Journal of Medicinal Chemistry

pubs.acs.org/jmc Article

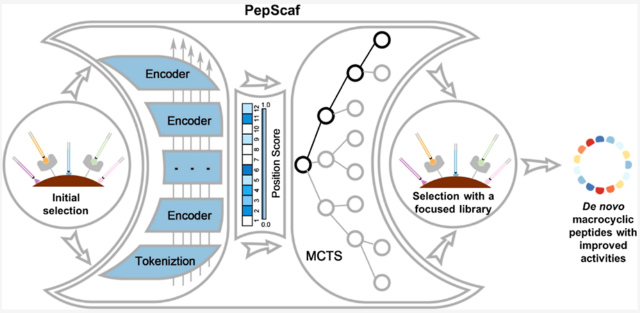
PepScaf: Harnessing Machine Learning with In Vitro Selection toward De Novo Macrocyclic Peptides against IL-17C/IL-17RE Interaction

Silong Zhai,[⊥] Yahong Tan,[⊥] Chengyun Zhang, Christopher John Hipolito, Lulu Song, Cheng Zhu, Youming Zhang, Hongliang Duan,* and Yizhen Yin*

Cite This: J. Med. Chem. 2023, 66, 11187−11200 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The combination of library-based screening and artificial intelligence (AI) has been accelerating the discovery and optimization of hit ligands. However, the potential of AI to assist in de novo macrocyclic peptide ligand discovery has yet to be fully explored. In this study, an integrated AI framework called PepScaf was developed to extract the critical scaffold relative to bioactivity based on a vast dataset from an initial in vitro selection campaign against a model protein target, interleukin-17C (IL-17C). Taking the generated scaffold, a focused macrocyclic peptide library was rationally constructed to target IL-17C, yielding over 20 potent peptides that effectively inhibited IL-17C/IL-17RE interaction. Notably, the top two peptides displayed exceptional potency with IC_{50} values of 1.4 nM. This approach presents a viable methodology for more efficient macrocyclic peptide discovery, offering potential time and cost savings. Additionally, this is also the first report regarding the discovery of macrocyclic peptides against IL-17C/IL-17RE interaction.


PepScaf
Initial selection
Encoder
Encoder
. . .
Encoder
Tokenziton
Position Score
MCTS
Selection with a focused library
De novo macrocyclic peptides with improved activities

■ INTRODUCTION

Pharmaceutical research and development are estimated to cost hundreds of millions of dollars and take an average cycle longer than 10 years.[1,2] The emergence of library-based screening combined with artificial intelligence (AI) has been streamlining and accelerating the discovery and optimization of hit ligands.[3−5] Library-based screening strategies, e.g., DNA-encoded chemical library (DEL) technology, enable for the deep exploration of large chemical space, which affords simultaneous readout of millions to billions of compounds against targets of interest.[6−8] AI-driven drug discovery has also embraced a growing number of successes with recent advances in computing power and availability of large datasets. With the assistance of AI, a highly active, selective, and bioavailable inhibitor of discoidin domain receptor-1 (DDR1) was successfully identified within 21 days.[9] A SMILES-based recurrent neural network (RNN) model, trained on a large set of known bioactive compounds, was also utilized to generate the agonists of retinoid X receptor (RXR) and peroxisome proliferator-activated receptor (PPAR) by Merk.[8] Given that the library-based screening strategies are capable of covering a large chemical space, AI based on the vast datasets could be a valuable tool for further improving the hit discovery efficiency. Several cases of DEL integrated with AI have been reported. McCloskey et al. successfully performed machine learning modeling using the data obtained from DEL screening against the targets including sEH (a hydrolase), Erα (a nuclear receptor), and c-KIT (a kinase).[10] Another example came from Lim et al., who combined DEL and machine learning for efficient screenings against carbonic anhydrase (CAIX), soluble epoxide hydrolase (sEH), and sirtuin 2 (SIRT2).[11] The results of these studies highlight the significant potential for streamlining drug discovery through the utilization of library-based screening techniques in conjunction with AI.

In recent years, transformer models[12] have become more and more prevalent in the field of natural language processing (NLP). Based on the transformer model, Devlin et al. introduced a new language representation model called bidirectional encoder representations from transformers (BERT), which produced the state-of-the-art results in a variety of NLP tasks.[13] Nowadays, with the advent of large models and an equally large amount of pre-training being done, BERT has played an important role in various areas, including drug development.[14,15] The BERT model features

Received: April 6, 2023
Published: July 22, 2023

https://doi.org/10.1021/acs.jmedchem.3c00627

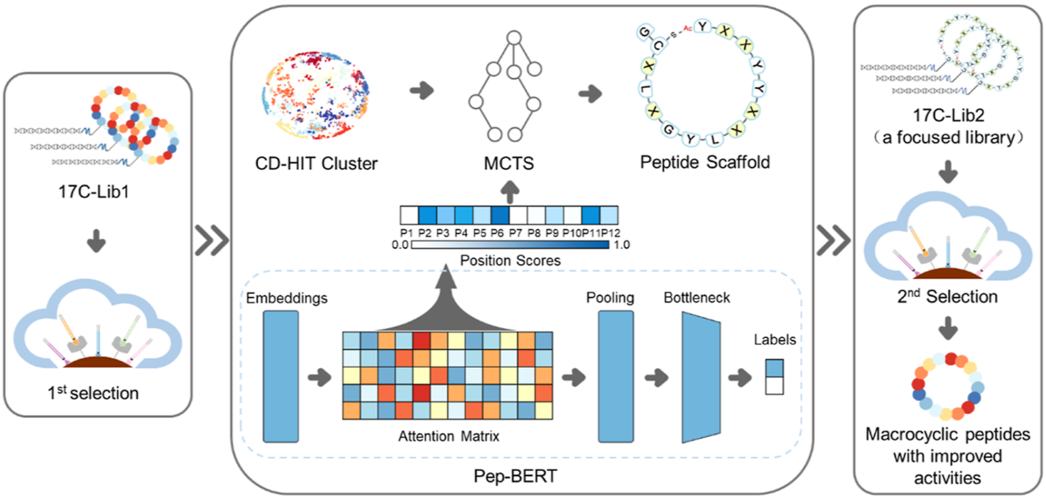


Figure 1. Schematic of the de novo macrocyclic peptides against IL-17C/IL-17RE interaction identified by the RaPID system integrated with the PepScaf framework. Left: the initial in vitro selection against IL-17C using 17C-Lib1. Middle: parallel clustering analysis and classification learning using Pep-BERT. Pep-BERT trained on the enrichment data obtained from the fourth round of 17C-Lib1 provided positional importance scores as the initial input for MCTS, which was employed in conjunction with clustering to generate the peptide scaffold. Right: the secondary selection against IL-17C was performed using 17C-Lib2 (a focused library that was constructed by taking the generated scaffold).

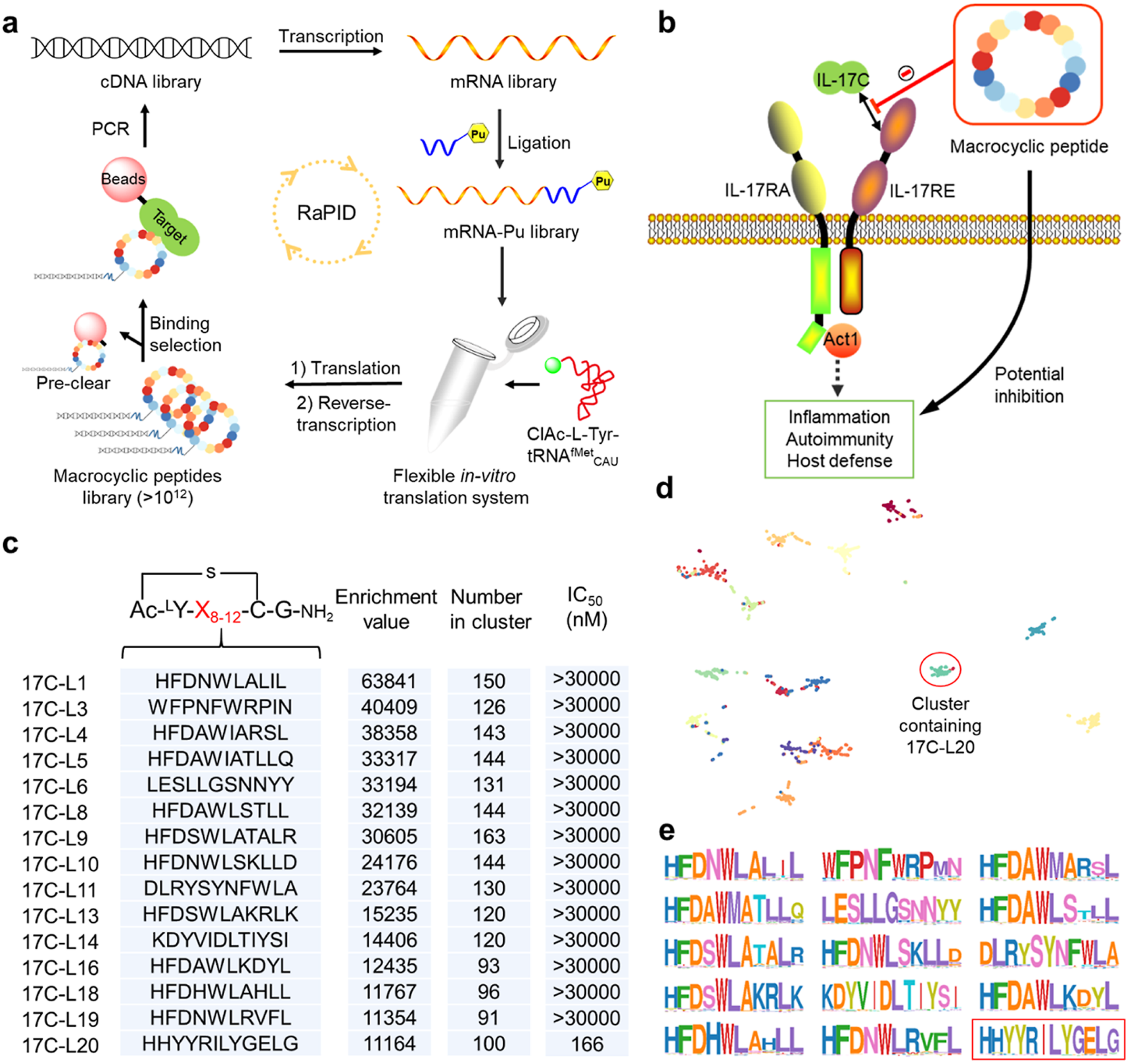
analysis on the basis of the mechanism of attention.[16−18] Attention weight, as a measure of importance, offers an easy and effective way to identify which elements are responsible for an output. According to the attention mechanism, a recent study proposed a program to search for mutant peptides with higher affinity for the target, human leukocyte antigen (HLA) allele.[19] While linear peptides have been the primary focus of previous studies, there is still significant untapped potential in the use of AI to assist in the discovery of macrocyclic peptide ligands, particularly through the integration of library-based screening methods.

Macrocyclic peptides have recently become an attractive modality for the development of therapeutics due to synthetic accessibility, high specificity, and tissue penetration, low toxicity, as well as the capability to block the protein−protein or protein−nucleic acid interaction.[20−25] Macrocyclic peptide libraries can be constructed by one-bead-one-compound (OBOC),[26] phage display,[27] split-intein circular ligation of peptides and proteins (SICLOPPS),[28] mRNA display,[29,30] etc. An excellent platform, referred to as the random non-standard peptide integrated discovery (RaPID) system, integrates the flexible in vitro translation (FIT) system[31,32] with mRNA display. It enables rapid selection of various de novo pseudo-natural peptide ligands from the thioether-closed macrocyclic peptide libraries with huge diversity (>10^{12} unique sequences) against the desired targets.[30,33,34] One desired target, interleukin-17C (IL-17C), a unique IL-17 cytokine family member, can specifically bind to interleukin-17 receptor E (IL-17RE) that is expressed on both epithelial cells and TH17 cells and signal by a heterodimeric receptor complex IL-17RA/RE. The downstream adaptor Act1 can then be recruited and induce the signaling pathways for autoimmunity, inflammation, host defense, etc.[35−39] Blocking the interaction between IL-17C and IL-17RE has shed light on the potential to treat autoimmune and inflammatory diseases, e.g., psoriasis and atopic dermatitis. However, there is no report of macrocyclic peptides that can block the IL-17C/IL-17RE interaction.

In this study, we report on the development of an integrated AI framework called PepScaf to direct the rational construction of a focused macrocyclic peptide library (17C-Lib2) by leveraging the vast dataset of sequence information generated from the initial in vitro selection campaign against IL-17C using a primary macrocyclic peptide library (17C-Lib1). A model based on BERT was adopted for training the enrichment data obtained from the fourth round of 17C-Lib1, aiming to obtain structural information about the activity and convert it into positional importance scores. Then, the Monte Carlo tree search (MCTS) algorithm was utilized to explore macrocyclic peptides in clustering data, for the purpose of generating a common scaffold found in a significant proportion of the peptide sequences. Using the scaffold that was generated, six critical positions in the middle of peptide sequences were fixed to construct the focused library, which was subsequently subjected to further selection against IL-17C by using the RaPID system. Finally, we obtained 20 peptides with IC_{50} values below 10 nM. In particular, the best two macrocyclic peptides exhibited their notable inhibitory activities against IL-17C/IL-17RE interaction with both IC_{50} values at 1.4 nM (Figure 1).

RESULTS AND DISCUSSION

In Vitro Selection of Macrocyclic Peptides against IL 17C/IL-17RE Interaction Using 17C-Lib1 by the RaPID System. In vitro selection by the RaPID system, which employs a macrocyclic peptide library containing over 10^{12} molecules, coupled with next-generation sequencing, can generate an extensive dataset of sequence information, precisely the kind of data required for machine learning. Thus, we first performed an in vitro affinity-based selection against a model target protein, IL-17C, by using the RaPID system (Figure 2a,b). An mRNA library consisting of AUG-$(NNK)_{8\text{-}12}$-UGC-$(GGC\text{-}AGC)_3$-UAG was designed, in which NNK was assigned to encode random sequences (N and K represent any of the four bases and U or G, respectively). The initiator AUG codon in the mRNA library was assigned to *N*-chloroacetyl-L-Tyr, which could simultaneously react with the cysteine (C) encoded by the downstream UGC to afford the thioether macrocyclic peptides. A glycine−serine (GS) triple-repeat peptide linker was encoded by the $(GGC\text{-}AGC)_3$ repeat



S
Ac-^{L}Y-X$_{8-12}$-C-G-NH$_2$

		Enrichment value	Number in cluster	IC$_{50}$ (nM)
17C-L1	HFDNWLALIL	63841	150	>30000
17C-L3	WFPNFWRPIN	40409	126	>30000
17C-L4	HFDAWIARSL	38358	143	>30000
17C-L5	HFDAWIATLLQ	33317	144	>30000
17C-L6	LESLLGSNNYY	33194	131	>30000
17C-L8	HFDAWLSTLL	32139	144	>30000
17C-L9	HFDSWLATALR	30605	163	>30000
17C-L10	HFDNWLSKLLD	24176	144	>30000
17C-L11	DLRYSYNFWLA	23764	130	>30000
17C-L13	HFDSWLAKRLK	15235	120	>30000
17C-L14	KDYVIDLTIYSI	14406	120	>30000
17C-L16	HFDAWLKDYL	12435	93	>30000
17C-L18	HFDHWLAHLL	11767	96	>30000
17C-L19	HFDNWLRVFL	11354	91	>30000
17C-L20	HHYYRILYGELG	11164	100	166

Figure 2. (a) Schematic of the RaPID system; (b) macrocyclic peptides against IL-17C/IL-17RE interaction; and (c) selected top 15 peptides: sequences, enrichment value in 17C-Lib1-4th, number in cluster, and IC$_{50}$ values as determined by ELISA. Each amino acid in the peptide sequence was labeled using the corresponding single-letter abbreviation, like elsewhere in the text. The IC$_{50}$ value (>30,000) suggests that the inhibitory activity of the macrocyclic peptide was weak, with less than half of the expected effect observed even at the highest concentration (30,000 nM). (d) Visualization of clustering in 17C-Lib1-4th by uniform manifold approximation and projection (UMAP). The target cluster containing 17C-L20 is highlighted and circled in red; Figure S2 provides a detailed representation of the clusters encompassing individual macrocyclic peptides. (e) Sequence logo of the amino acid frequency at each position of the selected top 15 peptide clusters (from top-left to bottom-right following the order decreasing enrichment). The target sequence logo is boxed in red.

codons, and the UAG stop codon was used to stall ribosome. After ligating a puromycin-CC-PEG-linker-DNA fragment with the mRNA library, the resulting mRNA-Pu library was subsequently added into the release factor 1 (RF1)-omitted FIT system to in vitro express a displayed macrocyclic peptide library, referred to as 17C-Lib1, in which the cognate genotype mRNA and the phenotype peptide were covalently linked via puromycin. The library was designed to boast a huge diversity (over 10^{12} unique sequences), allowing us to conduct affinity-based peptide screening against IL-17C immobilized on magnetic beads. The selection was performed for four rounds, and the enriched cDNA library in each round was separately analyzed by next-generation DNA sequencing (NGS).

Next, we built an enrichment dataset of the fourth round in 17C-Lib1 (17C-Lib1-4th), containing 365,680 valid macrocyclic peptides after data cleaning. According to the hit enrichment value of each macrocyclic peptide, the top 20 macrocyclic sequences were chosen for synthesis, and 15 peptides that met the desired purity criteria (>95%) were obtained (Figure 2c). Competitive enzyme-linked immunosorbent assay (ELISA) was then applied for determining the 50% inhibitory concentration (IC$_{50}$) values of the synthesized

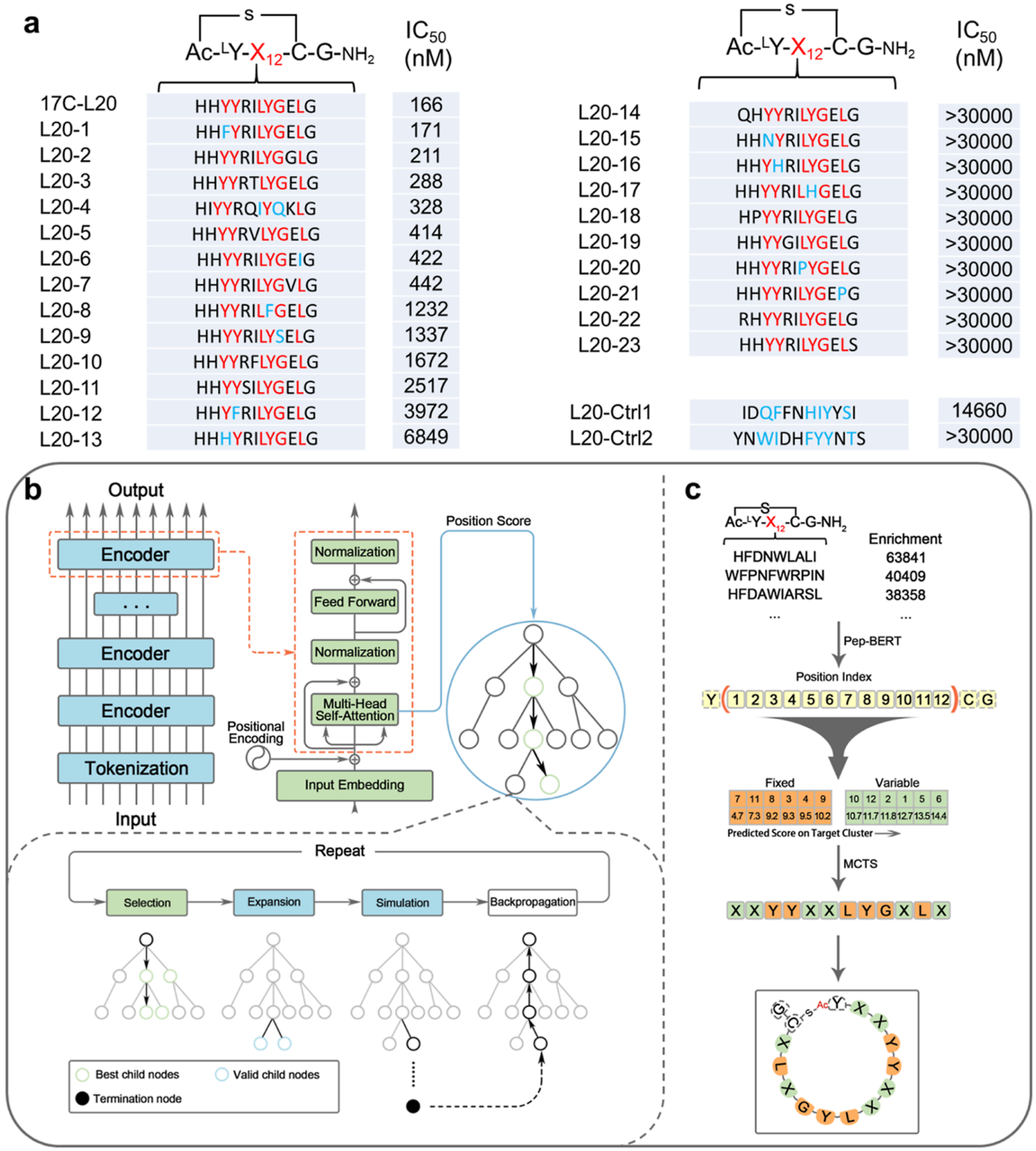


Figure 3. (a) Sequences and IC_{50} values of 17C-L20, selected top 23 peptides in the target cluster and the other two control peptides (L20-Ctrl1 and L20-Ctrl2) with distinct sequences. The amino acids generated by PepScaf are labeled in red, while distinct amino acids from this set are labeled in blue; (b) protocol of PepScaf: the architecture of Pep-BERT and the procedure of MCTS and (c) definition of the macrocyclic peptide scaffold. Pep-BERT performed importance ranking of 12 positions, with the first six positions selected as the fixed part and the subsequent six positions designated as the variable part based on the scores in this work. The MCTS generated corresponding amino acids at those positions, with "X" indicating variable amino acids (colored in green).

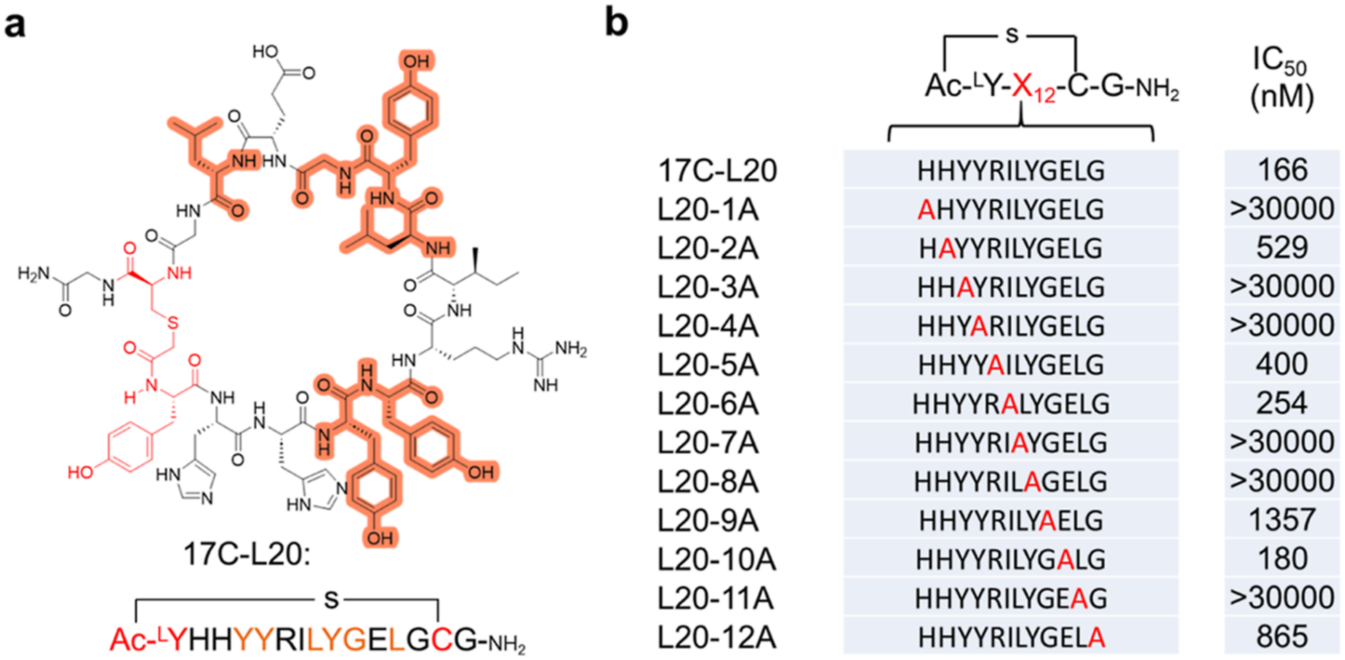
macrocyclic peptides against the interaction between IL-17C and IL-17RE. Unfortunately, only one macrocyclic peptide, 17C-L20, was demonstrated to be a potential inhibitor with an IC_{50} value of 166 nM. It was speculated that 17C-L20, despite its moderate potency, could precisely target the site responsible for disrupting the protein−protein interaction. In contrast, other peptides might bind to non-inhibitory locations. To obtain some peptides with improved bioactivity, we also tried to choose more peptides from 17C-Lib1-4th on the basis of the similarity to the top 15 hits. We first fished out all of the macrocyclic peptides with similar threshold (80%) to the selected top 15 peptides from 17C-Lib1-4th using the CD-HIT clustering tool[40] and subsequently gave a visualization of clustering by uniform manifold approximation and projection (UMAP).[41] It can be seen that the target cluster containing 17C-L20 (circled in red) is far away from the other clusters, suggesting certain structural specificity (Figures 2d and S2). In addition, the WebLogo tool[42] was used to statistic amino acid frequencies at every position of the clustered peptide sequences, and the sequence logos are shown in Figure 2e. It was found that the amino acid at each position of 17C-L20 had the highest frequency, suggesting that 17C-L20 might have already been selected as the most potential sequence in 17C-Lib1-4th against IL-17C/17-RE interaction. To check whether we could find other peptides with stronger inhibitory activities, we further chose, synthesized, and evaluated the other 23 peptides from the remaining 99 macrocyclic peptides in the same cluster as 17C-L20 (circled in red). While all of them exhibited lower activities than 17C-L20, two or three peptides demonstrated comparable activities to 17C-L20 as presented in Figure 3a.

Generation of the Critical Scaffold Related to the Bioactivity from the PepScaf Framework. With 17C-L20 as a hit in our hands, we next sought assistance from a focused library in order to further optimize the 17C-L20 hit ligand. A focused library constructed using the fragmented saturation mutagenesis approach by pooling together 49 sub-libraries, where each sub-library encodes the hit peptide interspersed with NNK codons, has been previously applied for peptide affinity maturation by the RaPID system.[43] This approach allowed for saturation mutagenesis at several different positions at the same time and finally afforded a peptide, PB1m6A9, which showed over 10-fold improvement in binding affinity for human PlxnB1 and also gained strong binding affinity to mouse PlxnB1.[43] In a different approach, we proposed to more rationally construct a focused library by integrating the power of AI tools using the dataset generated from the initial selection, under the context of lacking the cocrystal structures that could reveal how these molecules interact with protein surfaces. To do so, we proposed a framework termed PepScaf to direct the rational construction of a focused library. The scheme of PepScaf consisted of two main modules, a Pep-BERT classifier and the MCTS algorithm, which are depicted in Figure 3b.

The better and more intuitive way to find the most potential macrocyclic peptide hits is to build a quantitative structure−activity relationship (QSAR) model because it provides accurate predictions of measured end points instead of an independent ranking of biological activity.[36] Deep learning can effectively extract the desired chemical and physical features for a related task.[44] With respect to drug discovery, these features can include molecular structures, chemical properties, and other important characteristics. Additionally, models can search through large and complex chemical space,[45] which is particularly useful in the field of drug development where the optimization space can be discontinuous and challenging to navigate.[46] This means that deep learning can explore a wide range of possibilities and identify relationships between molecular structures and their biological activity, even in cases where the relationship is not immediately obvious. However, the success of such an approach heavily depends on the availability of a great amount of high-quality bioactivity data. Besides, the QSAR models lack the transferability due to the target-specific design of each approach.[47] Thus, it is difficult to build such an effective and practical QSAR model for a novel target by using a deep learning model. Alternatively, we could build a pseudo-QSAR model (enrichment values as metrics) according to the hypothesis that the top enriched peptides tend to more likely have high binding affinity.[48,49] The definition of the macrocyclic peptide scaffold is depicted in Figure 3c.

At first, we ranked the position index according to the position importance scores as provided by Pep-BERT, which was trained based on the enrichment data of each peptide in 17C-Lib1-4th. Since the start Y and the end CG were involved in ring formation, these three positions were not counted in the length calculation of the scaffold. Based on the scores at 12 positions as given by Pep-BERT, six positions with lower scores were specified as the fixed part (orange part in Figure 3c), and the variable part in green has positions with higher scores. The fixed positions and variable positions constituted a scaffold in a length of 12 (the position indexed from 1 to 12). Therefore, the valuable position information involving the activity was obtained after training on the vast enrichment data by a deep learning model, Pep-BERT.

Given the advantages of deep learning in feature extraction and the advantages of the algorithm as a chemical space exploration strategy, we further integrated the algorithm with deep learning in this work. A great number of tools[50−53] based on different classical algorithms such as the genetic algorithm[54] (GA) and the Monte Carlo (MC) method[55] have been applied for de novo molecular design over the past few decades. We first modified traditional algorithms including genetic algorithms to handle the discrete peptide sequences for the generation of the scaffold, but the results could not converge under limited time or computational resources. The possible reasons were speculated as follows: (1) the vast search space requires constraints to reduce the computation complexity, (2) data sensitivity of the traditional algorithms leads to the failing to handle the activity (enrichment) cliff[56] because the macrocyclic peptides have short and similar sequences (differ at only several amino acids), but the enrichment metrics can change greatly, and (3) the obscure characteristic of enrichment values because it is an indicator of confounding by multiple factors. We subsequently developed an effective and practical search-driven approach, leveraging the power of MCTS. Our approach focused on identifying amino acid residues that constitute the fixed part of the structure. To achieve this, we employed MCTS, guided by carefully designed growth rules aimed at maximizing the exploration of promising amino acids within the search space. These growth rules struck a delicate balance between exploration and exploitation, taking into account the generation of new amino acids as well as the utilization of those already generated. Through the incorporation of the reward function during backpropagation, the MCTS algorithm gradually learnt to prioritize peptide scaffolds that have a higher likelihood of satisfying a significant proportion of the peptides within the target cluster. Additionally, the efficacy of the scaffold could also be assessed by evaluating the number of peptides with low IC_{50} values that conform to the scaffold by adjusting its parameters and settings of MCTS. It was deduced that six positions comprising positions 3, 4, 7, 8, 9, and 11 seemed to be essential to the activity of 17C-L20, while the variable positions could be positions 1, 2, 5, 6, 10, and 12. Therefore, the scaffold



	$Ac\text{-}^{L}Y\text{-}X_{12}\text{-}C\text{-}G\text{-}NH_2$	IC_{50} (nM)
17C-L20	HHYYRILYGELG	166
L20-1A	AHYYRILYGELG	>30000
L20-2A	HAYYRILYGELG	529
L20-3A	HHAYRILYGELG	>30000
L20-4A	HHYARILYGELG	>30000
L20-5A	HHYYAILYGELG	400
L20-6A	HHYYRALYGELG	254
L20-7A	HHYYRIAYGELG	>30000
L20-8A	HHYYRILAGELG	>30000
L20-9A	HHYYRILYAELG	1357
L20-10A	HHYYRILYGALG	180
L20-11A	HHYYRILYGEAG	>30000
L20-12A	HHYYRILYGELA	865

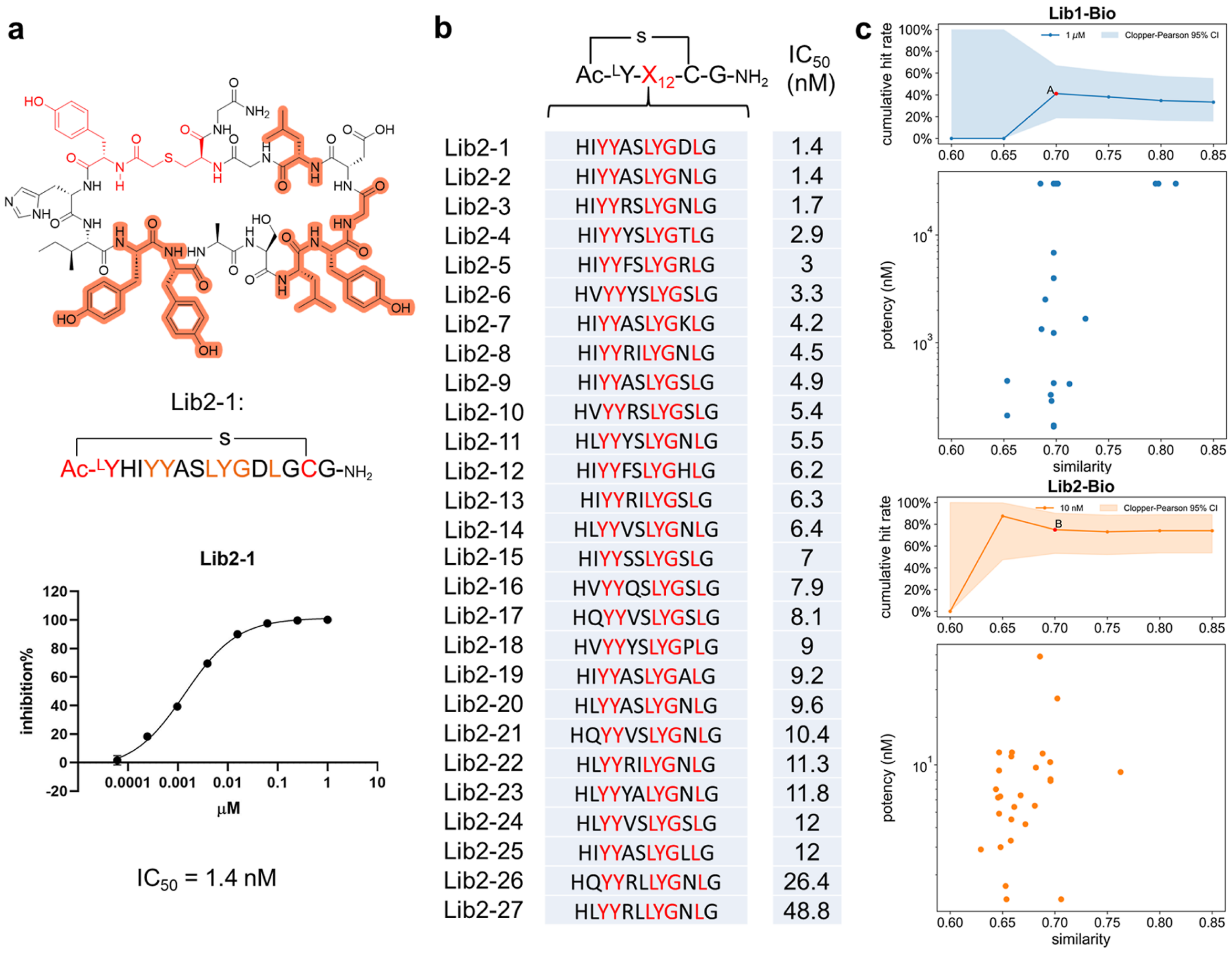
Figure 4. (a) Chemical structure of 17C-L20. The fixed amino acids generated by PepScaf are highlighted in orange. Amino acids involved in cyclization are labeled in red and (b) results of alanine scanning mutagenesis of 17C-L20.

(XXYYXXLYGXLX) was generated, where "X" represents a variable position that can be optimized.

Rationality Investigation of the Generated Critical Scaffold from the PepScaf Framework. Next, we investigated the rationality of the proposed scaffold by the PepScaf framework according to the laboratory results as listed in Figure 3a. With respect to the variable position, the replacement from *I* (17C-L20; IC_{50} = 166 nM) at position 6 to *T* (L20-3; IC_{50} = 288 nM) or *V* (L20-5; IC_{50} = 414 nM) showed only slightly decreased activities. This can also be seen by replacing position 10 (from *E* in 17C-L20 to *V* in L20-7). In contrast, position 3 was defined as a fixed position, at which the amino acid was replaced from *Y* (17C-L20) to *H* (L20-13; IC_{50} = 6849 nM) or *N* (L20-15; IC_{50} = 30,000 nM), causing the bioactivity cliffs. The activity cliffs can also be observed by replacing *Y* (17C-L20) at position 8 to *H* (L20-17; IC_{50} = 30,000 nM), *G* at position 9 to *S* (L20-9; IC_{50} = 1337 nM), and *L* at position 11 to *P* (L20-21; IC_{50} = 30,000 nM), suggesting the critical role of the amino acids in the fixed region to the bioactivity of 17C-L20. We additionally analyzed two peptides with distinct sequences from the scaffold, L20-Ctrl1 and L20-Ctrl2, whose IC_{50} values were 14,660 and 30,000, respectively. This also proved the effectiveness of the generated scaffold from the other aspect.

Before following the guidelines provided by this scaffold, we additionally apply alanine-scanning mutagenesis[57] to double-confirm the reliability of our scaffold. As can be seen from Figure 4b, the activity of 17C-L20 almost destroyed upon the replacement of amino acid at position 3, 4, 7, 8, or 11 to alanine (A), indicating that these positions are critical contributors to the overall activity. Apart from position 1, the substitution of residue position 2, 5, 6, 10, or 12 with A exhibited slight to moderate reduction of inhibitory activity. It should be noted that the replacement of the H residue at position 1 led to a poor inhibitory performance, while the module suggested that this position if variable. Despite the existence of misjudgment, the model results were almost consistent with the alanine scan results. This suggests the reliability of the PepScaf framework, which provides a "virtual alanine scanning" approach for identifying the crucial positions that can be used for constructing a focused library.[58] This approach holds potential for saving time and costs (refer to Tables S1−S3).

In Vitro Selection of Macrocyclic Peptides against IL-17C/IL-17RE Interaction Using a Focused Library (17C-Lib2) by the RaPID System and Data Analysis. Taking only one template containing the generated scaffold (Table S3), we constructed a focused library (17C-Lib2) and utilized the RaPID system once more to conduct in vitro selection against IL-17C. The remarkably higher recovery rates from first round to third round compared to the recovery rates of the equivalent rounds from the selection using the primary library also suggested the validity of the generated peptide scaffold by the PepScaf protocol (Figure S1). The isolated cDNA libraries were subjected to NGS, and 27 macrocyclic peptides were chosen based on their corresponding enrichment values and chemically synthesized (Figure 5a,b). The determined IC_{50} values are presented in Figure 5b. The IC_{50} values ranged from 1.4 to 48.8 nM, and 20 of 27 peptides exhibited notably improved inhibitory activities (IC_{50} < 10 nM) against IL-17C/IL-17RE interaction. The two most potent peptides, Lib2-1 and Lib2-2, exhibited over 100-fold improvements in activity compared to 17C-L20. Even the least active one in these 27 peptides, Lib2-27, was also demonstrated to be over 3 times more potent than 17C-L20. Overall, we concluded that the activity against IL-17C/IL-17RE was significantly improved by establishing a focused library under the guidance of AI. Our strategy to constructing a focused library with several critical residues, which is worth mentioning, might be more rational and efficient than pooling together 49 relatively random sub-libraries, as previously reported.[43] By adopting this new strategy, we could anticipate a significant reduction in the occurrence of undesired peptides within the focused library (Table S3). This holds great potential for enhancing the



	$Ac\text{-}^{L}Y\text{-}X_{12}\text{-}C\text{-}G\text{-}NH_2$	IC_{50} (nM)
Lib2-1	HIYYASLYGDLG	1.4
Lib2-2	HIYYASLYGNLG	1.4
Lib2-3	HIYYRSLYGNLG	1.7
Lib2-4	HIYYYSLYGTLG	2.9
Lib2-5	HIYYFSLYGRLG	3
Lib2-6	HVYYYSLYGSLG	3.3
Lib2-7	HIYYASLYGKLG	4.2
Lib2-8	HIYYRILYGNLG	4.5
Lib2-9	HIYYASLYGSLG	4.9
Lib2-10	HVYYRSLYGSLG	5.4
Lib2-11	HLYYYSLYGNLG	5.5
Lib2-12	HIYYFSLYGHLG	6.2
Lib2-13	HIYYRILYGSLG	6.3
Lib2-14	HLYYVSLYGNLG	6.4
Lib2-15	HIYYSSLYGSLG	7
Lib2-16	HVYYQSLYGSLG	7.9
Lib2-17	HQYYVSLYGSLG	8.1
Lib2-18	HVYYYSLYGPLG	9
Lib2-19	HIYYASLYGALG	9.2
Lib2-20	HLYYASLYGNLG	9.6
Lib2-21	HQYYVSLYGNLG	10.4
Lib2-22	HLYYRILYGNLG	11.3
Lib2-23	HLYYYALYGNLG	11.8
Lib2-24	HLYYVSLYGSLG	12
Lib2-25	HIYYASLYGLLG	12
Lib2-26	HQYYRLLYGNLG	26.4
Lib2-27	HLYYRLLYGNLG	48.8

Figure 5. (a) Chemical structure and dose−response data analysis of Lib2-1. The fixed amino acids generated by PepScaf are highlighted in orange. Amino acids involved in cyclization are labeled in red; (b) selected 27 peptides from the focused library (the amino acids colored in red are the fixed part) and their activities against IL-17C/IL-17RE interaction; and (c) cumulative hit rate plots and the scatter plot for bioactivity data. The Lib1-Bio data consisted of 17C-L20 and L20-1 to 23 (24 data points). The Lib2-Bio data are consisted of Lib2-1 to 27 (27 data points). The cumulative hit rate of Lib1-Bio at 1 μM is 41.2% (point A), and the cumulative hit rate of Lib2-Bio at 10 nM with a similarity of 0.7 is 75.0% (point B).

overall quality and specificity of the focused library, thereby facilitating the discovery of more specific macrocyclic peptides.

To better illustrate and compare the data of the bioactivities in Lib1-Bio (Figure 3a; bioactivities of 17C-L20, L20-1 to 23) and Lib2-Bio (Figure 5b; bioactivities of Lib2-1 to 27), we scattered these data points by Sokal similarity to the center of bioactivity data, as shown in Figure 5c. The center was virtually estimated by the *k*-means algorithm[59] according to the data of Lib1-Bio and Lib2-Bio. The plots showed the cumulative hit rate and potency of the peptides as a function of similarity (*x*-axis). More specifically, the above cumulative hit rate plots showed the hit rates of peptides with less than or equal to a given similarity and potency. The cumulative hit rate of Lib1-Bio at 1 μM is 41.2% (point A, 17 peptides tested), while the cumulative hit rate of Lib2-Bio at 10 nM with the same similarity (0.7) is 75.0% (point B, 24 peptides tested). We concluded that Lib2-Bio had a higher percentage of hit rates and tighter bioactivity restrictions as compared to Lib1-Bio. Since the data from both of Lib1-Bio and Lib2-Bio are distributed around a similarity of 0.7, it might be concluded that the focused library kept the fundamental structure (~70%) but mutated at some other critical positions, thereby leading to the significant improvements in bioactivity.

In addition to the distributions of the potency data of the peptide sequences in 17C-Lib1-4th and 17C-Lib2-3rd, we also analyzed the frequencies of 20 amino acids appeared at 12 positions (Figure 6a). In 17C-Lib1-4th, the three amino acids including L, T, and S were shown in dark blue, suggesting them as the three most frequent amino acids. In contrast, the dark blue areas in 17C-Lib2-3rd were lumps rather than rows, showing eight amino acids including 1H, 3Y, 4Y, 7L, 8Y, 9G, 11L, and 12G as the most frequent amino acids. This indicated that 1H and 12G at variable positions were also conserved and critical in 17C-Lib2-3rd apart from the other six fixed positions. Thus, it encouraged us to further analyze the amino acid frequencies at the six variable positions. As shown in Figure 6b, amino acids at variable positions 2, 5, 6, and 10 in 17C-Lib2-3rd were evenly distributed, and the top 10% macrocyclic peptides ranked by enrichment values exhibited a similar distribution to the entire peptide sequences in 17C-

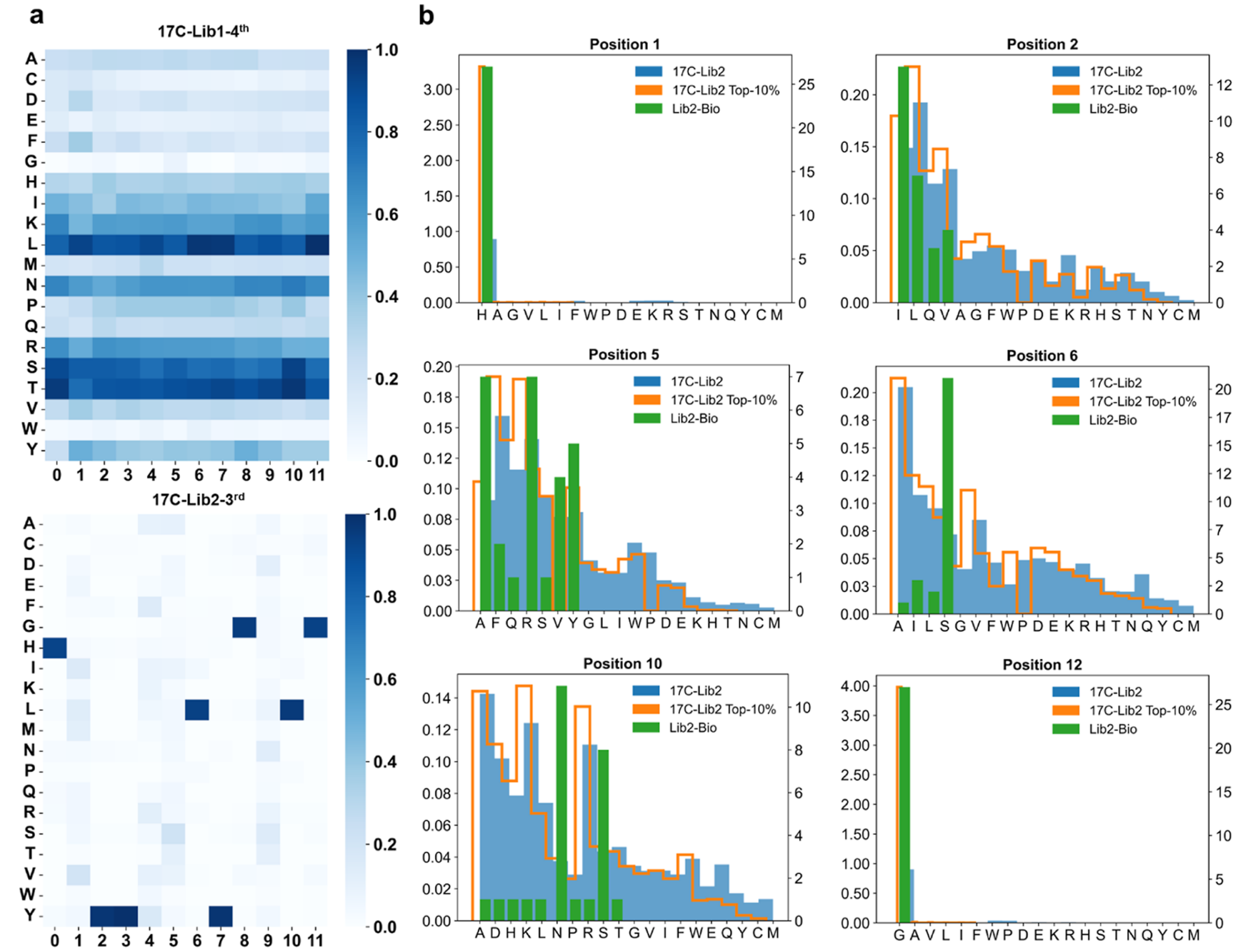


Figure 6. (a) Heat maps of amino acid frequencies at 12 positions of the peptides in 17C-Lib1-4th and 17C-Lib2-3rd. The min−max scaling technique was used to rescale the values within the range of 0−1 and (b) amino acid mutations at six variable positions. The histogram with blue fill or the step curve in orange graphs the probability density on the left *y*-axis for 17C-Lib2-4th or the top 10% of 17C-Lib2-3rd, respectively. The histogram with green fill graphs the frequencies of amino acids on the right *y*-axis.

Lib2-3rd. To a certain extent, the enrichment data of the top 10% peptides might be representative of the entire library. Besides, 1H and 12G remained unchanged during the selection process, although these two positions were proposed as variable. This discrepancy might arise from the fact that the correlation between the enrichment value and biological activity is not fully conclusive, even though there exists a certain level of correlation between them.[48,49] While the enrichment value serves as an indicator, it may not encompass all the nuances associated with biological activity. It is intriguing that these two amino acids reside in the neighborhood of the residues required for macrocyclic backbone formation. The replacements of these two amino acids might also induce a large conformation change of the peptide backbone, resulting in the reduction of activity. Taken together, four variable positions (2, 5, 6, and 10) in 17C-Lib2-3rd were amenable to mutation in the evolution for more potent peptides and consistent with our finding that the fundamental structure (∼70%) was retained during the selection process.

■ CONCLUSIONS

In conclusion, we developed an integrated AI framework called PepScaf, allowing us to obtain the critical scaffold relative to the bioactivity on the basis of the vast dataset from the initial in vitro selection campaign against a model protein target, IL-17C. Based on the scaffold generated by PepScaf, a focused library was constructed and applied for macrocyclic peptide selection against IL-17C again by using the RaPID system. This afforded us with 20 biologically active macrocycles against IL-17C/IL-17RE interaction with IC_{50} values below 10 nM, of which the best two cyclic peptides exhibited their notable inhibitory activities with both IC_{50} values at 1.4 nM. It is expected that the AI-based PepScaf framework might offer time and cost savings as compared to some other approaches, e.g., alanine scanning, providing a feasible and more efficient methodology toward macrocyclic peptide ligand screening. Moreover, this is the first report regarding the discovery of macrocyclic peptides against IL-17C/IL-17RE interaction. Additionally, our methodology not only relied on the enrichment data from the filtering pool but also combined wet experimental data with AI to find potential macrocyclic peptide ligands. While our testing was limited to a model

https://doi.org/10.1021/acs.jmedchem.3c00627

protein (IL-17C), this ligand/sequence-based (target-independent) strategy may hold broader applicability for other targets and diverse methodologies, including phage display. Thus, our methodology seems to be more in accordance with real-world scenarios as compared to the AI-only approach. The combination of wet and dry experiments may be the future trend to make AI tools become more mature and available to scientists.

■ EXPERIMENTAL SECTION

General Information. The enzymes and kits used for RaPID selections, including T4 RNA ligase and PURExpress kit (New England Biolabs; NEB), M-MLV reverse transcriptase RNase H Minus (Promega), Taq DNA polymerase and RNase inhibitor (ABclonal; Wuhan, China), and T7 RNA polymerase (NovoBiotechnology Co., Ltd.; Beijing, China), were purchased and used as received. Amino acids used for ribosomal synthesis were supplied by Sangon Biotech (Shanghai, China). Primers were synthesized by Genscript Biotech Corporation (Nanjing, China). Dynabeads M-280 Streptavidin was purchased from Thermo Fisher Scientific. Biotinylated IL-17C and IL-17RE were purchased from Acrobiosystems. Bovine serum albumin (BSA) was purchased from Beyotime. Streptavidin-conjugated horseradish peroxidase (streptavidin-HRP) and the 3,3′,5,5′-tetramethylbenzidine (TMB) substrate were supplied by Thermo Fisher Scientific. Rink Amide MBHA resin used for the chemical synthesis of macrocyclic peptides was purchased from Sunresin New Materials. Fmoc-protected amino acids (Fmoc-AA-OH) were supplied by Bide Pharmatech (Shanghai, China). The reagents, such as 1-hydroxybenzotriazole (HOBT), *N*,*N*′-diisopropylcarbodiimide (DIC), ethyldiisopropylamine (DIPEA), dimethylformamide (DMF), dichloromethane (DCM), diethylether (Et_2O), trifluoroacetic acid (TFA), acetonitrile (MeCN), dimethyl sulfoxide (DMSO), triisopropylsilane (TIS), piperidine triethylamine (Et_3N), and 1,2-ethanedithiol (EDT), were purchased from Highfine Biotech (Suzhou, China), Qiangsheng Chemical (Shanghai, China), Aladdin, or Macklin Biochemical Technology (Shanghai, China).

Selections against IL-17C by the RaPID System. Prior to the selections, ClAc-^{L}Tyr-tRNA$^{fMet}{}_{CAU}$ was prepared as previously reported.[31,32,60] The initial in vitro selection of macrocyclic peptides against IL-17C was conducted using a ClAc-^{L}Tyr-initiated macrocyclic peptide library (17C-Lib1) by the RaPID system. For the first round, a library of mRNA (AUG-$(NNK)_{8\text{-}12}$-$(GGC\text{-}AGC)_3$-UAG) was ligated with a DNA linker with puromycin (DNA-PEG-CC-Pu) using T4 RNA ligase to give the mRNA-Pu library, which was directly used for the subsequent translation in the RF1 (release factor 1) and Met (methionine)-deleted FIT (flexible in vitro translation) system containing 50 μM ClAc-^{L}Tyr-tRNA$^{fMet}{}_{CAU}$. The resulting solution (25 μL) was successively incubated at 37 °C for 2 h, 25 °C for 15 min, and 37 °C for 30 min after adding 5 μL of 100 mM EDTA (pH 8.0). Subsequently, a reverse transcription solution (25 mM Tris−HCl pH 8.3, 15 mM $Mg(OAc)_2$, 10 mM KOH, 0.25 mM dNTPs, 2 μM CGS3an13.R39, and 50 U M-MLV reverse transcriptase RNase H Minus) containing the RNase inhibitor was prepared by mixing the above reaction mixture and further incubated at 42 °C for 60 min. After adding the equal amount of blocking solution, the resulting solution was rotationally incubated at 4 °C for 60 min in the presence of IL-17C immobilized Dynabeads M-280 Streptavidin. Then, the solution was removed, and the beads were washed thrice with selection buffer. The cDNAs enriched with IL-17C binding sequences on the beads were eluted with 1× PCR mix by heating at 95 °C for 6 min, and the collected cDNAs in the supernatant were analyzed by the QuantStudio real-time PCR system (Applied Biosystems). Finally, amplification by PCR afforded the cDNAs, which were used for the in vitro transcription to produce the mRNA library for the second round. Starting from second round, the translation scale could be reduced to 5 μL, and pre-clearing steps were performed six times using the beads without IL-17C to remove the undesired bead binders.

To select macrocyclic peptides against IL-17C using a focused library, an mRNA library (AUG-$(NNK)_2$-TATTAT-$(NNK)_2$-CTTTATGGT-NNK-CTT-NNK-TGC-$(GGC\text{-}AGC)_3$-UAG) was prepared based on the generated scaffold in this study. The RaPID technology, described earlier, was employed for the selection process. The cDNAs enriched with IL-17C binding sequences were subjected to deep sequencing using the NovaSeq 6000 system (Ilumina). The recovery rate histograms of selections against IL-17C using 17C-Lib1 and 17C-Lib2 are shown in Figure S1.

Chemical Synthesis of the Selected Macrocyclic Peptides. The selected thioether-cyclized macrocyclic peptides were synthesized by the standard Fmoc solid-phase peptide synthesis (SPPS). Rink amide MBHA resin (0.5 g), suitable for C-amide peptide synthesis, was suspended in a freshly prepared solution (0.3 mmol of Fmoc-Gly-OH and 0.3 mmol of HOBT in 8 mL of DMF, along with 0.5 mL of DIC). The reaction proceeded for 1.5 h under nitrogen gas bubbling conditions. After filtration, the resin was washed with DMF and DCM at least three times, and the Fmoc group was removed using 20% piperidine in DMF. The coupling reactions of the subsequent amino acids were conducted by adding a freshly prepared solution (0.9 mmol Fmoc-AA-OH and 0.9 mmol HOBT in 10 mL of DMF, along with 1 mL of DIC). The reaction proceeded for 1 h under nitrogen gas bubbling conditions. The desired peptide length was achieved by repeating the deprotection and coupling steps. Next, the bromoacetyl group was attached to the free N-terminal α-amino group of the peptides on the resin.

The synthesized peptides on the resin were then cleaved from the resin, precipitated with Et_2O, and redissolved in DMSO. The thioether-cyclized peptides were obtained by adjusting the solution to pH 8.0, followed by incubation for 1 h. Finally, the solution was adjusted to pH 3−4 using TFA and purified by reverse phase HPLC with a mobile phase consisting of a 0.1% TFA aqueous solution and MeCN containing 0.1% TFA, under linear gradient conditions. The purity of the peptides was confirmed by an LC-2020 (Shimazu), and the mass spectra were recorded using an LCMS-2020 (Shimazu). All macrocyclic peptides exhibited a purity of >95% according to HPLC analysis, and the HPLC traces for all compounds have been included in the Supporting Information.

Evaluation of the Selected Macrocyclic Peptides against the IL-17C/IL-17RE Interaction. Competitive enzyme-linked immunosorbant assay (ELISA) was used for determining the IC_{50} (50% inhibitory concentration) values of the selected macrocyclic peptides against the interaction between IL-17C and IL-17RE. In brief, IL-17RE was first coated on the 96-well ELISA plate by adding 80 μL of IL-17RE (1 μg/mL) to each well, followed by overnight incubation at 4 °C. Following immobilization, the wells were separately washed four times with 150 μL of 1× PBST buffer and then blocked for 1 h at RT with 100 μL of 1× PBST buffer containing 2% BSA. After washing the wells four times again using 150 μL 1× PBST buffer, 100 μL of a freshly prepared mix of biotinylated IL-17C (0.5 nM) and each macrocyclic peptide at eight different concentrations was added to the separate wells of the IL-17RE-coated plate and incubated for another 1.5 h at RT. The wells were then washed four times with 150 μL of 1× PBST buffer and incubated with 150 μL of streptavidin-HRP solution (1:1000 dilution in 1× PBS) for 1 h at RT. After another round of washing, 100 μL of 3,3′,5,5′-tetramethylbenzidine (TMB) solution was added to each well, and the color development was allowed to proceed for 10 min at RT. The reaction was finally quenched by ELISA stop solution (Absin) and spectrophotometrically measured at 450 nm using a Tecan Spark multimode reader. The IC_{50} values were calculated by fitting the inhibition (%) data at each concentration of the macrocyclic peptides by GraphPad Prism 6 software.

Data Preprocessing. In the 17C-Lib1 dataset, a total of 761,445 raw experiment datapoints were generated. To ensure the quality and validity of the data, certain criteria were applied during the filtering process. In the experimental setting, it was required that macrocyclic peptide sequences begin with "M" and end with "CGSGSGSamber" in order to be considered valid. Any peptides deviating from this specified format were excluded from the dataset.

By library-based screening and implementing these filtering steps, we obtained 365,682 valid data (17C-Lib1-4th) generated for machine learning modeling. The minimum enrichment value (identical to DNA sequencing reads) of macrocyclic peptides in 17C-Lib1-4th is 1, which is interpreted as negligible binding ability. The peptides with low enrichment values (<10) take up to 99.4%. For the convenience of analysis, we focused on peptides consisting of 12 residues between ClAc-^{L}Y and C because the hit peptide has a length of 12. To balance the dataset, we randomly generated negative data points and divided it into train and test dataset at a ratio of 9:1 for the training of Pep-BERT. In our work, the peptide sequences with enrichment values were considered as positive data. To get the target cluster of peptides, the CD-HIT[40] was used to cluster the peptide sequences according to the similarity of amino acid frequency where all positive samples in 17C-Lib1 were clustered at a threshold of 80% similarity.

Design of PepScaf. Intuitively, the residue and its position of a macrocyclic peptide contributed to the performance in binding to IL-17C. If we fix those residues contributing the most to binding, the peptide is more likely to retain the affinity to the target. In this work, the length of the scaffold was set to 12, with six positions designated as the fixed part and the other six as the variable part. Namely, we assumed that the macrocyclic peptides containing those six residues in fixed positions could retain the inhibitory activity. Each of the other six positions could be replaced with one of the 20 native amino acid residues, following the hypothesis that the changes of variable positions are tolerated, and the activity will be maintained. The lengths of the fixed and variable parts can be adjusted based on the researcher's experience and preference. However, it is important to consider that increasing the length of the fixed part will result in a smaller library size, and extending the length of the variable part may alter the binding site of the macrocyclic peptide.

The attention score revealed the key amino acid sites of peptide sequence that were essential for binding or non-binding to the target. Also, it determined which part each position belongs to. Initialized with this attention score, MCTS was trained on the target cluster dataset. Depending on the recall back reward, MCTS first calculated the priority level of different positions and decided six crucial positions that might affect the foundational binding ability of macrocyclic peptides against IL-17C. Then, it determined the residue types on those positions depending on the contribution of 20 native amino acids. Finally, MCTS outputted a text format of the target scaffold, which was regarded as a paradigm that is viable to be explored as a potential macrocyclic peptide with high binding ability.

It should be noted that the length of the target scaffold and the number of the fixed positions can be simply altered to fit different tasks. After adding an additional null amino acid character/token to align the peptides, PepScaf becomes capable of processing peptides of different lengths using the modified encoding methods. Our AI-based method can be applied to other similar tasks, guaranteeing the generality of our scheme.

Definition of the Peptide Scaffold. In our task, each position provides additional information about its variability. To capture this variability, we divided all positions into three parts, with two of them used for our task. The variability of each position was determined by its position score, which was calculated from the attention matrix. The visualizations of the attention matrix are presented in Figure S2.

$$V_{\text{pos}} = \text{Norm}\left(\sum^{N_{\text{layer}}}\sum^{N_{\text{head}}}\sum^{N_{\text{token}}}(\mathbf{x}_{\text{pos}} + \mathbf{y}_{\text{pos}})\right)$$

We denoted a vocabulary of 20 canonical native amino acids using the symbol $\mathcal{A}$.

$$\mathcal{A} = \{L, M, P, K, \cdots, A\} \to \{a_1, a_2, a_3, \cdots, a_{20}\}$$

where $L, M, P, ..., A$ represent the single-letter abbreviations for the 20 amino acids, and $\mathcal{A}_m (m \le 20)$ represents a subset consisting of m amino acids.

The generation of scaffolds can be formulated mathematically using the Markov Decision Process (MDP)[61] since the generated molecule depends on the molecule being modified. Therefore, we began the generation of scaffolds using MCTS, which is well-suited for this MDP case. A peptide scaffold can then be defined as follows:

$$\begin{aligned}
\text{Scaf}_{v,o,f} &= \text{Scaf}_v, \text{Scaf}_o, \text{Scaf}_f \\
\text{Scaf}_v &= \text{Set}(V_{i,\mathcal{A}_{20}}) \qquad i \in [0, v - 1] \\
\text{Scaf}_o &= \text{Set}(O_{j,\mathcal{A}_m}) \qquad j \in [v, v + o - 1] \\
\text{Scaf}_f &= \text{Set}(F_{k,\mathcal{A}_1}) \qquad k \in [v + o, v + o + k - 1]
\end{aligned}$$

where $\boldsymbol{v}$, $\boldsymbol{o}$, and $\boldsymbol{f}$ are the lengths of variable part, optional part, and fixed part, respectively. $\boldsymbol{V}$, $\boldsymbol{O}$, and $\boldsymbol{F}$ are the building blocks of the peptide scaffold, which includes position and content information. The first parameter specifies the position, while the second parameter specifies the content (residue) at that position. In the variable part, any of the 20 native amino acids can be used as candidates. In the optional part, a specific number of amino acids must be used. In the fixed part, only a single amino acid can be used as its content.

As an example, we considered the target cyclic peptide (HHYYRILYGELG) and set the length of the variable part as 6, the length of the optional part as 0, and the length of the fixed part as 6. The scaffold can be defined as follows:

$$\begin{aligned}
\text{Scaf}_{6,0,6} &= V_{\text{part}}, F_{\text{part}} \\
&= V_0, V_1, \cdots, V_5, F_6, F_7, \cdots, F_{11}
\end{aligned}$$

The scaffold spans a peptide space with a scale of 20^6, while the original scale is 20^{12}. In addition, the scaffold is visualized in Figure 3c.

Pep-BERT. Our sequence-based Pep-BERT model was used to extract the position scores of macrocyclic peptides by analyzing the attention matrix and primarily address the following two tasks: (1) a binary classification task to predict the interaction between peptides and the target and (2) a position ranking task to determine the most contributing position from the inputted peptide cluster. Pep-BERT was modified from BERT. To better adapt to the task of peptide classification, we cut down the parameters of the original BERT mainly by reducing the number of layers and the size of vocabulary. Details on other hyperparameters can be found in the code.

Encodings. The unknown amino acid was substituted with the "[UNK]" character, which is not included in the amino acid alphabet. The encoding dictionary also included the special tokens in the original BERT such as "[PAD]", "[CLS]", "[SEP]", and "[MASK]". In deep learning applications, discrete amino acids are generally represented as continuous vectors through an embedding matrix. Here, we utilized relative positional embedding[12] to encode the position of the amino acid in the sequence.

Training. The dataset used for training consisted of 200 K macrocyclic peptides, which were balanced and divided for supervised learning. By training on this balanced dataset, Pep-BERT was able to learn the latent information about the diverse structures of macrocyclic peptides. The resulting test accuracy of the model was 80%, and the area under the receiver operating characteristic (AUROC) score was 88% (refer to the accuracy and AUROC curves shown in Figure S4). To optimize the model during training, cross-entropy loss was utilized.

MCTS. MCTS serves as the foundation for AlphaGo[62] and has proven to be instrumental in numerous successful AI applications.[63,64] It has also demonstrated satisfactory performance in the drug field.[65] As a heuristic algorithm, MCTS employs Monte Carlo simulation[66] to generate value estimates and guide searches toward rewarding trajectories in the search tree. In essence, MCTS prioritizes plausible amino acid nodes at specific positions, instead of exhaustively exploring all possibilities, resulting in a significant reduction in computing costs.

In our work, MCTS was used to combine the knowledge obtained from Pep-BERT to generate a scaffold of macrocyclic peptides. As illustrated in Figure 3b, MCTS comprises repeated iterations of four steps: selection, expansion, rollout/simulation, and backpropagation: (1) selection: starting at the root node, successive child nodes are recursively selected until a leaf node is reached. (2) Expansion: following a policy, candidate child nodes are enumerated until a leaf node ends this generation. Then, the best node is chosen from those child nodes. (3) Simulation: the final reward can be gained after finishing a generation that is one random rollout from a node. (4) Backpropagation: from the root node, the reward is recursively accumulated to update the information of visited nodes along the trace.

It is worth noting that the selection step, the core of MCTS, tends to deep exploration in the process of generating the scaffold, which requires much computing time. Nonetheless, MCTS can terminate when it reaches a specified computational limit (iteration times) or temporal limit (searching time). Here, the upper confidence bound[67] (UCB1) was introduced to MCTS for moderating the conflict between the exploitation of deep variants after moves with high average win rate and the exploration of moves with few simulations.

UCB1 balances between exploitation and exploration to avoid being trapped in local optimums. In this study, UCB1 is a scalar and maximization picks the node with the largest value. The Algorithm S5 gives sufficient information about the MCTS algorithm and the UCB1 formula.

Molecular Fingerprint Similarity of Macrocyclic Peptides. The Sokal similarity[68] is a widely used measure for quantifying the similarity between two binary datasets, frequently applied in fields like biology, chemistry, and information retrieval. It enables comparison of the presence or absence of specific attributes or features. Figure 5c shows that the shared *X*-axis was organized based on peptide similarity, with molecular fingerprint similarity calculated using the RDKit tool[69] utilizing the Sokal similarity metric. Additionally, the center peptide sequence of the data (a combined datasets of Lib1-Bio and Lib2-Bio) was identified using the *k*-means algorithm. By employing this approach, we guarantee that the peptide data points in Figure 5c are well distributed and easily discernible.

Probability Density. Figure 6b presents a series of statistics on the amino acid probability density at different positions, describing the distribution of optimized amino acids at the six variable positions. For instance, in the first plot, the blue-filled histogram illustrates the occurrence of 20 amino acids at position 1 (12-mer cyclic peptide) in the primary library (17C-Lib1). The occurrences are then normalized to obtain the probability density distribution of different amino acids.

Since the primary library contains a large number of peptides with an enrichment value of 1, we further analyzed the top 10% of peptides with higher enrichment values (orange step curve histogram). The distribution is similar to that in the blue histogram, indicating that the highly enriched peptides have better data quality. Due to limited activity data and the possibility of incomplete occurrence of all the 20 amino acids at certain positions, the right axis describes the frequency of amino acid occurrence at different positions in the library.

■ ASSOCIATED CONTENT

Data Availability Statement

The source Python code for data analysis, trained model weights for Pep-BERT and MCTS, and instructions to reproduce the work can be found at https://github.com/hongliangduan/PepScaf.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.3c00627.

> Recovery rate histograms of selections against IL-17C using 17C-Lib1 and 17C-Lib2 by means of the RaPID system; detailed representation of the clusters encompassing individual macrocyclic peptides; visualizations of multi-head attention matrix for Pep-BERT; receiver operating characteristic curve of 17C-Lib1 and cluster dataset; algorithm; simplified algorithm diagram illustrating the core steps of the MCTS algorithm; direct comparison of the PepScaf framework with alanine scanning mutagenesis used for identifying the crucial positions; direct comparison of the PepScaf framework with a fragmented saturation mutagenesis approach used for constructing a focused library; direct comparison of the templates used for constructing a focused library by employing the scaffold generated from PepScaf or fragmented saturation mutagenesis approach (L20-O1 to O48); and characterization of the synthesized macrocyclic peptides by ESI-MS and analytical HPLC (S13) (PDF)
>
> Molecular formula strings and the associated biological data (CSV)

■ AUTHOR INFORMATION

Corresponding Authors

Hongliang Duan − *School of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China*; Email: hduan@zjut.edu.cn

Yizhen Yin − *State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China; Shandong Research Institute of Industrial Technology, Jinan 250101, China*; orcid.org/0000-0003-2466-2167; Email: yizhenyin.1987@sdu.edu.cn

Authors

Silong Zhai − *School of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China*

Yahong Tan − *State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China*

Chengyun Zhang − *School of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China*

Christopher John Hipolito − *Screening & Compound Profiling, Quantitative Biosciences, Merck & Co., Inc., Kenilworth, New Jersey 07033, United States*

Lulu Song − *State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China*

Cheng Zhu − *School of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China*

Youming Zhang − *State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.3c00627

Author Contributions

$^{\perp}$S.Z. and Y.T. contributed equally to this work. H.D. and Y.Y. conceived and supervised the research project. S.Z., C.Z. (Chengyun Zhang) and C.Z. (Cheng Zhu) performed the dry experiments and analyzed the data; Y.T. and L.S. performed the wet experiments, e.g., RaPID selections. S.Z., Y.T., H.D., and Y.Y. wrote the manuscript. C.J.H. and Y.Z. proofread and discussed the manuscript. All of the authors have edited and reviewed the manuscript.

Funding

This work was financially supported by the National Natural Science Foundation of China (NSFC) (nos. 22277065 and

81903425), the Natural Science Foundation of Zhejiang Province (LD22H300004), the Taishan Scholar Program of Shandong Province (no. tsqn202103148), and the Qilu Youth Scholar Startup Funding of Shandong University (to Yizhen Yin).

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

AI, artificial intelligence; AUROC, area under the receiver operating characteristic; BERT, bidirectional encoder representations from transformers; BSA, bovine serum albumin; CAIX, carbonic anhydrase; DCM, dichloromethane; DDR1, discoidin domain receptor-1; DEL, DNA-encoded chemical library; DIC, *N*,*N*′-diisopropylcarbodiimide; DIPEA, ethyldiisopropylamine; DMF, dimethylformamide; DMSO, dimethyl sulfoxide; EDT, 1,2-ethanedithiol; ELISA, enzyme-linked immunosorbent assay; Erα, estrogen receptor alpha; Et_2O, diethylether; Et_3N, piperidine triethylamine; FIT, flexible in vitro translation; Fmoc-AA-OH, Fmoc-protected amino acids; GA, genetic algorithm; HLA, human leukocyte antigen; HOBT, 1-hydroxybenzotriazole; IC_{50}, the 50% inhibitory concentration; IL-17C, interleukin-17C; IL-17RE, interleukin-17 receptor E; MC, Monte Carlo; MCTS, Monte Carlo tree search; MeCN, acetonitrile; NGS, next-generation DNA sequencing; NLP, natural language processing; OBOC, one-bead-one-compound; Pep-BERT, BERT model for peptide classification; PPAR, peroxisome proliferator-activated receptor; QSAR, quantitative structure−activity relationships; RaPID, random non-standard peptide-integrated discovery; RF1, release factor 1; RNN, recurrent neural networks; RXR, retinoid X receptor; sEH, soluble epoxide hydrolase; SICLOPPS, split-intein circular ligation of peptides and proteins; SIRT2, sirtuin 2; SPPS, solid-phase peptide synthesis; streptavidin-HRP, streptavidin-conjugated horseradish peroxidase; TFA, trifluoroacetic acid; TIS, triisopropylsilane; TMB, 3,3′,5,5′-tetramethylbenzidine; UCB, upper confidence bound; UMAP, uniform manifold approximation and projection

■ REFERENCES

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203−214.

(2) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20−33.

(3) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; Zvonok, N.; Jain, M. K.; Savych, O.; Radchenko, D. S.; Nikas, S. P.; Petasis, N. A.; Moroz, Y. S.; Roth, B. L.; Makriyannis, A.; Katritch, V. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature* **2022**, *601*, 452−459.

(4) Li, Z.; Li, X.; Huang, Y.-Y.; Wu, Y.; Liu, R.; Zhou, L.; Lin, Y.; Wu, D.; Zhang, L.; Liu, H.; Xu, X.; Yu, K.; Zhang, Y.; Cui, J.; Zhan, C.-G.; Wang, X.; Luo, H.-B. Identify Potent SARS-CoV-2 Main Protease Inhibitors via Accelerated Free Energy Perturbation-Based Virtual Screening of Existing Drugs. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 27381−27387.

(5) Gentile, F.; Yaacoub, J. C.; Gleave, J.; Fernandez, M.; Ton, A.-T.; Ban, F.; Stern, A.; Cherkasov, A. Artificial Intelligence−Enabled Virtual Screening of Ultra-Large Chemical Libraries with Deep Docking. *Nat. Protoc.* **2022**, *17*, 672−697.

(6) Harris, P. A.; King, B. W.; Bandyopadhyay, D.; Berger, S. B.; Campobasso, N.; Capriotti, C. A.; Cox, J. A.; Dare, L.; Dong, X.; Finger, J. N.; Grady, L. C.; Hoffman, S. J.; Jeong, J. U.; Kang, J.; Kasparcova, V.; Lakdawala, A. S.; Lehr, R.; McNulty, D. E.; Nagilla, R.; Ouellette, M. T.; Pao, C. S.; Rendina, A. R.; Schaeffer, M. C.; Summerfield, J. D.; Swift, B. A.; Totoritis, R. D.; Ward, P.; Zhang, A.; Zhang, D.; Marquis, R. W.; Bertin, J.; Gough, P. J. DNA-Encoded Library Screening Identifies Benzo[b] [1,4]Oxazepin-4-Ones as Highly Potent and Monoselective Receptor Interacting Protein 1 Kinase Inhibitors. *J. Med. Chem.* **2016**, *59*, 2163−2178.

(7) Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D. DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space. *Nat. Rev. Drug Discovery* **2017**, *16*, 131−147.

(8) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform.* **2018**, *37*, 1700153.

(9) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(10) McCloskey, K.; Sigel, E. A.; Kearnes, S.; Xue, L.; Tian, X.; Moccia, D.; Gikunju, D.; Bazzaz, S.; Chan, B.; Clark, M. A.; Cuozzo, J. W.; Guié, M.-A.; Guilinger, J. P.; Huguet, C.; Hupp, C. D.; Keefe, A. D.; Mulhern, C. J.; Zhang, Y.; Riley, P. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. *J. Med. Chem.* **2020**, *63*, 8857−8866.

(11) Lim, K. S.; Reidenbach, A. G.; Hua, B. K.; Mason, J. W.; Gerry, C. J.; Clemons, P. A.; Coley, C. W. Machine Learning on DNA-Encoded Library Count Data Using an Uncertainty-Aware Probabilistic Loss Function. *J. Chem. Inf. Model.* **2022**, *62*, 2316−2331.

(12) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in neural information processing systems*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. *30*.

(13) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. **2018**, arXiv preprint arXiv:1810.04805.

(14) Charoenkwan, P.; Nantasenamat, C.; Hasan, M. M.; Manavalan, B.; Shoombuatong, W. BERT4Bitter: A Bidirectional Encoder Representations from Transformers (BERT)-Based Model for Improving the Prediction of Bitter Peptides. *Bioinformatics* **2021**, *37*, 2556−2562.

(15) Zhang, X.-C.; Wu, C.-K.; Yang, Z.-J.; Wu, Z.-X.; Yi, J.-C.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings Bioinf.* **2021**, *22*, bbab152.

(16) Clark, K.; Khandelwal, U.; Levy, O.; Manning, C. D. What Does BERT Look at? An Analysis of BERT's Attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*; Association for Computational Linguistics: Florence, Italy, 2019; pp 276−286.

(17) van Aken, B.; Winter, B.; Löser, A.; Gers, F. A. How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management; CIKM '19*; Association for Computing Machinery: New York, NY, USA, 2019; pp 1823−1832.

(18) Wiegreffe, S.; Pinter, Y. Attention Is Not Not Explanation. 2019, arXiv preprint arXiv:1908.04626. https://doi.org/10.48550/arXiv.1908.04626 (accessed Sept 5, 2019).

(19) Chu, Y.; Zhang, Y.; Wang, Q.; Zhang, L.; Wang, X.; Wang, Y.; Salahub, D. R.; Xu, Q.; Wang, J.; Jiang, X.; Xiong, Y.; Wei, D.-Q. A Transformer-Based Model to Predict Peptide−HLA Class I Binding and Optimize Mutated Peptides for Vaccine Design. *Nat. Mach. Intell.* **2022**, *4*, 300−311.

(20) Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in Peptide Drug Discovery. *Nat. Rev. Drug Discovery* **2021**, *20*, 309−325.

(21) Vinogradov, A. A.; Yin, Y.; Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J. Am. Chem. Soc.* **2019**, *141*, 4167−4181.

(22) Hosseinzadeh, P.; Bhardwaj, G.; Mulligan, V. K.; Shortridge, M. D.; Craven, T. W.; Pardo-Avila, F.; Rettie, S. A.; Kim, D. E.; Silva, D.-A.; Ibrahim, Y. M.; Webb, I. K.; Cort, J. R.; Adkins, J. N.; Varani, G.; Baker, D. Comprehensive Computational Design of Ordered Peptide Macrocycles. *Science* **2017**, *358*, 1461−1466.

(23) Bhardwaj, G.; O'Connor, J.; Rettie, S.; Huang, Y.-H.; Ramelot, T. A.; Mulligan, V. K.; Alpkilic, G. G.; Palmer, J.; Bera, A. K.; Bick, M. J.; Di Piazza, M.; Li, X.; Hosseinzadeh, P.; Craven, T. W.; Tejero, R.; Lauko, A.; Choi, R.; Glynn, C.; Dong, L.; Griffin, R.; van Voorhis, W. C.; Rodriguez, J.; Stewart, L.; Montelione, G. T.; Craik, D.; Baker, D. Accurate de Novo Design of Membrane-Traversing Macrocycles. *Cell* **2022**, *185*, 3520−3532.e26.

(24) Sohrabi, C.; Foster, A.; Tavassoli, A. Methods for Generating and Screening Libraries of Genetically Encoded Cyclic Peptides in Drug Discovery. *Nat. Rev. Chem.* **2020**, *4*, 90−101.

(25) Li, X.; Craven, T. W.; Levine, P. M. Cyclic Peptide Screening Methods for Preclinical Drug Discovery. *J. Med. Chem.* **2022**, *65*, 11913−11926.

(26) Lam, K. S.; Lebl, M.; Krchňák, V. The "One-Bead-One-Compound" Combinatorial Library Method. *Chem. Rev.* **1997**, *97*, 411−448.

(27) Heinis, C.; Rutherford, T.; Freund, S.; Winter, G. Phage-Encoded Combinatorial Chemical Libraries Based on Bicyclic Peptides. *Nat. Chem. Biol.* **2009**, *5*, 502−507.

(28) Tavassoli, A. SICLOPPS Cyclic Peptide Libraries in Drug Discovery. *Curr. Opin. Chem. Biol.* **2017**, *38*, 30−35.

(29) Guillen Schlippe, Y. V.; Hartman, M. C. T.; Josephson, K.; Szostak, J. W. In Vitro Selection of Highly Modified Cyclic Peptides That Act as Tight Binding Inhibitors. *J. Am. Chem. Soc.* **2012**, *134*, 10469−10477.

(30) Huang, Y.; Wiedmann, M. M.; Suga, H. RNA Display Methods for the Discovery of Bioactive Macrocycles. *Chem. Rev.* **2019**, *119*, 10360−10391.

(31) Murakami, H.; Ohta, A.; Ashigai, H.; Suga, H. A Highly Flexible tRNA Acylation Method for Non-Natural Polypeptide Synthesis. *Nat. Methods* **2006**, *3*, 357−359.

(32) Goto, Y.; Katoh, T.; Suga, H. Flexizymes for Genetic Code Reprogramming. *Nat. Protoc.* **2011**, *6*, 779−790.

(33) Goto, Y.; Suga, H. The RaPID Platform for the Discovery of Pseudo-Natural Macrocyclic Peptides. *Acc. Chem. Res.* **2021**, *54*, 3604−3617.

(34) Peacock, H.; Suga, H. Discovery of De Novo Macrocyclic Peptides by Messenger RNA Display. *Trends Pharmacol. Sci.* **2021**, *42*, 385−397.

(35) Ramirez-Carrozzi, V.; Sambandam, A.; Luis, E.; Lin, Z.; Jeet, S.; Lesch, J.; Hackney, J.; Kim, J.; Zhou, M.; Lai, J.; Modrusan, Z.; Sai, T.; Lee, W.; Xu, M.; Caplazi, P.; Diehl, L.; de Voss, J.; Balazs, M.; Gonzalez, L.; Singh, H.; Ouyang, W.; Pappu, R. IL-17C Regulates the Innate Immune Function of Epithelial Cells in an Autocrine Manner. *Nat. Immunol.* **2011**, *12*, 1159−1166.

(36) Chang, S. H.; Reynolds, J. M.; Pappu, B. P.; Chen, G.; Martinez, G. J.; Dong, C. Interleukin-17C Promotes Th17 Cell Responses and Autoimmune Disease via Interleukin-17 Receptor E. *Immunity* **2011**, *35*, 611−621.

(37) Song, X.; Zhu, S.; Shi, P.; Liu, Y.; Shi, Y.; Levin, S. D.; Qian, Y. IL-17RE Is the Functional Receptor for IL-17C and Mediates Mucosal Immunity to Infection with Intestinal Pathogens. *Nat. Immunol.* **2011**, *12*, 1151−1158.

(38) Song, X.; Gao, H.; Lin, Y.; Yao, Y.; Zhu, S.; Wang, J.; Liu, Y.; Yao, X.; Meng, G.; Shen, N.; Shi, Y.; Iwakura, Y.; Qian, Y. Alterations in the Microbiota Drive Interleukin-17C Production from Intestinal Epithelial Cells to Promote Tumorigenesis. *Immunity* **2014**, *40*, 140−152.

(39) Vandeghinste, N.; Klattig, J.; Jagerschmidt, C.; Lavazais, S.; Marsais, F.; Haas, J. D.; Auberval, M.; Lauffer, F.; Moran, T.; Ongenaert, M.; Van Balen, M.; Dupont, S.; Lepescheux, L.; Garcia, T.; Härtle, S.; Eyerich, K.; Fallon, P. G.; Brys, R.; Steidl, S. Neutralization of IL-17C Reduces Skin Inflammation in Mouse Models of Psoriasis and Atopic Dermatitis. *J. Invest. Dermatol.* **2018**, *138*, 1555−1563.

(40) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **2012**, *28*, 3150−3152.

(41) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.

(42) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14*, 1188−1190.

(43) Bashiruddin, N. K.; Hayashi, M.; Nagano, M.; Wu, Y.; Matsunaga, Y.; Takagi, J.; Nakashima, T.; Suga, H. Development of Cyclic Peptides with Potent in Vivo Osteogenic Activity through RaPID-Based Affinity Maturation. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 31070−31077.

(44) Yang, S.-Q.; Ye, Q.; Ding, J.-J.; Yin, M.-Z.; Lu, A.-P.; Chen, X.; Hou, T.-J.; Cao, D.-S. Current Advances in Ligand-Based Target Prediction. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1504.

(45) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(46) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.

(47) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(48) Olson, C. A.; Nie, J.; Diep, J.; Al-Shyoukh, I.; Takahashi, T. T.; Al-Mawsawi, L. Q.; Bolin, J. M.; Elwell, A. L.; Swanson, S.; Stewart, R.; Thomson, J. A.; Soh, H. T.; Roberts, R. W.; Sun, R. Single-Round, Multiplexed Antibody Mimetic Design through mRNA Display. *Angew. Chem., Int. Ed.* **2012**, *51*, 12449−12453.

(49) Hammond, P. W.; Alpin, J.; Rise, C. E.; Wright, M.; Kreider, B. L. In Vitro Selection and Characterization of Bcl-XL-binding Proteins from a Mix of Tissue-specific mRNA Display Libraries. *J. Biol. Chem.* **2001**, *276*, 20898−20906.

(50) Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: A Soft Tool for 3D Drug Design of Protein Targets by Artificial Intelligence and Classical Algorithm. *Briefings Bioinf.* **2020**, *22*, bbaa161.

(51) Durrant, J. D.; Amaro, R. E.; McCammon, J. A. AutoGrow: A Novel Algorithm for Protein Inhibitor Design. *Chem. Biol. Drug Des.* **2009**, *73*, 168−178.

(52) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461.

(53) Chéron, N.; Jasty, N.; Shakhnovich, E. I. OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. *J. Med. Chem.* **2016**, *59*, 4171−4188.

(54) Pegg, S. C.-H.; Haresco, J. J.; Kuntz, I. D. A Genetic Algorithm for Structure-Based de Novo Design. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 911−933.

(55) Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De Novo Design of Enzyme Inhibitors by Monte Carlo Ligand Generation. *J. Med. Chem.* **1995**, *38*, 466−472.

(56) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **2014**, *19*, 1069−1080.

(57) Morrison, K. L.; Weiss, G. A. Combinatorial Alanine-Scanning. *Curr. Opin. Chem. Biol.* **2001**, *5*, 302−307.

(58) Lei, Y.; Li, S.; Liu, Z.; Wan, F.; Tian, T.; Li, S.; Zhao, D.; Zeng, J. A Deep-Learning Framework for Multi-Level Peptide−Protein Interaction Prediction. *Nat. Commun.* **2021**, *12*, 5465.

(59) Ahmed, M.; Seraj, R.; Islam, S. M. S. The K-Means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronicsweek* **2020**, *9*, 1295.

(60) Morimoto, J.; Hayashi, Y.; Suga, H. Discovery of Macrocyclic Peptides Armed with a Mechanism-Based Warhead: Isoform-Selective Inhibition of Human Deacetylase SIRT2. *Angew. Chem., Int. Ed.* **2012**, *51*, 3423−3427.

(61) Kaelbling, L. P.; Littman, M. L.; Cassandra, A. R. Planning and Acting in Partially Observable Stochastic Domains. *Artif. Intell.* **1998**, *101*, 99−134.

(62) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484−489.

(63) Batra, R.; Loeffler, T. D.; Chan, H.; Srinivasan, S.; Cui, H.; Korendovych, I. V.; Nanda, V.; Palmer, L. C.; Solomon, L. A.; Fry, H. C.; Sankaranarayanan, S. K. R. S. Machine Learning Overcomes Human Bias in the Discovery of Self-Assembling Peptides. *Nat. Chem.* **2022**, *14*, 1427−1435.

(64) Bryant, P.; Pozzati, G.; Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. Predicting the Structure of Large Protein Complexes Using AlphaFold and Monte Carlo Tree Search. *Nat. Commun.* **2022**, *13*, 6028.

(65) Sun, M.; Xing, J.; Meng, H.; Wang, H.; Chen, B.; Zhou, J. MolSearch: Search-Based Multi-Objective Molecular Generation and Property Optimization. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; KDD '22*; Association for Computing Machinery: New York, NY, USA, 2022; pp 4724−4732.

(66) Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; Colton, S. A Survey of Monte Carlo Tree Search Methods. *IEEE Trans. Comput. Intell. AI Games* **2012**, *4*, 1−43.

(67) Auer, P.; Cesa-Bianchi, N.; Fischer, P. Finite-Time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* **2002**, *47*, 235−256.

(68) Sokal, R. R. Distance as a Measure of Taxonomic Similarity. *Syst. Zool.* **1961**, *10*, 70−79.

(69) Landrum, G.; Tosco, P.; Kelley, B.; sriniker; gedeck; NadineSchneider; Vianello, R.; Ric; Dalke, A.; Cole, B.; AlexanderSavelyev; Swain, M.; Turk, S.; N, D.; Vaucher, A.; Kawashima, E.; Wójcikowski, M.; Probst, D.; Cosgrove, D.; Pahl, A. J. P.; Berenger, F.; JLVarjo; O'Boyle, N.; Fuller, P.; Jensen, J. H.; Sforna, G. *DoliathGavid. Rdkit/Rdkit: 2020_03_1 (Q1 2020) Release*, 2020.