METALORIAN: *De Novo* GENERATION OF HEAVY METAL-BINDING PEPTIDES WITH CLASSIFIER-GUIDED DIFFUSION SAMPLING

Yinuo Zhang,[1,3] **Divya Srijay,**[1] **Pranam Chatterjee**[1,2,†]

[1]Department of Biomedical Engineering, Duke University
[2]Department of Computer Science, Duke University
[3]Centre for Computational Biology, Duke-NUS Medical School, Singapore

[†]Corresponding author: pranam.chatterjee@duke.edu

ABSTRACT

We present **Metalorian**, a conditional diffusion model tailored to generate *de novo* heavy metal-binding peptides. Our approach leverages the embedding space of MetaLATTE, a multi-label classifier fine-tuned on known metal-binding sequences, to guide the generation of peptides with specific metal-binding capabilities. The model utilizes a co-evolving diffusion framework that simultaneously handles continuous protein embeddings and discrete metal-binding properties, allowing for focused generation of shorter, economically-viable peptides. We demonstrate the effectiveness of our approach by generating peptide binders for copper, cadmium, and cobalt binding. Our results show that the generated peptides maintain key properties such as charge and hydrophobicity while significantly reducing sequence length and molecular weight compared to known metal-binding proteins. Co-folding and binding energy analysis using molecular dynamics further validate the potential binding capacities of these novel sequences. Finally, we experimentally demonstrate that Metalorian-generated peptides effectively bind to cobalt resin via phage display. Overall, our work solidifies a foundational platform for designing heavy metal-binding peptides for targeted bioremediation campaigns, and further motivates utilization of well-trained, continuous latent spaces for diffusion-based *de novo* peptide design.

1 MEANINGFULNESS STATEMENT

We focused on bioremediation as an underexplored application of protein design due to the complexity of metal ion binding both *in vitro* and *in silico*. Our work demonstrates how deep protein representations can uncover essential metal-binding motifs hidden within larger proteins, extracting only the critical components needed for ion chelation. By standing on the shoulders of well-trained protein language models and extending them to this environmentally crucial domain, Metalorian creates meaningful biological representations by distilling nature's metal-binding strategies into minimal, functional peptides that can address pressing environmental challenges.

2 INTRODUCTION

Metals are fundamental to biological systems, serving as enzyme co-factors and catalysts essential for cellular functions. While elements like zinc, copper, and iron are critical for maintaining biochemical processes, others such as cadmium, lead, and mercury pose toxicity risks at elevated concentrations Kostenkova et al. (2022); Klug (2010); Tchounwou et al. (2012); Chen et al. (2023). Understanding and manipulating metal-binding properties are therefore crucial for applications in environmental remediation and pharmaceutical development Dixit et al. (2015).

Recent advances in protein language models (pLMs) have transformed our ability to extract functional insights directly from protein sequences without relying on structural data Lin et al. (2023); Elnaggar et al. (2021). Building on this progress, MetaLATTE, a multi-label classifier fine-tuned on ESM-2 embeddings, predicts metal-binding probabilities with high accuracy, utilizing contrastive and focal loss functions to address the challenges of imbalanced datasets and multi-label classification Zhang et al. (2024); Permyakov (2021). While MetaLATTE excels in metal-binding prediction Zhang et al. (2024), the growing demand extends beyond prediction to the actual generation of novel metal-binding biomolecules.

Peptides, due to their structural simplicity, functional versatility, and favorable biochemical properties, emerge as ideal candidates for metal-binding applications. Their smaller size (180–5000 Da), diverse functional groups, and ease of synthesis make them highly suitable for selective metal chelation, environmental remediation, and therapeutic interventions Luo et al. (2024); Akbarian et al. (2022). To address this need, we introduce **Metalorian**, a co-evolving conditional diffusion model designed to generate metal-binding peptides de novo. Metalorian leverages the metal-binding knowledge embedded within MetaLATTE's fine-tuned representation space to produce peptides with specific metal affinities.

Our model integrates two interacting diffusion processes: a continuous diffusion model operating in the ESM embedding space and a discrete diffusion model capturing metal-binding labels. Contrastive learning between these components ensures that generated sequences retain strong specificity to their target metal-binding properties Lee et al. (2023); Schroff et al. (2015). To experimentally validate Metalorian's predictions, we developed a novel phage assay, demonstrating successful cobalt binding using Metalorian-generated peptides.

The key innovations of our approach include:

1. Adaptation of the co-evolving diffusion framework Lee et al. (2023) to protein sequence generation, enabling simultaneous handling of continuous metal-sensitive protein embeddings and discrete metal-binding properties.
2. Integration of contrastive learning Lee et al. (2023); Schroff et al. (2015) in both continuous and discrete diffusion processes to enhance binding specificity.
3. Development of a novel phage assay to experimentally validate metal-binding activity, demonstrated through successful cobalt binding with Metalorian-generated peptides.

Through this methodology, we aim to generate *de novo*, shorter, and economically viable peptides with specific metal-binding capabilities, including those targeting underrepresented and toxic metals, using protein sequence information alone.

3 METHODS

3.1 DIFFUSION MODEL

Metalorian Our model, Metalorian, adapts the co-evolving conditional diffusion framework Lee et al. (2023) to protein sequence generation by incorporating both continuous (protein embeddings) and discrete (metal-binding labels) components (Figure S1). The model consists of two interacting diffusion processes:

Continuous Diffusion Model (DiffusionProteinModel) We operate continuous diffusion model on ESM-2 protein embeddings ($x_0^C \in \mathbb{R}^{B\times L\times 1280}$) that was fine-tuned with metal-binding proteins Zhang et al. (2024). During diffusion model training, the ESM-2 backbone have the last 10 layers unfrozen. The forward process follows:

$$q(x_t^C|x_0^C) = \mathcal{N}(x_t^C; \sqrt{\bar{\alpha}_t}x_0^C, (1-\bar{\alpha}_t)I) \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$.
The reverse process is parameterized by ϵ_{θ_C} and conditioned on discrete variables:

$$p_{\theta_C}(x_{t-1}^C|x_t^C, x_t^D) = \mathcal{N}(x_{t-1}^C; \mu_{\theta_C}(x_t^C, t|x_t^D), \sigma_t^2 I) \tag{2}$$

Discrete Diffusion Model (MultinomialDiffusion) We operate discrete diffusion model on metal-binding labels ($x_0^D \in \mathbb{R}^{B\times 15}$) using a multinomial diffusion process. The model uses TabularUnet as its backbone denoising network. The forward process uses categorical distributions:

$$q(x_t^D|x_{t-1}^D) = \mathcal{C}(x_t^D; (1-\beta_t)x_{t-1}^D + \beta_t/K) \tag{3}$$

where K is the number of classes.
The reverse process is conditioned on continuous embeddings:

$$p_{\theta_D}(x_{t-1}^D|x_t^D, x_t^C) = \sum q(x_{t-1}^D|x_t^D, x_0^D)p_{\theta_D}(x_0^D|x_t^D, x_t^C) \tag{4}$$

The two models are trained jointly with combined loss:

$$L_{\text{total}} = \underbrace{L_{\text{diff}_C}(\theta_C) + \lambda_C L_{\text{CL}_C}(\theta_C)}_{\text{continuous}} + \underbrace{L_{\text{diff}_D}(\theta_D) + \lambda_D L_{\text{CL}_D}(\theta_D)}_{\text{discrete}} \tag{5}$$

where L_{diff_C}, L_{diff_D} are diffusion losses for continuous and discrete components; L_{CL_C}, L_{CL_D} are contrastive learning losses with negative sampling; λ_C, λ_D are weighting coefficients.

The contrastive learning loss uses a triplet formulation with positive and negative samples:

$$L_{\text{CL}} = \sum_{i=0}^{S} \max\{d(A_i, P_i) - d(A_i, N_i) + m, 0\} \tag{6}$$

where A is the anchor, P is a positive sample, N is a negative sample, d is a distance metric, m is the margin, and S is the number of samples. This approach encourages the model to learn the true correlation between continuous embeddings and discrete labels while being robust to mismatched conditions.
For positive sample generation, we generate $\hat{x}_0^{C+}$ conditioned on matching discrete label x_t^D and $\hat{x}_0^{D+}$ conditioned on matching continuous embedding x_t^C. For negative sample generation, within a minibatch, we generate $\hat{x}_0^{C-}$ using mismatched discrete label x_t^{D-} and $\hat{x}_0^{D-}$ using mismatched continuous embedding x_t^{C-}.

3.2 SAMPLING

Progressive Verification Sampling We propose two complementary sampling approaches for protein sequence generation with metal-binding properties. The first approach, Progressive Verification Sampling (Algorithm 1), is designed for well-represented metal classes in our training data. During phase one (steps T to T_c), it follows standard diffusion sampling:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) \tag{7}$$

In phase two (steps T_c to 0), we introduce a verification mechanism that ensures label alignment through multiple sampling attempts. At each timestep, we evaluate both predictor probability $P(\mathbf{x}_t)[y_{\text{target}}]$ from the MetaLATTE classification model and discrete label alignment $\arg\max(\mathbf{x}_t^D) = y_{\text{target}}$. Sampling continues until success criteria are met: predictor confidence exceeds threshold τ and labels align. At each timestep, the continuous and discrete models inform each other's sampling process:

$$\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t = f_\theta^C(\mathbf{x}_t^C|t, \mathbf{x}_t^D) \tag{8}$$

$$\mathbf{x}_{t-1}^D = f_\theta^D(\mathbf{x}_t^D|t, \mathbf{x}_{t-1}^C) \tag{9}$$

where f_θ^C conditions the continuous sampling on the current discrete state, and f_θ^D updates the discrete state based on both previous discrete state and new continuous sample.

Gradient-Guided Sampling For metal classes with limited training examples or complex binding patterns, we introduce Gradient-Guided Label Sampling (Algorithm 2). This approach extends classifier guidance Dhariwal & Nichol (2021) with dynamic scaling Dinh et al. (2024) and label alignment. At each timestep t, we compute:

$$s = \exp(P(\mathbf{x}_t)[y_{\text{target}}]) - \beta \sum_{i \neq y_{\text{target}}} \exp(P(\mathbf{x}_t)[i]) \tag{10}$$

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t + \lambda_t \nabla_{\mathbf{x}_t} s, \boldsymbol{\sigma}_t) \tag{11}$$

where λ_t doubles the base guidance scale λ when the predictor's confidence for the target class falls below threshold τ. This co-evolution of continuous and discrete states, combined with gradient guidance, ensures that both sequence generation and label prediction remain consistent throughout the sampling process. The key difference here is that in gradient-guided sampling, this cross-model interaction works alongside the gradient guidance, while in progressive verification sampling it works with the verification mechanism.

3.3 EVALUATION

We assessed both the physicochemical properties and structural stability of our generated metal-binding peptides. Initial characterization using Biopython package (v1.78) Cock et al. (2009) and molecular dynamics system was prepared using Ambertools24 Case et al. (2023) and MCPB.py Li & Merz Jr (2016) for specific metal-protein interactions. Other details can be found in (Appendix A.2).

3.4 EXPERIMENTAL SCREENING PLATFORM

Phagemids were generated using a phage display vector from Dr. Anoop Patel's lab at Duke University. The vector was digested with SapI (NEB), and binder sequences were cloned via Gibson Assembly (NEB). Recombinant constructs were electroporated into TG1 electrocompetent cells (Biosearch Technologies), grown in 2XYT medium (Sigma-Aldrich) with 2% glucose and 100 μg/mL ampicillin at 37°C overnight, then infected with M13KO7 helper phage (NEB) and incubated at 37°C for 1 hour. Infected cells were pelleted, resuspended in 2XYT with 100 μg/mL ampicillin and kanamycin, and incubated at 30°C for 16 hours. Phage particles were precipitated with PEG buffer (20% PEG, 2.5 M NaCl), filtered (0.45 μm), and titrated by infecting TG1 cells, plating on LB agar with ampicillin, and incubating at 30°C for 16 hours. Samples were Sanger sequenced (Genewiz), and remaining phage was stored in 20% glycerol at -80°C. For phage elution, 1×10^{10} phage were incubated with HisPur™ Cobalt Resin (ThermoFisher) in PBS at 4°C for 30 minutes, washed with TBST (0.5% Tween 20), and eluted with 0.5 M EDTA (pH 8.0, ThermoFisher). The eluate was used to infect TG1 cells, plated on LB agar with ampicillin, incubated at 30°C for 16 hours, and colonies were counted to assess phage enrichment. A schematic for our experimental screening pipeline can be found in Figure S2.

4 RESULTS

4.1 *In silico* GENERATION OF HEAVY METAL BINDERS

Using our Metalorian diffusion model guided by the MetaLATTE multi-label metal binding classifier Zhang et al. (2024), we generated metal-binding peptide sequences with lengths between 30 and 80 residues. This length range was chosen to ensure sequences remain within the peptide range suitable for *in silico* analysis while maintaining sufficient length for stable folding. We focused our generation on copper, zinc, and cobalt binding peptides due to their significance in environmental remediation and the availability of metal ion binding simulations.

The comparison between known metal-binding proteins from the MbPA database Li et al. (2023) and our generated peptides revealed successful optimization of several key properties (Figure 1). Our sequences demonstrated reduced length and molecular weight while maintaining similar charge distributions to natural proteins. The slight increase in hydrophobicity scores in our generated sequences likely reflects the increased presence of cysteine residues, which although classified as hydrophobic in standard scales, are polar amino acids crucial for metal binding (Figure S3). Analysis of sequence diversity through Shannon entropy (Figure 1) showed that our generated sequences exhibit lower Shannon entropy compared to known metal-binding proteins, indicating reduced randomness and diversity. This indicates that our model has captured essential amino acid patterns characteristic of metal-binding proteins, potentially concentrating on residues more actively involved in metal coordination with shorter lengths. Furthermore, our observations aligns with those in metallothioneins, where specific amino acids like cysteines and histidines predominantly determine binding specificity and affinity Calatayud et al. (2021); Permyakov (2021).

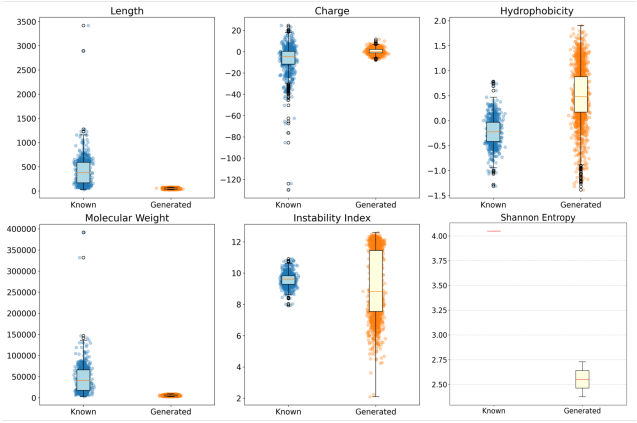


Figure 1: Distribution comparison of physical properties. Length, charge, hydrophobicity, molecular weights, instability index, and Shannon's entropy for real copper, cadmium, cobalt ion binding proteins from the training dataset compared to peptides generated via Metalorian.

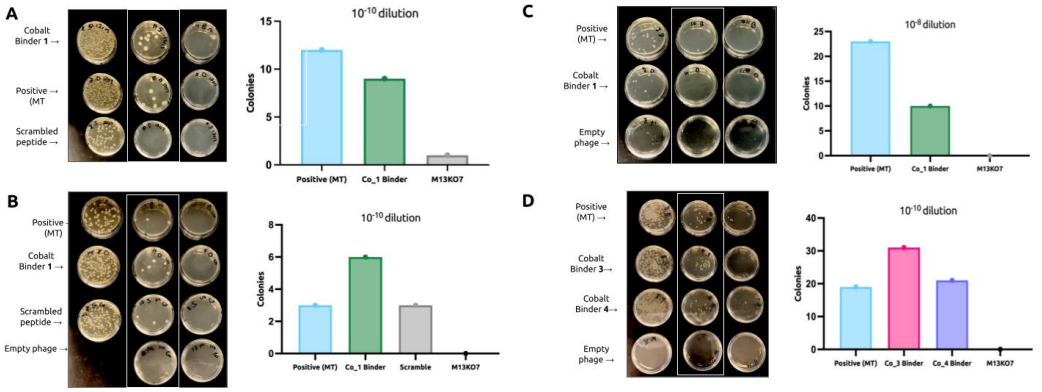


Figure 2: *In vitro* validation of cobalt binding. A-C) **Co_1 binding**: Enrichment between positive (MT) control and binder compared to negative controls in eluted phage. D) **Additional Co Binders:** Enrichment of additonal binders, Co_3 and Co_4.

To validate the structural stability and metal-binding capability of our generated sequences, we performed detailed molecular dynamics simulations. Our analysis of the Cu^{2+}-binding and Cd^{2+}-binding peptides (Figure S4 S5) revealed stable structural characteristics with low backbone RMSD values and well-maintained coordination geometry. The binding site analysis confirmed the design's stability in holding the metal ions. Free energy landscape analysis further confirmed a well-defined global minimum state with low RMSD and radius of gyration, indicating stable conformational preferences.

4.2 *In vitro* EVALUATION OF HEAVY METAL BINDERS

While our computational analysis initially focused on Cu and Cd to establish the generalizability of our approach across different transition metals, we sought to validate our generative pipeline with an independent experimental system (Figure S2). Given practical considerations of phage display experiment, we focused our *in vitro* validation on cobalt-binding peptides. Our phage selection experiments confirmed that the Metalorian-generated sequences successfully enriched for Co binders (Figure 2), reinforcing the effectiveness of our computationally designed sequences in capturing metal-specific binding properties. Future work will focus on establishing high throughput metal binder screens, extending to multi-metal binding scenarios, further refining our ability to generate custom metal-binding peptides with high specificity.

5 CONCLUSION

In this work, we present Metalorian, a conditional diffusion model that leverages the MetaLATTE classifier embedding space Zhang et al. (2024) to generate *de novo* metal-binding peptides. Through our co-evolving diffusion framework and contrastive learning approach, we successfully generated peptides with significantly reduced sequence lengths while maintaining essential physicochemical properties and metal-binding capabilities, as validated through molecular dynamics simulations. While current *in silico* tools like AlphaFold3 Abramson et al. (2024) have limitations in evaluating diverse metal-binding peptides, here, we have developed novel experimental validation methods, including high-throughput phage display, to directly assess binding affinities and specificities. Overall, our model provides a promising platform for designing selective metal-binding peptides, with potential applications in environmental remediation. Future work will focus on extending our experimental validation to additional heavy metals, evaluating multi-metal binding scenarios, and developing a user-friendly platform where researchers can specify target metals and desired peptide lengths for custom sequence generation.

ACKNOWLEDGEMENTS

We thank Lauren Hong for the Metalorian LOGO design. We thank Mark III Systems for providing computational support for this project. This work was funded by a Garden Grant from the Homeworld Collective.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Mohsen Akbarian, Ali Khani, Sara Eghbalpour, and Vladimir N Uversky. Bioactive peptides: Synthesis, sources, applications, and proposed mechanisms of action. *International journal of molecular sciences*, 23(3):1445, 2022.

Ariel A Aptekmann, J Buongiorno, D Giovannelli, M Glamoclija, Diego U Ferreiro, and Yana Bromberg. mebipred: identifying metal-binding potential in protein sequence. *Bioinformatics*, 38 (14):3532–3540, 2022.

Sara Calatayud, Mario Garcia-Risco, Veronika Pedrini-Martha, Douglas J Eernisse, Reinhard Dallinger, Òscar Palacios, Mercè Capdevila, and Ricard Albalat. Modularity in protein evolution: modular organization and de novo domain evolution in mollusk metallothioneins. *Molecular biology and evolution*, 38(2):424–436, 2021.

David A Case, Hasan Metin Aktulga, Kellon Belfon, David S Cerutti, G Andrés Cisneros, Vinícius Wilian D Cruzeiro, Negin Forouzesh, Timothy J Giese, Andreas W Gotz, Holger Gohlke, et al. Ambertools. *Journal of chemical information and modeling*, 63(20):6183–6191, 2023.

Xingqi Chen, Yuanchun Zhao, Yuqing Zhong, Jiajia Chen, and Xin Qi. Deciphering the functional roles of transporter proteins in subcellular metal transportation of plants. *Planta*, 258(1):17, 2023.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11): 1422, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL `https://arxiv.org/abs/2105.05233`.

Anh-Dung Dinh, Daochang Liu, and Chang Xu. Rethinking conditional diffusion sampling with progressive guidance. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruchita Dixit, X Wasiullah, Deepti Malaviya, Kuppusamy Pandiyan, Udai B Singh, Asha Sahu, Renu Shukla, Bhanu P Singh, Jai P Rai, Pawan Kumar Sharma, et al. Bioremediation of heavy metals from soil and aquatic environment: an overview of principles and criteria of fundamental processes. *Sustainability*, 7(2):2189–2212, 2015.

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

William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019. URL https://github.com/Lightning-AI/lightning.

Aaron Klug. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annual review of biochemistry*, 79:213–231, 2010.

Peter A Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, 33(12):889–897, 2000.

Kateryna Kostenkova, Gonzalo Scalese, Dinorah Gambino, and Debbie C. Crans. Highlighting the roles of transition metals and speciation in chemical biology. *Current Opinion in Chemical Biology*, 69:102155, August 2022. ISSN 1367-5931. doi: 10.1016/j.cbpa.2022.102155. URL http://dx.doi.org/10.1016/j.cbpa.2022.102155.

Chaejeong Lee, Jayoung Kim, and Noseong Park. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pp. 18940–18956. PMLR, 2023.

Jinzhao Li, Xiang He, Shuang Gao, Yuchao Liang, Zhi Qi, Qilemuge Xi, Yongchun Zuo, and Yongqiang Xing. The metal-binding protein atlas (mbpa): An integrated database for curating metalloproteins in all aspects. *Journal of Molecular Biology*, 435(14):168117, 2023.

Pengfei Li and Kenneth M Merz Jr. Mcpb. py: A python based metal center parameter builder, 2016.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Yingyong Luo, Yunfeng Zhang, Zhuang Xiong, Xiaodie Chen, Ajia Sha, Wenqi Xiao, Lianxin Peng, Liang Zou, Jialiang Han, and Qiang Li. Peptides used for heavy metal remediation: A promising approach. *International Journal of Molecular Sciences*, 25(12):6717, 2024.

Mary CM O'Brien and CC Chancey. The jahn–teller effect: An introduction and current review. *American Journal of Physics*, 61(8):688–697, 1993.

Eugene A Permyakov. Metal binding proteins. *Encyclopedia*, 1(1):261–292, 2021.

Daniel R Roe and Thomas E Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pp. 145–158. Springer, 2011.

Paul B Tchounwou, Clement G Yedjou, Anita K Patlolla, and Dwayne J Sutton. Heavy metal toxicity and the environment. *Molecular, clinical and environmental toxicology: volume 3: environmental toxicology*, pp. 133–164, 2012.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Yinuo Zhang, Phil He, Ashley Hsu, and Pranam Chatterjee. Metalatte: Metal binding prediction via multi-task learning on protein language model latents. *bioRxiv*, pp. 2024–06, 2024.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 DATA

Training data for metal-binding proteins was sourced from the MbPA database Li et al. (2023), focusing on 14 transition and heavy metals (Ag, Cd, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, Pt, V, W, Zn) with at least 6 samples per metal class. Non-binding proteins from the mebipred database Aptekmann et al. (2022) were included as negative samples to balance the dataset. The data was split into 80% training and 20% validation sets using balanced stratification Sechidis et al. (2011) to maintain label distribution across stages.

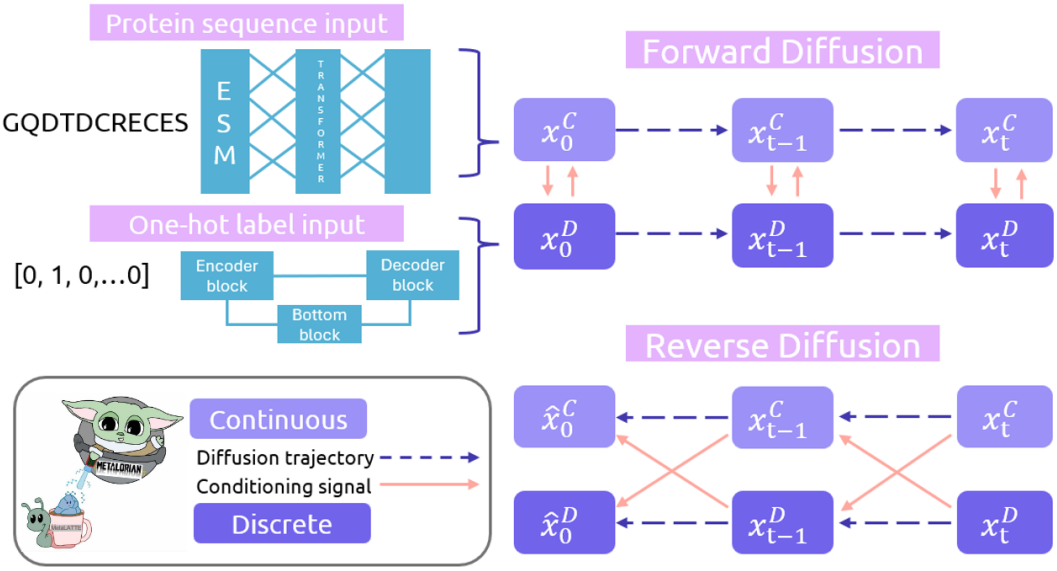


Figure S1: Metalorian Schematic.

A.2 EVALUATION

Initial characterization using Biopython package (v1.78) Cock et al. (2009) focused on key biophysical properties. For detailed structural and dynamics analysis, we performed molecular dynamics simulations on randomly selected sequences from our generated peptide pools. The simulation pipeline began with structural prediction using AlphaFold3 Abramson et al. (2024), followed by metal ion docking using Autodock VINA Trott & Olson (2010) to ensure consistent binding pose generation across different metal ions. The molecular dynamics system was prepared using Ambertools24 Case et al. (2023) and MCPB.py Li & Merz Jr (2016) for specific metal-protein interactions. Trajectory analysis was conducted using CPPTRAJ Roe & Cheatham III (2013), and binding stability was assessed through free energy calculations using MM-PBSA Kollman et al. (2000). Final structural visualization was performed using Pymol Schrödinger, LLC (2015) (v 3.1), where the residues in the protein targets with polar contacts to the peptide binder with distances closer than 2.8 Å are annotated.

A.3 MODEL ARCHITECTURES AND TRAINING DETAILS

We adapted the original Co-evolving Contrastive Diffusion Model with the original ESM model as the continuous diffusion part and tabular Unet for the discrete diffusion part. Metalorian was trained on in-house 7xA100 Nvidia GPUs. The model's configuration encompassed a batch size of 140 and a learning rate set at 2×10^{-4}. The AdamW optimizer Loshchilov & Hutter (2019) was employed for optimization. A length restriction of with a range of 30 to 80 is applied during training and sampling,

thus generated sequences are controlled within a range. The entire implementation and parallelization is performed with the PyTorch Lightning framework Falcon & The PyTorch Lightning team (2019). The specific hyperparameters of our model are given below.

Table S1: Co-evolving Conditional Diffusion Model Architecture

Hyperparameter	**Value**
ESM Model Base	`ESM2_t33_650M_UR50D`
Max Sequence Length	1280
Training Batch Size	140
Diffusion Steps (T)	50
Continuous Model (DiffusionProteinModel)	
Input Dimension	1280
Output Dimension	1280
Time Embedding Dimension	1280
Condition Projection	$15 \rightarrow 1280$
Output Projection	1280
Discrete Model (TabularUnet)	
Input Dimension	15
Condition Projection	$1280 \rightarrow 15$
Output Dimension	15
Encoder Dimensions	[64, 128, 256]
Decoder Dimensions	[256, 128, 64]
Output Layer	$64 \rightarrow 15$

A.4 Experimental Design

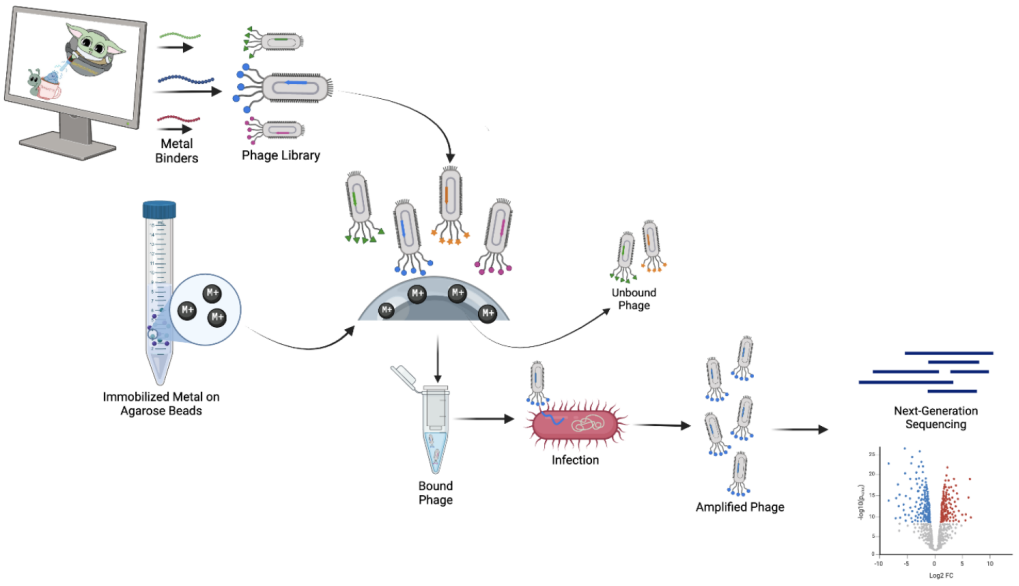


Figure S2: Schematic of the experimental pipeline for screening Metalorian-generated metal-binding peptides. Binder sequences were cloned into a phage display vector, expressed in TG1 cells, and displayed on phage particles. Phage libraries were incubated with immobilized metal on agarose beads, and bound phages were eluted, amplified, and analyzed using next-generation sequencing to assess binding enrichment.

A.5 Amino Acid Composition Analysis

Figure S3: Amino Acid Composition Analysis between the database and the generated peptides. UP: histograms of direct comparison of the frequency of amino acid distribution. Bottom: histograms of the discrepancy. Green indicates increase of the frequency, and red indicates the opposite.

A.6 Extended Molecular Dynamics Analysis of Cu^{2+} and Cd^{2+} Binding Peptides

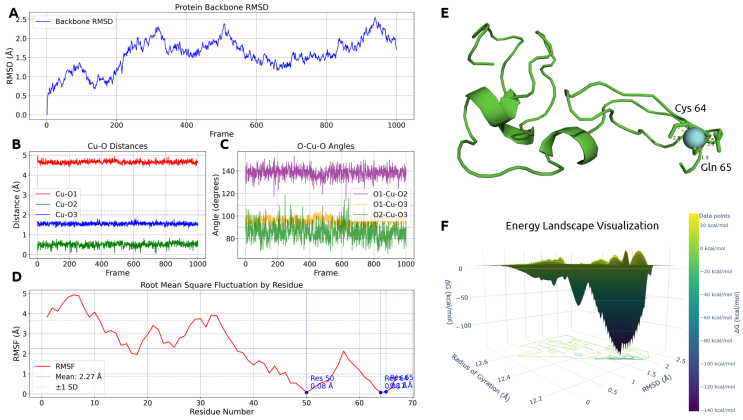


Figure S4: Structural stability and coordination geometry of the Cu^{2+} binding site analysis based on the trajectory from the molecular dynamics results. A) **Root Mean Square Deviation (RMSD)**: The average RMSD of the protein backbone was 1.577 ± 0.419 Å, reflecting global structural stability. B) **Cu-O Distances**: Distances between Cu^{2+} and its coordinating oxygen atoms (O1, O2, O3) are presented, with O3 showing the most consistent interaction at an average distance of 1.552 ± 0.073 Å. C) **O-Cu-O Angles**: Angles between Cu^{2+} and coordinating oxygen atoms reveal a distorted square planar geometry, consistent with Cu^{2+}'s coordination preferences O'Brien & Chancey (1993). D) **Root Mean Square Fluctuation (RMSF)**: The binding site residues (50, 64, and 65) showed minimal flexibility (<0.2 Å), indicating a rigid and stable coordination environment. E) **Pymol visualization of the examined peptide**. F) **Energy landscape visualization**. Free energy landscape of Cu^{2+} binding showing favorable interaction energy dominated by electrostatic and solvation effects, calculated from molecular dynamics simulations using direct interaction terms.

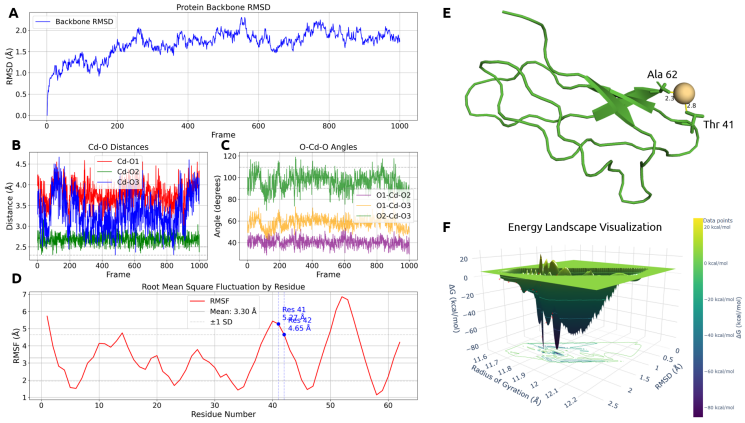


Figure S5: Structural stability and coordination geometry of the Cd^{2+} binding peptide based on molecular dynamics simulations. A) **Root Mean Square Deviation (RMSD)**: The average backbone RMSD was 1.686 ± 0.306 Å, demonstrating overall structural stability. B) **Cd-O Distances**: Coordination distances (Cd-O1: 3.792 ± 0.276 Å, Cd-O2: 2.690 ± 0.125 Å, Cd-O3: 3.331 ± 0.437 Å) suggest a dynamic metal-binding environment. C) **O-Cd-O Angles**: The coordination angles (O1-Cd-O2: 40.43° ± 4.00°, O1-Cd-O3: 58.11° ± 5.21°, O2-Cd-O3: 96.17° ± 8.08°) indicate flexibility in the binding geometry, with some distances exceeding typical Cd-O coordination bonds (2.2-2.5 Å). D) **Root Mean Square Fluctuation (RMSF)**: The binding site maintains overall stability despite slight flexibility in the coordination environment. E) **Molecular visualization**: PyMOL representation of the examined peptide. F) **Free energy landscape**: The energy landscape of Cd^{2+} binding shows a well-defined global minimum with favorable interaction energy dominated by electrostatic and solvation effects.

A.7 SAMPLING PSEUDOCODE

Algorithm 1 Progressive Verification Sampling

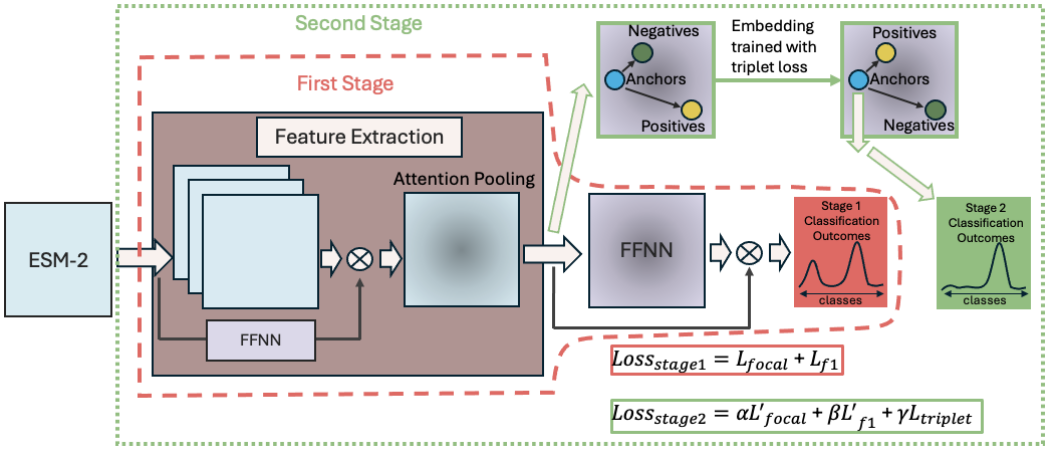
Require: Continuous model f_θ^C, Discrete model f_θ^D, Predictor P, Target label idx y_{target}, Control start step T_c, Total steps T, Success threshold τ

Initialize $\mathbf{x}_T^C \sim \mathcal{N}(0, 0.9\mathbf{I})$
Let (l_{min}, l_{max}) be sequence length bounds
$L \leftarrow \text{Uniform}(l_{min}, l_{max})_B$ ▷ *sample sequence lengths for batch*
$M \leftarrow [i < L_b]_{B \times T}$ ▷ *create length mask matrix*
$\mathbf{x}_T^D \leftarrow \log(\text{OneHot}(y_{\text{target}})) \in \mathbb{R}^{B \times C}$ ▷ *C-dim class embedding*
Phase 1: Standard diffusion from T to T_c
for $t = T, \ldots, T_c + 1$ **do**
  $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t \leftarrow f_\theta^C(\mathbf{x}_t^C, t, \mathbf{x}_t^D)$
  $\boldsymbol{\epsilon} \leftarrow \begin{cases} \mathcal{N}(0, \mathbf{I}) & \text{if } t > 0 \\ \mathbf{0} & \text{otherwise} \end{cases}$
  $\mathbf{x}_{t-1}^C \leftarrow (\boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}) \odot M$
  $\mathbf{x}_{t-1}^D \leftarrow f_\theta^D(\mathbf{x}_t^D, t, \mathbf{x}_{t-1}^C)$
end for
Phase 2: Controlled denoising from T_c to 0
$s_{\text{best}} \leftarrow -1$
$(\mathbf{x}_{\text{best}}^C, \mathbf{x}_{\text{best}}^D) \leftarrow$ None
for $t = T_c, \ldots, 0$ **do**
  for $k = 1, \ldots, K$ **do** ▷ *K attempts per timestep*
    $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t \leftarrow f_\theta^C(\mathbf{x}_t^C, t, \mathbf{x}_t^D)$
    $\boldsymbol{\epsilon} \leftarrow \begin{cases} \mathcal{N}(0, (1 - k/K)\mathbf{I}) & \text{if } t > 0 \\ \mathbf{0} & \text{otherwise} \end{cases}$ ▷ *noise schedule*
    $\mathbf{x}_{t-1}^{C*} \leftarrow (\boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}) \odot M$ ▷ *candidate continuous*
    $\mathbf{x}_{t-1}^{D*} \leftarrow f_\theta^D(\mathbf{x}_t^D, t, \mathbf{x}_{t-1}^{C*})$ ▷ *candidate discrete*
    $\mathbf{h} \leftarrow f_\theta^C\text{decode}(\mathbf{x}_{t-1}^{C*})$
    $\mathbf{h} \leftarrow \mathbf{h} \odot M$
    $\mathbf{z} \leftarrow \arg\max(\mathbf{h}, \text{dim} = -1)$
    $\mathbf{p}_P \leftarrow P(\mathbf{z})$ ▷ *predictor probabilities*
    $\mathbf{p}_D \leftarrow \exp(\mathbf{x}_{t-1}^{D*})$ ▷ *discrete probabilities*
    $s \leftarrow \frac{1}{2}(\mathbf{p}_P[y_{\text{target}}] + I[\arg\max(\mathbf{p}_D) = y_{\text{target}}])$ ▷ *score*
    if $s > s_{\text{best}}$ **then**
      $s_{\text{best}} \leftarrow s$
      $(\mathbf{x}_{\text{best}}^C, \mathbf{x}_{\text{best}}^D) \leftarrow (\mathbf{x}_{t-1}^{C*}, \mathbf{x}_{t-1}^{D*})$
    end if
    if $\arg\max(\mathbf{p}_P) = y_{\text{target}} \wedge \arg\max(\mathbf{p}_D) = y_{\text{target}} \wedge \mathbf{p}_P[y_{\text{target}}] > \tau$ **then**
      $(\mathbf{x}_t^C, \mathbf{x}_t^D) \leftarrow (\mathbf{x}_{t-1}^{C*}, \mathbf{x}_{t-1}^{D*})$ ▷ *successful candidate found*
      break
    end if
  end for
  if no successful candidate and $\mathbf{x}_{\text{best}}^C \neq$ None **then**
    $(\mathbf{x}_t^C, \mathbf{x}_t^D) \leftarrow (\mathbf{x}_{\text{best}}^C, \mathbf{x}_{\text{best}}^D)$
  end if
end for
return $\text{Decode}(\mathbf{x}_0^C), \mathbf{x}_0^D, P(\text{Decode}(\mathbf{x}_0^C))$

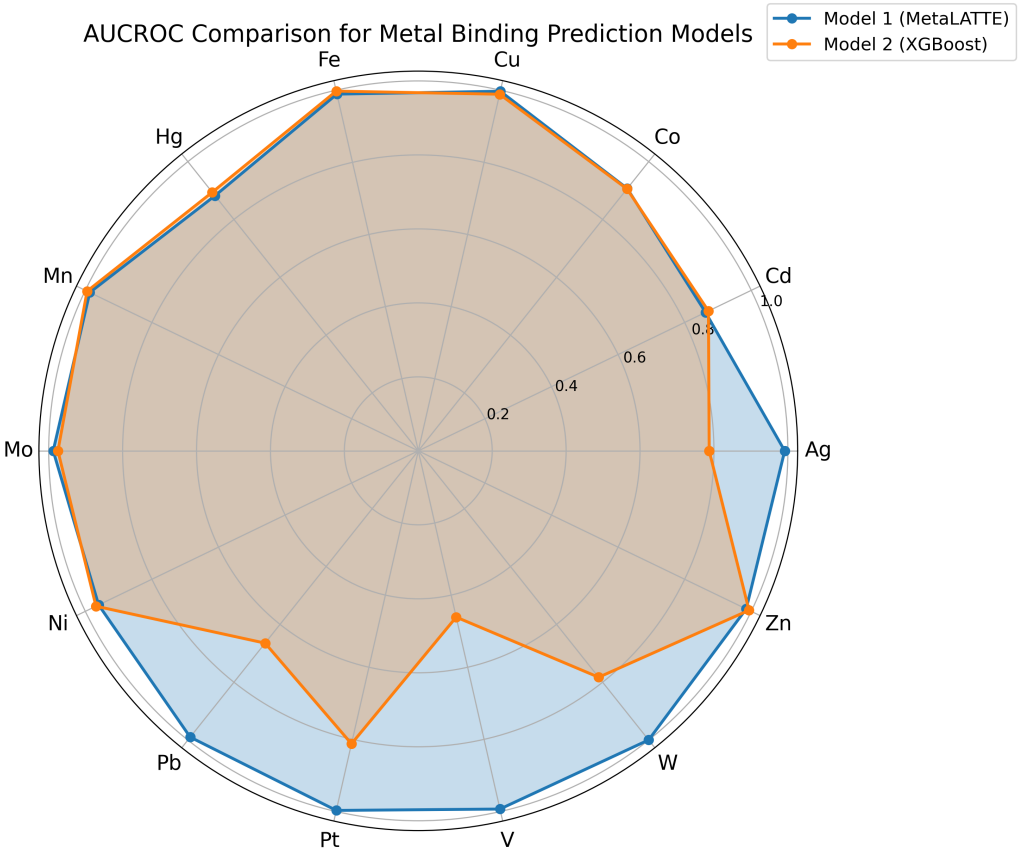
Algorithm 2 Gradient-Guided Label Sampling

Require: Continuous model f_θ^C, Discrete model f_θ^D, Predictor P, Target label idx y_{target}, Steps T, Guidance scale λ,

Initialize $\mathbf{x}_T^C \sim \mathcal{N}(0, 0.9\mathbf{I})$
$L \leftarrow \text{Uniform}(l_{\text{min}}, l_{\text{max}})_B$
$M \leftarrow [i < L_b]_{B \times T}$
$\mathbf{x}_T^D \leftarrow \log(\text{OneHot}(y_{\text{target}})) \in \mathbb{R}^{B \times C}$
$(\mathbf{x}_{\text{best}}^C, \mathbf{x}_{\text{best}}^D) \leftarrow (\mathbf{x}_T^C, \mathbf{x}_T^D)$
for $t = T, \ldots, 0$ **do**
  $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t \leftarrow f_\theta^C(\mathbf{x}_t^C, t, \mathbf{x}_t^D)$
  $\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_t \odot M$
  for $k = 1, \ldots, K$ **do** ▷ *K updates per timestep*
    $\mathbf{h} \leftarrow f_\theta^C \text{decode}(\mathbf{x}_t^C)$
    $\mathbf{h} \leftarrow \mathbf{h} \odot M$
    $\mathbf{z} \leftarrow \arg\max(\mathbf{h}, \text{dim} = -1)$
    $\mathbf{p} \leftarrow P(\mathbf{z})$
    $s \leftarrow \exp(\mathbf{p}[y_{\text{target}}]) - 0.1 \sum_{i \neq y_{\text{target}}} \exp(\mathbf{p}[i])$ ▷ *target score*
    $\nabla \leftarrow \frac{\partial s}{\partial \mathbf{x}_t^C}$
    $\lambda_t \leftarrow \lambda \cdot (2 \cdot I[\mathbf{p}[y_{\text{target}}] < 0.5])$
    $\mathbf{g} \leftarrow \lambda_t \cdot \nabla \odot M$ ▷ *scaled gradient*
    $\alpha \leftarrow 1 - t/T$
    $\sigma_{\text{noise}} \leftarrow 0.3(1-\alpha)^2$ ▷ *noise schedule*
    $\boldsymbol{\epsilon} \leftarrow \begin{cases} \mathcal{N}(0, \sigma_{\text{noise}}\mathbf{I}) & \text{if } t > 0 \\ \mathbf{0} & \text{otherwise} \end{cases}$
    $\mathbf{x}_t^C \leftarrow \boldsymbol{\mu}_t + \mathbf{g} + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}$
  end for
  $\mathbf{x}_t^D \leftarrow f_\theta^D(\mathbf{x}_t^D, t, \mathbf{x}_t^C)$
  $\mathbf{x}_t^D \leftarrow \alpha \mathbf{x}_t^D + (1-\alpha)\log(\text{OneHot}(y_{\text{target}}))$
  Update $(\mathbf{x}_{\text{best}}^C, \mathbf{x}_{\text{best}}^D)$ if improved based on $P(\text{Decode}(\mathbf{x}_t^C))[y_{\text{target}}]$
end for
return $\text{Decode}(\mathbf{x}_{\text{best}}^C), \mathbf{x}_{\text{best}}^D, P(\text{Decode}(\mathbf{x}_{\text{best}}^C))$



(a) MetaLATTE pipeline.



(b) Performance of MetaLATTE classifier compared to XGBoost. MetaLATTE's prediction is more balanced with the underrepresentative datasets.

Figure S6: MetaLATTE performance and pipeline overview. A more detailed description of MetaLATTE results can be found in Zhang, et al. Zhang et al. (2024)