De Novo Short Linear Motif (SLiM) Discovery With AlphaFold-Multimer

Theodore Sternlieb
Dyno Therapeutics
Watertown, MA 02472
theodore.sternlieb@dynotx.com

Abhishaike Mahajan
Dyno Therapeutics
Watertown, MA 02472
abhishaike.mahajan@dynotx.com

Davian Ho*
Dyno Therapeutics
Watertown, MA 02472
davian@berkeley.edu

Jeffrey Chan
Dyno Therapeutics
Watertown, MA 02472
jeffrey.chan@dynotx.com

Abstract

Short Linear Motifs (SLiMs) are short, disordered peptide fragments, which mediate a large class of protein-protein interactions (PPIs). SLiM-mediated interactions are often dynamic, low affinity interactions, which play a crucial role in cell regulation and signal transduction. Despite their importance to cell function, challenges in experimental throughput and manual aggregation of information across numerous experiments pose significant bottlenecks in fully characterizing SLiMs, including their binding partners, diversity, and consolidation into a unified dataset [10][11]. As a result, only a minuscule fraction of the estimated hundreds of thousands of SLiMs have been identified [16]. The prospect of employing computational SLiM discovery methods to prioritize SLiM-protein interactions for experimental validation, thus accelerating our comprehension of SLiMs, continues to be intriguing. SLiM discovery methods are typically divided into two classes: (1) *Instance Detection*: which focuses on discovering novel instances of known SLiMs and (2) *De Novo Discovery*: which focuses on discovering unknown SLiMs. Unfortunately, up until now, *de novo* SLiM discovery has been too challenging to serve as a useful tool to aid experimental characterization and has only been applied in limited settings. However, recent progress in protein structure prediction has translated to significant progress across many applications, so we posit that improved protein structure resolution may make *de novo* SLiM discovery tractable. In this work, we curate a SLiM discovery benchmark dataset, devise an AlphaFold-Multimer-based SLiM discovery method, and demonstrate settings in which our method can accurately perform *de novo* SLiM discovery.

1 Introduction

Short linear motifs (SLiMs) are patterns of short (often consecutive) amino acids found throughout the eukaryotic proteome that mediate protein-protein interactions (PPIs) critical for cellular function such as signaling, localization, and degradation. Although SLiMs are embedded in larger proteins, often just three to ten amino acids within a short disordered region drive preferential binding affinity [2]. These SLiM-mediated interactions are often characterized as transient with low binding affinity and promiscuous recognition. Hundreds of thousands of such interactions are estimated in the human

*Work performed during internship at Dyno.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

proteome with mutations in SLiMs implicated in several known diseases [16]. Furthermore, the ability to modulate the regulatory activity of SLiM-mediated interactions is of great interest for therapeutics. Discovery and detailed characterization of SLiMs and their binding partners would unlock many biological and clinical applications.

The current state-of-the-art resource for previously discovered and characterized SLiMs is the Eukaryotic Linear Motif (ELM) database, which contains hand-curated SLiMs and their biological functions [11]. The ELM database holds over 300 different SLiM classes (a grouping of SLiMs which share similar biological function and sequence) as well as around 2,400 SLiM-mediated PPI pairs and corresponding motif/binding domain annotations. The ELM database also contains handcrafted regular expressions (regexs) defining evolutionary conserved, semi-conserved and degenerate positions within the SLiM. Despite the progress made with this dataset, experimental SLiM discovery workflows remain challenging, hampered by the scale, accuracy, and cost of experimental characterization. Computational SLiM discovery methods developed thus far focus on two tasks: (1) *SLiM instance discovery* which involves discovery of new instances of known SLiMs such as distinct SLiM-protein interactions, and (2) *De Novo discovery* which looks for motif enrichment across homologs or sequences with similar functional classifications[7]. However, *de novo* methods often generate false positives owing to spurious evolutionary conservation and fail to model any semblance of structural interaction[4]. We seek to address the shortcomings of (2) by forgoing notions of motif enrichment, and instead rely on recent advances in protein structure prediction.

In recent years, breakthroughs in protein structure prediction such as AlphaFold[9][1] have enabled highly accurate *in silico* resolution of both monomeric and multimeric protein structure. In the wake of these achievements, a number of studies [8][5][17] were published using model confidence scores to determine binding affinity of protein-protein complexes predicted by these models across a wide range of applications[12]. Despite these results, translation to SLiM mediated interactions with far smaller affinities is not obvious, although one recent study [12] suggested the ability of AlphaFold-Multimer to discriminate between short, SLiM-containing peptides and non-binders. Still, the ability of AlphaFold2 to do discovery or design of SLiMs for a target protein where the SLiM's flanking peptide context is unknown remains unstudied. In this work, we seek to tackle the *de novo* SLiM discovery problem directly by curating a benchmark SLiM discovery dataset for evaluation, developing a structure-based *de novo* SLiM discovery method, and demonstrating our ability to discover SLiMs. In Section 2, we define the *de novo* SLiM discovery task. In Section 3, we detail the creation of a benchmark SLiM discovery dataset to evaluate performance. We then describe our SLiM discovery method in Section 4, and detail the full results in Section 5. Finally, we identify future directions enabled by this work in Section 6.

2 *De Novo* SLiM Discovery Task

The goal of the *de novo* SLiM discovery task is to discover SLiM-mediated interactions without reliance on experimental assay data. More concretely, given some globular protein $\mathcal{P}$ and target SLiM length n, we would like to find the set of n-mers which bind to $\mathcal{P}$ with some sufficiently high affinity. Note that while the flanking regions of the SLiM may impact binding affinity through steric or physiochemical mediation[3], the high entropy of those positions suggest that characterizing only the SLiM itself is sufficient. However without access to the complete characterization of binding affinities for each n-mer, we instead benchmark methods for the *de novo* SLiM discovery task by seeking to maximize the rank of experimentally validated SLiMs against other n-mers.

3 Benchmark Dataset Curation

In order to benchmark AlphaFold's performance on SLiM discovery, we design a dataset using 20 validated SLiM-protein pairs which aims to simulate the *de novo* SLiM discovery task. For each protein, we construct a set of decoy SLiMs to determine how well AlphaFold can distinguish a true SLiM from a decoy. True SLiM-protein pairs are derived from the ELM database[11]. We primarily focusing on interactions mediated by smaller SLiMs (≤ 5 amino acids) as they are often more difficult to extract using homology detection tools. Although interactions mediated by larger, more complex classes of SLiMs exist, we leave a thorough analysis of these to future work. Additionally, we note that we have simplified the problem by only comparing true SLiMs against decoys of equal length. We do this to avoid the complications arising from comparing AlphaFold2 metrics across binders

of different lengths[15]. Additionally, we BLAST the protein of each SLiM-protein pair against the PDB to determine whether relevant structural data implicating the SLiM and protein might exist in AlphaFold's training data. Selected SLiM-protein interactions, as well as annotations denoting whether relevant solved SLiM-protein complexes exist, are available in Table 1.

We design a set of random and rationally designed decoys following a similar approach to [12]. The random decoys were designed by generating random n-mers of equal length for each SLiM. The rationally designed SLiMs are designed to evaluate the sensitivity of discovery methods to mutations. We select a conserved amino acid in the SLiM and substitute it with a chemically similar (putatively positive SLiMs) or distant amino acids (putatively negative SLiMs) by using Miyata distances[14]. Although there is no guarantee that putative positives or negatives designed in this way are true positive or negatives, many positions in different SLiM classes are robust to substitutions of chemically similar amino acids, and point substitutions to chemically distant amino acids are likely to impair binding[6]. To determine whether AlphaFold's predictions favor substitutions to chemically similar amino acids, we include sequences generated by randomly mutating 1 or 2 amino acids of the true SLiM in order to compare against. Finally to ensure that AlphaFold is not merely picking up on the poor evolutionary plausibility of random decoys, we include a number of random protein fragments. We summarize the decoy types and their counts in the final benchmark dataset in Table 2.

4 SLiM Discovery Method

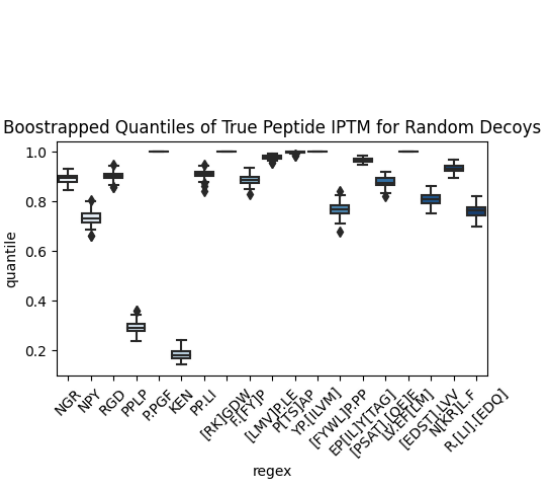
Our SLiM discovery method utilizes AlphaFold-Multimer[5] to discover SLiMs for a target protein. We perform the discovery task by searching the design space of possible n-mers and fold them, along with modified flanking context, in complex with the target protein using ColabFold[13]7. We first determine the optimal flanking context as well as output confidence metric that maximizes the ability of AlphaFold-Multimer's various AlphaFold to discriminate between positive SLiMs and random decoys for a given protein. For each selected protein, we compare the true SLiM against a set of 20 random n-mer decoys with equal length to create a pool of 21 SLiMs per protein. We then test multiple SLiM-containing constructs for each SLiM by adding no context, 5 flanking glycines and 5 flanking alanines on either side of the SLiM. We observe that adding both glycine and alanine flanking regions improves the discriminative ability of AlphaFold, with glycine flanking context performing marginally better. To further optimize the context, we examine the effect of varying the number of flanking residues by testing 3 through 7 glycines. We find that while performance varies across proteins, 3 flanking glycines on either side consistently improves the performance of all considered metrics, with ipTM showing the best performance2. We find that both binding location SLiM conformation are both affected by additional flanking region. For all following experiments, we use 3 flanking glycines on each side as our standard input into AlphaFold-Multimer and evaluate binding affinity based on the ipTM metric.

5 Results

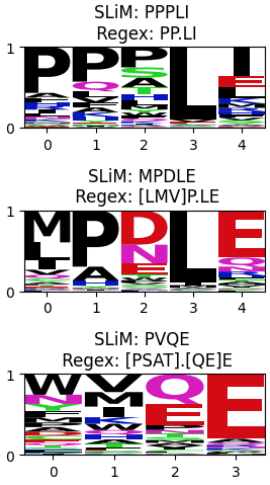
We predict the structure for the full dataset of roughly 20,000 SLiM-protein pairs described in 3 using the flanking context procedure described above. We find that our SLiM discovery method had heterogeneous performance based on the target protein, with AlphaFold accurately ranking SLiMs for some proteins and ranking randomly for others as shown in Figure 3. To ensure that the heterogeneity isn't solely explained by overfitting to the AlphaFold training set, we find that while almost all of the target proteins with relevant SLiM-protein complexes in the PDB saw good SLiM discovery performance, good ranking accuracy was also achieved for roughly half of the target proteins which were not represented in the PDB. Additionally, we note that our approach highly ranks the true SLiM along with many of the putative positives regardless of whether the original complex was in the PDB. For two SLiMs with poor performance, RGD and [RK]GDW, we suspect that the target protein often occurs in the form of an integrin heterodimer while we only modeled a single monomer. Based on the ability of our method to also extract putative, previously unseen positives, we hypothesized that our method was able to leverage AlphaFold's ability to extract the varying degrees of evolutionary conservation across positions within the SLiM. To further investigate this, we analyzed the subset of well-predicted target proteins and calculate a log fold-change for each amino acid at each position to determine relative enrichment. We examine three SLiM instances, whose SLiM classes are described by the regexes PP.LI, [LMV]P.LE and [PSAT].[QE]E, plot their fold change in Figure 1b and examine

concordance with the corresponding regexes for each SLiM. We find high entropy over the degenerate positions (represented by '.' in the regexs) in PP.LI and [LMV]P.LE, as well as high enrichment for correct amino acids at semi-conserved positions in the case of [LMV] and [QE], although this is not true at the first position of [PSAT].[QE]E. Hence for many target proteins, AlphaFold is able to partially recapitulate the conserved, semi-conserved and degenerate positions of SLiM classes.

While fully characterizing the possible sequence space for the *de novo* SLiM discovery task provides an accurate assessment of the precision and recall of our method, computational budget constraints makes this impractical. Rather, to evaluate our method's performance on the *de novo* SLiM discovery task under computational constraints, we bootstrap the predicted quantile of the true SLiM against the set of random decoys and find that for 13 out the the 20 target proteins tested, the true SLiM is at or above the 80th quantile as shown in Figure 1a. Its worth noting that due to SLiMs exhibiting putative positives from edit distance 1 Miyata substitutions, no SLiM discovery method should rank the true SLiM at the 100th percentile without some train-test leakage. However, it is clear that the method does not perform well consistently for each target protein, showing remarkably poor performance in certain settings,showing random or worse than random performance such as accurately ranking the SLiMs PPLP or F.[FY]P.



(a) Bootstrapped Quantile



(b) Detecting conserved patterns

Figure 1: *Left*: We calculate bootstrapped predicted quantile for positive SLiMs by performing 100 trials in which we sample 300 SLiMs from the original dataset for each protein and compute the quantile of the true SLiM's ipTM. The distributions of quantiles are then plotted. *Right*: Logo plots for top sequences from three well predicted complexes. To generate the plots, amino acid by position count matrices were generated for both the top 80 sequences and the rest of the dataset. Next, log fold change was calculated by dividing the 2 matrices and taking the log of the result. Finally, the softmax function was applied to log-fold change values at each position to generate a position weight matrix which was then plotted.

6 Conclusions

We have designed an AlphaFold-Multimer-based method for *de novo* discovery of short linear motifs, conditioned on a target protein. We evaluate the performance of our method on our curated benchmark dataset containing randomized and rationally designed decoys. For half of the target proteins, our method is capable of not only discriminating SLiM binders from negative decoys, but also accurately ranking the neighborhood around the true SLiM. Such a method can be used for discovery of SLiMs-mediated interactions as well as design of SLiM-based therapeutics. Additionally, we find that AlphaFold is able to partially recapitulate the degeneracy and conservation of different positions in SLiMs in accordance with their associated SLiM class's regex. Given these findings, we find in reasonable to conclude that our approach is able to identify candidate SLiMs for a target domain. However, if we wish to find a novel n-mer SLiM for a given target domain, we are left with

the need to enumerate and fold the space of all n-mers, which becomes prohibitively expensive for $n > 4$. Screening all 3-mers again a 200 amino acid long domain costs roughly $4,500. In order to overcome this limitation, we identify two possible avenues for improving efficiency. First, a more sophisticated SLiM proposal method would avoid the need to sample the full n-mer space by only proposing biologically plausible n-mers. Second, our results showing some level of smoothness for ipTM scores over n-mer space suggests that iterative optimization methods may be able to speed up search. We leave these directions open to future work.

References

[1] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K Rathinaswamy, Udit Dalwadi, Calvin K Yip, John E Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a 3-track network. *Science*, 2021. Preprint available at https://doi.org/10.1101/2021.06.14.448402.

[2] Norman E Davey, Martha S Cyert, and Alan M Moses. Short linear motifs - ex nihilo evolution of protein regulation. *Cell communication and signaling : CCS*, 13:43, Nov 2015. PMID: 26589632.

[3] Norman E Davey, Kim Van Roey, Robert J Weatheritt, Grischa Toedt, Bora Uyar, Brigitte Altenberg, Aidan Budd, Francesca Diella, Holger Dinkel, and Toby J Gibson. Attributes of short linear motifs. *Molecular bioSystems*, 8(1):268–281, Jan 2012.

[4] Richard J Edwards and Nicholas Palopoli. Computational prediction of short linear motifs from protein sequences. *Methods in Molecular Biology*, 1268:89–141, 2015.

[5] 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. *Journal Name*, Volume Number:Page Numbers, Year.

[6] Toby J. Gibson, Holger Dinkel, Kim Van Roey, and Francesca Diella. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Communication and Signaling*, 13(1):42, 11 2015.

[7] Peter Hraber, Paul E O'Maille, Andrew Silberfarb, Katie Davis-Anderson, Nicholas Generous, Benjamin H McMahon, and Jeanne M Fair. Resources to discover and use short linear motifs in viral proteins. *Trends in Biotechnology*, 38(1):113–127, 2020. PMID: 31427097.

[8] Isak Johansson-Åkhe and Björn Wallner. Improving peptide-protein docking with alphafold-multimer using forced sampling. *Frontiers in Bioinformatics*, 2:959160, 2022.

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

[10] Izabella Krystkowiak and Norman E Davey. Slimsearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Research*, 45(W1):W464–W469, 2017.

[11] Manjeet Kumar, Sushama Michael, Jesús Alvarado-Valverde, Bálint Mészáros, Hugo Sámano-Sánchez, András Zeke, Laszlo Dobson, Tamas Lazar, Mihkel Örd, Anurag Nagpal, Nazanin Farahi, Melanie Käser, Ramya Kraleti, Norman E Davey, Rita Pancsa, Lucía B Chemes, and Toby J Gibson. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Research*, 50(D1):D497–D508, 10 2021.

[12] Chop Yan Lee, Dalmira Hubrich, Julia K. Varga, Christian Schäfer, Mareen Welzel, Eric Schumbera, Milena Đokić, Joelle M. Strom, Jonas Schönfeld, Johanna L. Geist, Feyza Polat, Toby J. Gibson, Claudia Isabelle Keller Valsecchi, Manjeet Kumar, Ora Schueler-Furman, and Katja Luck. Systematic discovery of protein interaction interfaces using alphafold and experimental validation. *bioRxiv*, 2023.

[13] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022.

[14] Takashi Miyata, Sanzo Miyazawa, and Teruo Yasunaga. Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, 12(3):219–236, 1979.

[15] Vivian Monzon, Daniel H Haft, and Alex Bateman. Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinformatics Advances*, 2(1):vbab043, 01 2022.

[16] Peter Tompa, Norman E Davey, Toby J Gibson, and M Madan Babu. A million peptide motifs for the molecular biologist. *Molecular cell*, 55(2):161–169, 2014.

[17] Björn Wallner. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*, 39(9):btad573, 09 2023.

7 Appendix

Linear Motif	Regex	Binder Protein ID	Target Protein ID	Resolved Structure
LNGR	NGR	P11276	P06756	True
NPY	NPY	C6UYL8	Q9UQB8	True
RGD	RGD	P21404	P18564	False
PPLP	PPLP	Q64213	Q9R1C7	False
PAPGF	P.PGF	Q9NRY6	O75340	False
KEN	KEN	Q08981	P53197	True
PPPLI	PP.LI	O75376	Q06455	True
KGDW	[RK]GDW	P22827	P08514	False
FNFP	F.[FY]P	P28562	P28482	False
MPDLE	[LMV]P.LE	Q15185	Q9GZT9	False
PSAP	P[TS]AP	B5TVE8	Q99816	False
YPKI	YP.[ILVM]	P33400	Q12033	False
FPPPP	[FYWL]P.PP	P18206	Q8N8S7	True
EPLYA	EP[IL]Y[TAG]	Q5QT02	P41240	False
PVQE	[PSAT].[QE]E	P18347	B5DFH7	False
LVAEFL	LV.EF[LM]	P50542	O75381	True
DILVV	[EDST].LVV	Q13137	Q9BXW4	True
NRLNF	N[KR]L.F	P36094	P24869	False
RSLCE	R.[LI].[EDQ]	Q8I2C7	Q8I6Z5	False
LPLPP	[FYWL]P.PP	Q702N8	Q9UI08	False

Table 1: **Table of SLiM-protein complexes.** SLiMs were selected based on a number of criteria including length, specificity of the Regex, lack of non flanking wildcards and presence in pdb

Decoy Type	Count per Protein
negative	8
positive_miyata_ed_1	8
positive_miyata_ed_2	18
positive_miyata_ed_3	18
brute_ed_1	50
brute_ed_2	50
protein_fragment	100
random	800

Table 2: **Decoy types and their respective counts per protein.** Brute_ed_n correspond to decoys n point substitutions away from the true SLiM. Positive_miyata_ed_n refer to decoys with n point substitutions to maximally similar amino acids, where as negative refers to decoys with a single maximally dissimilar point substitution. Protein_fragments are slices from other SLiMs and random decoys refer to random n-mers

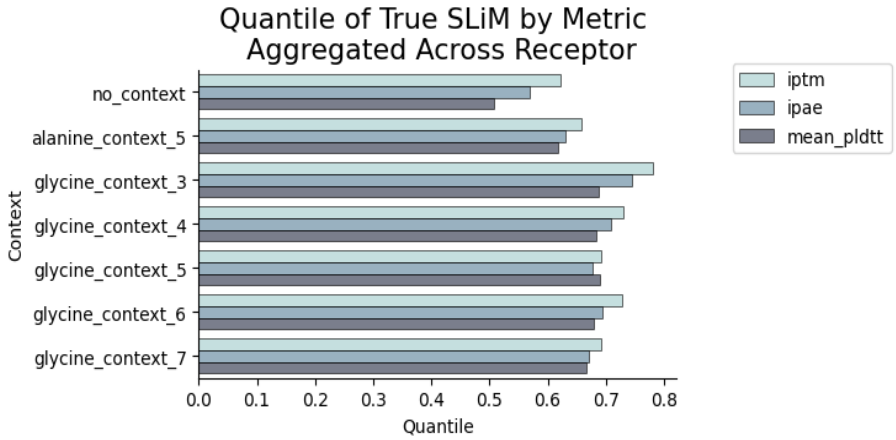


Figure 2: **Optimizing flanking context.** Percent Rank by ipTM metric for each protein and each context strategy. Although performance is variable, 3 flanking glycines on either side of the SLiM most frequently outperforms other metrics

AlphaFold-Multimer Parameters

- AlphaFold-Multimer Version: 2.3
- use_templates: False
- max_num_recycles: True
- early_stopping: True
- MSA_method: mmseqs2_uniref_env
- num_predictions_per_model: 1
- num_models: 5

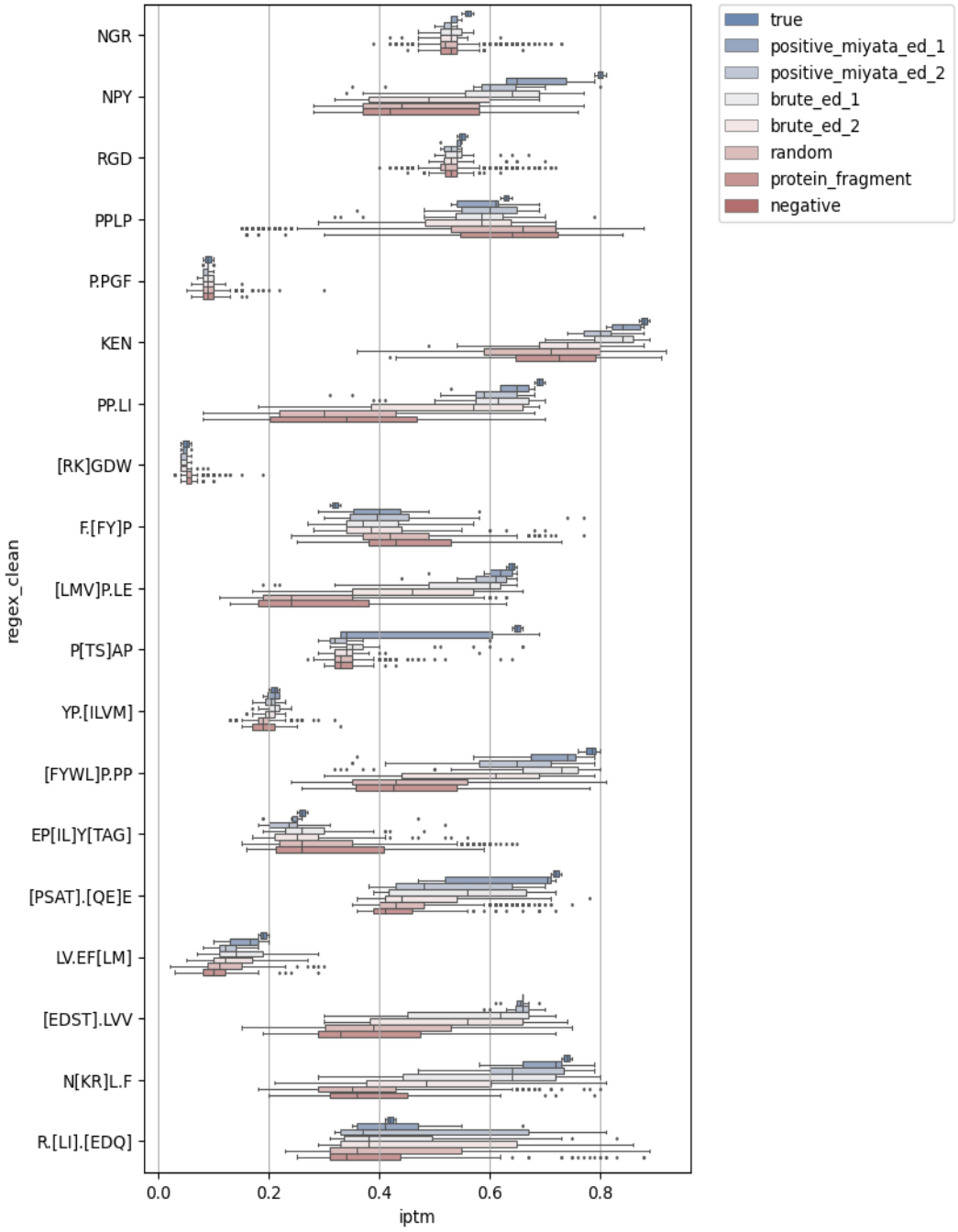


Figure 3: ipTM scores for each SLiM-protein complex broken down by decoy type. The ipTM values for true decoys are duplicated with $\pm.01$ so that they are visible.

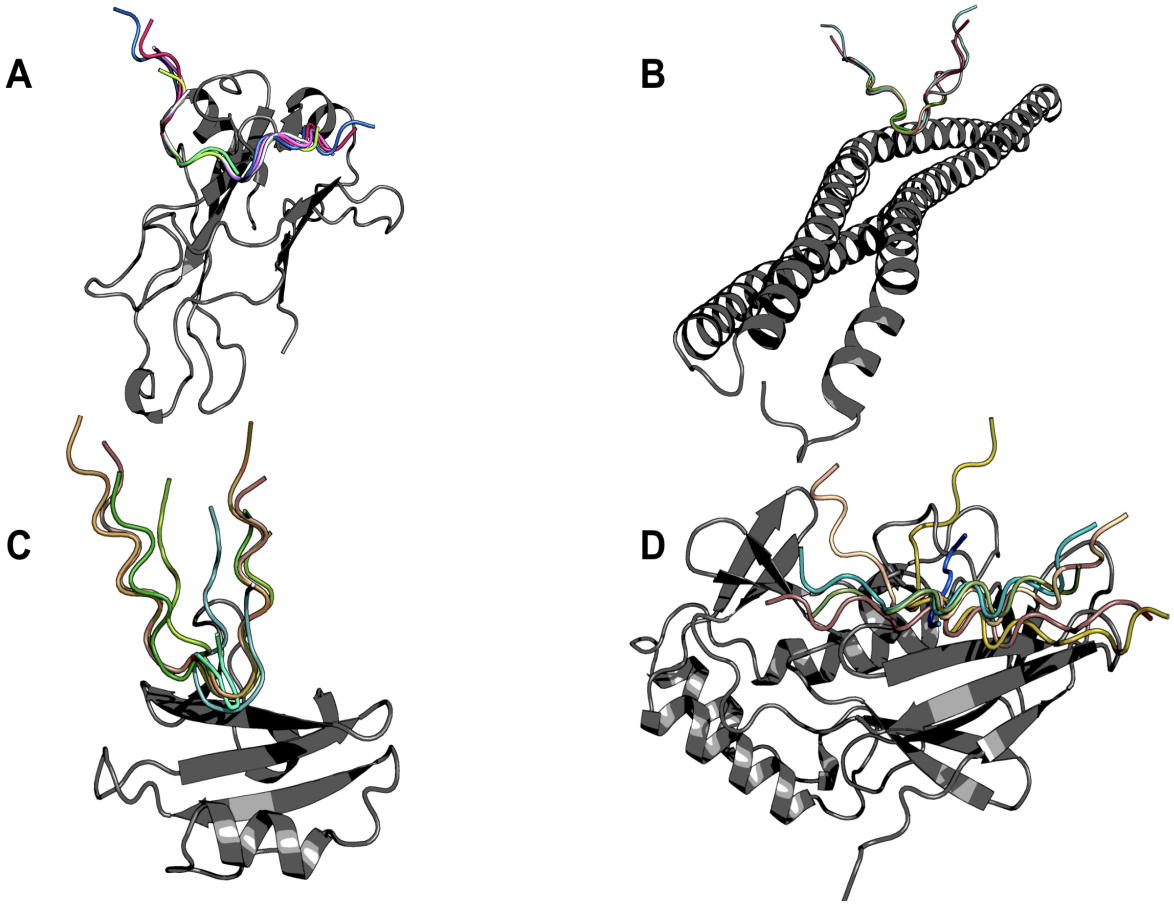


Figure 4: **Sample Structures**. Positive SLiMs with varying numbers of glycine flanking context in complex with their binding partners. SLiMs in Order: [PSAT].[QE], NPY, EP[IL]Y[TAG], R.[LI].[EDQ]. In both A and B, binding pose remains comparable across different flanking contexts. For C, binding orientation of each SLiM remain constant, but the loop conformation changes. For D, binding orientation varies