HARP-pSAR: A pLM-based Framework for Protein Fitness Landscapes in Practical Drug Discovery

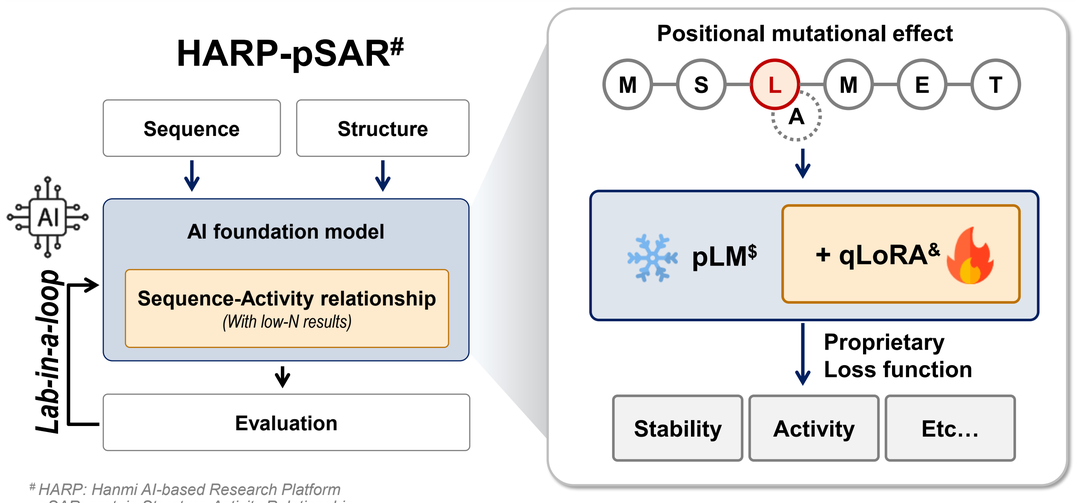
Poster D-008

Seungsu Han, Yohan Kim, Yoon Keun Cho, Seon Myeong Lee, Jung Kuk Kim, Sungmin Bae, and In Young Choi.
Hanmi Pharm. Co., Ltd., Seoul, Republic of Korea

Introduction

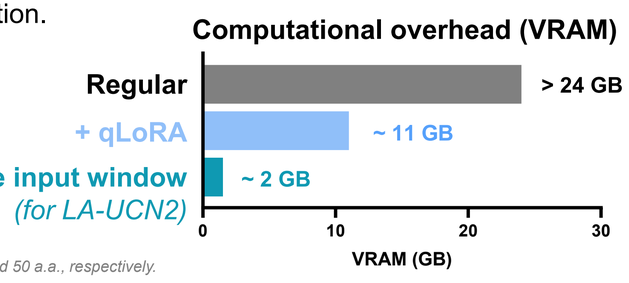
Artificial Intelligence (AI) accelerates drug discovery, but moving beyond AlphaFold's structural data to accurately predict protein fitness, such as distinguishing on-target and off-target effects, is essential for rapid candidate evaluation. BERT-based Protein Language Models (pLMs) address this by extracting intrinsic information directly from sequence data. However, acquiring experimental fitness labels is expensive and time-consuming, making reliable fitness landscape construction with 'low-N' data a major technical hurdle. To address this challenge, we introduce HARP-pSAR (Hanmi AI-driven Research Platform-Protein Sequence-Activity Relationship), one of Hanmi's AI research platform tools which is a pLM-based platform designed for fitness landscape learning under severe low-N data constraints. By applying LoRA (Low-Rank Adaptation) quantization for efficient fine-tuning, the platform significantly reduces computational overhead, allowing it to be operated directly on local PCs.

HARP-pSAR: Fitness prediction with low-N experiment data



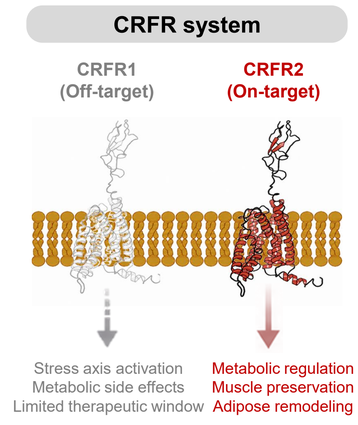
HARP: Hanmi AI-based Research Platform
pSAR: protein Structure-Activity Relationship
$ pLM: Protein Language Model
& qLoRA: Quantized Low-Rank Adaption

- **AI-based SAR for peptide/protein.**
- **Work with low-N results.** (30+ results)
- **Enhanced with lab-in-a-loop method.**
  1) Feasible candidate generation.
  2) Evaluation.
  3) Re-training AI model.
  → Generation again.

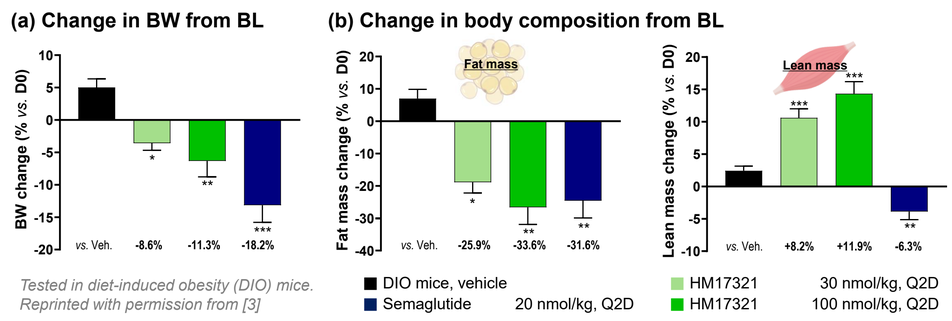


ESM-2 650M used as a foundation model.
Regular and LA-UCN2 input window is set to 200 a.a. and 50 a.a., respectively.

Case Study: LA-UCN2 (Long-acting Urocortin 2 analog)



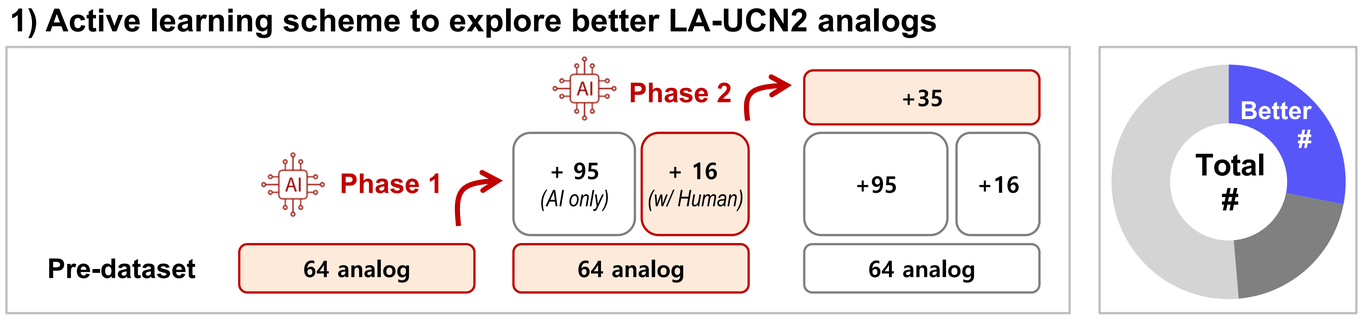
HM17321 is designed for obesity medication with lean mass maintenance.
Hanmi's LA-UCN2 (a novel long-acting CRFR2-selective UCN2 analog)



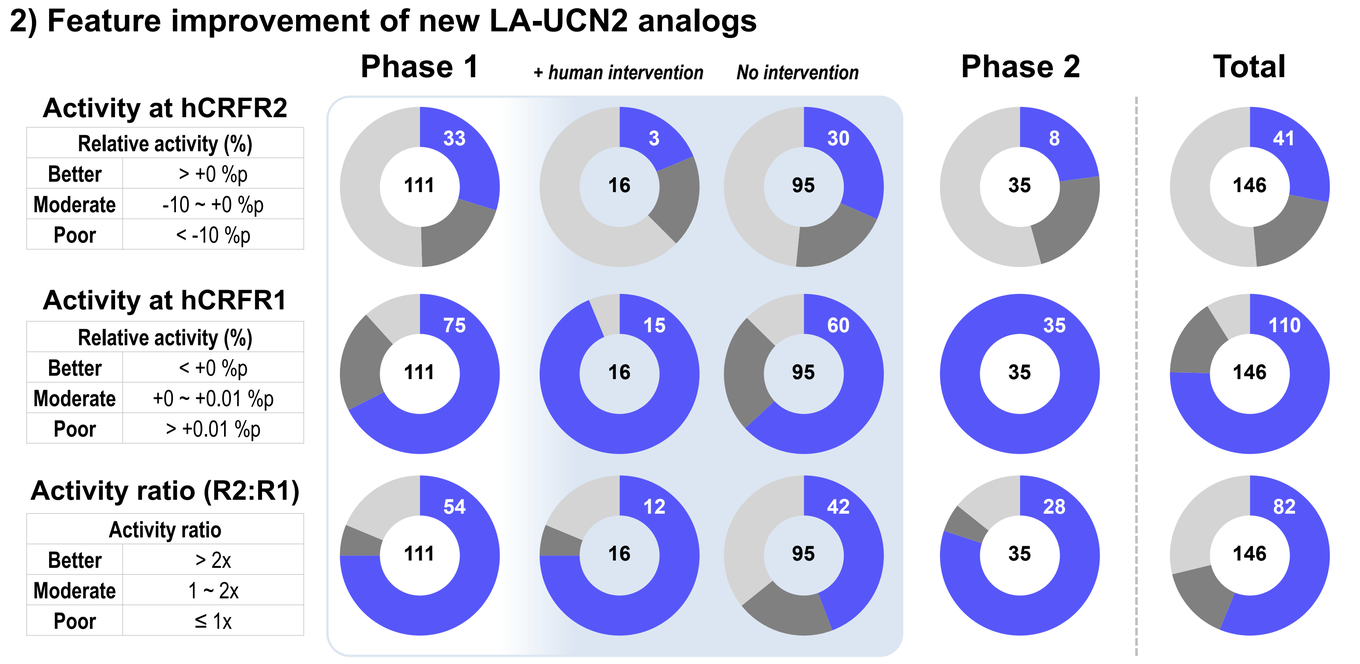
Tested in diet-induced obesity (DIO) mice.
Reprinted with permission from [3]

LA-UCN2: Active learning toward out-of-distribution (OOD) landscape

1) Active learning scheme to explore better LA-UCN2 analogs



2) Feature improvement of new LA-UCN2 analogs



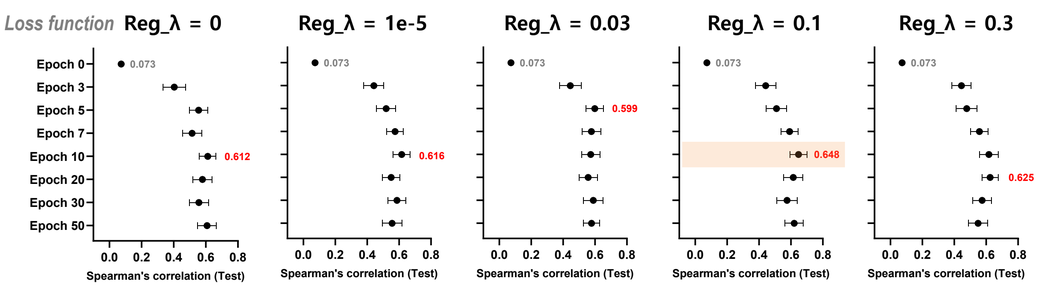
vs. Hanmi's LA-UCN2 representative analog

Framework Configuration for the case study

Feature example: CRFR1-mediated cAMP activity of LA-UCN2s

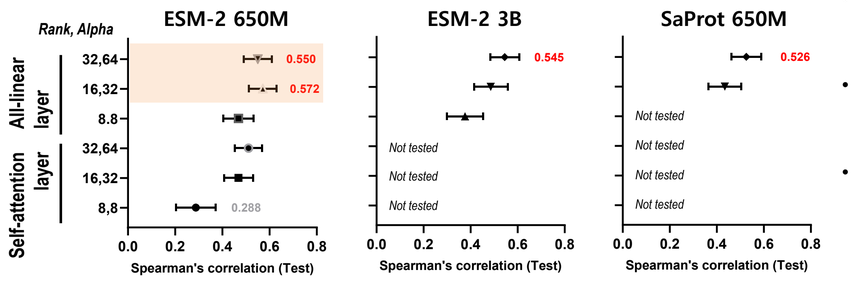
1) Training hyperparameters

Mean w/ S.E.M. (n = 30)



Partial dataset used (28 for training, 10 for test). 100-runs with distinct sampling seed.
ESM-2 650M used as a foundation model. qLoRA (4-bit quantization) was applied with r = 16 and alpha = 32 to all-linear layers.

2) Foundation model & LoRA hyperparameters

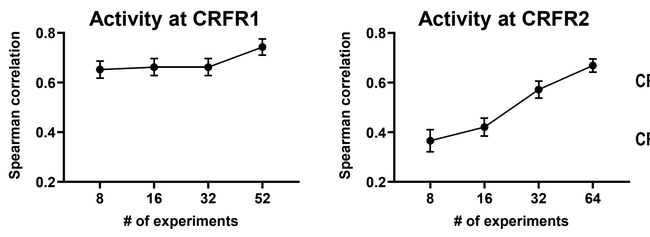


- **ESM-2 650M** *Sequence-based, developed by Meta*
- **ESM-2 3B** *Sequence-based, developed by Meta*
- **SaProt 650M** *Structure-aware, developed by Westlake University*

Partial dataset used (28 for training, 10 for test). 100-runs with distinct sampling seed.
Training epoch and loss function regularization lambda were set to 10 and 0.1, respectively.

Mean ± 95% CI (n = 100)

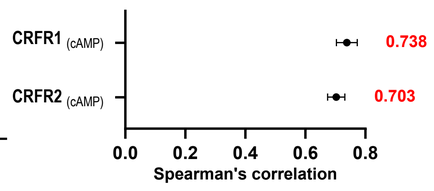
'Lab-in-a-loop' effect



Mean ± 95% CI (n = 100)

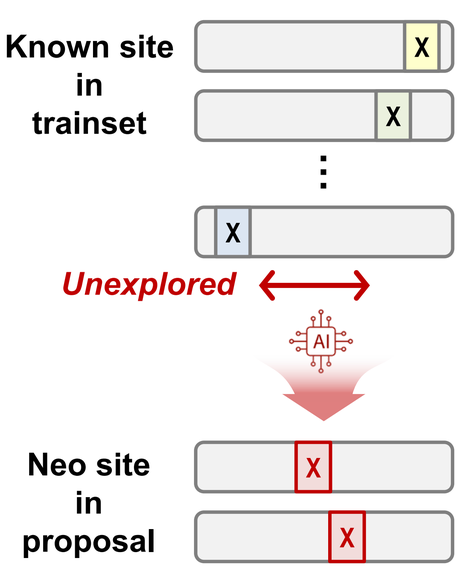
Model performance (Configured)

ESM-2 650M, the optimized hyperparameters.
52-shots (CRFR1) and 64-shots (CRFR2)



Mean ± 95% CI (n = 100)

LA-UCN2: Neo site proposal by HARP-pSAR



Activity at hCRFR2

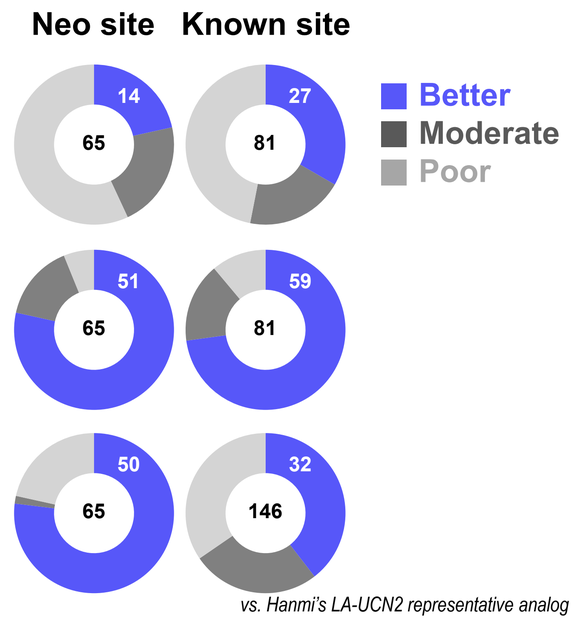
Relative activity (%)	
Better	> +0 %p
Moderate	-10 ~ +0 %p
Poor	< -10 %p

Activity at hCRFR1

Relative activity (%)	
Better	< +0 %p
Moderate	+0 ~ +0.01 %p
Poor	> +0.01 %p

Activity ratio (R2:R1)

Activity ratio	
Better	> 2x
Moderate	1 ~ 2x
Poor	≤ 1x



vs. Hanmi's LA-UCN2 representative analog

References

1. Lin et al., *Science*. 2023 Mar 17;379(6637):1123-1130.
2. Su et al., *bioRxiv*. 2024, doi: 10.1101/2023.10.01.560349.
3. Lee et al., ObesityWeek 2024 poster-504.

Concluding Remarks

- **HARP-pSAR is a pLM platform designed for low-N fitness landscape learning. LoRA quantization minimizes computational overhead, enabling high-performance AI drug discovery directly on local PCs.**
- **In study case of CRFR2-selective UCN2 analogs, HARP-pSAR demonstrated strong OOD generalization, identifying a novel, experimentally validated neo-residue site.**
- **This validates HARP-pSAR as a cost-effective, deployment-ready tool. We will extend this framework to additional pipeline assets to accelerate Hanmi's therapeutic development strategies.**