous regions of top-performing sequences that drive the target property, as well as non-salient regions. This information guides *interpretable mutations*, where non-salient regions are selectively replaced or modified using motifs from critical regions, enabling more informative exploration of the sequence space compared to random mutations. Consequently, `IDEAS` accelerates sequence optimization while maintaining biological plausibility and interpretability.

The key contributions of this work are threefold: (1) we propose `IDEAS`, an interpretable evolutionary framework that leverages explainable models (XAI) to guide mutations and accelerate exploration of the biological sequence design space; (2) we conduct a comprehensive empirical evaluation of `IDEAS` against seven state-of-the-art baselines across six biological design tasks and multiple oracle evaluation budgets, using design acceleration, diversity, and novelty as evaluation metrics; and (3) through extensive ablation studies, we analyze the impact of different mutation strategies within `IDEAS` and demonstrate that a wide range of faithful XAI methods can be seamlessly integrated as plug-and-play components.

2 Problem Formulation

In this work, given an initial dataset $\mathcal{D}_0 = \{(\mathbf{x}_i, y_i)\}_{i=0}^{N_0-1}$, where y_i denotes the continuous property value of sequence $\mathbf{x}_i$, we aim to design biological sequences $\mathbf{x} \in \mathcal{V}^L$ that exhibit desired properties. Here, $\mathcal{V}$ denotes the vocabulary (e.g., amino acids or nucleotides), and L denotes the sequence length. Designed sequences are evaluated using a black-box oracle model $f : \mathcal{V}^L \to \mathbb{R}$. These oracle models are accurate but expensive, such as wet-lab experiments or high-fidelity computational simulations.

Given the vast search space of biological sequences ($|\mathcal{V}|^L$), exhaustive oracle evaluation is infeasible. However, recent advances in experimental techniques and their reduced turnaround times have made active learning a practical approach for iterative design with feedback at each round. In this work, we perform ten rounds of iterative design, querying B new sequences per iteration to discover sequences with high $f(\mathbf{x})$, diversity, and novelty. In practice, the number of allowable oracle evaluations is often limited. To reflect this constraint, we compare different design methods under oracle budgets of $B \in \{20, 50, 100\}$.

3 IDEAS: INTERPRETABILITY DRIVEN EVOLUTIONARY APPROACH FOR BIOLOGICAL SEQUENCES

To address the poor scalability of traditional evolutionary methods to long sequences and the large data requirements of generative model–based approaches, we propose a novel active learning framework for biological sequence design. Our method leverages an explainable model (XAI) to quantify position-level attribution scores, which are then used to guide interpretable and informed mutations, enabling sample-efficient exploration of the design space under constrained oracle budgets.

Definition 1 (Position-wise attribution scores)**.** *Given a biological sequence $x_i \in \mathcal{V}^L$ with property $y_i = f(x_i)$, a position-wise attribution score $\phi_i \in \mathbb{R}^L$ is defined as qualitative contribution of each position $j \in [0, L-1]$ in x_i towards y_i.*

Our explainability-driven design method, IDEAS, consists of four distinct steps. Since the design process proceeds iteratively, we describe the operations performed at the t-th iteration in detail below.

Step 1 (Selecting top sequences): The top B sequences from $\mathcal{D}_{t-1}$ (dataset at $t-1$ iteration) is selected as: $\mathcal{D}^{\text{top}} = \{(\mathbf{x}_i, y_i \in \mathcal{D}_{t-1} : y_i \text{ ranks in top } B\}$.

Step 2 (Training XAI model): Using $\mathcal{D}_{t-1}$, an XAI model (g) is trained such that

$$\hat{y}_i = g(x_i; \theta) \tag{1}$$

Once the θ is optimized, XAI can quantify the attribution score of each position in the sequence as

$$\phi_i = g_{att}(x_i, \theta^*, \hat{y}_i), \text{ where, } \phi_i \in \mathbb{R}^L \tag{2}$$

Step 3 (Extracting Motifs): In this study, we define motif s_j as the contiguous sub-segment of length m within a biological sequence such that $s_j \in \mathcal{V}^m$. The attribution score, I_j, of s_j in x_i is calculated using ϕ_i in Equation. 2 as:

$$I_j = \frac{1}{m} \sum_{k=j}^{j+m-1} \phi_i[k], \; j \in [0, L-m] \tag{3}$$

The motif size m in Equation 3 is a tunable hyperparameter that can be selected based on domain expertise. In this work, we set $m = 1$ for all datasets for the reasons discussed in Section 5.3.3. As shown in Appendix 5.3.3, increasing m can improve design performance, but the gains plateau rapidly. All s_j's and their corresponding I_j's extracted from x_i are stored in a list as $\mathcal{M}_i = \{(s_j, I_j) : j \in [0, L-m]\}$. Similarly, motifs and their corresponding attribution scores are extracted from all top B sequences in $\mathcal{D}^{\text{top}}$ and aggregated as $\mathbf{M} = [\mathcal{M}_0 | \mathcal{M}_1 | \ldots | \mathcal{M}_{B-1}]$, $|\mathbf{M}| = n_m$. It is important to note that the repeating motifs s_j in $\mathbf{M}$ are collapsed into one entry by averaging their corresponding I_j values.

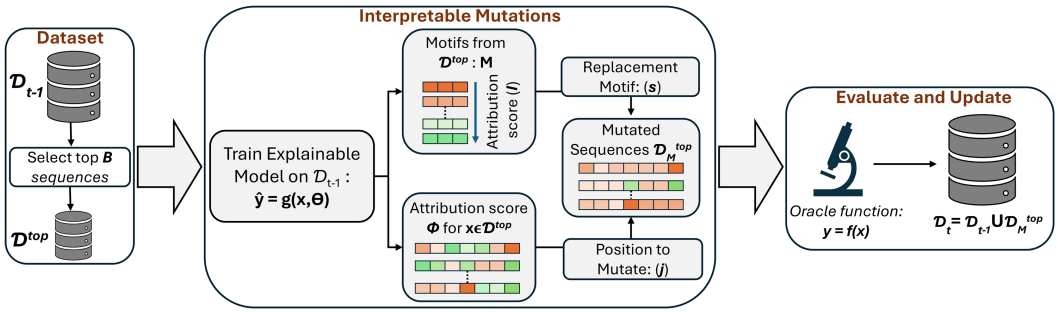

Figure 1: The active learning process of IDEAS. At iteration t, the top-B sequences from $\mathcal{D}_{t-1}$ form $\mathcal{D}^{\text{top}}$. An explainable model $g(\theta)$ trained on $\mathcal{D}_{t-1}$ extracts motifs $\mathbf{M}$ and position-wise attribution scores ϕ_i for sequences in $\mathcal{D}^{\text{top}}$. Contiguous positions with low ϕ_i (green) are probabilistically selected for mutation, while high-attribution motifs I (red) are selected as replacements. The resulting mutated sequences $\mathcal{D}_M^{\text{top}}$ are evaluated by the oracle f and added to $\mathcal{D}_{t-1}$ to form $\mathcal{D}_t$.

Step 4 (Interpretable Mutations): In this step, each sequence in $\mathcal{D}^{top}$ is mutated with motifs from $\mathbf{M}$ to maximize y. For mutations, firstly, the attribution scores ϕ_i at the position/monomer level are converted to the motif level using

$$\psi_i[j] = \sum_{k=j}^{j+m-1} \phi_i[k],\ j \in [0, L-m] \tag{4}$$

such that $\psi_i \in \mathbb{R}^{L-m+1}$. The scores $\psi_i[j]$ are then transformed into a probability distribution as follows:

$$p_i[j] = \frac{\exp(-\psi_i[j]/\tau_1)}{\sum_{k=0}^{L-m} \exp(-\psi_i[k]/\tau_1)}, \quad j = 0, \ldots, L-m, \tag{5}$$

where the higher $p_i[j]$ values are assigned to motifs with smaller attribution score. The temperature parameter τ_1 controls the concentration of the probability distribution: a smaller τ_1 amplifies differences in ψ_i values, leading to sharper probabilities that favor positions with low importance scores, while a larger τ_1 leads to more uniform assignment of probability scores. Similarly, we convert the motif attribution scores in $\mathbf{M}$ into probabilities using a temperature parameter τ_2:

$$p_f[s] = \frac{\exp(I_s/\tau_2)}{\sum_{k=0}^{n_m-1} \exp(I_k/\tau_2)}, \quad s = 0, \ldots, n_m - 1, \tag{6}$$

where I_s denotes the attribution score of the s-th motif in $\mathbf{M}$. Higher $p_f[s]$ values correspond to motifs with larger attribution scores. For each mutation in x_i, we select: (i) a starting position j in sequence $\mathbf{x}_i$ to mutate, and (ii) the index s of replacement motif from $\mathbf{M}$, as follows:

$$j \sim \text{Categorical}(p_i[0], p_i[1], \ldots, p_i[L-m]) \tag{7a}$$

$$\mathbf{s} \sim \text{Categorical}(p_f[0], p_f[1], \ldots, p_f[n_m - 1]) \tag{7b}$$

Using the above equations, we perform **two mutations** for each sequence $\mathbf{x}_i \in \mathcal{D}^{\text{top}}$: (i) an **exploitative mutation** with $\tau_1 = \tau_2 = 1$, and (ii) an **exploratory mutation** with $\tau_1 = \tau_2 = 1000$. The resulting mutated sequence is denoted by $\mathbf{x}_i^M$. Each mutated sequence is evaluated using the oracle and collected into $\mathcal{D}_M^{\text{top}} = \left\{(\mathbf{x}_i^M, f(\mathbf{x}_i^M))\right\}_{i=0}^{B-1}$, which satisfies $|\mathcal{D}_M^{\text{top}}| = B$. The dataset for the next iteration is then updated as $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \mathcal{D}_M^{\text{top}}$. One iteration ($t$) of the design process is illustrated in Figure 1.

The mutation strategy described between Equations 3–7 represents just one possible approach. Depending on the temperature parameters τ_1 and τ_2, `IDEAS` can implement a variety of mutation strategies. Through an ablation study, we empirically demonstrate that the current strategy yields the most effective design performance.

Table 1: Area under the curve (AUC) comparison across six datasets and eight methods. Best results are shown in **bold** and second best are underlined.

	Aliphatic			GRAVY			α-helix			TF1			TF2			TF3		
Method	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100
BO	580.90 (30.97)	551.45 (11.43)	469.95 (28.55)	12.23 (2.28)	15.27 (1.44)	12.18 (0.54)	2.66 (0.72)	1.91 (0.08)	1.06 (0.39)	2.93 (0.19)	2.92 (0.13)	2.07 (0.02)	3.20 (0.13)	2.88 (0.06)	2.87 (0.03)	3.24 (0.13)	3.47 (0.04)	3.56 (0.08)
CMA-ES	450.62 (67.90)	689.86 (19.16)	839.00 (39.78)	17.09 (1.03)	20.82 (0.70)	22.19 (0.75)	2.69 (0.31)	3.42 (0.21)	3.37 (0.46)	4.09 (0.44)	4.12 (0.27)	4.12 (0.30)	4.33 (0.34)	4.65 (0.28)	4.85 (0.19)	3.62 (0.44)	4.00 (0.47)	4.39 (0.23)
AdaLead	556.11 (97.02)	996.55 (110.77)	1341.10 (146.38)	13.49 (1.02)	26.90 (3.27)	31.00 (2.40)	**4.79** (0.43)	5.71 (0.26)	5.15 (0.20)	4.30 (0.46)	4.67 (0.42)	4.40 (0.29)	4.40 (0.49)	5.20 (0.30)	5.43 (0.11)	3.61 (0.80)	4.13 (0.79)	4.70 (0.29)
Cbas	456.59 (27.65)	447.97 (47.89)	437.65 (60.27)	6.69 (1.62)	10.67 (0.38)	10.30 (0.57)	2.22 (0.04)	2.22 (0.28)	1.91 (0.36)	3.24 (0.22)	3.03 (0.02)	3.11 (0.11)	2.81 (0.88)	2.73 (0.18)	2.79 (0.07)	3.29 (0.13)	3.25 (0.10)	3.41 (0.02)
Dbas	509.73 (63.71)	508.31 (25.60)	509.40 (40.09)	6.89 (0.38)	9.92 (0.51)	9.60 (0.84)	1.81 (0.25)	2.39 (0.57)	1.90 (0.18)	3.09 (0.11)	3.19 (0.14)	3.03 (0.09)	2.76 (0.11)	2.87 (0.15)	2.80 (0.09)	3.28 (0.18)	3.17 (0.13)	3.35 (0.10)
GFN-AL	198.07 (57.15)	217.18 (47.58)	207.21 (75.68)	9.64 (0.77)	9.39 (0.39)	9.56 (0.64)	1.40 (0.33)	1.44 (0.36)	1.48 (0.36)	3.75 (0.17)	3.57 (0.15)	3.43 (0.13)	5.08 (0.32)	4.77 (0.38)	4.38 (0.34)	**4.89** (0.24)	4.58 (0.25)	4.29 (0.20)
DynaPPO	97.78 (22.30)	188.91 (11.39)	432.46 (58.98)	10.58 (0.26)	15.72 (0.96)	18.64 (0.26)	1.88 (0.33)	2.23 (0.38)	2.99 (0.29)	3.34 (0.07)	3.30 (0.02)	3.36 (0.08)	3.98 (0.14)	4.08 (0.02)	3.80 (0.02)	3.99 (0.08)	3.90 (0.02)	3.99 (0.05)
`IDEAS`	**910.14** (106.55)	**1639.84** (164.34)	2024.83 (153.63)	**22.56** (2.18)	**37.36** (2.45)	39.28 (1.05)	4.75 (0.46)	5.83 (0.21)	**6.55** (0.13)	**4.72** (0.33)	**4.72** (0.37)	4.85 (0.22)	**5.31** (0.37)	**5.82** (0.24)	5.98 (0.23)	4.59 (0.53)	4.84 (0.60)	5.19 (0.39)
`IDEAS-X`	712.44 (85.62)	1592.69 (190.97)	**2182.40** (146.41)	17.96 (0.74)	35.10 (0.94)	**40.53** (1.12)	4.65 (0.33)	**5.96** (0.24)	6.42 (0.13)	4.60 (0.83)	4.68 (0.40)	**4.95** (0.45)	4.79 (0.48)	5.77 (0.49)	**6.13** (0.13)	4.39 (0.66)	**5.31** (0.25)	**5.24** (0.41)

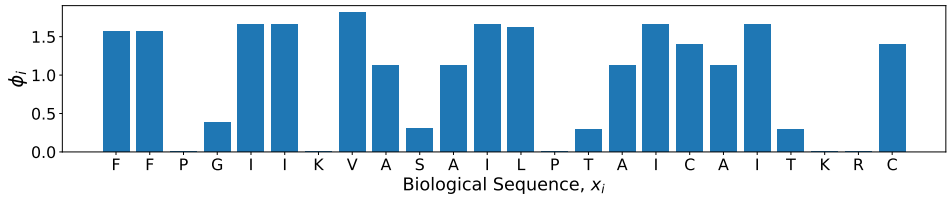

Figure 2: Position-wise attribution scores ϕ_i for a sample sequence x_i from the GRAVY index dataset (Appendix B).

3.1 XAI Model

As shown in Equations 1 and 2, an explainable model (XAI) is required to compute attribution scores ϕ_i. Two classes of methods can provide these scores: (i) post-hoc explainability techniques such as DeepLift (Shrikumar et al., 2017), Integrated Gradients (Sundararajan et al., 2017), or GradientSHAP (Lundberg & Lee, 2017), and (ii) architecturally interpretable models such as attention-based Transformers (Vaswani et al., 2017; Wu et al., 2020). Any method that produces faithful (i.e., reality-consistent) attribution scores (Dasgupta et al., 2022; Pandey et al., 2025) is compatible with the `IDEAS` framework. As shown empirically in Section 5.3.4, the choice of XAI model does not affect design performance as long as the resulting attribution scores are comparably faithful.

Based on the work of Pandey et al. (2025) on the XAI model, we adopt COLOR as our XAI. COLOR segments input sequences x_i of length L into $L - m + 1$ overlapping motifs of size m, learns their latent representations, and models motif interactions through linear transformations in latent space. This architecture enables efficient computation of position-wise attribution scores ϕ_i. We select COLOR for two key advantages: (i) faithful (aligns with the ground-truth) attribution scores, and (ii) less trainable parameters that ensure sample-efficient learning from limited data. More details on COLOR are provided in Appendix A.

In Figure 2, we visualize attribution scores ϕ_i for a representative GRAVY-index sequence. Amino acids V, I, and F receive the highest attributions, consistent with the analytical form of the GRAVY index (Appendix B), indicating that COLOR yields faithful explanations.

3.2 Metrics to quantify the performance

We compare sequence design methods based on their ability to generate high-quality, diverse, and novel sequences (Bengio et al., 2021). We reformulate the quality metric to emphasize sample efficiency by measuring which method generates high-property sequences with fewer oracle evaluations, quantified through the Area Under the Curve (AUC):

$$\text{AUC} = \int_0^T \frac{1}{B} \sum_{\mathbf{x} \in \mathcal{D}^{\text{top}}(t)} f(\mathbf{x})\, dt, \tag{8}$$

where T denotes the total number of design iterations (set to $T = 10$ in our experiments). Note that this AUC is unbounded above and typically lies in $[0, +\infty)$.

While we adopt the diversity and novelty definitions from Bengio et al. (2021), we standardize their evaluation using a common fitness threshold y_c to ensure fair comparison across methods. After all design iterations, we identify for each method the iteration t at which the mean fitness $\langle y_t \rangle$ of generated sequences reaches y_c, where $\mathbf{y}_t \in \mathbb{R}^{B \times 1}$ denotes the fitness values of the B generated sequences at iteration t, and denote the corresponding sequences as $\mathbf{x}_t \in \mathbb{R}^{B \times L}$. Diversity is then computed as the mean pairwise edit distance among sequences in $\mathbf{x}_t$, and novelty as the mean edit distance between $\mathbf{x}_t$ and the shared initial population $\mathbf{x}_0$. Using a common y_c ensures all methods are evaluated at equivalent fitness levels, enabling consistent and unbiased comparison.

4 Related Work

Deep learning models have been widely combined with post-hoc explainable AI (XAI) methods, such as DeepLIFT and Integrated Gradients, to identify critical motifs in biological sequences, particularly DNA. These approaches have been applied to tasks like binding affinity prediction, where attribution scores highlight sequence regions driving model predictions (de Almeida et al., 2022;

Shrikumar et al., 2018; Avsec et al., 2021; Horton et al., 2023; Seitz et al., 2024), and several studies provide experimental validation that XAI methods can reliably distinguish functionally critical positions from non-salient regions (de Almeida et al., 2022). More recently, Seitz et al. (2025) used explainable models with a cluster summary matrix (CSM) to systematically identify essential positions and positions where mutations strongly affect functional outcomes. Meanwhile, Transformer-based architectures (Vaswani et al., 2017) have enabled powerful sequence-to-property prediction models, and studies have explored whether attention weights correspond to biologically important positions (Vig et al., 2020; Liu et al., 2024; Karimi et al., 2020). However, recent evidence suggests that attention scores are often poorly aligned with ground-truth attribution and may not be reliable as faithful explanations in biological sequence analysis (Pandey et al., 2025).

Despite the substantial progress in using explainable models to identify critical and non-salient motifs, existing work has largely focused on *post-hoc explainability* and biological insight. To the best of our knowledge, no prior method has systematically leveraged attribution scores from XAI models to *actively guide the biological sequence design process*. This is exactly the gap that motivates our proposed method, `IDEAS`.

5 Experimental Results

We evaluate `IDEAS` against **seven baseline methods** spanning evolutionary algorithms, reinforcement learning, and generative models on **six continuous-property optimization tasks**. Performance is measured using the AUC metric (Equation 8), which quantifies how efficiently each method discovers high-property sequences under limited oracle queries. We further analyze diversity–performance and novelty–performance trade-offs using AUC-Diversity and AUC-Novelty plots, characterizing each method's balance between exploitation and exploration.

Datasets and Oracle Functions. The six optimization tasks span multiple biological sequence types. The first two tasks optimize the **Aliphatic Index** and **GRAVY Index** of Anti-Cancer Peptides (ACPs), consisting of sequences over 20 amino acids ($|\mathcal{V}| = 20$) with lengths between 20 and 97. These properties admit closed-form analytical expressions (Appendix B), enabling exact oracle evaluations and qualitative analysis of the optimization process.

The third task optimizes the α**-helix ratio** of protein primary sequences, defined as the fraction of residues forming α-helical secondary structure. We use 3,655 protein sequences curated by Gupta & Zou (2019b) from UniProt (uni, 2017), with lengths ranging from 20 to 50, and employ PSIPRED (Buchan et al., 2013; Gupta & Zou, 2019b) as the oracle.

The final three tasks focus on optimizing transcription factor (TF) binding affinity to DNA sequences of length 8 over the nucleotide alphabet ($|\mathcal{V}| = 4$). Using experimentally measured binding affinities for all possible 8-mers compiled by Barrera et al. (2016b), we randomly select three TFs, denoted **TF1**, **TF2**, and **TF3**. Additional details are provided in Appendix B.

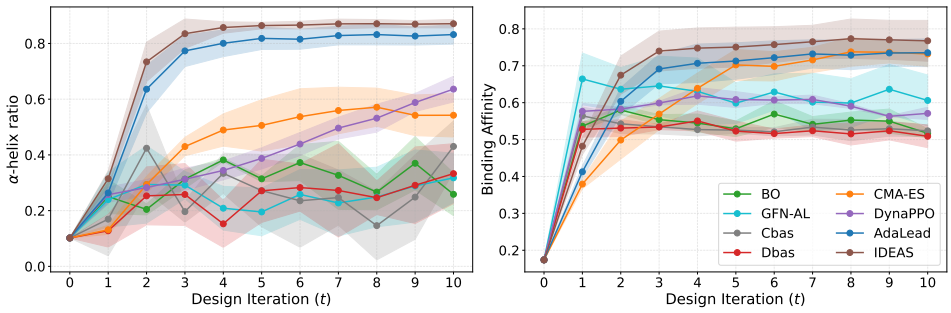

Figure 3: Evolution of the property y as a function of design iteration (t) for a) the α-helix ratio and b) the Binding affinity (TF1) datasets for $B = 100$.

Baselines. We compare `IDEAS` against a range of evolutionary, reinforcement learning, and generative baselines. For evolutionary methods, we include **AdaLead** (Sinai et al., 2020), which performs greedy hill climbing in sequence space, and **CMA-ES** (Hansen, 2006), which optimizes a continuous embedding of one-hot sequence encodings while adaptively learning a full covariance matrix over sequence dimensions. As an RL-based baseline, we evaluate **DynaPPO** (Angermueller et al.,

2019), which applies on-policy proximal policy optimization to sequence design. We also include **GFN-AL** (Jain et al., 2022), a GFlowNet-based approach (Bengio et al., 2021) that learns a stochastic generative policy and samples diverse high-property sequences via flow-matching objectives. In addition, we compare against **Bayesian Optimization (BO)** adapted to large biological sequence spaces following Sinai et al. (2020). Finally, we include two VAE-based generative models, **Dbas** (Brookes & Listgarten, 2018) and **Cbas** (Brookes et al., 2019), which use adaptive probabilistic sampling to progressively bias generation toward the objective function.

Evaluation Strategy. For all datasets, we first curate an initial dataset $D_0 = \{x_i \mid y_i < y_{\text{cut}}\}$, where the cutoff values y_{cut} for each dataset are provided in Table 5 in Appendix B. Starting from $\mathcal{D}_0$, we perform 10 design iterations (t) for each of three oracle function budgets, $B \in \{20, 50, 100\}$. For each method and budget, we conduct 10 independent trials and report results as the mean and standard deviation across these trials.

To isolate the effect of exploration in IDEAS, as discussed in Section 3, we also evaluate a restricted variant, denoted IDEAS−X, which performs only a single exploitative mutation with $\tau_1 = \tau_2 = 1$ and disables exploratory mutations.

5.1 IDEAS OFFERS ACCELERATION IN SEQUENCE DESIGN

Starting from the initial dataset $\mathcal{D}_0$, we perform ten design iterations to optimize the target properties. Figures 3 and Appendix C.1 show the evolution of the mean property value of the generated sequences as a function of the design iteration t. Across all datasets and oracle budgets B, IDEAS consistently achieves faster improvements than competing baselines. We quantify this advantage using the area under the curve (AUC) of the corresponding trajectories, with results summarized in Table 1. **Overall, IDEAS outperforms all baselines by an average of 19%**, demonstrating its effectiveness for biological sequence optimization. Comparing IDEAS with its restricted variant IDEAS−X, we observe that their mean AUC values differ by less than 2%, indicating that exploitative mutations account for the majority of the performance gains in IDEAS. The contribution of exploratory mutations is therefore not reflected in the AUC metric and is instead examined in our analysis of diversity and novelty.

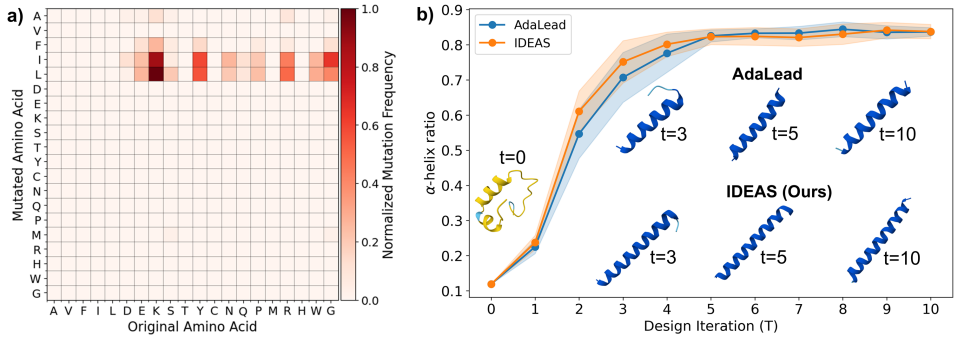

Figure 4: (a) Normalized frequency of all possible 20 × 20 mutations for the GRAVY index. (b) Structural evolution on the α-helix ratio dataset comparing IDEAS with AdaLead (second-best).

To investigate the source of IDEAS's superior performance, we analyze the normalized frequency of all exploitative mutations for the GRAVY index, shown in Figure 4a. The figure reveals that the majority of mutations favor substitutions to amino acids I and L, with E→I and E→L among the most frequent transitions. This pattern is consistent with the analytical form of the GRAVY index presented in Appendix B, which assigns positive contributions to I and L and a negative contribution to E. Figure 4b shows the evolution of the α-helix ratio as a function of t, along with the structures of the best sequences at selected iterations. We compare IDEAS with AdaLead, the second-best method. After $t = 5$, the best sequence produced by both the methods is largely helical. Moreover, **IDEAS achieves a best sequence with a 10% higher α-helix ratio than AdaLead**, highlighting its ability to accelerate the sequence design process.

5.2 IDEAS PERFORMANCE LIES ON THE PARETO CURVE

As noted by Jain et al. (2022), biological sequence design requires balancing AUC, diversity, and novelty. Accordingly, Figure 5 presents the AUC–Diversity trade-off for an oracle budget of $B = 50$,

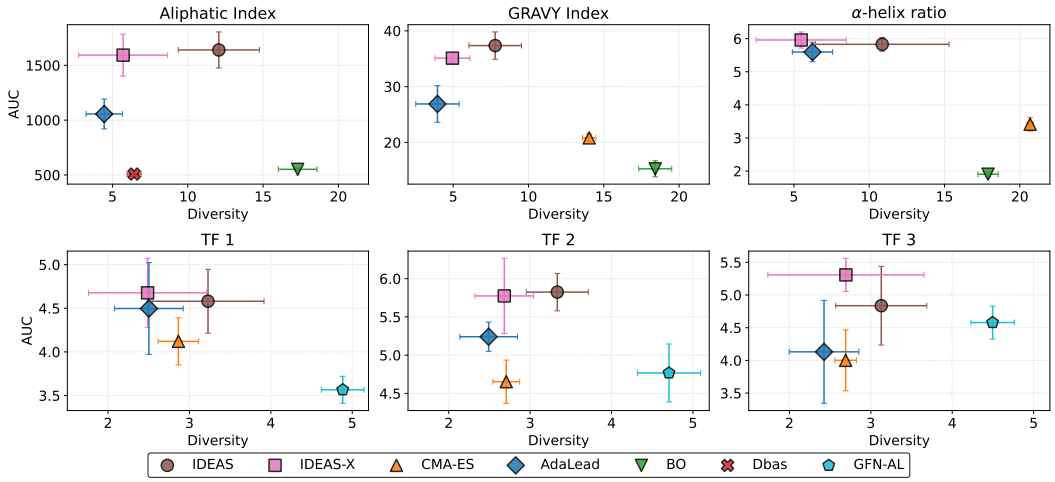

Figure 5: Trade-off between diversity and AUC across six datasets with $B = 50$. IDEAS consistently lies on the Pareto curve.

while results for $B \in \{20, 100\}$ are shown in Figure 14 in the Appendix. The plots report the top five methods ranked by AUC in Table 1. Across budgets, IDEAS lies on the Pareto frontier, indicating a favorable balance between AUC and diversity. Although IDEAS and IDEAS−X achieve comparable AUC, IDEAS attains 41% higher diversity, attributable to its exploratory mutation strategy.

Figure 15 in the Appendix shows the AUC–Novelty trade-off, where either IDEAS or IDEAS−X lies on the Pareto frontier with less than 2% difference in novelty. Overall, the AUC, Diversity, and Novelty analyses demonstrate that IDEAS achieves a favorable balance across all metrics while improving AUC by 19% over the baselines.

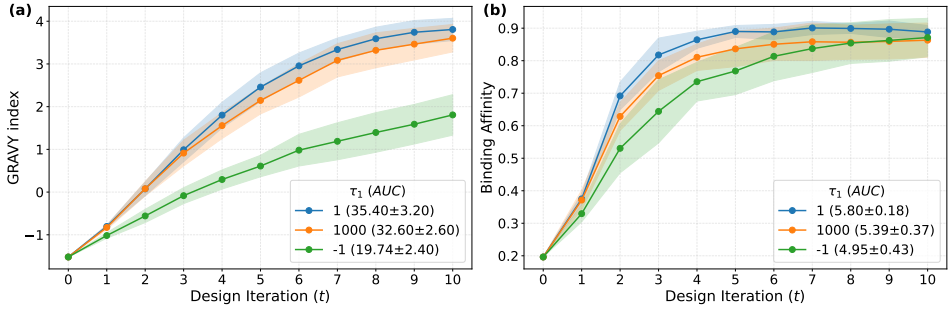

Figure 6: Impact of the position-selection temperature τ_1 (Equation 4) on design performance for (a) GRAVY index and (b) Binding affinity (TF1).

5.3 Ablation Study

Having established the performance of IDEAS relative to baselines in terms of AUC, diversity, and novelty, we next examine the contribution of individual components, including the temperature parameters (τ_1, τ_2) and the number of mutations. For clarity in the ablation study, we disable the exploratory component of IDEAS to isolate the effects of these parameters.

5.3.1 Effect of temperature on design

Equations 4–7 show that the mutation position (j) and replacement motif (s) are controlled by temperature parameters τ_1 and τ_2, respectively; we therefore analyze their impact on design performance.

Effect of τ_1: Figures 6a and 6b show the effect of τ_1 on optimizing the GRAVY index and binding affinity. Setting $\tau_1 = 1$ biases mutations toward positions with smaller ψ_i (Equation 4), while $\tau_1 = -1$ favors positions with larger ψ_i, and $\tau_1 = 1000$ corresponds to uniform random position

selection. In this study, we fix $\tau_2 = 1$ to bias IDEAS toward replacement motifs with higher attribution scores. Across both tasks, $\tau_1 = 1$ achieves average improvements of 8% and 48% over the $\tau_1 = 1000$ and $\tau_1 = -1$ settings, respectively, demonstrating the importance of informed mutation position selection. Diversity and novelty differ by less than 1% across all settings.

Effect of τ_2: Analogous to the study of τ_1, we evaluate $\tau_2 \in \{1, 1000, -1\}$ while fixing $\tau_1 = 1$. Setting $\tau_2 = 1$ biases IDEAS toward selecting replacement motifs with higher attribution scores, whereas $\tau_2 = -1$ favors motifs with lower attribution scores. Figures 7a and 7b show the effect of τ_2 on the GRAVY index and binding affinity, respectively. Across both tasks, $\tau_2 = 1$ achieves average improvements of 11% and 22% over the $\tau_2 = 1000$ and $\tau_2 = -1$ settings, underscoring the importance of replacement motif selection during mutation.

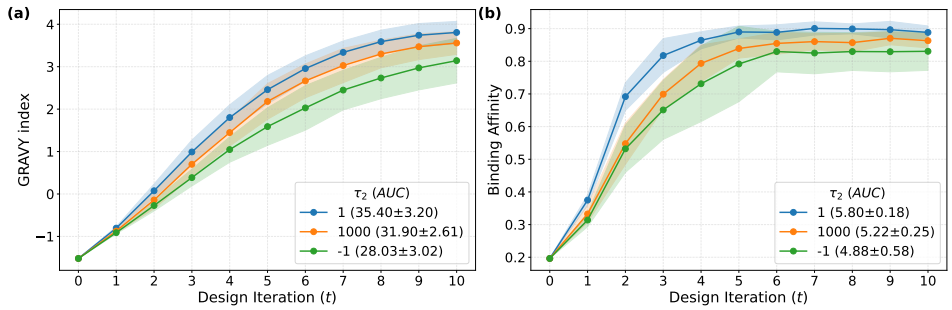

Figure 7: Impact of the replacement-motif selection temperature τ_2 (Equation 6) on design performance for (a) GRAVY index and (b) Binding affinity (TF1).

For the GRAVY index, setting $\tau_1 = 1$ improves performance by 79% over $\tau_1 = -1$, while $\tau_2 = 1$ yields a 26% improvement over $\tau_2 = -1$. In contrast, for binding affinity, $\tau_1 = 1$ provides a 17% improvement, whereas $\tau_2 = 1$ provides a 19% improvement over their respective negative settings. These results indicate that mutation position selection (controlled by τ_1) has a larger impact on GRAVY optimization, while replacement motif selection (controlled by τ_2) plays a comparatively greater role in binding affinity optimization. Overall, neither component alone dominates sequence design performance; prioritizing one over the other can lead to a decelerated design process.

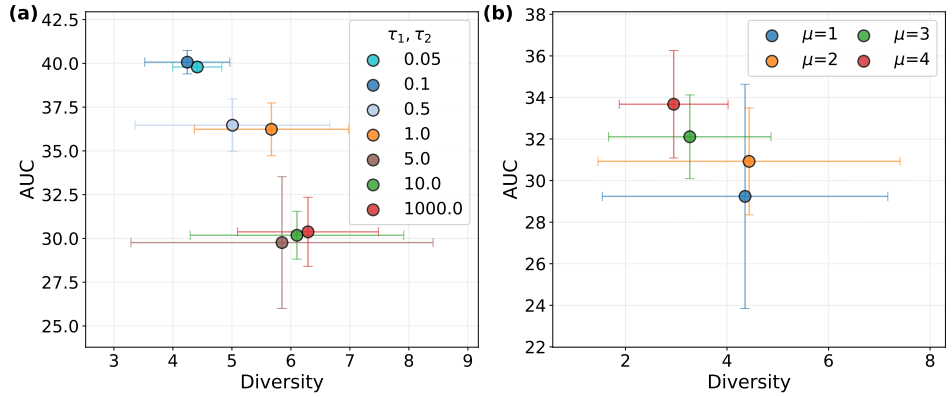

Figure 8: Impact of a) Temperature τ_1, τ_2, and b) the number of mutations (μ) on design performance.

Varying τ_1 and τ_2 together: Having examined the effects of τ_1 and τ_2 independently, we next study how design performance varies as the mutation strategy transitions from highly exploitative ($\tau_1, \tau_2 = 0.05$) to highly exploratory ($\tau_1, \tau_2 = 1000$). The AUC–Diversity trade-off in Figure 8a shows that AUC decreases as temperature increases and plateaus for $\tau_1, \tau_2 \geq 5$, while the maximum AUC is reached around $\tau_1, \tau_2 \approx 0.1$. In contrast, diversity increases with temperature, reaching its minimum near $\tau_1, \tau_2 \approx 0.1$ and saturating around $\tau_1, \tau_2 \approx 5$. These results indicate that increased exploration slows optimization while promoting diversity, highlighting the trade-off between design efficiency and diversity.

5.3.2 EFFECT OF NUMBER OF MUTATIONS OF DESIGN

Based on the description of IDEAS in Section 3, we perform a single exploitative mutation by default; for ablation studies, the exploratory component is disabled. In this experiment, we examine the effect of the number of exploitative mutations (μ) on design performance while fixing $\tau_1 = \tau_2 = 1$. For $\mu > 1$, we sample μ unique mutation positions from x_i using Equation 7a without replacement, and select μ replacement motifs using Equation 7b with replacement. This choice reflects the fact that multiple positions can be mutated using the same motif. Figure 8b presents the resulting AUC-Diversity trade-off as a function of μ. As μ increases, AUC improves while diversity decreases, indicating a trade-off between optimization performance and diversity. Mutating multiple positions with similar motifs reduces sequence variability, leading to lower diversity.

5.3.3 EFFECT OF MOTIF SIZE

Figure 9 illustrates the effect of the motif size m in Equation 3 on design performance. Increasing the motif size from $m = 1$ to $m = 2$ leads to substantial gains, with AUC and Diversity improving by 23% and 36%, respectively. Beyond this point, both metrics exhibit diminishing returns and largely plateau as m increases further. Novelty also shows a modest increase with larger motifs, though the improvement is comparatively limited. Overall, this analysis indicates that while an initial increase in motif size can significantly enhance design performance, larger motifs provide marginal additional benefits.

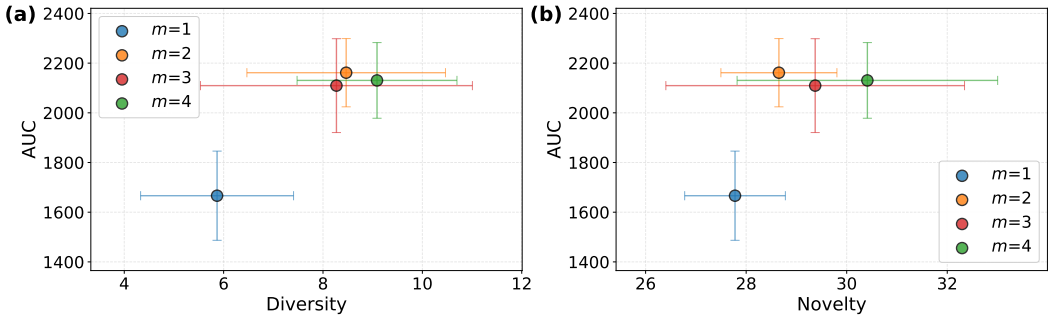

Figure 9: Effect of motif size (m) shown on Aliphatic index dataset using a) AUC-Diversity and b)AUC-Novelty plot.

We select the motif size m based on the COLOR (Pandey et al., 2025) model, which partitions a sequence into non-overlapping motifs of length m and learns a global representation by aggregating their latent embeddings. In practice, m is chosen to maximize predictive performance (e.g., R^2, MAE, MSE) on a held-out validation set during COLOR training, and this optimal m is then used within IDEAS. That said, m need not depend solely on COLOR and can also be guided by domain knowledge. For instance, in collagen, a motif size of $m = 3$ arises naturally from its repeating Gly–X–Y structural unit.

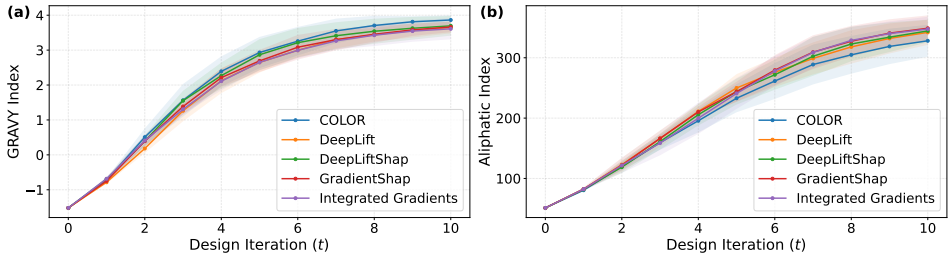

Figure 10: Effect of different XAI models with similar faithful attribution scores on optimizing (a) the GRAVY index and (b) the Aliphatic index.

5.3.4 IDEAS CAN BE INTEGRATED WITH DIFFERENT XAI MODELS

To demonstrate that IDEAS is agnostic to the choice of XAI, we integrate five representative explainability methods—COLOR, DeepLift, DeepLiftShap, GradientSHAP, and Integrated Gradients—into the IDEAS framework and evaluate their performance on optimizing the GRAVY and Aliphatic indices. These properties are well-suited for this analysis because their oracle functions $f(x_i)$ are analytical expressions, enabling direct assessment of the faithfulness of the resulting attribution scores ϕ_i. In this setting, the attribution scores produced by all five explainability methods are equally faithful to the ground-truth contributions. As shown in Figure 10, all methods achieve nearly identical optimization performance, with average AUC values differing by less than 2%. These results indicate that IDEAS is robust to the choice of XAI model, provided the attribution method yields faithful explanations.

5.3.5 ROBUSTNESS TO ATTRIBUTION NOISE

We evaluate the sensitivity of IDEAS to attribution quality ϕ by injecting noise into training labels of the GRAVY index dataset:

$$Y_{\text{noisy}} = Y + \mathcal{N}\left(0, (\sigma(Y) \cdot \eta)^2\right), \tag{9}$$

where η controls the noise level and ϕ quality is measured via Pearson correlation (PCC) with ground truth attribution scores. The results demonstrate two key robustness properties: (1) robust to attribution noise: a 66% drop in PCC ($\eta = 0 \rightarrow 10$) leads to only an 8% reduction in AUC; (2) robust to noisy training data: even at large noise ($\eta = 10$), AUC remains above the second-best baseline (AdaLead, AUC $= 26.9$). Performance degrades to AdaLead's level only when ϕ is effectively random ($\eta = 50$).

Table 2: Effect of label noise η on attribution quality (PCC) and design performance (AUC) on the GRAVY index dataset.

η	PCC	AUC (mean $\pm$ std)
0	0.99	37.10 ± 1.97
0.5	0.97	36.77 ± 1.71
1	0.91	37.04 ± 2.00
2	0.87	35.73 ± 2.55
5	0.76	36.42 ± 2.68
10	0.60	34.22 ± 2.29
50	-0.1	26.44 ± 4.79

6 CONCLUSION

In this work, we introduced IDEAS, an interpretable evolutionary framework for biological sequence design that integrates explainable models (XAI) with evolutionary optimization to accelerate design under constrained oracle budgets. By leveraging attribution-informed mutations, IDEAS replaces random perturbations with interpretable mutations, resulting in substantially improved sample efficiency over conventional evolutionary methods. Moreover, IDEAS achieves faster design progress than reinforcement learning– and generative model–based approaches by effectively exploiting attribution scores without requiring large training datasets. Across six design tasks and multiple oracle budgets, IDEAS achieves an average 19% acceleration over seven competitive baselines while consistently operating on the Pareto frontier of AUC–Diversity and AUC–Novelty trade-offs.

Limitations and Future Work. While IDEAS achieves strong design acceleration, its diversity and novelty are lower than those of RL- and generative model–based approaches due to the inherently local nature of evolutionary mutations. A promising direction for future work is to integrate attribution-guided signals into RL or generative frameworks, combining the sample efficiency and interpretability of IDEAS with the global exploration capabilities of learned sequence generators.

LLM USAGE STATEMENT

The usage of LLMs in this work is limited to paper writing support, language refinement, and experimental data processing. Specifically, LLMs assisted in improving the clarity and coherence of the manuscript, generating LaTeX tables, and formatting results for presentation. Importantly, LLMs were not involved in the design of algorithms, the development of theoretical results, or the execution of experiments, ensuring that all core scientific contributions remain entirely the work of the authors.

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

Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2017.

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *International conference on learning representations*, 2019.

Frances H. Arnold. Design by directed evolution. *Accounts of Chemical Research*, 31(3):125–131, 1998. doi: 10.1021/ar960017f.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, 53(3):354–366, 2021.

Luis A Barrera, Anastasia Vedenko, Jesse V Kurland, Julia M Rogers, Stephen S Gisselbrecht, Elizabeth J Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, et al. Survey of variation in human transcription factors reveals prevalent dna binding changes. *Science*, 351 (6280):1450–1454, 2016a.

Luis A Barrera, Anastasia Vedenko, Jesse V Kurland, Julia M Rogers, Stephen S Gisselbrecht, Elizabeth J Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, et al. Survey of variation in human transcription factors reveals prevalent dna binding changes. *Science*, 351 (6280):1450–1454, 2016b.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

Jesse D. Bloom and Frances H. Arnold. In the light of directed evolution: Pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences*, 106(supplement_1):9995–10000, 2009. doi: 10.1073/pnas.0901522106. URL https://www.pnas.org/doi/abs/10.1073/pnas.0901522106.

David Brookes, Hahnbeom Park, and Jennifer Listgarten. Conditioning by adaptive sampling for robust design. In *International conference on machine learning*, pp. 773–782. PMLR, 2019.

David H Brookes and Jennifer Listgarten. Design by adaptive sampling. *arXiv preprint arXiv:1810.03714*, 2018.

Daniel W. A. Buchan, Federico Minneci, Tim C. O. Nugent, Kevin Bryson, and David T. Jones. Scalable web services for the psipred protein analysis workbench. *Nucleic Acids Research*, 41 (W1):W349–W357, 06 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt381. URL https://doi.org/10.1093/nar/gkt381.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, pp. 4794–4815. PMLR, 2022.

Bernardo P de Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature genetics*, 54(5):613–624, 2022.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

Jacob J Graham, Shri V Subramani, Xinyan Yang, Timothy M Russell, Fuzhong Zhang, and Sinan Keten. Charting the envelope of mechanical properties of synthetic silk fibers through predictive modeling of the drawing process. *Science Advances*, 11(10):eadr3833, 2025.

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

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

Nikolaus Hansen. The cma evolution strategy: a comparing review. *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms*, pp. 75–102, 2006.

Connor A Horton, Amr M Alexandari, Michael GB Hayes, Emil Marklund, Julia M Schaepe, Arjun K Aditham, Nilay Shah, Peter H Suzuki, Avanti Shrikumar, Ariel Afek, et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science*, 381(6664): eadd1250, 2023.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

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.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *CoRR*, abs/1906.02691, 2019. URL http://arxiv.org/abs/1906.02691.

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.

Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Pierce J Ogden, Eric D Kelsic, Sam Sinai, and George M Church. Comprehensive aav capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science*, 366(6469):1139–1143, 2019.

Akash Pandey, Wei Chen, and Sinan Keten. Color: A compositional linear operation-based representation of protein sequences for identification of monomer contributions to properties. *Journal of Chemical Information and Modeling*, 65(9):4320–4333, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Evan E Seitz, David M McCandlish, Justin B Kinney, and Peter K Koo. Interpreting cis-regulatory mechanisms from genomic deep neural networks using surrogate models. *Nature machine intelligence*, 6(6):701–713, 2024.

Evan E Seitz, David M McCandlish, Justin B Kinney, and Peter K Koo. Uncovering the mechanistic landscape of regulatory dna with deep learning. *bioRxiv*, pp. 2025–10, 2025.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416*, 2018.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D Kelsic. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Olena Tokareva, Matthew Jacobsen, Markus Buehler, Joyce Wong, and David L. Kaplan. Structure–function–property–design interplay in biopolymers: Spider silk. *Acta Biomaterialia*, 10(4):1612–1626, 2014. ISSN 1742-7061. doi: https://doi.org/10.1016/j.actbio.2013.08.020. URL https://www.sciencedirect.com/science/article/pii/S1742706113004121. Biological Materials.

Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. *CoRR*, abs/2107.06882, 2021. URL https://arxiv.org/abs/2107.06882.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Zhengxuan Wu, Thanh-Son Nguyen, and Desmond C. Ong. Structured self-attention weights 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 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.24. URL https://aclanthology.org/2020.blackboxnlp-1.24/.

Sihyun Yu, Sungsoo Ahn, Le Song, and Jinwoo Shin. Roma: Robust model adaptation for offline model-based optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4619–4631. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/24b43fb034a10d78bec71274033b4096-Paper.pdf.

Marc Zimmer. Green fluorescent protein (gfp): applications, structure, and related photophysical behavior. *Chemical reviews*, 102(3):759–782, 2002.

APPENDIX

A COLOR: XAI MODEL

COLOR, developed by Pandey et al. (2025), computes position-wise attribution scores using an **absolute** value operation. This approach assigns equal positive importance to positions that contribute positively or negatively to the property of interest. For sequence design, this is suboptimal, as it is crucial to distinguish between positively and negatively contributing positions and motifs. In the present study, we replace the absolute operation with a **ReLU** function, enabling `IDEAS` to focus exclusively on positively attributing positions and motifs.

All datasets considered in this work consist of continuous properties; therefore, we use mean squared error (MSE) as the loss function and train the COLOR model using the Adam optimizer. Model training is stopped based on the validation loss. The number of training parameters in the model for each dataset is given in Table 3.

Dataset	Number of parameters in COLOR model
Aliphatic index	10,260
GRAVY index	10,260
α-helix ratio	14,440
TF1, TF2, TF3	9,981

Table 3: Number of training parameters in the COLOR model for different datasets.

B DATASET DETAILS

B.1 ALIPHATIC INDEX

The aliphatic index is a physicochemical property of a protein sequence that quantifies the relative volume occupied by aliphatic side chains. It is defined as the weighted sum of the mole fractions of aliphatic amino acids—alanine (A), valine (V), isoleucine (I), and leucine (L)—in the sequence. Formally, for a sequence x of length L, the aliphatic index is given by

$$\mathrm{AI}(x) = \chi_A + a\,\chi_V + b\,(\chi_I + \chi_L), \tag{10}$$

where $\chi_\cdot$ denotes the mole fraction of the corresponding amino acid in x, and $a = 2.9$, $b = 3.9$ are empirically determined coefficients. Due to its analytical formulation, the aliphatic index admits an exact oracle function, making it well-suited for evaluating the faithfulness of attribution scores and the effectiveness of sequence design methods.

B.2 GRAVY INDEX

GRAVY Index. The grand average of hydropathy (GRAVY) index measures the overall hydrophobicity of a protein sequence and is computed as the average of hydropathy values of its constituent amino acids. For a sequence x of length L, the GRAVY index is given by $\mathrm{GRAVY}(x) = \frac{1}{L}\sum_{i=1}^{L} h(x_i)$, where $h(\cdot)$ (Table 4) denotes the hydropathy score of an amino acid. Similar to the aliphatic index, the GRAVY index admits an analytical oracle, enabling exact evaluation and faithful attribution analysis.

C ADDITIONAL RESULTS

C.1 EVOLUTION OF PROPERTY AS A FUNCTION OF DESIGN ITERATIONS

Figure 12 and Figure 13 compare the evolution of the target property y as a function of design iterations across eight methods on six different datasets for oracle budget B =100, 50, and 20, respectively.

Table 4: Hydropathy values for amino acids.

Amino Acid	**Hydropathy Value**
Alanine (A)	1.8
Arginine (R)	-4.5
Asparagine (N)	-3.5
Aspartic acid (D)	-3.5
Cysteine (C)	2.5
Glutamine (Q)	-3.5
Glutamic acid (E)	-3.5
Glycine (G)	-0.4
Histidine (H)	-3.2
Isoleucine (I)	4.5
Leucine (L)	3.8
Lysine (K)	-3.9
Methionine (M)	1.9
Phenylalanine (F)	2.8
Proline (P)	-1.6
Serine (S)	-0.8
Threonine (T)	-0.7
Tryptophan (W)	-0.9
Tyrosine (Y)	-1.3
Valine (V)	4.2

Table 5: Dataset parameters showing cutoff values (y_{cut}) and initial sample sizes (N_0).

Dataset	Cutoff (y_{cut})	N_0
Aliphatic index	20	50
GRAVY index	0.14	100
α-helix ratio	0.15	200
TF1, TF2, TF3	0.2	250

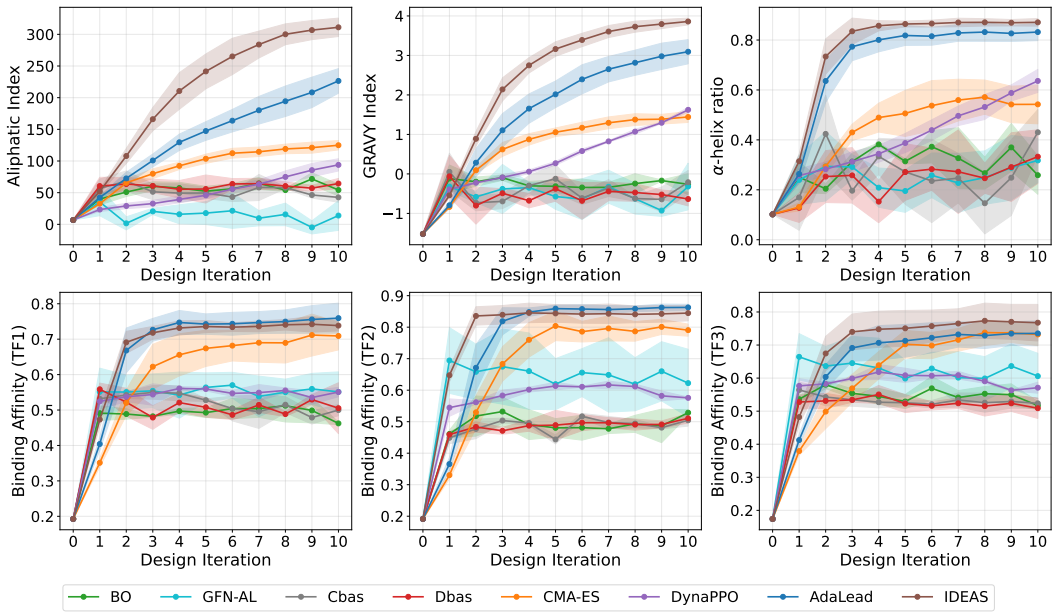

Figure 11: Evolution of property y shown as a function of design iteration (t) for all the datasets and $B = 100$.

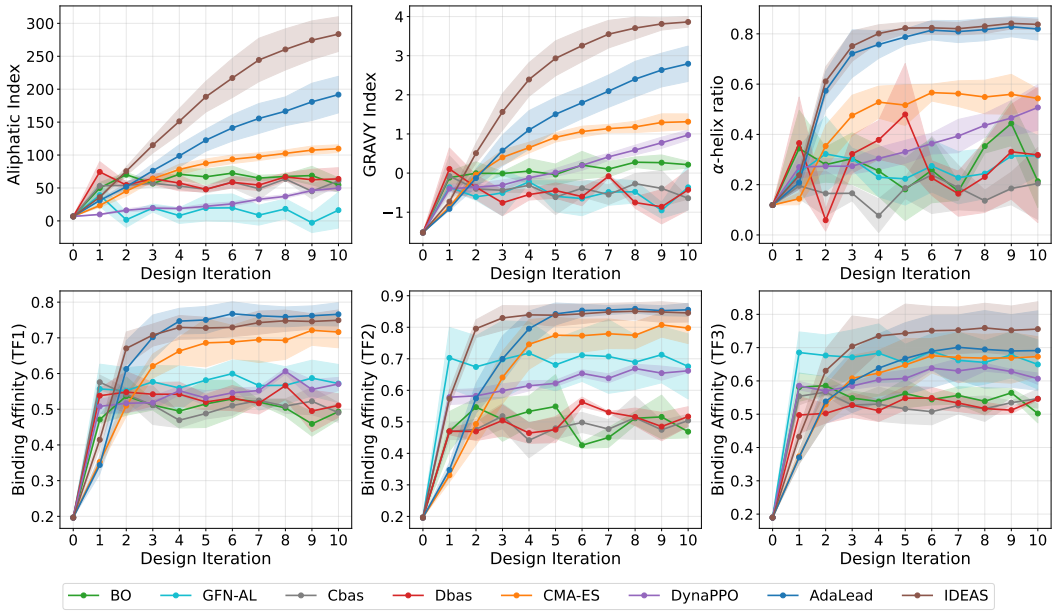

Figure 12: Evolution of property y shown as a function of design iteration (t) for all the datasets and $B = 50$.

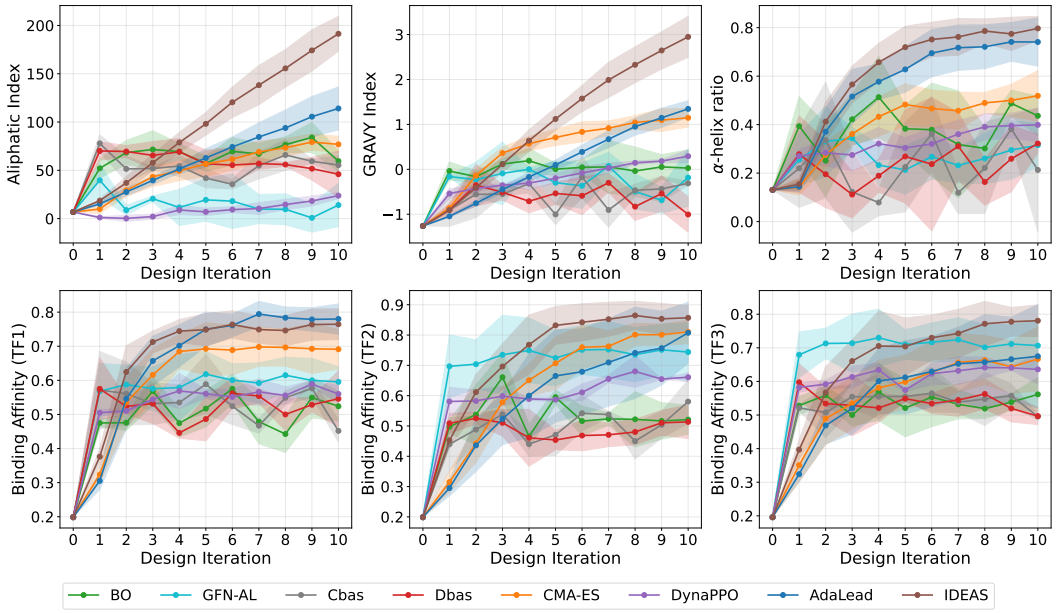

Figure 13: Evolution of property y shown as a function of design iteration (t) for all the datasets and $B = 20$.

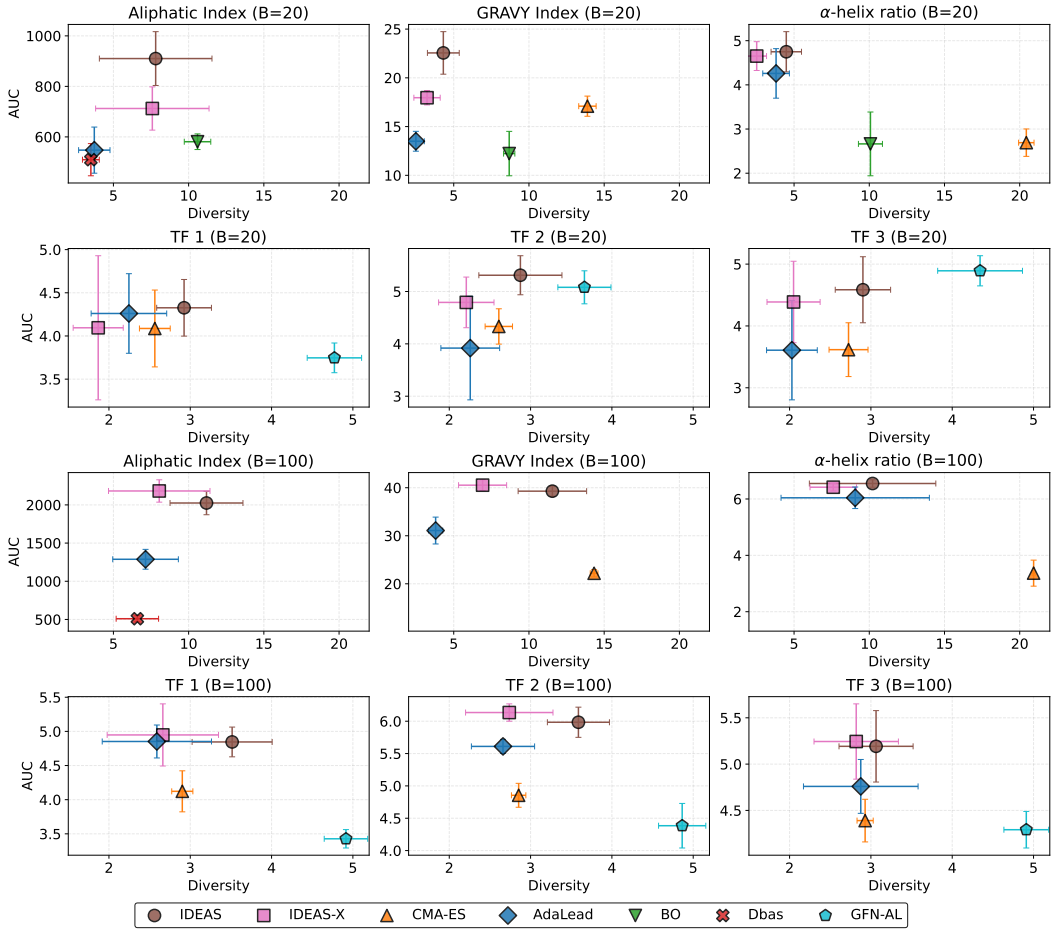

Figure 14: Trade-off between Diversity and performance (AUC) across six different dataset for B=20, 50.

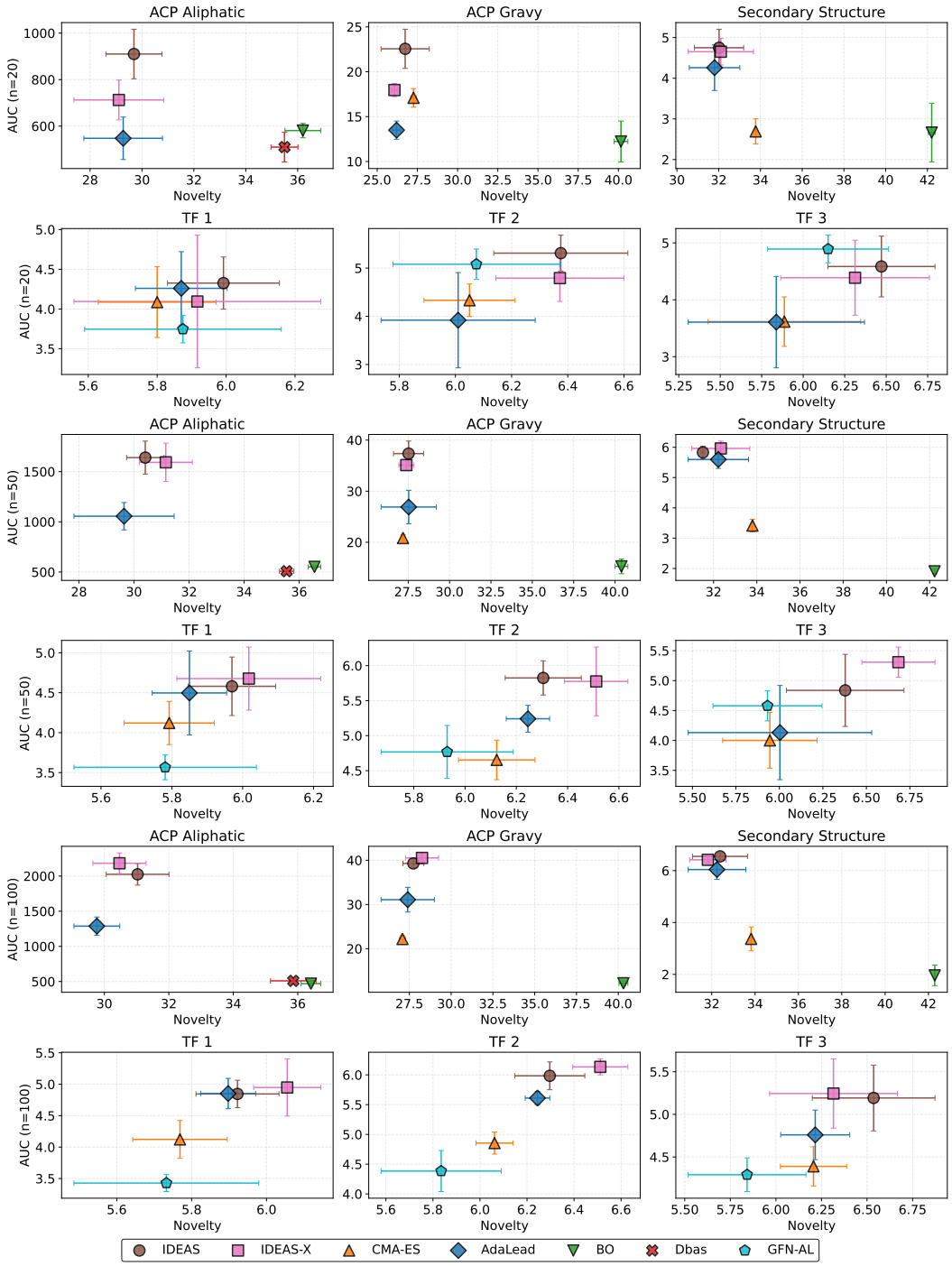

Figure 15: Trade-off between Novelty and performance (AUC) across six different dataset for B=20, 50 , and 100.