FROM MINIMAL DATA TO MAXIMAL INSIGHT: A MACHINE-LEARNING GUIDED PLATFORM FOR PEPTIDE DISCOVERY

Pouriya Bayat, Spencer Perkins, Sebastian Clancy, Sahil Swapnesh Patel, Richard Fei Yin, Kristof Bozovicar, Idorenyin Iwe, Mohammad Simchi, Ilan Yaniv Zeisler, Serena Sing, Vivian White, Matthew Xie, Sean Palter, Keith Pardee
Leslie Dan Faculty of Pharmacy
University of Toronto
Toronto, Ontario, Canada
{pouriya.bayat, kristof.bozovicar, ia.iwe, keith.pardee}@utoronto.ca
{spencer.perkins, sebastian.clancy, r.yin, m.simchi, ilan.zeisler, ser.singh, vivian.white, matthew.xie, s.palter}@mail.utoronto.ca
{sas.patel}@alumni.utoronto.ca

Peptides, typically ranging from 2 to 50 amino acid residues, represent a fundamental class of biomolecules with extraordinary versatility in biological systems. Their significance spans multiple domains, from cellular signaling and hormonal regulation to potential therapeutic interventions and antibacterial mechanisms. Unlike larger protein molecules, peptides offer distinct advantages such as enhanced cellular penetration (Thapa & Sullivan, 2018) and lower immunogenicity (Real-Fernandez et al., 2023). Given their multifaceted roles and enormous demand for peptide biologics, computational strategies, particularly those based on machine learning (ML), have become important tools for accelerating peptide design and discovery (Goles et al., 2024; Nielsen et al., 2024) with examples such as RFpeptide (Rettie et al., 2024), Pepflow (Abdin & Kim, 2024), PepMLM (Chen et al., 2024), and Peptune (Tang et al., 2025) for *de novo* design. Although ML-guided discovery workflows can be extremely powerful tools, their effectiveness is often hindered by the scarcity of extensive datasets required for training robust models (Alzubaidi et al., 2023).

Recently, techniques such as biophysics-based protein language models (Gelman et al., 2025), deep learning-based (Biswas et al., 2021), and few-shot learning (Zhou et al., 2024), have been developed for single proteins to aid in data-scarce scenarios. While these techniques have been successful in addressing data scarcity for single proteins, biologics usually do not work in isolation. Thus, there is a clear need for models that can consider both binder partners, as peptide-protein binding involves complex structural and functional dependencies that single-protein models fail to capture effectively.

Here we describe Minimal Data Maximal Insight (MDMI), a novel computational method for peptide discovery which identifies novel binders from a limited dataset through a structural-based approach. In MDMI, we work with approximately 100 well-characterized peptide sequences to a specific target protein. We focus on a split Green Fluorescent Protein (GFP) system with GFP11, a 16-amino acid fragment as our peptide model, and its complementary larger fragment, GFP1-10 (217 amino acids) as our target protein proxy. When GFP1-10 and GFP11 successfully interact, they reconstitute into a functional GFP complex, emitting fluorescence, which serves as an optical readout of functionality (Figure. 1a).

Our pipeline combines a structure-based predictive model with a generative model. For the sequence-agnostic predictive model (Figure 1b.), we began by simulating 3D structures of the GFP1-10/GFP11 complex for each mutant using AlphaFold Multimer (Evans et al., 2022). These structures were scored using SPserver (Aguirre-Plans et al., 2021) for statistical potentials and PyRosetta (Chaudhury et al., 2010) for physics-informed evaluation. Next, an ensemble model was trained based on these scores. Together, these components form the foundation for our peptide binder prediction model. Next, a genetic algorithm, implemented via PyGad, introduces genetic diversity and broadens the search space for peptide sequence variations (Figure 1c). Starting with an initial randomized population of sequences, each sequence's functional potential is assessed, focusing on brightness when bound to GFP1-10. Sequences with higher potential are used to generate offspring, combining traits through crossover operations. Top candidates, exhibiting high solubility, are selected for experimental testing. Using our homemade dataset and upon validation of the predictive

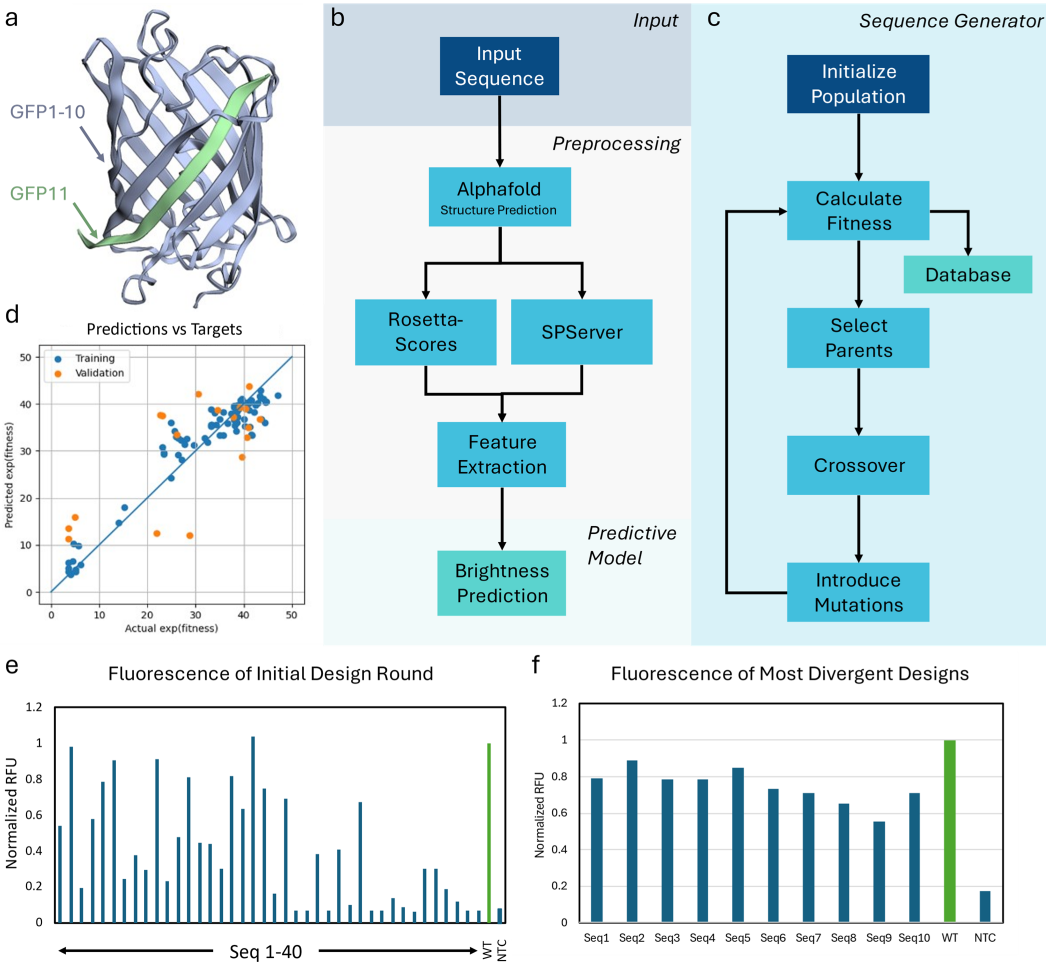

Figure 1: **a)**The GFP1-10 fragment (gray) and GFP11 fragment (green) interact to reconstitute a functional Green Fluorescent Protein (GFP) **b)** Schematic representation of the predictive model in MDMI pipeline **c)** Schematic representation of the genetic algorithm workflow **d)** Scatter plot comparing the predicted brightness values axis) to experimentally measured brightness (x-axis) for training and validation datasets. **e)** Distribution of normalized fluorescence (RFU) for sequences with 19–38% mutation rates compared to the wild-type (WT) sequence. **f)** Performance of highly mutated GFP11 variants (up to 62% sequence difference)

model (Figure 1d), we employed the genetic algorithm to identify novel sequences with 3-6 mutations, corresponding to approximately 19-38% mutation rates. The genetic algorithm evaluated 1317 sequences, from which 40 sequences with the highest predicted functionality and high solubility were selected for wet lab validation. As illustrated in Figure 1e, the sequences within the designed batch exhibited significantly higher signals than the control group. Specifically, 63% of the designed sequences exhibited more than 20% activity relative to the wild-type, 28% showed over 60% activity, and 20% demonstrated greater than 75% activity compared to the wild-type indicating successful enhancement through structural modeling and genetic algorithm as well as employing a well-curated dataset. Next, to further understand the underlying motifs contributing to successful binding and functionality, we analyzed the sequences that exhibited more than 60% activity. By identifying frequently occurring amino acids within these sequences, we assembled new variants incorporating these residues. This approach led to sequences with 10 amino acid differences, representing significant deviations from the wild type (62% sequence difference) corresponding to new and distant peaks on the sequence-function landscape. The results of the screening, shown in Figure 1f, revealed that all of the designed sequences displayed a high degree of fluorescence compared to the wild-type GFP11, despite 10 out of the 16 amino acids being different.

By decoupling peptide design from large datasets, MDMI democratizes access to advanced computational tools for labs with limited resources. With only one round of screening, we were able to train a machine learning algorithm capable of predicting sequences with high degree of functionality. MDMI's ability to engineer highly divergent yet functional sequences opens doors to *de novo* peptide therapeutics with reduced immunogenicity and enhanced stability. Demonstrated on GFP11, MDMI's pipeline is adaptable to any peptide-protein system, offering a blueprint for accelerating therapeutic peptide discovery (e.g., antimicrobials, targeted drug delivery).

REFERENCES

Osama Abdin and Philip M. Kim. Direct conformational sampling from peptide energy landscapes through hypernetwork-conditioned diffusion. *Nature Machine Intelligence 2024 6:7*, 6:775–786, 6 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00860-4. URL https://www.nature.com/articles/s42256-024-00860-4.

Joaquim Aguirre-Plans, Alberto Meseguer, Ruben Molina-Fernandez, Manuel Alejandro Marín-López, Gaurav Jumde, Kevin Casanova, Jaume Bonet, Oriol Fornes, Narcis Fernandez-Fuentes, and Baldo Oliva. Spserver: split-statistical potentials for the analysis of protein structures and protein–protein interactions. *BMC Bioinformatics*, 22:1–13, 12 2021. ISSN 14712105. doi: 10.1186/S12859-020-03770-5/TABLES/1. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03770-5.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, A. S. Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A. Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H. Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin Abbosh, and Yuantong Gu. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data 2023 10:1*, 10:1–82, 4 2023. ISSN 2196-1115. doi: 10.1186/S40537-023-00727-2. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00727-2.

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-n protein engineering with data-efficient deep learning. *Nature Methods 2021 18:4*, 18:389–396, 4 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01100-y. URL https://www.nature.com/articles/s41592-021-01100-y.

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. Pyrosetta: A script-based interface for implementing molecular modeling algorithms using rosetta, 2010. ISSN 13674803.

Tianlai Chen, Madeleine Dumas, Rio Watson, Sophia Vincoff, Christina Peng, Lin Zhao, Lauren Hong, Sarah Pertsemlidis, Mayumi Shaepers-Cheu, Tian Zi Wang, Divya Srijay, Connor Monticello, Pranay Vure, Rishab Pulugurta, Kseniia Kholina, Shrey Goel, Matthew P. DeLisa, Ray Truant, Hector C. Aguilar, and Pranam Chatterjee. Pepmlm: Target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling, 2024. URL https://arxiv.org/abs/2310.03842.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034. URL https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034.

Sam Gelman, Bryce Johnson, Chase Freschlin, Arnav Sharma, Sameer D'Costa, John Peters, Anthony Gitter, and Philip A. Romero. Biophysics-based protein language models for protein engineering. *bioRxiv*, 2025. doi: 10.1101/2024.03.15.585128. URL https://www.biorxiv.org/content/early/2025/01/14/2024.03.15.585128.

Montserrat Goles, Anamaría Daza, Gabriel Cabas-Mora, Lindybeth Sarmiento-Varón, Julieta Sepúlveda-Yañez, Hoda Anvari-Kazemabad, Mehdi D. Davari, Roberto Uribe-Paredes, Álvaro

Olivera-Nappa, Marcelo A. Navarrete, and David Medina-Ortiz. Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides. *Briefings in Bioinformatics*, 25, 5 2024. ISSN 14774054. doi: 10.1093/BIB/BBAE275. URL https://dx.doi.org/10.1093/bib/bbae275.

Jens Christian Nielsen, Claudia Hjorringgaard, Mads Morup Nygaard, Anita Wester, Lisbeth Elster, Trine Porsgaard, Randi Bonke Mikkelsen, Silas Rasmussen, Andreas Nygaard Madsen, Morten Schlein, Niels Vrang, Kristoffer Rigbolt, and Louise S. Dalboge. Machine-learning-guided peptide drug discovery: Development of glp-1 receptor agonists with improved drug properties. *Journal of Medicinal Chemistry*, 67:11814–11826, 7 2024. ISSN 15204804. doi: 10.1021/ACS.JMEDCHEM.4C00417/SUPPL_FILE/JM4C00417_SI_002.CSV. URL https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.4c00417.

Feliciana Real-Fernandez, Fosca Errante, Andrea Di Santo, Anna Maria Papini, and Paolo Rovero. Therapeutic proteins immunogenicity: a peptide point of view. *Open Exploration 2019 1:5*, 1: 377–387, 10 2023. ISSN 2836-7677. doi: 10.37349/EDS.2023.00025. URL https://www.explorationpub.com/Journals/eds/Article/100825.

Stephen A. Rettie, David Juergens, Victor Adebomi, Yensi Flores Bueso, Qinqin Zhao, Alexandria N. Leveille, Andi Liu, Asim K. Bera, Joana A. Wilms, Alina Üffing, Alex Kang, Evans Brackenbrough, Mila Lamb, Stacey R. Gerben, Analisa Murray, Paul M. Levine, Maika Schneider, Vibha Vasireddy, Sergey Ovchinnikov, Oliver H. Weiergräber, Dieter Willbold, Joshua A. Kritzer, Joseph D. Mougous, David Baker, Frank DiMaio, and Gaurav Bhardwaj. Accurate de novo design of high-affinity protein binding macrocycles using deep learning. *bioRxiv*, 2024. doi: 10.1101/2024.11.18.622547. URL https://www.biorxiv.org/content/early/2024/11/18/2024.11.18.622547.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion, 2025. URL https://arxiv.org/abs/2412.17780.

Raj Kumar Thapa and Millicent O. Sullivan. Gene delivery by peptide-assisted transport. *Current Opinion in Biomedical Engineering*, 7:71–82, 9 2018. ISSN 2468-4511. doi: 10.1016/J.COBME.2018.10.002.

Ziyi Zhou, Liang Zhang, Yuanxi Yu, Banghao Wu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications 2024 15:1*, 15:1–13, 7 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49798-6. URL https://www.nature.com/articles/s41467-024-49798-6.