Freeze, Diffuse, Decode: Task-Aware Adaptation of Transformer Embeddings for Antimicrobial Peptide Design

Pankhil Gawade* GAWADE.PANKHIL@HELMHOLTZ-MUNICH.DE
Institute of AI for Health, Helmholtz Zentrum Munchen
Technical University of Munich, TUM School of Computation, Information and Technology

Adam Izdebski*
Institute of AI for Health, Helmholtz Zentrum Munchen
Technical University of Munich, TUM School of Computation, Information and Technology

Myriam Lizotte
Mila – Quebec AI Institute, Montreal, Canada
Department of Mathematics and Statistics, Université de Montreal

Kevin R. Moon
Dept. of Mathematics & Statistics Utah State University, Logan, UT, USA

Jake S. Rhodes
Dept. of Statistics Brigham Young University, Provo, UT, USA

Guy Wolf
Mila – Quebec AI Institute, Montreal, Canada
Department of Mathematics and Statistics, Université de Montreal

Ewa Szczurek EWA.SZCZUREK@HELMHOLTZ-MUNICH.DE
Institute of AI for Health, Helmholtz Zentrum Munchen
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Abstract

Pretrained transformers provide general-purpose molecular embeddings for downstream tasks. While these embeddings provide task-agnostic structural patterns, they lack task-specific alignment, limiting downstream performance. Here, we introduce FREEZE, DIFFUSE, DECODE (FDD), a diffusion-based framework that adapts pretrained transformer embeddings to downstream tasks by building a task-supervised diffusion geometry over the frozen embeddings, without any backbone training. Applied to antimicrobial peptide design, FDD yields low-dimensional, predictive, and interpretable representations that support property prediction, retrieval, and latent-space interpolation.

Keywords: molecular representation learning, manifold learning, diffusion maps, antimicrobial peptide design

*. Equal contribution.

1. Introduction

Molecular representation learning is fundamental to modern drug discovery. Embedding molecules into high-dimensional features that capture general structural properties enables downstream tasks such as molecular property prediction, molecule retrieval, and chemical space exploration. Recently, pretrained transformers have been reshaping molecular machine learning by providing embeddings that are increasingly leveraged across downstream tasks (Chithrananda et al., 2020; Zhou et al., 2023; Lin et al., 2023).

However, pretrained embeddings underperform on downstream tasks. In particular, probing pretrained embeddings for molecular property prediction is often outperformed by hand-crafted descriptors (Praski et al., 2025; Adamczyk et al., 2025). A possible reason is that pretrained embeddings are organized around general structural similarity, lacking task-specific alignment. While fine-tuning offers strong task-specific alignment, it inflates the representation's intrinsic dimensionality, a correlate of poorer generalization (Ansuini et al., 2019; Valeriani et al., 2023), and is sample-inefficient, offering a poor fit for low-data regimes typical of antimicrobial peptide design (van Tilborg et al., 2024).

To address this gap, we introduce Freeze, Diffuse, Decode (FDD), a framework based on diffusion geometry (Coifman and Lafon, 2006) that adapts frozen transformer embeddings to a downstream task without fine-tuning. FDD constructs a task-supervised diffusion geometry over the embeddings rather than raw inputs (Rhodes et al., 2023a; Aumon et al., 2025), then trains an encoder to extend it to unseen molecules. The resulting representation is low-dimensional and task-aware, improving over the frozen embeddings and narrowing the gap to fine-tuning. Our contributions are:

- We introduce FDD, which adapts frozen transformer embeddings to a downstream task through supervised diffusion geometry, without retraining the backbone.
- We show that FDD yields low-dimensional, predictive, and interpretable representations for molecular property prediction and peptide retrieval.
- We show that FDD supports latent-space interpolation whose decoded peptides gradually acquire physicochemical traits of antimicrobial activity.

2. Freeze, Diffuse, Decode

We propose Freeze, Diffuse, Decode (FDD), a framework for adapting frozen embeddings of a pretrained molecular transformer to downstream tasks via supervised diffusion geometry. Our framework runs in three stages. We first *freeze* a pretrained transformer and extract its embeddings (Section 2.1). We then *diffuse* by building a supervised diffusion geometry over the frozen embeddings and training an encoder to map any molecule into low-dimensional geometry-aware representations (Section 2.2). We finally *decode* these representations into task-specific outputs, fitting a lightweight predictor for property prediction and retrieval or decoding back to sequence space for generation (Section 2.3).

2.1. Freeze

The *Freeze* stage extracts embeddings from a pretrained transformer. Let $\mathcal{V}$ be a vocabulary of tokens, for example amino acids or the characters of a SMILES string. A molecule

$\mathbf{x} = (x_1, \ldots, x_T) \in \mathcal{V}^T$ is a finite sequence of such tokens, and the design space $\mathcal{X} = \mathcal{V}^*$ collects all finite-length sequences over $\mathcal{V}$. A frozen, pretrained transformer

$$f \colon \mathcal{X} \to \mathbb{R}^D \tag{1}$$

maps each molecule to a D-dimensional embedding $\mathbf{h} = f(\mathbf{x})$, where D is typically large.

2.2. Diffuse

The *Diffuse* stage adapts the frozen embeddings to a downstream task by learning a mapping

$$g_\theta \colon \mathbb{R}^D \to \mathbb{R}^d, \qquad d \ll D, \tag{2}$$

so that $\mathbf{z} := (g_\theta \circ f)(\mathbf{x}) \in \mathbb{R}^d$ is low-dimensional, yet informative for the downstream task. We construct a label-aware diffusion geometry on the training embeddings, then fit a parametric autoencoder that extends that geometry to molecules unseen during training.

Supervised diffusion geometry We measure similarity between frozen embeddings with RF-GAP proximities (Rhodes et al., 2023b), which treat two points as close when a random forest trained to predict y routes them to the same leaves. Specifically, we fit a random forest to the labeled pairs $\mathcal{D} = \{(\mathbf{h}_n, y_n)\}_{n \in [N]}$, where $\mathbf{h}_n = f(\mathbf{x}_n)$ is the embedding of molecule $\mathbf{x}_n$ and $[N] = \{1, \ldots, N\}$. Each tree t is grown on a bootstrap sample $B(t) \subseteq \mathcal{D}$ of size N, drawn with replacement. We write $c_j(t)$ for the number of times $\mathbf{h}_j$ is drawn into $B(t)$, $S_i = \{t : \mathbf{h}_i \notin B(t)\}$ and $\bar{S}_i = \{t : \mathbf{h}_i \in B(t)\}$ for the sets of trees in which $\mathbf{h}_i$ is out-of-bag and in-bag, respectively, and $J_i(t)$ for the in-bag points that fall in the same leaf as $\mathbf{h}_i$ in tree t. The leaf holding $\mathbf{h}_i$ in tree t contains $|M_i(t)| = \sum_j c_j(t)\, \mathbb{1}(\mathbf{h}_j \in J_i(t))$ in-bag points in total, counting each $\mathbf{h}_j$ as many times as it is drawn into $B(t)$. For every ordered pair $(i, j) \in [N] \times [N]$, we define RF-GAP proximity as

$$p_{\mathrm{GAP}}(i,j) = \begin{cases} \dfrac{1}{|\bar{S}_i|} \displaystyle\sum_{t \in \bar{S}_i} \frac{c_i(t)}{|M_i(t)|}, & i = j, \\[2ex] \dfrac{1}{|S_i|} \displaystyle\sum_{t \in S_i} \frac{c_j(t)\, \mathbb{1}(\mathbf{h}_j \in J_i(t))}{|M_i(t)|}, & i \neq j. \end{cases} \tag{3}$$

Restricting the off-diagonal sum to out-of-bag trees makes RF-GAP proximities geometry- and accuracy-preserving, in the sense that a proximity-weighted nearest-neighbor predictor reproduces the random forest's out-of-bag predictions (Rhodes et al., 2023b).

Next, we collect pairwise embedding proximities into a matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ with entries $\mathbf{W}_{ij} = p_{\mathrm{GAP}}(i, j)$ for $i, j \in [N]$. We make it symmetric, $\frac{1}{2}(\mathbf{W} + \mathbf{W}^\top)$, and row-normalize the result into a row-stochastic diffusion operator $\mathbf{P} \in \mathbb{R}^{N \times N}$ that defines a Markov chain over $\mathcal{D}$. Thus, each entry $\mathbf{P}_{ij}$ is the one-step transition probability from $\mathbf{h}_i$ to $\mathbf{h}_j$, and each row $\mathbf{P}_i$ captures the local geometry around $\mathbf{h}_i$.

Finally, following RF-PHATE (Rhodes et al., 2023a), we diffuse the operator $\mathbf{P}$ by powering it to a diffusion time t, which accumulates every length-t transition path and so integrates local proximities into global structure (Coifman and Lafon, 2006). We choose t large enough to smooth away local noise yet small enough to preserve the manifold's structure, deferring the selection criterion to Section C. We apply the potential transform

$U_t = -\log \mathbf{P}^t$ elementwise, defining the *potential distance* $V_t(i,j) = \|\mathbf{u}_i - \mathbf{u}_j\|_2$ between its rows $\mathbf{u}_i$ and $\mathbf{u}_j$ and embed these distances into $\mathbb{R}^d$ by metric multidimensional scaling, yielding supervised target coordinates $\mathbf{z}_i^G \in \mathbb{R}^d$ (Moon et al., 2019).

Landmarks The potential-distance MDS over all N training points dominates the cost of building the supervised targets, scaling as $O(N^3)$ in time and $O(N^2)$ in memory. To scale to large N, we follow PHATE (Moon et al., 2019) and compute the diffusion geometry on $M \ll N$ landmarks, embedding the landmarks with MDS and extending the coordinates to the remaining points through their transition probabilities to the landmarks. This lowers the dominant cost to $O(M^3 + NM)$ time and $O(NM)$ memory.

Geometry-regularized autoencoder To extend supervised target coordinates $\mathbf{z}_i^G$ out-of-sample, i.e., beyond the training dataset, we follow RF-AE (Aumon et al., 2025) and train an autoencoder that reconstructs transition vectors $p_i \in \Delta^{N-1}$ encoding the supervised geometry. The encoder $\mathrm{Enc}_\theta(p_i) = \mathbf{z}_i \in \mathbb{R}^d$ compresses a transition vector $p_i \in \Delta^{N-1}$, and the decoder $\mathrm{Dec}_\phi(\mathbf{z}_i) = \hat{p}_i \in \Delta^{N-1}$ returns a transition distribution through a softmax output layer, while a geometric term anchors the bottleneck to the precomputed RF-PHATE coordinates (Duque et al., 2023). We fit (θ, ϕ) to minimize

$$\mathcal{L}(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}\Big[(1-\lambda)\, D_{\mathrm{KL}}(p_i \,\|\, \hat{p}_i) + \lambda \,\|\mathbf{z}_i - \mathbf{z}_i^G\|_2^2\Big], \tag{4}$$

where the reconstruction term is the Kullback–Leibler divergence between the transition distributions p_i and $\hat{p}_i$, and the geometric term regresses $\mathbf{z}_i$ onto its RF-PHATE target $\mathbf{z}_i^G$. The weight $\lambda \in [0,1]$ interpolates between a plain autoencoder ($\lambda = 0$) and pure regression onto the RF-PHATE embedding ($\lambda = 1$).

2.3. Decode

The *Decode* stage maps latent representations $\mathbf{z}$ to a downstream task-specific output.

Prediction and retrieval For prediction and retrieval, we probe $\mathbf{z}$ by k-NN or fitting a linear prediction head on the latent representations, mapping a molecule to its target $\hat{y}$.

Generation The autoencoder decoder Dec_ϕ maps a latent representation to a reconstructed proximity vector $\hat{p} = \mathrm{Dec}_\phi(\mathbf{z}) \in \mathbb{R}^N$; a learned map $m_\xi \colon \mathbb{R}^N \to \mathbb{R}^D$ then lifts these proximities to the frozen embedding space, $\hat{\mathbf{h}} = m_\xi(\hat{p})$, and the frozen transformer's own decoder maps $\hat{\mathbf{h}}$ back to a sequence in $\mathcal{X}$.

Latent space interpolation via neural ODEs To explore continuous interpolation in the latent space, we define an ordinary differential equation (ODE) over the diffusion coordinates $\boldsymbol{z}(t) \in \mathbb{R}^d$

$$\frac{d\boldsymbol{z}(t)}{dt} = v_\psi(\boldsymbol{z}(t), t),$$

where $v_\psi \colon \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a learnable velocity field parameterized by a neural network with weights ψ, trained with Sinkhorn loss (Cuturi, 2013; Genevay et al., 2018). Solving the above ODE produces smooth interpolation paths $\{\boldsymbol{z}(t) \mid t \in [0,1]\}$ in the latent space, enabling transformation of latent representations from one region to the other. Specifically,

at each training iteration, we sample a mini-batch $\{\boldsymbol{z}_m(0)\}_{m=1}^M$ from one region in the latent space and a target mini-batch $\{\boldsymbol{z}_m(1)\}_{m=1}^M$ from the other. Next, we compute trajectory points $\boldsymbol{z}_m(t)$ via the ODE solver and evaluate the accelerated Sinkhorn Loss $\mathcal{L}_{\text{Sinkhorn}}$, with $\varepsilon = 0.1$ regularization for stable gradients, given by

$$\mathcal{L}_{\text{Sinkhorn}} = \text{OT}_\varepsilon(\mu, \nu) \; - \; \tfrac{1}{2}\Big[\text{OT}_\varepsilon(\mu, \mu) + \text{OT}_\varepsilon(\nu, \nu)\Big], \tag{5}$$

where OT_ε is the entropy-regularized optimal transport term

$$\text{OT}_\varepsilon(\mu, \nu) = \min_{P \in \Pi(\mu,\nu)} \left[\sum_{i,j} P_{ij} \, \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 + \varepsilon \, \text{KL}(P \| \mu \otimes \nu)\right], \tag{6}$$

where μ, ν are respectively the source and target marginal probability distributions (in our case, uniform distributions over the sample trajectory points $\{\boldsymbol{z}_m(0)\}$ or the targets $\{\boldsymbol{z}_m(1)\}$ respectively), $\{\boldsymbol{z}_i\} \sim \mu$ and $\{\boldsymbol{z}_j\} \sim \nu$ are samples from the two probability distributions, P is a matrix representing a transport plan subject to marginal constraints defined by Π, and $\text{KL}(P \| \mu \otimes \nu)$ denotes the KL-divergence of P with respect to the product measure $\mu \otimes \nu$. Finally, we jointly optimize the neural network parameters by backpropagating through both the ODE integration (adjoint sensitivity) and the Sinkhorn computation. Once trained, latent codes $\boldsymbol{z}_m(t)$ are passed through our decoding pipeline, yielding novel peptide sequences.

2.4. Inference

At inference, we route a molecule's frozen embedding $\mathbf{h} = f(\mathbf{x})$ through the fitted random forest, applying Equation (3) with every tree being out-of-bag, to obtain a transition vector $\mathbf{p} \in \Delta^{N-1}$ that the encoder maps to a latent representation $\mathbf{z} = \text{Enc}_\theta(\mathbf{p})$. The task-specific head then decodes $\mathbf{z}$ to a desired output.

3. Experiments

To demonstrate the utility of adapting pretrained transformer embeddings to downstream tasks in antimicrobial peptide (AMP) discovery, we benchmark FDD on property prediction (Section 3.1 and 3.2), peptide retrieval (Section 3.3), and latent-space interpolation (Section 3.4). Our results indicate that FDD yields task-aware representations that improve over the pretrained embeddings across all three settings. Additionally, we report all experimental details, ablations and additional results in Appendix C, D and E, respectively.

3.1. FDD outperforms pretrained embeddings on property prediction

To evaluate whether the task-aware representations produced by FDD improve downstream prediction, we consider binary antimicrobial activity classification across *ESKAPE* bacterial species from Pirtskhalava et al. (2021) (see Appendix A). We compare embeddings extracted from two pretrained backbones: Hyformer (Izdebski et al., 2025) (33M parameters) and ESM-2 (Lin et al., 2023) (35M parameters). For each backbone we probe the pretrained and FDD embeddings with a linear classifier; end-to-end fine-tuning of the backbone serves as an

Table 1: Predictive performance on antimicrobial activity classification across ESKAPE bacterial species. Among non-fine-tuned methods, the best embedding per species is marked **bold**; fine-tuning is reported separately as an upper-bound reference. Mean (std) across 10 seeds; pretrained is deterministic.

	Bacterial Species, AUC (↑)					
Method	*A. baumannii*	*E. coli*	*E. faecium*	*K. pneumoniae*	*P. aeruginosa*	*S. aureus*
Hyformer						
Pretrained	0.843	0.882	0.800	0.826	0.842	0.823
AE	0.780 (0.006)	0.826 (0.004)	0.752 (0.016)	0.774 (0.009)	0.747 (0.003)	0.764 (0.005)
FDD (ours)	**0.871 (0.013)**	**0.864 (0.005)**	**0.834 (0.014)**	**0.828 (0.010)**	**0.859 (0.004)**	**0.832 (0.005)**
Fine-tuning	0.859 (0.005)	0.885 (0.005)	0.821 (0.019)	0.841 (0.014)	0.854 (0.010)	0.846 (0.004)
ESM-2						
Pretrained	0.828	0.865	0.750	0.799	0.818	0.809
AE	0.802 (0.012)	0.816 (0.003)	0.803 (0.020)	0.753 (0.002)	0.744 (0.004)	0.766 (0.001)
FDD (ours)	**0.852 (0.008)**	**0.866 (0.006)**	**0.825 (0.016)**	**0.829 (0.012)**	**0.856 (0.003)**	**0.837 (0.007)**
Fine-tuning	0.840 (0.003)	0.898 (0.006)	0.832 (0.009)	0.853 (0.002)	0.856 (0.010)	0.854 (0.005)

upper-bound reference, and a vanilla autoencoder (AE) as an unsupervised dimensionality-reduction control.

FDD improves the predictive performance of raw pretrained embeddings for both Hyformer and ESM-2 (Table 1), raising the average AUC from 0.836 to 0.848 for Hyformer and from 0.812 to 0.844 for ESM-2, with a per-species gain in every case except for Hyformer on *E. coli*. In contrast, unsupervised autoencoder (AE) degrades performance across all species and backbones (down to 0.774 and 0.781 on average), showing that embedding compression without supervised diffusion geometry may discard useful structure. Fine-tuning attains the highest average AUC overall (0.851 for Hyformer, 0.856 for ESM-2), but FDD recovers most of this gap without updating the backbone. On a per-species basis FDD matches or exceeds fine-tuning on three of six species for Hyformer (*A. baumannii*, *E. faecium*, *P. aeruginosa*) and on (*A. baumannii* and *P. aeruginosa*) for ESM-2. Overall, FDD extracts task-relevant structure from frozen pretrained embeddings, recovering most of the fine-tuning expressivity.

3.2. FDD adapts without inflating intrinsic dimensionality

Intrinsic dimensionality (ID) measures the effective number of dimensions a representation uses, the dimension of the manifold on which embeddings concentrate and typically well below the ambient width D. It is a useful proxy for adaptation cost. Lower terminal-layer ID correlates with better generalization (Ansuini et al., 2019; Valeriani et al., 2023), while fine-tuning typically realigns the backbone by recruiting additional directions, inflating ID, precisely the kind of added complexity that drives overfitting in the low-data AMP regime. An ideal adaptation improves task alignment while keeping the manifold compact.

We estimate local ID via the participation ratio of the local PCA spectrum over active peptides (MIC $< 32\,\mu$M), in three per-species spaces: frozen pretrained Hyformer (PT), the

FDD projection of PT, and fine-tuned Hyformer (FT). Full estimator and subset details are in Appendix B.

Table 2: Estimated local intrinsic dimensionality (ID) across ESKAPE bacterial species. Avg. is the cross-species mean.

	Bacterial Species, ID (↓)						
Method	*A. baumannii*	*E. coli*	*E. faecium*	*K. pneumoniae*	*P. aeruginosa*	*S. aureus*	Avg.
Pretrained	6.10	6.14	6.25	6.05	6.18	6.33	6.18
FDD (ours)	5.62	6.44	4.83	5.72	6.67	6.46	5.96
Fine-tuning	7.43	8.24	6.57	7.04	7.84	8.28	7.57

Fine-tuning raises the average ID from 6.18 (Pretrained) to 7.57, inflating it across all species. FDD instead holds ID near the pretrained manifold (average 5.96) and below FT in all six cases. FDD thus improves downstream predictive performance without the dimensional inflation of fine-tuning, consistent with a compact, task-aware reorganization of the frozen geometry rather than a wholesale expansion of it.

3.3. FDD retrieves novel active peptides

We test whether FDD embeddings capture biological similarity well enough to retrieve novel actives. We fit a kNN classifier (k=5) on FDD embeddings of the Veltri et al. (2018) dataset and use it to rank peptides in the held-out classification set of Szymczak et al. (2023), none seen during training. We compare against state-of-the-art AMP classifiers, a random forest on physicochemical descriptors (RF), HydrAMP (Szymczak et al., 2023), AMPScanner (Yang et al., 2024), and AMPlify (Li et al., 2022) on the same held-out set, reporting retrieval precision P_{amp} (↑).

Table 3: AMP retrieval performance.

Method	P_{amp} (↑)
RF	0.80
HydrAMP	0.83
AMPScanner	0.89
AMPlify	0.89
FDD (ours)	**0.90**

FDD attains the highest P_{amp} (0.90; Table 3), on par with the strongest baselines while operating on representations not trained for classification. The embedding space separates active from inactive peptides (Figures 1 and 4), supporting accurate nearest-neighbor retrieval of novel actives.

3.4. FDD enables smooth latent-space interpolation

Finally, we investigate whether FDD enables smooth latent-space interpolation between AMP and non-AMP peptides. For this, we use the dataset from Veltri et al. (2018) and map all peptides into the FDD latent representation space. Specifically, we select a non-AMP latent vector $\mathbf{z}_0$ and integrate a continuous trajectory from $\mathbf{z}_0$ towards the AMP manifold using a Neural ODE (Chen et al., 2019) trained with Sinkhorn loss (Marino and Gerolin, 2020) to align the evolving distribution with that of active peptides (Figure 2). Intermediate points along the trajectory are decoded back to peptide sequences via a learned MLP mapping, enabling analysis of physicochemical descriptor changes along the path.

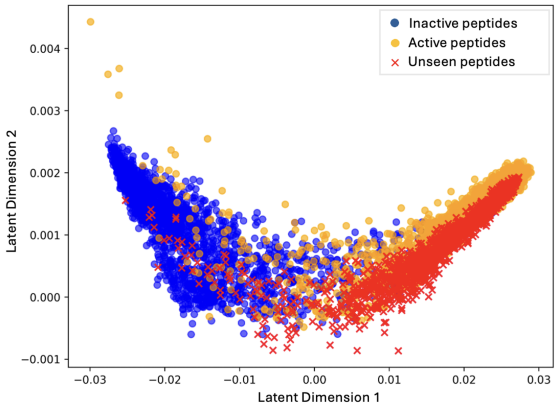

Figure 1: FDD for AMP retrieval. Red points are held-held-out active peptide embeddings projected into the FDD latent space.

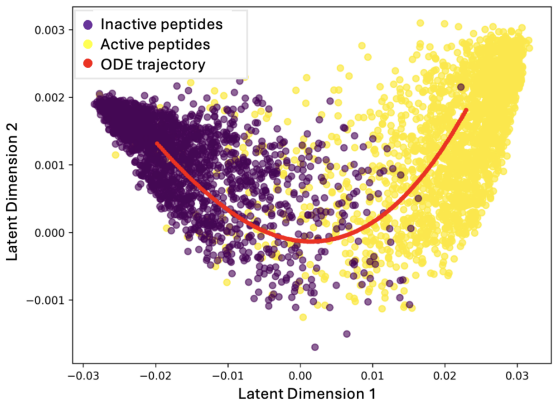

Figure 2: Interpolated trajectory in the FDD latent space between non-AMP and AMP regions.

Peptides decoded along the learned interpolation path exhibit a smooth and monotonic transition in physicochemical descriptors that correlate with antimicrobial properties of peptides (Figure 3). As an example, net charge increases from near-neutral to cationic (Figure 3 (a)), isoelectric point rises accordingly (Figure 3 (b)), and the hydrophobic amino-acid ratio grows, reflecting stronger membrane affinity typical of decoded AMPs (Figure 3 (c)). These results indicate that the FDD latent space encodes a biologically meaningful manifold, where interpolation traces a gradual acquisition of antimicrobial properties.

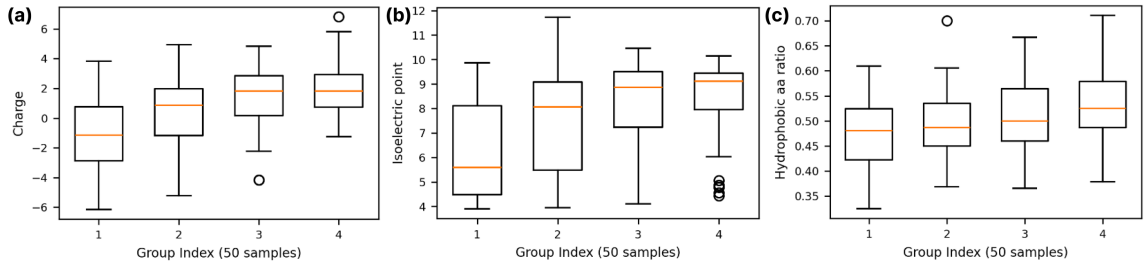

Figure 3: Evolution of physicochemical descriptors along the latent interpolation path: **(a)** net charge, **(b)** isoelectric point, and **(c)** hydrophobic amino-acid ratio. All descriptors steadily increase across the interpolation path, indicating a smooth transition from non-AMP to AMP-like regions in latent space.

4. Related Work

We relate FDD to prior work on adapting pretrained embeddings and on diffusion-geometric representation learning.

Adapting pretrained embeddings Two paradigms dominate transfer from pretrained embeddings. Fine-tuning updates the backbone and is expressive, but reshapes the pretrained geometry, degrades out-of-distribution robustness, and risks catastrophic forgetting (Kumar et al., 2022; Andreassen et al., 2021; Rajaee and Pilehvar, 2021). Parameter-

efficient variants such as adapters (Houlsby et al., 2019) and LoRA (Hu et al., 2021) reduce cost but still modify the backbone. Probing instead freezes the backbone and fits a shallow predictor on top, preserving the geometry but lacking the expressivity for nonlinear, task-specific structure (Belinkov, 2022; Hewitt and Manning, 2019; White et al., 2021). FDD keeps the backbone frozen and adapts its embeddings through supervised diffusion, recovering fine-tuning expressivity.

Diffusion-geometric representation learning Diffusion Maps (Coifman and Lafon, 2006) build a diffusion operator from sample affinities and extract coordinates that preserve global structure while denoising local neighborhoods. PHATE (Moon et al., 2019) extends this to a diffusion-potential representation stabilizing local and global structure. Supervised variants inject task information into the kernel. RF-GAP and RF-PHATE (Rhodes et al., 2023b,a) derive label-aware affinities from a random forest, biasing the manifold toward task-relevant directions, and random forest autoencoders (RF-AE) (Aumon et al., 2025) instantiate the geometry-regularized autoencoder framework (Duque et al., 2023) with a parametric encoder that extends the geometry to unseen samples.

5. Conclusions

We introduced Freeze, Diffuse, Decode (FDD), a framework based on diffusion geometry that adapts pretrained transformer embeddings to downstream tasks without backbone fine-tuning. We showed that FDD yields low-dimensional, predictive, and interpretable representations that improve over the pretrained embeddings on property prediction, retrieval and latent-space interpolation. As building the diffusion geometry scales cubically in the number of training points, large datasets require a landmark approximation, making FDD best suited to the low-data regimes typical of antimicrobial peptide design.

Acknowledgments This project has received funding from the European Research Council (ERC) under the European Funding Union's Horizon 2020 research and innovation programme (grant agreement No 810115–DOG-AMP) [E.S.]; a FRQNT doctoral scholarship and Mitacs-Globalink travel award [M.L.]; a Canada CIFAR AI chair (via Mila) and Humboldt research fellowship [G.W.]; NSF grant 221232 [K.M.]; and an IVADO visiting scholar award [K.M.].

Disclosure of Interests Projects in Szczurek lab at the University of Warsaw are co-funded by Merck KGaA [E.S.].Part of this work was supported by the Helmholtz International Lab Causal Cell Dynamics (InterLabs-0029) - Grant support from the Initiative and Networking Fund of the Hermann von Helmholtz-Association Deutscher Forschungszentren e.V.[G.W.]

References

Jakub Adamczyk, Piotr Ludynia, and Wojciech Czech. Molecular fingerprints are strong models for peptide function prediction, 2025. URL https://arxiv.org/abs/2501.17901.

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning, 2021. URL https://arxiv.org/abs/2106.15831.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks, 2019. URL https://arxiv.org/abs/1905.12784.

Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*, 399(10325):629–655, February 2022.

Adrien Aumon, Shuang Ni, Myriam Lizotte, Guy Wolf, Kevin R. Moon, and Jake S. Rhodes. Random forest autoencoders for guided representation learning, 2025. URL https://arxiv.org/abs/2502.13257.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations, 2019. URL https://arxiv.org/abs/1806.07366.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013. URL https://arxiv.org/abs/1306.0895.

Andrés F. Duque, Sacha Morin, Guy Wolf, and Kevin R. Moon. Extendable and invertible manifold learning with geometry regularized autoencoders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7381–7394, 2023. doi: 10.1109/TPAMI.2022.3222104.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262. URL https://www.biorxiv.org/content/early/2017/11/12/214262.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84

of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/genevay18a.html.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4129–4138, 2019. doi: 10.18653/v1/N19-1419.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. URL https://arxiv.org/abs/1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Adam Izdebski, Jan Olszewski, Pankhil Gawade, Krzysztof Koras, Serra Korkmaz, Valentin Rauscher, Jakub M. Tomczak, and Ewa Szczurek. Synergistic benefits of joint molecule generation and property prediction, 2025. URL https://arxiv.org/abs/2504.16559.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Chenkai Li, Darcy Sutherland, S Austin Hammond, Chen Yang, Figali Taho, Lauren Bergman, Simon Houston, René L Warren, Titus Wong, Linda M N Hoang, Caroline E Cameron, Caren C Helbing, and Inanc Birol. AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics*, 23(1):77, January 2022.

Zeming Lin, Halil Akin, Roshan Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *bioRxiv*, 2023.

Simone Di Marino and Augusto Gerolin. Optimal transport losses and sinkhorn algorithm with general convex regularization, 2020. URL https://arxiv.org/abs/2007.00976.

Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12): 1482–1492, 2019. doi: 10.1038/s41587-019-0336-3.

J. O'Neill. Tackling drug-resistant infections globally: final report and recommendations. page 84 pp., 2016.

Malak Pirtskhalava, Anthony A Amstrong, Maia Grigolava, Mindia Chubinidze, Evgenia Alimbarashvili, Boris Vishnepolsky, Andrei Gabrielian, Alex Rosenthal, Darrell E Hurt, and Michael Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids*

Research, 49(D1):D288–D297, 01 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa991. URL https://doi.org/10.1093/nar/gkaa991.

Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning, 2025. URL https://arxiv.org/abs/2508.06199.

Sara Rajaee and Mohammad Taher Pilehvar. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Jake S. Rhodes, Adrien Aumon, Sacha Morin, Marc Girard, Catherine Larochelle, Boaz Lahav, Elsa Brunet-Ratnasingham, Wei Zhang, Adele Cutler, Anhong Zhou, Daniel E. Kaufmann, Stephanie Zandee, Alexandre Prat, Guy Wolf, and Kevin R. Moon. Gaining biological insights through supervised data visualization. *bioRxiv*, 2023a. doi: 10.1101/2023.11.22.568384.

Jake S. Rhodes, Adele Cutler, and Kevin R. Moon. Geometry- and accuracy-preserving random forest proximities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10947–10959, 2023b. doi: 10.1109/TPAMI.2023.3263774.

Louis B Rice. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *J. Infect. Dis.*, 197(8):1079–1081, April 2008.

Simon Steshin. Lo-hi: Practical ml drug discovery benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Paulina Szymczak, Marcin Możejko, Tomasz Grzegorzek, Radosław Jurczak, Marta Bauer, Damian Neubauer, Karol Sikora, Michał Michalski, Jacek Sroka, Piotr Setny, Wojciech Kamysz, and Ewa Szczurek. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat. Commun.*, 14(1):1453, March 2023.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models, 2023. URL https://arxiv.org/abs/2302.00294.

Derek van Tilborg, Helena Brinkmann, Emanuele Criscuolo, Luke Rossen, Rıza Özçelik, and Francesca Grisoni. Deep learning for low-data drug discovery: Hurdles and opportunities. *Current Opinion in Structural Biology*, 86:102818, 2024. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2024.102818. URL https://www.sciencedirect.com/science/article/pii/S0959440X24000459.

Daniel Veltri, Uday Kamath, and Amarda Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16):2740–2747, August 2018.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 132–138, 2021. doi: 10.18653/v1/2021.naacl-main.12.

Cheng-Hong Yang, Yi-Ling Chen, Tin-Ho Cheung, and Li-Yeh Chuang. Multi-objective optimization accelerates the de novo design of antimicrobial peptide for staphylococcus aureus. *International Journal of Molecular Sciences*, 25:13688, 12 2024. doi: 10.3390/ijms252413688.

Gengmo Zhou, Yu Luo, Yi Li, et al. Uni-mol: A universal 3d molecular representation learning framework. *Nature Communications*, 14:4074, 2023.

Appendix A. ESKAPE Dataset Details

Antimicrobial Resistance (AMR) is among the most pressing threats to global health. As per (Antimicrobial Resistance Collaborators, 2022), AMR was responsible for an estimated 1.27 million deaths and associated with a further 4.95 million, exceeding the toll of HIV/AIDS or malaria. AMR-related infections are estimated to kill around 10 million people annually by the year 2050 (O'Neill, 2016). This burden is compounded by a stagnant discovery pipeline, with few mechanistically novel antibiotic classes reaching the clinic in decades while resistance to existing drugs continues to spread.

A disproportionate share of this burden traces to a small group of organisms. The acronym **ESKAPE** *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* denotes pathogens that most readily *escape* the action of available antibiotics through multidrug resistance (Rice, 2008). These species dominate hospital-acquired infections and a large fraction of attributable AMR mortality: in the 2019 estimates, six leading pathogens (largely overlapping with this group) accounted for roughly 73% of deaths directly attributable to resistance (Antimicrobial Resistance Collaborators, 2022), and they remain top priorities for new antibacterials. We evaluate antimicrobial activity on this panel, with one adjustment reflecting the mortality ranking: we substitute *Escherichia coli*, the single largest contributor to AMR-attributable deaths (Antimicrobial Resistance Collaborators, 2022), for *Enterobacter* spp., yielding six clinically prioritized targets spanning Gram-positive (*E. faecium*, *S. aureus*) and Gram-negative (*K. pneumoniae*, *A. baumannii*, *P. aeruginosa*, *E. coli*) organisms. Antimicrobial peptides offer a promising modality against these pathogens, but designing them requires predicting potency measured as the minimum inhibitory concentration (MIC) from sequence, and labeled MIC data are scarce on a per-species basis, precisely the low-data regime our method targets. Each peptide is labeled as *active* if its measured MIC is below $32\,\mu$M. Sequences are split 80/20 (train/test) using stratified sampling (fixed seed), with no duplicate sequences present in the data. The full dataset comprises **14,843** unique peptides (11,872 train / 2,971 test across all species).

Table 4: Per-species statistics for the ESKAPE MIC dataset. N is the number of unique peptides. Peptides with MIC $< 32\,\mu$M are labeled active.

Species	N	N_{tr}	N_{te}	**Active tr/te**	**% Act.**
E. faecium	293	234	59	161 / 40	69%
S. aureus	4,264	3,411	853	1,985 / 497	58%
K. pneumoniae	1,384	1,107	277	611 / 153	55%
A. baumannii	1,177	941	236	656 / 165	70%
P. aeruginosa	3,101	2,480	621	1,393 / 349	56%
E. coli	4,624	3,699	925	2,440 / 610	66%
Total	**14,843**	**11,872**	**2,971**	–	–

Appendix B. Local Intrinsic Dimensionality

A representation that resolves a task with fewer effective dimensions exposes a simpler, smoother manifold along which to interpolate, retrieve neighbors, and fit a head from few

labels. ID quantifies the geometric cost of adaptation: fine-tuning recruits additional directions to realign the backbone, and in the low-data AMP regime this added complexity is what drives overfitting.

For each species we restrict to *active* peptides (MIC $< 32\,\mu$M, y=1) after removing duplicate sequences, so that ID characterizes the geometry of the functionally relevant region of the embedding space. We estimate ID locally via the participation ratio (PR; Gao et al. (2017)) of the local PCA spectrum, a continuous effective-rank measure. For each active point x_i we take its k=20 nearest neighbors under exact (ℓ_2, brute-force) distance within the active subset, center them by their mean to form $N_i \in \mathbb{R}^{k\times d}$, and compute the local covariance

$$C_i = \frac{1}{k-1} N_i^\top N_i.$$

The local effective dimensionality is

$$\mathrm{PR}_i = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2},$$

which equals 1 for a strictly one-dimensional structure and d for a perfectly isotropic d-dimensional cloud, where λ_j are the eigenvalues of C_i. The ID of an embedding space is the mean over active points, $\widehat{\mathrm{ID}} = \frac{1}{n_{\mathrm{act}}} \sum_{i=1}^{n_{\mathrm{act}}} \mathrm{PR}_i$.

Embedding spaces ID is computed in three spaces per species:

- **PT**: mean-pooled frozen pretrained Hyformer embeddings;
- **FDD**: the FDD projection of the PT embeddings;
- **FT**: mean-pooled embeddings from the fine-tuned Hyformer.

Validity of the local scale PR estimates *local, linear* dimensionality, the effective rank of the tangent-space covariance, and is bounded above by $k-1=19$. The reported values ($\sim$5–8) sit well inside this range, so k is large enough to resolve the relevant scale while remaining local enough to approximate the tangent space.

Appendix C. Experimental Details

This appendix collects the hyperparameters and implementation choices deferred from Section 2.

Random forest and proximities We fit a random forest with 100 trees, unlimited maximum depth, and a minimum of 1 sample per leaf to the frozen training embeddings $\{(\mathbf{h}_n, y_n)\}$ without feature standardization (random forests are invariant to monotone input transformations). RF-GAP proximities (Equation (3)) are symmetrized as $\frac{1}{2}(\mathbf{W}+\mathbf{W}^\top)$ and row-normalized into the diffusion operator $\mathbf{P}$.

RF-PHATE targets We raise $\mathbf{P}$ to diffusion time $t = \texttt{auto}$, selected by the von Neumann entropy criterion of PHATE (Moon et al., 2019), apply the potential transform, and embed the potential distances into $\mathbb{R}^d$ with $d \in \{16, 32, 64\}$ (selected by cross-validated AUC; $d = 32$ fixed for all visualizations) using landmark metric MDS with an SGD solver. Up to $L = \min(n_{\mathrm{train}} - 1,\, 2000)$ landmarks are used to make the diffusion graph tractable; the diffusion operator includes a teleportation term with $\beta = 0.9$.

Geometry-regularized autoencoder The encoder and decoder are 3-layer multilayer perceptrons of widths $[800, 400, 100]$ with ELU activations; the decoder ends in a log-softmax over the L landmark proximities. We sweep the loss weight $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ in Equation (4) (cross-validated; $\lambda = 10^{-2}$ is typically best) and optimize with AdamW (learning rate 10^{-3}, weight decay 10^{-5}, batch size 64, 500 epochs).

Latent-space interpolation The neural-ODE velocity field v_ψ is a 2-layer network trained with the Sinkhorn divergence at regularization $\varepsilon = 0.1$, integrated with `rk4` and the adjoint method. We give the full entropic optimal-transport objective in Section 2.3.

C.1. Embedding Extraction Details

All the embeddings are extracted from the frozen backbone. For sequences of length T, we pool the final-layer per-token hidden states $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T]$ using a binary attention mask $m_t \in \{0, 1\}$ that excludes padding. We use two pooling operators:

$$\text{CLS:} \quad \mathbf{h}_0, \qquad \text{Mean:} \quad \bar{\mathbf{h}} = \frac{\sum_t m_t \, \mathbf{h}_t}{\sum_t m_t}.$$

Hyformer (HYF) Embeddings We load a pretrained Hyformer: an 8-layer transformer with d=512, 8 attention heads, and an amino-acid vocabulary of 34 tokens. Sequences are tokenized with the ESM amino-acid tokenizer (padded or truncated to 50 tokens) and passed through the backbone to obtain hidden states $\mathbf{H} \in \mathbb{R}^{T \times 512}$. We derive two representations: **PT-CLS**, the hidden state $\mathbf{h}_0 \in \mathbb{R}^{512}$ at position 0 (the task/CLS token), and **PT-Mean**, the mask-weighted mean $\bar{\mathbf{h}} \in \mathbb{R}^{512}$.

ESM2 Embeddings We load ESM2-35M (12 layers, d=480) from Hugging Face. Sequences are tokenized with the ESM2 tokenizer (padded or truncated to 128 tokens, sufficient for antimicrobial peptides) and passed through the backbone. To mirror the Hyformer PT-Mean protocol, we use mean pooling only, yielding $\bar{\mathbf{h}} \in \mathbb{R}^{480}$ over the final-layer hidden states.

Table 5: Frozen-embedding extraction configuration.

Backbone	Layers	d	Max tokens	Pooling	Batch
Hyformer-33M	8	512	50	CLS, Mean	64
ESM2-35M	12	480	128	CLS, Mean	16

Downstream classifiers In both cases, the frozen embeddings serve as input to a logistic regression classifier with ℓ_2 regularization, fit on the 80 % training split. The regularization strength C is selected via 5-fold cross-validation over a log-uniform grid $[10^{-4}, 10^4]$. Features are standardized (zero mean, unit variance) before fitting, with the scaler fitted on training data only. For the fine-tuning conditions (**FT**), the full model backbone (Hyformer or ESM2) plus a linear head is end-to-end trained using AdamW (lr=10^{-4}, β=$(0.9, 0.95)$, cosine schedule with 10 % linear warm-up, early stopping on validation AUC with patience equal to 15).

Appendix D. Ablations

FDD latent dimension sensitivity All reported results use a fixed latent dimension d=32 and regularization λ=10^{-2} (Table 1); this section varies each in turn to assess sensitivity. We sweep $d \in \{2, 8, 16, 32, 64, 128\}$ at fixed λ=10^{-2} on frozen Hyformer mean embeddings (d_{in}=512); results are mean (std) over 5 seeds (Table 6).

Performance is flat in d above a small threshold: for the five larger species it saturates by d=16, with the fixed default d=32 within one standard deviation of the per-species best in every case (e.g. *A. baumannii* and *P. aeruginosa* peak exactly at d=32; *S. aureus*, *E. coli*, and *K. pneumoniae* differ by ≤0.003 AUC). The sole exception is *E. faecium*, our smallest dataset (N_{tr}=234), which peaks at d=8 (0.853 vs. 0.834 at d=32); the optimal latent dimension shrinks with the available data, consistent with the low-data regime FDD targets. Performance drops sharply only at d=2, indicating FDD needs a handful of dimensions but no more. We therefore fix d=32 as a robust default across species.

Table 6: FDD dimensionality sweep on Hyformer embeddings across ESKAPE bacterial species (ROC-AUC). The best d per species is marked **bold**. Mean (std) across 5 seeds; λ=10^{-2} fixed throughout.

Species	d=2	d=8	d=16	d=32	d=64	d=128
A. baumannii	0.696 (0.016)	0.846 (0.015)	0.870 (0.011)	**0.871 (0.013)**	0.869 (0.017)	0.866 (0.013)
E. faecium	0.848 (0.029)	**0.853 (0.013)**	0.843 (0.012)	0.834 (0.014)	0.833 (0.017)	0.836 (0.018)
K. pneumoniae	0.785 (0.002)	0.821 (0.009)	0.829 (0.007)	0.828 (0.010)	0.828 (0.010)	**0.830 (0.011)**
P. aeruginosa	0.689 (0.006)	0.843 (0.004)	0.852 (0.003)	**0.859 (0.004)**	0.857 (0.004)	0.853 (0.006)
S. aureus	0.717 (0.003)	0.817 (0.003)	**0.835 (0.006)**	0.832 (0.009)	0.832 (0.008)	0.830 (0.006)
E. coli	0.763 (0.010)	0.858 (0.008)	**0.868 (0.005)**	0.867 (0.004)	0.867 (0.008)	0.866 (0.005)

FDD geometric regularization sensitivity Holding latent dimensionality d=32 fixed, we sweep the geometric regularization hyperparameter $\lambda \in \{0.0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ (Table 7; PT is the frozen 512-d Hyformer baseline). FDD improves over the frozen baseline on five of six species, the exception being *E. coli* (PT 0.882 vs. FDD 0.867), where the raw embeddings are already highly discriminative (consistent with Table 1). Sensitivity to λ is uniformly low: the full grid moves AUC by ≤0.015 for every species, and typically by ≤0.005, so any $\lambda \in [10^{-3}, 10^{-1}]$ performs within noise. We fix λ=10^{-2} throughout.

Appendix E. Additional Results

E.1. FDD representations remain competitive on out-of-distribution molecular property prediction

To evaluate whether our framework generalizes beyond antimicrobial peptides, we benchmark FDD on out-of-distribution molecular property prediction tasks from (Steshin, 2023). We probe frozen embeddings of Hyformer with a k-nearest neighbors (kNN) classifier. Following Steshin (2023), we compare FDD against molecular fingerprints (ECFP4, MACCS), frozen Hyformer embeddings, and a label-frequency dummy baseline. Full end-to-end fine-tuning serves as an empirical upper bound.

Figure 4: UMAP of CLS embeddings per ESKAPE bacterial species, colored by antimicrobial activity (red: active, MIC $< 32\,\mu$M; blue: inactive). For each species we show frozen pretrained embeddings and their FDD projection (32-d), for both backbones. Each panel is an independent UMAP fit. Only the within-panel arrangement of active vs. inactive points is meaningful, not positions across panels.

Table 7: FDD ROC-AUC across five regularization strengths λ at fixed latent dimensionality d=32, evaluated within each ESKAPE bacterial species. PT is the frozen Hyformer mean-embedding baseline. The best λ per species is marked **bold**. Mean (std) across 5 seeds.

Species	**PT**	λ=0	λ=10^{-3}	λ=10^{-2}	λ=10^{-1}	λ=1	**Best λ**
A. baumannii	0.843	0.841 (0.010)	0.856 (0.011)	**0.871 (0.013)**	0.869 (0.014)	0.863 (0.009)	10^{-2}
E. faecium	0.800	0.799 (0.035)	**0.842 (0.021)**	0.834 (0.014)	0.828 (0.019)	0.833 (0.023)	10^{-3}
K. pneumoniae	0.826	0.806 (0.009)	0.824 (0.012)	0.828 (0.010)	**0.830 (0.009)**	0.828 (0.013)	10^{-1}
P. aeruginosa	0.842	0.818 (0.004)	0.854 (0.008)	**0.859 (0.004)**	0.856 (0.003)	0.853 (0.003)	10^{-2}
S. aureus	0.823	0.816 (0.002)	0.831 (0.010)	0.832 (0.009)	**0.835 (0.007)**	0.835 (0.009)	10^{-1}
E. coli	0.882	0.864 (0.004)	0.866 (0.009)	0.867 (0.004)	**0.870 (0.005)**	0.868 (0.002)	10^{-1}

Table 8: kNN classification performance on molecular property prediction. Among kNN methods, the best result per dataset is marked **bold**. Fine-tuning is reported separately as an upper-bound reference. Mean (std) across three seeds.

	DATASET, AUPRC (↑)			
METHOD	DRD2-HI	HIV-HI	KDR-HI	SOL-HI
DUMMY BASELINE	0.677 (0.061)	0.040 (0.014)	0.609 (0.081)	0.215 (0.008)
ECFP4	0.706 (0.047)	0.067 (0.029)	0.646 (0.048)	0.426 (0.022)
MACCS	0.702 (0.042)	**0.072 (0.036)**	0.610 (0.072)	0.422 (0.009)
HYFORMER	0.700 (0.057)	0.052 (0.018)	0.646 (0.071)	0.484 (0.040)
FDD (OURS)	**0.747 (0.040)**	0.049 (0.016)	**0.663 (0.058)**	**0.494 (0.054)**
FINE-TUNING	0.784 (0.082)	0.158 (0.128)	0.701 (0.022)	0.640 (0.036)

The empirical results demonstrate that FDD yields robust, competitive molecular representations, establishing the top frozen-feature performance on three of the four tasks. Broadly, these findings confirm that FDD remains competitive with established baselines outside the primary peptide domain.

E.2. Sample Efficiency Under Label Scarcity

Fine-tuning attains the best full-data AUC (Table 1), which invites the question of why one would adapt frozen embeddings rather than fine-tune. The answer is that fine-tuning is unreliable when labels are scarce. For each ESKAPE species we subsample the training split to $N \in \{20, 50, 100\}$ with balanced classes (fixed test split, 10 seeds) and compare frozen pretrained embeddings (PT), FDD, and end-to-end fine-tuning (FT) under an identical linear probe; we test the FT-versus-PT gap with a one-sided Wilcoxon signed-rank test over paired seeds (Figure 5).

In every species FT is the weakest method at small N and converges to the frozen methods only by $N = 100$, trailing by up to $\approx$0.14 AUC at $N = 20$ (*E. faecium*); the Wilcoxon test confirms FT $<$ PT at low-N points, though fine-tuning's large variance at $N = 20$ leaves several gaps short of significance. FDD tracks PT throughout, and in five of six species the two are statistically indistinguishable. The exception is *E. faecium*, our

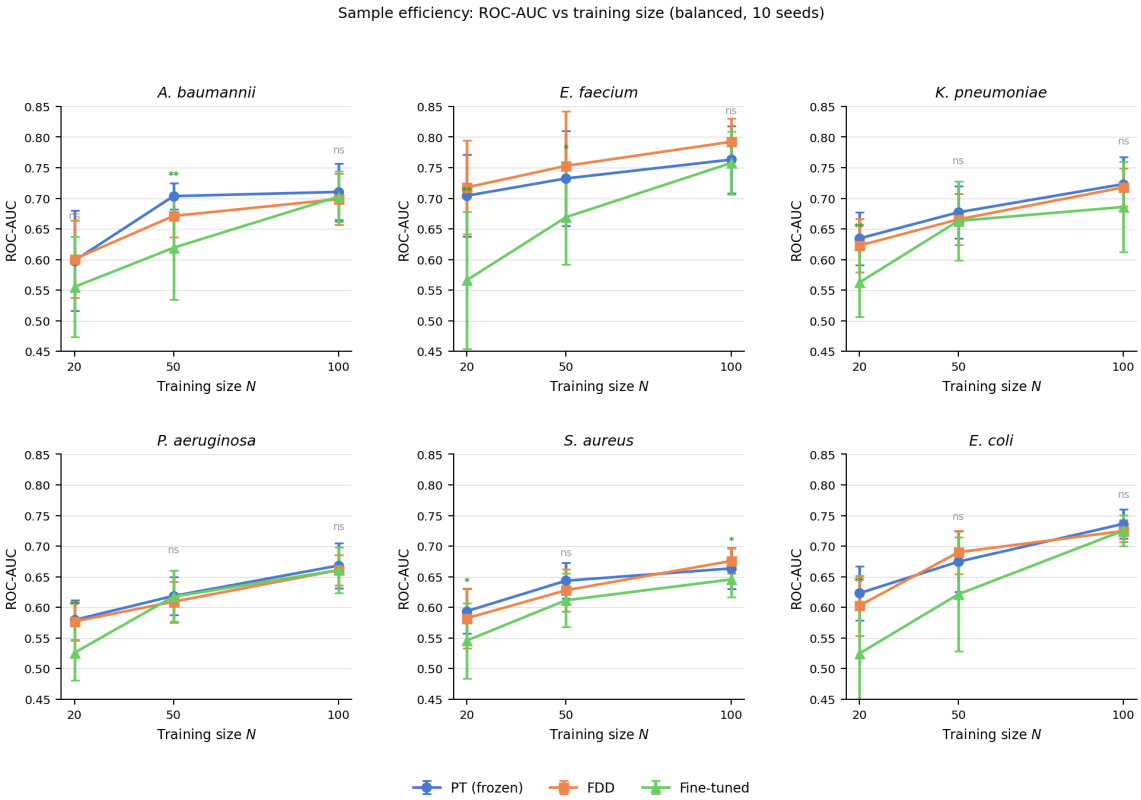

Figure 5: Sample efficiency across ESKAPE bacterial species. Mean ROC-AUC versus balanced training size N over 10 seeds (error bars ± 1 s.d., shared axes). Markers report a Wilcoxon test of FT vs. PT ($^{*}\,p<0.05$, $^{**}\,p<0.01$, $^{***}\,p<0.001$, ns = not significant).

smallest set, where FDD separates from both, and its margin over PT widens with N, matching Table 1.

This isolates one point: the method that is strongest at full data is the least reliable under scarcity, motivating frozen adaptation. It is *not* evidence that FDD beats frozen embeddings at small N, where they are comparable; FDD's edge is a full-data, geometric effect (Tables 1, 2), achieved while matching the frozen baseline's sample efficiency.