

HydrogelFinder: A Foundation Model for Efficient Self-Assembling Peptide Discovery Guided by Non-Peptidal Small Molecules

Xuanbai Ren, Jiaying Wei, Xiaoli Luo, Yuansheng Liu, Kenli Li, Qiang Zhang, Xin Gao, Sizhe Yan, Xia Wu, Xingyue Jiang, Mingquan Liu, Dongsheng Cao, Leyi Wei, Xiangxiang Zeng,* and Junfeng Shi*

Self-assembling peptides have numerous applications in medicine, food chemistry, and nanotechnology. However, their discovery has traditionally been serendipitous rather than driven by rational design. Here, HydrogelFinder, a foundation model is developed for the rational design of self-assembling peptides from scratch. This model explores the self-assembly properties by molecular structure, leveraging 1,377 self-assembling non-peptidal small molecules to navigate chemical space and improve structural diversity. Utilizing HydrogelFinder, 111 peptide candidates are generated and synthesized 17 peptides, subsequently experimentally validating the self-assembly and biophysical characteristics of nine peptides ranging from 1–10 amino acids—all achieved within a 19-day workflow. Notably, the two de novo-designed self-assembling peptides demonstrated low cytotoxicity and biocompatibility, as confirmed by live/dead assays. This work highlights the capacity of HydrogelFinder to diversify the design of self-assembling peptides through non-peptidal small molecules, offering a powerful toolkit and paradigm for future peptide discovery endeavors.

1. Introduction

Driven by supramolecular interactions (e.g., hydrogen bonding, hydrophobic interactions, and electrostatic interactions), peptides can self-assemble in water to form ordered nanostructures, such as nanofibers, which, in turn, form three-dimensional networks, ultimately leading to supramolecular hydrogelation.^[1–5] Supramolecular hydrogels resemble extracellular matrices in tissues in that they both have a highly water content and fibrils that function similarly to cytoskeleton. These properties have led to their extensive study as emerging potential biomaterials for tissue engineering,^[6] drug delivery,^[7,8] cancer cell inhibition,^[9,10] regenerative medicine,^[11] or antibacterial applications.^[12] Despite these advances in supramolecular hydrogels,^[13–18] designing self-assembling peptides based solely on molecular structure remains challenging for chemists.^[19]

X. Ren, X. Luo, Y. Liu, K. Li, M. Liu, X. Zeng
College of Information Science and Engineering
Hunan University
Changsha 410003, China
E-mail: xzeng@hnu.edu.cn

J. Wei, S. Yan, X. Wu, X. Jiang, J. Shi
State Key Laboratory of Chemo/Bio-Sensing and Chemometrics, School of
Biomedical Sciences
Hunan University
Changsha 410003, China
E-mail: Jeff-Shi@hnu.edu.cn

Q. Zhang
ZJU-Hangzhou Global Scientific and Technological Innovation Center
Hangzhou 311200, China

Q. Zhang
College of Computer Science and Technology
Zhejiang University
Hangzhou 310013, China

X. Gao
Computational Bioscience Research Center (CBRC), Computer, Electrical
and Mathematical Sciences and Engineering Division
King Abdullah University of Science and Technology (KAUST)
Thuwal 23955-6900, Saudi Arabia

D. Cao
Xiangya School of Pharmaceutical Sciences
Central South University
Changsha 410003, China

L. Wei
School of Software
Shandong University
Jinan 250100, China

L. Wei
Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR)
Shandong University
Jinan 250100, China

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adv.202400829>

© 2024 The Authors. Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: [10.1002/adv.202400829](https://doi.org/10.1002/adv.202400829)

Numerous classic self-assembling molecules have been discovered unintentionally rather than through rational design.^[20–23] For instance, Weiss et al.^[23] serendipitously found that cholesteryl 4-(2-anthryloxy) butyrate could form a hydrogel while studying its photochemistry. Zhang et al. discovered that a class of amphiphilic peptides derived from the yeast protein, *Zuotin*, could self-assemble in physiological buffer (e.g., Dulbecco modified Eagle's medium) to form an “insoluble macroscopic membrane”.^[22,24] Similarly, Xu et al. observed that Fmoc-D-Ala-D-Ala, an intermediate in peptide synthesis, could form a hydrogel through hydrogen bonding and hydrophobic interactions.^[25] Many other reports have also described the unexpected discovery of peptides that can self-assemble into hydrogels under various conditions.^[23,26–28] However, the rational design of self-assembling peptides using traditional methods faces formidable challenges, particularly in accurately modeling the intricate interactions between water molecules and peptides, and achieving a delicate balance of hydrophilicity and hydrophobicity.^[29] These further complicate the design of self-assembly peptides.

In response to these challenges, recent advancements in machine learning have profoundly impacted fields such as chemistry,^[30,31] materials science,^[32,33,34] and biomedical research.^[35–42] Machine learning, as an invaluable tool, has become crucial in deciphering the complexities of peptide self-assembly. For example, Sankaranarayanan et al.^[43] have ingeniously harnessed Monte Carlo tree search (MCTS) alongside coarse-grained molecular dynamics (CGMD) simulations to discover novel pentapeptides. Wang et al.^[44] have deftly combined support vector machines (SVM) with CGMD to predict peptides aggregation propensity (AP) and identify potent tetrapeptides. Li et al.^[45] have employed a robust deep learning framework, along with CGMD, to predict the self-assembly properties of a vast peptide library, successfully forecasting the AP of both pentapeptides and decapeptides. Despite these advancements, challenges persist in the application of machine learning to peptide self-assembly discovery. Current methods rely on costly CGMD simulations to derive AP values for peptides in training sets, resulting in a time-consuming and labor-intensive process with limitations in generalizing beyond the training data. Moreover, traditional machine learning approaches often focus solely on amino acid sequences, neglecting the significant impact of peptide modifiers on self-assembly processes within the vast and diverse chemical space of peptides.

In this work, we propose HydrogelFinder, an innovative foundation model comprising three key modules: HydrogelFinder-mining for literature data mining, HydrogelFinder-GPT employing a deep generative model, HydrogelFinder-predict as a virtual screening tool (Figure 1). This integrated system facilitates the rational design, rapid production, and screening of self-assembling peptides. To navigate the complex chemical space of peptides and discover a diverse array of potential self-assembled peptide candidates, we focus on the perspective of molecular structure, rather than merely through amino acid sequences, to delve into the self-assembly characteristics of peptides. Leveraging HydrogelFinder-mining, we construct a molecular library comprising 2669 self-assembled molecules, encompassing both peptides and non-peptidal small molecules. Non-peptidal small molecules play a pivotal role in guiding the exploration of chem-

ical space and enhancing structural diversity. As a proof of concept, utilizing HydrogelFinder-GPT with this comprehensive library as a training set, we identified 2000 compounds for screening, which yielded 111 previously unreported candidates. Experimental characterization of 17 structurally diverse peptides from this set revealed that nine peptides, ranging from 1–10 amino acids in length, exhibited ability to self-assemble into hydrogels with diverse properties. Notably, two randomly selected peptides showed low cytotoxicity toward human cell lines. An additional highlight is the identification of the shortest self-assembling lipid-peptide compound documented to date that does not require a metal ion. This work establishes HydrogelFinder as a highly effective foundation model for AI-based design and generation of self-assembling peptides.

2. Results and Discussion

2.1. Overview of HydrogelFinder

In pursuit of efficiently sampling structurally diverse self-assembled peptides within the expansive chemical space, our approach comprises three integral modules. First, HydrogelFinder-mining engages in literature mining to construct an effective and chemically diverse training dataset. Following this, HydrogelFinder-GPT employs a deep generative model to model the relationship between molecular structural features and aggregation propensity, and HydrogelFinder-predict facilitates virtual screening to evaluate candidates.

Specifically, HydrogelFinder-mining compiles a set of molecular graphs related to self-assembly, converting these images into SMILES representations.^[46] This process facilitates the construction of a HYDROGEL-POSITIVE training dataset, enhancing chemical diversity with a collection of 1292 self-assembling peptides and 1377 self-assembling small molecules, totaling 2669 entries (Figure 1A). Additionally, we establish a publicly accessible self-assembling molecules database for the broader research community at <http://hydrogelfdb.com>. For the rational generation of self-assembling peptide, we proposed an automated deep generative model, HydrogelFinder-GPT (Figure 1B). This model, utilizing a transformer architecture, learns the rules of self-assembly from molecule sequence strings.^[47] Beginning with pretraining on a comprehensive collection of ChEMBL small molecules to grasp molecular grammar, the model undergoes fine-tuning on an autonomously constructed HYDROGEL-POSITIVE dataset (Table S2, Supporting Information). This fine-tuning process refines the model's understanding towards self-assembly properties. Candidates generated by HydrogelFinder-GPT undergo evaluation through HydrogelFinder-predict before experimental validation.

2.2. Generation of Self-Assembling Molecules by HydrogelFinder-GPT

To evaluate the efficacy of self-assembling molecule generation by HydrogelFinder-GPT, we conducted a comprehensive evaluation of the model's performance based on validity, uniqueness, novelty, and activity under various training strategies. The

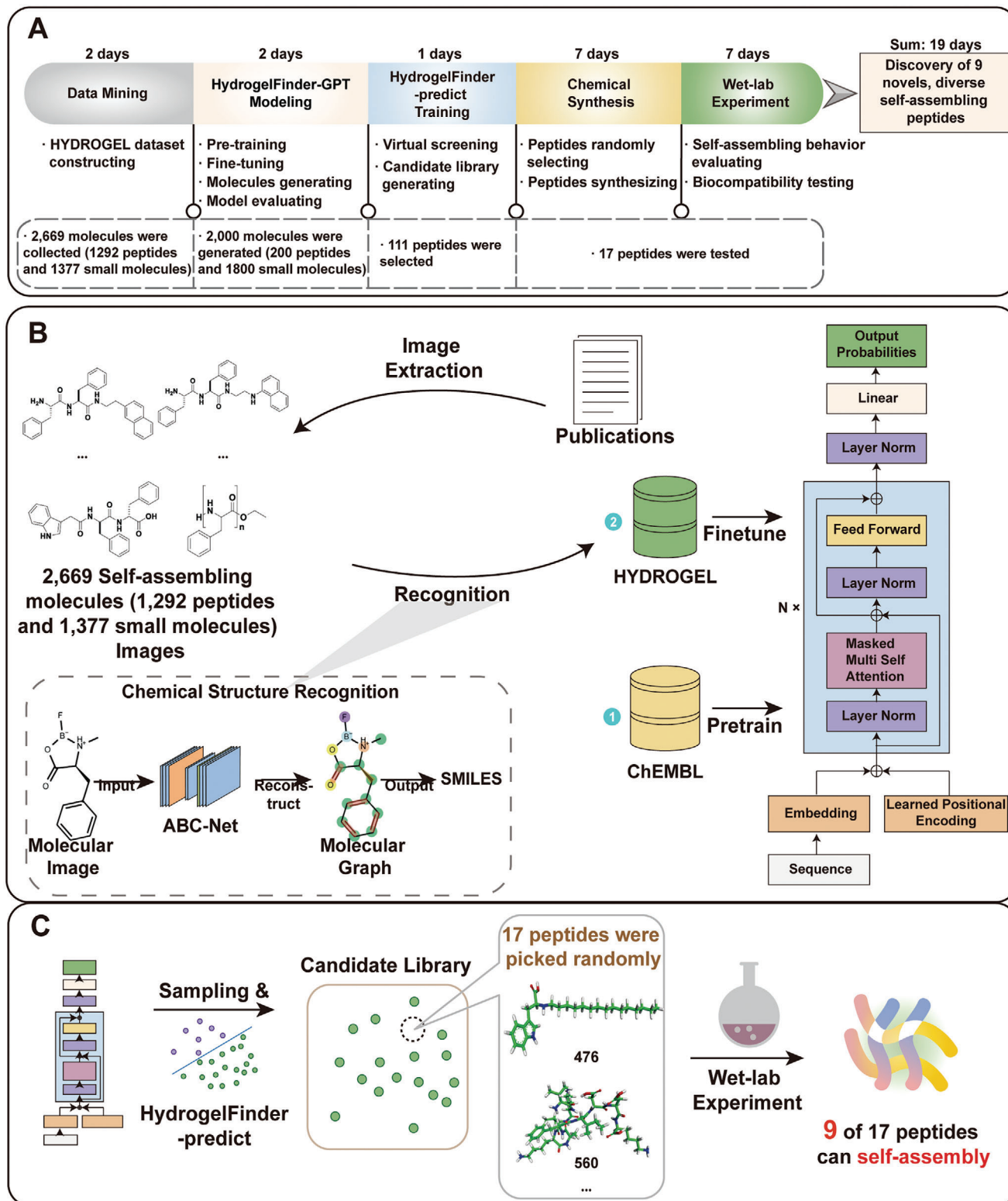


Figure 1. A) Workflow of the artificial intelligence framework for self-assembling peptide design and discovery. B) The general workflow for the design of self-assembling peptide using HydrogelFinder-GPT C) We sampled molecules from the fine-tuned network, and performed high-throughput prediction with the HydrogelFinder-predict to build a candidate library. We randomly selected 17 peptides from the candidate library, 9 of which can self-assemble under specific conditions.

Table 1. Performance of generative neural network (Figure 2A Data Supplement).

Metrics	Pre-training	W/O Self-assembling Small Molecule	W/O Pre-training	HydrogelFinder-GPT
Valid (t)	100%	100%	100%	99.95%
Unique (t)	99.95%	86.05%	97.40%	94.95%
Novel (t)	99.62%	90.41%	94.87%	92.15%
Active (t)	1.99%	72.05%	72.83%	73.98%

model-generated sequences were subjected to evaluation using HydrogelFinder-predict, with the active rate, representing the ratio of potential self-assembly, serving as a key metric. As shown in Figure 2A, HydrogelFinder-GPT achieved the highest active rate at 73.98%, surpassing other training strategies. Additionally, HydrogelFinder-GPT exhibited impressive figures of 94.95% uniqueness and 92.15% novelty, signifying its capability to produce de novo and valid molecules even after the finetuning process (see details in Section 4.2 and Table 1).

Having established the model's competence in generating self-assembling molecules, we next investigated properties associated with gelation ability, such as hydrogen bond acceptors (HBA) and donors (HBD), number of basic groups (nBase), Ghose–Crippen LogKow (LogP), topological polar surface area (TPSA), and molecular weight (Mol.wt),^[48] these values were determined through RDKit calculations. We randomly sampled chemicals of the same size (2000 sequences) from both the HYDROGEL-POSITIVE and HYDROGEL-NEGATIVE datasets, plotting the distributions for each parameter (Figure S1, Supporting Information). The analysis revealed that the distribution of de novo candidates for each property closely mirrored that of the HYDROGEL-POSITIVE dataset, further substantiating the effectiveness of HydrogelFinder-GPT in designing self-assembly-like compounds (More details in the Supporting Information).

2.3. Exploration of Structurally Diverse Self-Assembling Peptides by HydrogelFinder-GPT

To achieve structurally diverse self-assembling peptides, our strategy involved the guidance of chemical space exploration using non-peptidal small molecules. To assess the diversity of candidates generated by the model under different training strategies, we employed the Tanimoto similarities as a metric (More details in Methods). As shown in Figure 2B, HydrogelFinder-GPT generated candidates with a high diversity score of 0.803. This notable diversity is attributed to the incorporation of non-peptidal small molecules during training, its removal resulted in a decreased diversity of the generated candidates to 0.757. Additionally, we visualized the model output features using uniform manifold approximation and projection (UMAP) plots.^[49,50] As shown in Figure 2C, the candidates generated by HydrogelFinder-GPT exhibited a wider chemical space distribution compared to the training set without small molecules. This result underscores the significant enhancement in our model's performance with the addition of non-peptidal small molecules data to the training set.

As a proof of concept, we present the length and decoration statistic for 111 peptide-based candidates. In Figure 2D, our model demonstrated the ability to generate sequences with

lengths ranging from 1 to 14 amino acids, surpassing the length of pentapeptides reported in previous studies^[19,43–45,51] The Orange bars represent the self-assembling peptides confirmed in subsequent wet-lab experiments, while the green bars represent those that did not exhibit self-assembly (details in Section 2.6). Specifically, we highlight the successful self-assembly of nine peptides, comprising amino acid sequences of seven distinct lengths (Figure 2E). Notably, we reported the discovery of the shortest self-assembling lipid peptide, gel 476, which comprises only a single amino acid with a long alkyl chain.

2.4. Discovery of Self-Assembling Peptide Derivatives by HydrogelFinder

An often-overlooked challenge in the development of self-assembling peptides lies in understanding the influence of chemical modification. To address this, we selected gel 476 and gel 133 as case studies, representing successful instances of self-assembling peptides with chemical modification identified through our studies. In our study, we compared activity scores of gel 476 and gel 133 with and without modification (obtained by HydrogelFinder-predict and RDKit). Additionally, we scrutinized a range of properties associated with gelation ability, as summarized in Table 2. The results showed that a significant decrease in the active rate, LogP, and molecular weight of the peptides after the removal of modification. This decline can be attributed to the fact that these modifications can alter the hydrophilic and hydrophobic nature of the peptides. For example, the addition of 9-fluorenylmethyl carbamate (Fmoc) group increases hydrophobicity, potentially facilitating the self-assembly of peptides under certain conditions. Moreover, the modification group may introduce new intermolecular interactions such as hydrogen bonding, hydrophobic interactions, which are critical for the self-assembly.

To further demonstrate the model's proficiency in discovering structurally diverse self-assembling peptides, we computed the Tanimoto similarities of gel 476 and gel 133 to the HYDROGEL-POSITIVE dataset (Figure 3). The majority of compounds in the training set exhibited substantially dissimilar from gel 476 and gel 133, with mean Tanimoto similarities of 0.19 and 0.18, respectively. This serves as additional evidence of HydrogelFinder's ability to identify structurally diverse self-assembling peptides with modification.

2.5. Evaluation of Gelation Behavior

As shown in Figure 1C, we next sought to experimentally validate the performance of our AI method. To this end, we randomly selected 17 peptides for synthesis, spanning a range of

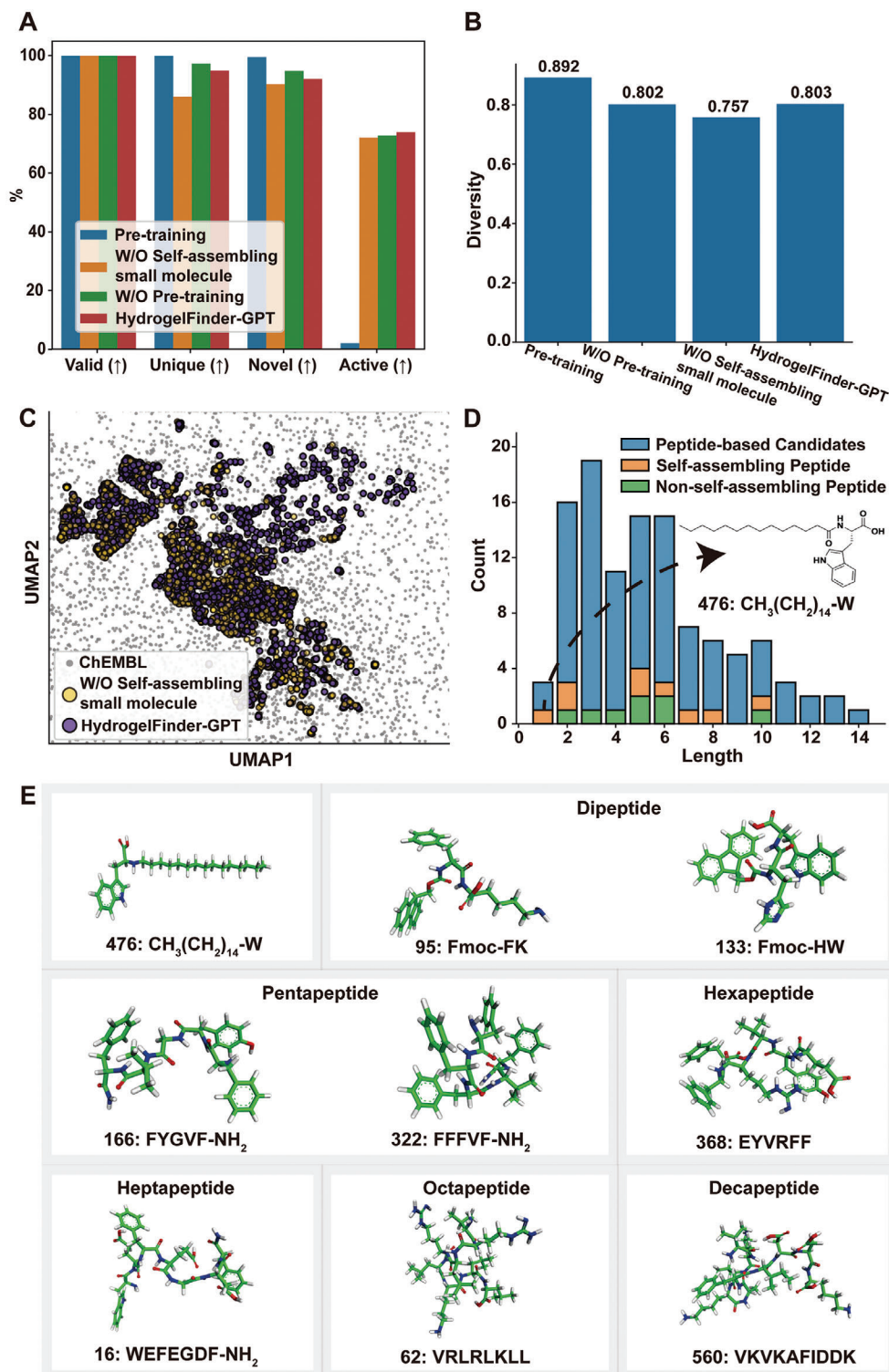


Figure 2. Performance evaluation of HydrogelFinder-GPT and chemical structures of self-assembling peptides. A) Performance comparison of generative models with different training strategies over valid, unique, novel and active on generation task (Details see Section 4.2 and Table 1). B) The generative capacity of model structural diversity under different training strategies. C) UMAP visualization of the chemical space distribution of candidates generated by HydrogelFinder-GPT and generated by training set without self-assembling small molecules. D) The statistical distributions of peptide-based candidates. The peptide-based candidates have 111 sequences (blue), of which 9 molecules can self-assembly (Orange), while 8 molecules failed (Green). E) Chemical structures and identification numbers (IDNs) of the nine peptides selected from the candidate library that are able to self-assemble.

Table 2. Computational characterization and activity of peptides with and without modifiers.

Sequence	Active	LogP	HBA	HBD	NBASE	TPSA	M.W
CH ₃ (CH ₂) ₁₄ -W	1	5.9809	2	3	0	82.19	414.288
W	0.703	1.1223	2	3	1	79.11	204.089
Fmoc-HW	1	4.1529	5	5	0	149.2	563.216
HW	0.003	0.5729	4	5	1	136.8	341.144

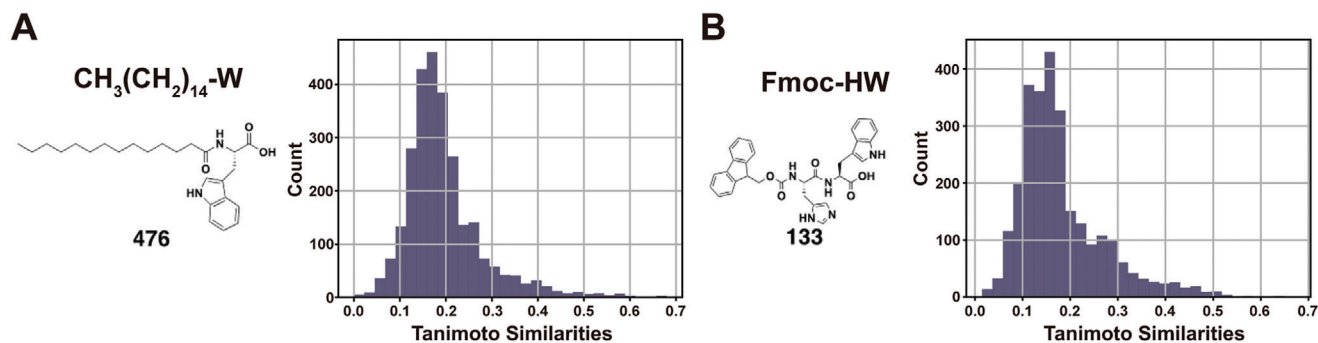


Figure 3. Structures of self-assembling peptides A) 476 and B) 133 with modification and corresponding distributions of Tanimoto similarities to HYDROGEL-POSITIVE data set.

HydrogelFinder-predicted scores from low to high. Notably, these peptides are entirely new and do not appear in the HYDROGEL-POSITIVE dataset. We subsequently evaluated their gelation capability (Figure S2, Supporting Information). After purification by HPLC, the obtained compounds were dissolved in water and the pH of the solution was adjusted to trigger self-assembly and hydrogelation. Non-self-assembling peptides formed either a precipitate or a low viscosity fluid irrespective of pH (2–12). Nonetheless, more than half of these peptides (9/17) could self-assemble in aqueous solution. This high proportion of self-assembling peptides confirmed the predictive accuracy of the HydrogelFinder model. The gelation properties of the nine self-assembling peptides are summarized in Table 3.

Observing gel behavior in an inverted test tube is a rapid and facile approach to determine whether a dissolved peptide formed a gel. Representative images of the nine candidates holding water and resisting flow, which together indicated gel formation, are shown Table 3. Interestingly, we observed that all peptides containing proline (P) failed to form a hydrogel (Table S2, Supporting Information). Additionally, only a few self-assembling peptides composed of proline were detected in the hydrogel-positive database. This suggested that the presence of proline reduces conformational flexibility, diminishing the likelihood of hydrogel formation. Obviously, the pH value of each hydrogel ranged widely, between 3.0 (highly acidic) to 8.0 (weakly basic), and these pH values did not exactly match their calculated isoelectric points (pI). It also warrants mentioned that molecules 95 and 133 shared the same Fmoc motif, which is commonly used to protect amino acids during solid-phase peptide synthesis. Additionally, peptide 368 formed an opaque hydrogel, likely attributable to large aggregates in the hydrogel matrix. Furthermore, previous reports demonstrated that aromatic-aromatic interactions between Fmoc moieties can promote the hydrogelation of small molecules.^[52]

Interestingly, as a lipid peptide, molecule 476, comprised of a single amino acid could also form a gel at pH 5.0, thus representing the shortest lipid peptide of which we are aware.

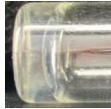
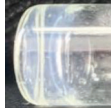
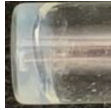


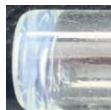

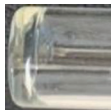

2.6. Biophysical Characterization of Self-Assembling Peptides

To further explore the gelation properties and self-assembly behavior of the candidate self-assembling molecules, we carried out a series of biophysical characterization experiments. The rheological properties of a hydrogel can vary in a manner dependent on their structure, and moreover, these properties are typically critical for their function in tissue engineering, drug delivery, or other applications. An oscillatory rheological analysis was performed to monitor hydrogel storage modulus (G' , a measure of the elastic response of the material) as a function of time (Figure 4A), revealing a gradual increase in gel 16, reaching equilibrium at 1254 Pa, and with a consistently higher storage modulus than loss modulus (G'' , a measure of viscosity), suggesting the formation of a robust hydrogel. Self-assembling peptide 62, 166, and 322 exhibited a similar trend, reaching equilibrium at 2,143 Pa, 1,0991 Pa, and 21,393 Pa, respectively, which were all consistently higher than their loss moduli, implying that these peptides could form relative rigid hydrogels.

TEM was then used to examine morphology of hydrogel matrix, which revealed the formation of 9–10 nm width nanofibers in gel 16, while the self-assembled nanofibers in gels 62, 166, and 322 were approximately 14 nm, 25 nm, and 14 nm, respectively. These nanofibers were entangled with one another, forming a 3D-network that comprised the hydrogel matrix (Figure 4B).

Circular Dichroism (CD) spectroscopy is a valuable technique for detecting secondary structures in the peptide assemblies. In the Far-UV range (190–260 nm), the majority of chromophores were peptide bonds.^[53] The CD spectrum of hydrogel 16 had a

Table 3. Hydrogelation properties of selected molecules.

IDN	Sequence	M.W.	Calc. pI ^{a)}	Conc.	pH	Buffer	Images
16	WEFEGDF-NH ₂	927.4	2.9	1.0 wt%	5.0	1X PBS	
62	VRLRLKLL	1009.7	12.4	1.0 wt%	8.0	1X PBS	
95	Fmoc-FK	515.2	10.1	2.0 wt%	3.0	Water	
133	Fmoc-HW	563.2	7.9	1.0 wt%	3.0	Water	
166	FYGVF-NH ₂	630.7	7.	0.5wt%	7.0	Water	
322	FFFVF-NH ₂	704.3	7.0	1.0 wt%	7.0	Water	
368	EYVRF	859.4	6.9	1.0 wt%	3.0	1X PBS	
476	CH ₃ (CH ₂) ₁₄ -W	414.3	2.5	3.0 wt%	5.0	Water	
560	VKVAFIDDK	1161.4	9.8	1.0 wt%	8.0	1X PBS	

^{a)} Calc. pI is obtained from: https://www.novopro.cn/tools/calc_peptide_property.html

broad negative band at 218 nm, which was typical of a β -sheet conformation. Similar features were detected in gels **62**, **166** and **322**, suggesting the prevalence of β -sheets in the self-assembling peptides, which agreed well with fact that the majority of self-assembling molecules adopted β -sheet conformation.^[54] To further explore the molecular configuration of the assembled structures, we used Fourier-transformed infrared spectroscopy (FT-IR) to examine each gel.^[55,56] In gel **16**, a peak was detected at 1617 cm⁻¹ which was assigned as an amide I band, suggesting the formation of a β -sheet. Similarly, a peak at 1627 cm⁻¹ was observed in gel **62**, a strong peak at 1621 cm⁻¹ was present in gel **166** spectra, and gel **322** also had a peak at 1637 cm⁻¹, all of which in-

dicated that a β -sheet conformation was adopted by the peptides during gelation. Overall, these data consistently supported the likelihood that these peptides self-assembled into ordered nanostructures which formed β -sheets that comprised nanofibers in a gel matrix.

2.7. Biocompatibility of Self-Assembling Peptides

Supramolecular hydrogels can serve as scaffolds for cell culture because of their high similarity with an extracellular matrix.^[57] However, this application requires high biocompatibility. Thus, we chose to evaluate cytotoxicity of peptides **16** and **62** using

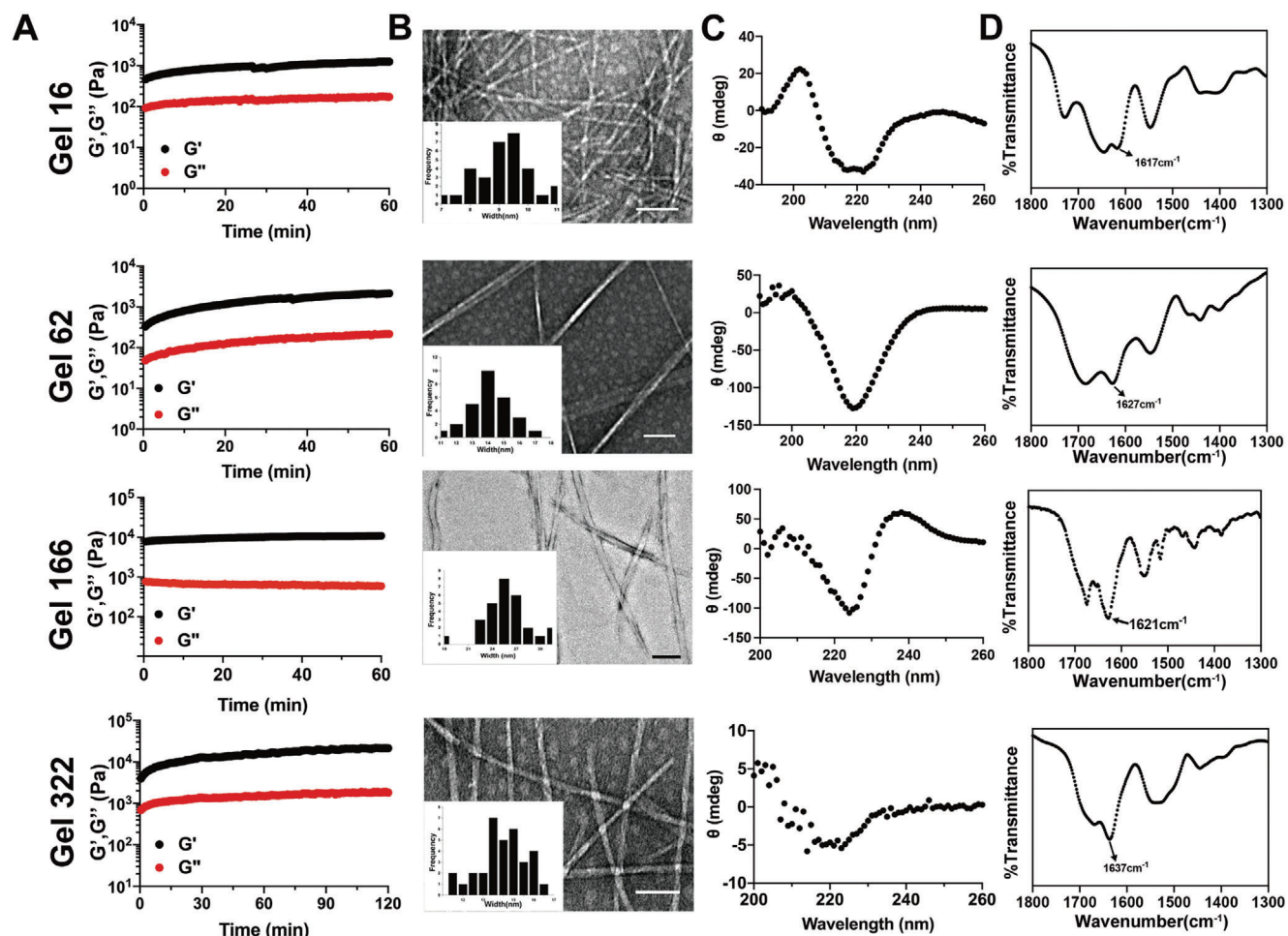


Figure 4. Biophysical characterization of hydrogel 16, 62, 166, 322. A) Rheological dynamic time sweep monitors the storage modulus (G') and loss modulus (G'') of different hydrogels as a function of time B) Representative TEM images reveal the morphologies of selected hydrogels, insets are distribution of nanofibers width (scale bar = 100 nm). C) CD spectra of hydrogels suggested the formation of β -sheet structure. D) FT-IR spectra of hydrogels further confirmed their specific secondary structures.

MTT assays. As common cell lines, human dermal fibroblasts (NHDF) and stem cells from human exfoliated deciduous teeth (SHED) were chosen as model cells for evaluating the cytotoxicity of peptides,^[69,70] peptides 16 and 62 exhibited high cell compatibility at concentrations as high as 500 μM (Figure 5A). Live/dead assays using calcein AM to stain for viable cells and propidium iodide (PI) to stain for compromised or dead cells cultured on the gel surface showed that the vast majority of SHED cells were positive for calcein AM staining, while few or no PI stained cells could be detected (Figure 5B). These results implied that either molecule itself or the bulk gel was exhibited limited toxicity, suggesting its potential use in cell culture applications.

3. Discussion

We developed HydrogelFinder, a foundation model integrating literature mining module, a generative language model, and machine learning for de novo design of self-assembling peptides. This model was complemented by comprehensive experimental validation, confirming gelation behavior and biocompatibility.

In essence, HydrogelFinder was devised to construct a candidate library and establish a publicly accessible database for the hydrogel research community (<http://hydrogeldb.com>). It leverages non-peptidic small molecules to guide the *in-silico* discovery of structurally diverse self-assembling peptides. Through experimental characterization of 17 randomly selected, structurally diverse candidate peptides, we identified nine molecules capable of forming hydrogels in water under different conditions. These peptides displayed distinct self-assembly behaviors and length distributions spanning from 1 to 10 amino acids, with two modifications influencing peptide self-assembly. Biophysical characterization revealed the formation of ordered nanostructures, such as nanofibers, within the hydrogel matrix. Notably, two hydrogels demonstrated low cytotoxicity *in vitro*, suggesting potential applications as scaffolds for cell culture or drug delivery.

Furthermore, the extensive hydrogel candidate library generated in this work remains largely unexplored, holding the potential to address ongoing and future research questions. This includes robust determination of the common structural features defining self-assembling molecules. Recognizing that the

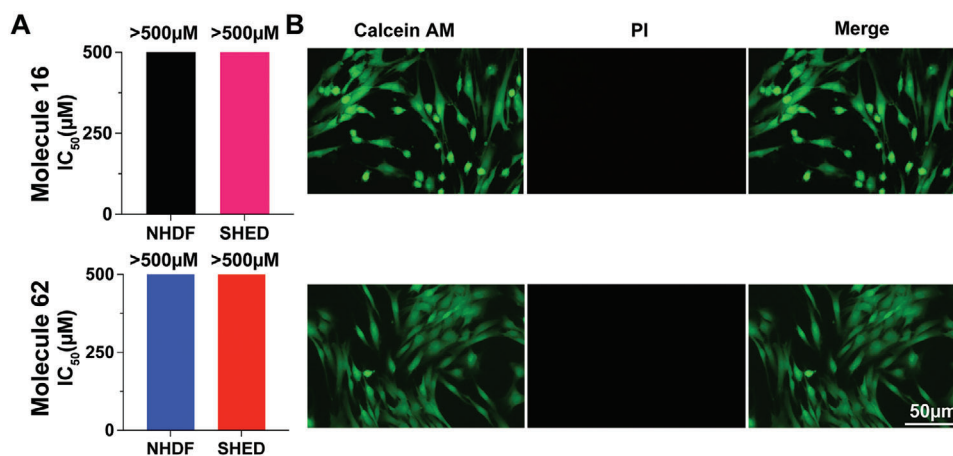


Figure 5. Cell compatibility analysis of compound 16 and 62. A) IC₅₀ of NHDF and SHED cells treated with molecule 16 (top panel) and 62 (bottom panel) for 24 h. B) Live/dead assays of cells grown on the surface of hydrogel 16 (right of top panel) or 62 (right of bottom panel) for 24 h showing high biocompatibility with SHED cultures. Living cells show positive staining with calcein AM (green), while dead or compromised cells are stained with propidium iodide (PI, red). Each column represents three individual experiments ($n = 3$).

Transformer-based architecture excels in extracting 1D sequence information, yet falls short in capturing the 3D complexities of peptide chain's secondary, tertiary and quaternary structures. To address this issue, we are actively developing geometric deep learning models. These models aim to efficiently extract 3D features and fuse multimodal information, enhancing the accuracy and efficiency of candidate molecule generation.

This work thus establishes a powerful framework for designing self-assembling peptides, accompanied by a sizable database for public exploration and the experimental characterization of several hydrogels. Beyond a proof-of-concept demonstration, HydrogelFinder stands poised for use in the design and screening of new peptide crucial for urgent biomedical applications.

4. Experimental Section

Datasets: The original data used in this work were originated from a few pre-available sources (ChEMBL, CPPsite, and ZINC). Data from ChEMBL was used to pre-training generative neural network (HydrogelFinder-GPT pre-training network). Literature data mining by HydrogelFinder-mining to construct HYDROGEL-POSITIVE dataset for HydrogelFinder-GPT fine-tuning. CPPsites and ZINC was used to construct a HYDROGEL-NEGATIVE dataset. Training the HydrogelFinder-predict model using HYDROGEL-POSITIVE and HYDROGEL-NEGATIVE dataset.

ChEMBL. ChEMBL^[58] was a large molecules database. The pre-training network was trained with a subset of ChEMBL version 25. Initially, the complete dataset was standardized with the MolVS Python module using the super parent setting, which standardizes fragment, charge, isotope, stereochemistry and tautomeric states. Molecules were filtered to only contain the atoms [H, C, N, O, F, S, Cl, Br] and heavy atoms that were fewer than 50 in number. In general, a subset of ChEMBL was constructed for pre-training procedure with 300000 molecules.

CPPsite. CPPsite^[59–61] was an updated version of manually curated database (CPPsite) of cell-penetrating peptides (CPPs). The current version holds around 1850 peptide entries, including their predicted tertiary structure of cell-penetrating peptides. CPPsite also maintains information on cell-penetrating peptide properties and abilities to delivery different cargo in model systems (in vitro and in vivo). In this work, CPPsite was only used as a part of negative samples in the HYDROGEL dataset.

ZINC. ZINC^[62] was a free database of commercially-available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds one can search for analogs in under a minute. In this work, “aggregator” was used as the filter key to search this database, and used the result as another part of negative samples in the HYDROGEL dataset. It was worth to note that “aggregates” typically denotes the non-specific, often irreversible association of molecules, which frequently leads to the formation of amorphous structures. In contrast, “self-assembly” refers specifically to the process by which molecules organize themselves into ordered, functional structures through non-covalent molecular interaction.^[63,64] While there can be overlap between aggregation and self-assembly in some cases, known aggregators were chosen to use as a negative training set for HydrogelFinder-predict to ensure that the model prioritizes the prediction of self-assembly behavior.

HYDROGEL dataset. For the HYDROGEL-POSITIVE dataset construction, a thorough search of the Pubmed database was conducted using the “hydrogel” as a query, which yielded 25082 publications. Embedded images in relevant papers were extracted for the training set using the pdf2image python library. Each image was then enumerated, and the molecular graph within each image was identified and translated into SMILES strings using ABC-Net,^[65] an advanced model for chemical structure recognition tasks developed by our group. According to the findings in the papers, compounds used to successfully generate hydrogels were classified as HYDROGEL-POSITIVE samples, while other compounds were classified as HYDROGEL-NEGATIVE samples. All molecules in the HYDROGEL-POSITIVE dataset were thus considered self-assembling molecules, which enable hydrogel formation. Apart from the data collected in publications, the HYDROGEL-NEGATIVE dataset also included compounds obtained from publicly accessible databases, such as ZINC and CPPsites. Data filtering was conducted using the RDKit python library to exclude compounds that failed to transform into a corresponding molecular graph in the HYDROGEL dataset (Supporting Information). In total, the HYDROGEL dataset contained 2669 positive samples and 16761 negative samples after removing duplication, 70% of which were then randomly selected for the training set, while the remaining 10% was used as the testing set. During the training process, 20% of the training set was set aside to serve as the validation set (Table 3).

HydrogelFinder-GPT: A Transformer decoder model was used as architecture for our training, taking the sequence as input using one-hot encoding. The method consists of two parts: pre-training and fine-tuning.

Model pre-training: The decoder module was used as our basic attention module (Figure 1B). In general, this network contains 18 decoder modules, each of which has 48 dimensional states, 12 attention heads and

Table 4. Performance of generative neural network (the up-arrows denote that the higher scores are considered better, the down-arrows mean that lower scores are better).

Metrics	Pre-training	W/O Self-assembling Small Molecule	W/O Pre-training	HydrogelFinder-GPT
FCD (↓)	33.7017	7.1341	5.3607	4.9586
SNN (↑)	0.2153	0.5936	0.4730	0.5149
Frag (↑)	0.6705	0.9952	0.9947	0.9957
Scaff (↑)	0.1388	0.5576	0.4866	0.4821

a position-wise feed-forward network with 512 dimensional inner states. Here, the chemical generation task was modeled as a text generation task in the general natural language processing domain. Specifically, given an unsupervised dataset of compounds $\nu = \{c_0, \dots, c_m\}$, a pair of tokens were inserted ([START] and [END] token) for each compound, and concatenated them to get our pretraining corpus $\mathcal{U} = \{u_0, \dots, u_n\}$, where m denotes the size of the dataset and n was the number of tokens in the corpus. A standard language modeling objective was used to maximize the following likelihood:

$$L(u) = \sum_i \log P(u_i | u_{-k}, \dots, u_{-1}; \Theta) \quad (1)$$

where k was the size of the context window, and the conditional probability P was modeled using a neural network with parameters Θ . In this work, k was set to 128. The model trained on 40GB NVIDIA V100 in 36 hours.

Model Performance Evaluation: The testing set with 271 samples of HYDROGEL-POSITIVE dataset was used for the final performance assessment of HydrogelFinder-GPT. The metrics, included validity,^[66] uniqueness,^[66] novelty,^[66] FCD,^[67] nearest neighbor similarity (SNN),^[66] Fragment similarity (Frag),^[68] scaffold similarity (Scaff),^[69] Active and diversity. Among them, the valid percentage of molecular strings that can be translated back into molecular graphs; the unique percentage of non-duplicated molecular strings; and the novel percentage of chemicals that were not present in the training set. FCD measures the similarity of chemical structures and bioactivities between a testing set and the generated chemicals according to features extracted by a well-trained deep neural network.^[67] Frag and Scaff were cosine distances between the vectors of fragments or scaffold frequencies correspondingly to a generated distribution and the distribution of a testing set. SNN was the average similarity of generated chemicals to the nearest chemical from a testing set distribution. The four metrics were implemented by the MOSES.^[67] A sample scored higher than 0.5 in the HydrogelFinder-predict was considered as a candidate capable to form a hydrogel. The active metric represents the average score between the active proportion of candidates in novel chemicals and in total generated chemicals. The diversity of a set of molecules were define as the average pairwise Tanimoto similarities between them, where Tanimoto similarities $dist(X, Y) = 1 - sim(X, Y)$.

The performance was assessed of the model using several metrics, including SNN, Frag, FCD, and Scaff, to gauge the similarity between the generated chemicals and the hydrogel data in the testing set of the HYDROGEL-POSITIVE dataset from different angles (Table 4). Notably, HydrogelFinder-GPT significantly outperforms the pre-training network across these metrics. Only in cases where small molecules were absent from the training set did the SNN and Scaff metrics favor the pre-training network. However, overall, HydrogelFinder-GPT exhibited a more balanced performance. The data generated by HydrogelFinder-GPT closely resembled the testing set, implying that the model effectively learned to characterize the distribution of hydrogels within the chemical space and can generate chemically similar yet completely new compounds.

High-throughput Prediction HydrogelFinder-predict Model: To evaluate the performance of the generative neural network to generate potential self-assembling compounds, a probabilistic SVM classification model was used (More details in Supporting Information). The model was trained to discriminate active compounds that could self-assemble to form hy-

drogels from inactive ones according to their 2048-bit-radius extended connectivity fingerprint (ECFP) representations. Given the size of the HYDROGEL-POSITIVE datasets and the HYDROGEL-NEGATIVE datasets were highly imbalanced (Table S2, Supporting Information), Sampling was performed up to resample positive samples with respect to the negative ones until they reach the same size (See method in the Supporting Information). The model with $C = 10$ and $\gamma = 0.01$ was considered to have the highest AUROC (0.9862) toward the testing set of the HYDROGEL dataset.

Peptides Synthesis: All peptides were synthesized via standard solid-phase peptide synthesis using a CS136S peptide synthesizer, with Rink-AM resin or 2-Cl resin and activation by HCTU. The resin-bound peptides were cleaved using a cocktail of TFA/Triisopropylsilane/H₂O (95:2.5:2.5) for 3 h. The resin mixture was filtered and washed with excess TFA. Crude peptides were obtained by concentrating the filtrate and precipitating it with cold ether. The crude product was purified by reverse phase HPLC with a semi-preparative C18 column. HPLC solvents comprise solvent A (0.1% TFA in MilliQ water) and solvent B (0.1% TFA in 9:1 acetonitrile/water). All peptides were lyophilized after HPLC purification, and subsequently analyzed using analytical HPLC and MALDI-TOF MS.^[18]

Preparation of Hydrogel: All compounds were placed in a glass tube (diameter 10 mm) and first dissolved in D.I water, sonicated for 5 min and putted on ice for 30 min. Then added 2x PBS buffer (5.4 mM KCl, 20 mM Na₂HPO₄, 4 mM KH₂PO₄) or D.I water to reach the final concentration of 1.0 wt%. pH was adjusted with NaOH or HCl. The solutions were stored in 37°C incubator overnight. Gelation was confirmed by the inverting method. In this method, when peptide solution had already formed a gel at the bottom of sample vial, the vial was inverted, and the gel remained in place without falling or flowing.

Circular Dichroism Spectroscopy: Circular Dichroism spectra were collected on Jasco X spectropolarimeter (Jasco corp., Tokyo, Japan). CD wavelength spectra were measured from 260 to 200 nm using a 0.1 mm quartz cell. Wavelength scans were collected by scanning in 1 nm step intervals with a 3s averaging time.

Oscillatory Rheology: All rheological experiments were performed on an Anton Parr equipped with a steel 15 mm parallel geometry tool. In a typical time-dependent experiment, peptide solution was transferred to the rheometer stage and lower the geometry to 0.5 mm, then the temperature was increased to 37°C within 1.0 min. To avoid dehydration, a layer of silicon oil was applied around the edge of the sample at the start of the measurement. Dynamic strain sweep experiments were performed to ensure that the time-sweep data was collected in the linear regime of strain. The dynamic strain sweep was performed varying the strain from 0.1 to 100% at a constant frequency (6 rad s⁻¹).

Transmission Electron Microscopy: The sample was prepared by placing a drop of peptide solution on a 200-mesh copper grid (Electron Microscopy China) and allowed to stand for 1.0 min, then blotted with filter paper. Subsequently a drop of 1.0% Uranyl Formate was placed on the grid and allowed to stand for 1–2 min, then blotted with a piece of filter paper and left to air dry. Images were taken with a JEOL JEM-2100Plus at 80 kV accelerating voltage. By calculating the width of peptide, The image J was used to measure 30 times for 1 TEM picture, and to gather statistics with frequency.^[26]

Fourier Transform Infrared Spectroscopy (FTIR): Hydrogel sample was prepared for FTIR studies at a concentration of 1.0 wt% in PBS buffer or D.I water. Prepared hydrogel was lyophilized and dried hydrogel (xerogel)

powder was embedded in KBr pellet and analyzed in FTIR. The spectrum was collected using a Nicolet In MX microscopic infrared spectrometer (Thermo Scientific Co., USA) between the wavelengths 4000–400 cm^{-1} under 16 scans on an average. KBr thin film was used as blank control.^[70]

Cell Viability Assay: MTT assay was employed to assess cytotoxicity of all molecules. In a typical experiment, NHDF cells were seeded into 96-well plate at a density of 8000 cells/well, allowed to adhere overnight at 37°C, 5% CO_2 . The culture medium was replaced with fresh serum-free medium containing 0.1–500 μM peptides. Blank medium or DMSO was used as positive control and negative control, respectively. After 48 h incubation period, 100 μL of fresh serum-containing media was added into each well. 10 μL of (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyl-tetrazolium bromide (MTT, 5 mg mL^{-1} in PBS) was added to each well and samples incubated for 3–4 hours, then the medium was replaced with 100 μL DMSO and incubated at 37°C with shaking for 0.5–1 h to facilitate formazan crystal solubilization. Absorbance was recorded at 540 nm using a UV microplate reader (Molecular Devices, Spectra Max M5). The absorbance of the negative controls was subtracted from each sample as a blank, and the percent viability was calculated as follows: (Absorbance peptide-treated cells – Absorbance negative controls) / (Absorbance untreated cells – Absorbance negative controls) \times 100. IC_{50} was calculated using Graphpad Prism 9.0 software.^[71]

Biocompatibility Test: Following the protocol in previous work,^[13] hydrogel (1.0 wt%, 50 μL) was prepared in a 96 well plate. The plate was placed into an incubator at 37°C and 5% CO_2 and allowed to equilibrate for 24 h. Serum-free MEM- α media (Gibco) of 100 μL was added on the top of gel and equilibrate overnight. Stem cell from human exfoliated deciduous teeth was trypsinized and counted using a hemocytometer. The resulting suspension was diluted with serum containing DMEM, 100 μL cell suspension (8,000 cells mL^{-1}) was placed onto the top of hydrogel. After 24 h incubation, the medium was removed and washed gently with PBS to remove the serum proteins. Cell viability was evaluated by using a Live/dead assay. Typically, 100 μL assay buffer containing both 1 μL calcein AM and 1 μL PI was added into each well. The dye was allowed to incubator for 30 min before washing with PBS 3 times, after that 100 μL cell imaging solution was added into each well for imaging. Fluorescence images were taken on EVOS FL Auto.^[60]

Statistics: All quantitative statistical experiments were replicated at least three times ($n = 3$). Data were presented as mean \pm standard deviation ($X \pm \text{SD}$). Statistical analyses were performed in GraphPad Prism 9 software.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (21975068, 51903082, 62372159), Natural Science Foundation of Hunan Province (2022JJ10008, 2020RC3017), start-up package from Hunan University, and Office of Research Administration (ORA) at King Abdullah University of Science and Technology (KAUST) under award numbers FCC/1/1976-44-01, FCC/1/1976-45-01, URF/1/4663-01-01, REI/1/5202-01-01, REI/1/4940-01-01, and RGC/3/4816-01-01.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

X.R., J.W., and X.L. contributed equally to this work. J.S. and X.Z. conceived the original ideas and guided the project. X.L., X.R. and J.W. designed and

performed the experiments, and analyzed the data. X.G., K.L., Y.L., Q.Z., L.W. and D.C. guided the design of computing algorithm. S.Y. helped in sample collection. X.W. and X.J. did part of wet laboratory experiments. M.L. constructed the website. All authors provided critical feedback and helped to shape the research, analysis and manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

artificial intelligence, deep generative model, machine learning, self-assembly

Received: January 23, 2024

Revised: March 10, 2024

Published online: May 5, 2024

- [1] X. Du, J. Zhou, J. Shi, B. Xu, *Chem. Rev.* **2015**, *115*, 13165.
- [2] L. A. Estroff, A. D. Hamilton, *Chem. Rev.* **2004**, *104*, 1201.
- [3] J. Gao, J. Zhan, Z. Yang, *Adv. Mater.* **2020**, *32*, 1805798.
- [4] E. Gazit, *Chem. Soc. Rev.* **2007**, *36*, 1263.
- [5] R. V. Uljijn, A. M. Smith, *Chem. Soc. Rev.* **2008**, *37*, 664.
- [6] M. J. Webber, E. A. Appel, E. W. Meijer, R. Langer, *Nat. Mater.* **2016**, *15*, 13.
- [7] S. E. Miller, Y. Yamada, N. Patel, E. Suárez, C. Andrews, S. Tau, B. T. Luke, R. E. Cachau, J. P. Schneider, *ACS Cent. Sci.* **2019**, *5*, 1750.
- [8] P. Majumder, A. Singh, Z. Wang, K. Dutta, R. Pahwa, C. Liang, C. Andrews, N. L. Patel, J. Shi, N. de Val, S. T. R. Walsh, A. B. Jeon, B. Karim, C. D. Hoang, J. P. Schneider, *Nat. Nanotechnol.* **2021**, *16*, 1251.
- [9] Y. Kuang, J. Shi, J. Li, D. Yuan, K. A. Alberti, Q. Xu, B. Xu, *Angew. Chem., Int. Ed.* **2014**, *53*, 8104.
- [10] X. Li, Y. Wang, Y. Zhang, C. Liang, Z. Zhang, Y. Chen, Z. W. Hu, Z. Yang, *Adv. Funct. Mater.* **2021**, *31*, 2100729.
- [11] Z. Álvarez, A. N. Kolberg-Edelbrock, I. R. Sasselli, J. A. Ortega, R. Qiu, Z. Syrgiannis, P. A. Mirau, F. Chen, S. M. Chin, S. Weigand, E. Kiskinis, S. I. Stupp, *Science* **2021**, *374*, 848.
- [12] L. Schnaider, S. Brahmachari, N. W. Schmidt, B. Mensa, S. Shaham-Niv, D. Bychenko, L. Adler-Abramovich, L. J. W. Shimon, S. Kulusheva, W. F. DeGrado, E. Gazit, *Nat. Commun.* **2017**, *8*, 1365.
- [13] J. Shi, G. Fichman, J. P. Schneider, *Angew. Chem., Int. Ed.* **2018**, *57*, 11188.
- [14] C. G. Pappas, R. Shafi, I. R. Sasselli, H. Siccardi, T. Wang, V. Narang, R. Abzalimov, N. Wijerathne, R. V. Uljijn, *Nat. Nanotechnol.* **2016**, *11*, 960.
- [15] C. Li, A. Iscen, H. Sai, K. Sato, N. A. Sather, S. M. Chin, Z. Álvarez, L. C. Palmer, G. C. Schatz, S. I. Stupp, *Nat. Mater.* **2020**, *19*, 900.
- [16] J. Boekhoven, W. E. Hendriksen, G. J. M. Koper, R. Eelkema, J. H. v. Esch, *Science* **2015**, *349*, 1075.
- [17] I. Yoshimura, Y. Miyahara, N. Kasagi, H. Yamane, A. Ojida, I. Hamachi, *J. Am. Chem. Soc.* **2004**, *126*, 12204.
- [18] K. Jian, C. Yang, T. Li, X. Wu, J. Shen, J. Wei, Z. Yang, D. Yuan, M. Zhao, J. Shi, *J. Nanobiotechnol.* **2022**, *20*, 201.
- [19] P. W. J. M. Frederix, G. G. Scott, Y. M. Abul-Hajja, D. Kalafatovic, C. G. Pappas, N. Javid, N. T. Hunt, R. V. Uljijn, T. Tuttle, *Nat. Chem.* **2015**, *7*, 30.
- [20] M. Aoki, K. Murata, S. Shinkai, *Chem. Lett.* **1991**, *20*, 1715.
- [21] Y. C. Lin, R. G. Weiss, *Macromolecules* **1987**, *20*, 414.

- [22] S. Zhang, T. Holmes, C. Lockshin, A. Rich, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 3334.
- [23] B. Xing, C.-W. Yu, K.-H. Chow, P.-L. Ho, D. Fu, B. Xu, *J. Am. Chem. Soc.* **2002**, *124*, 14846.
- [24] T. C. Holmes, S. de Lacalle, X. Su, G. S. Liu, A. Rich, S. G. Zhang, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6728.
- [25] Y. Zhang, H. Gu, Z. Yang, B. Xu, *J. Am. Chem. Soc.* **2003**, *125*, 13680.
- [26] J. Shi, X. Du, D. Yuan, J. Zhou, N. Zhou, Y. Huang, B. Xu, *Biomacromolecules* **2014**, *15*, 3559.
- [27] D. W. Choo, J. P. Schneider, N. R. Graciani, J. W. Kelly, *Macromolecules* **1996**, *29*, 355.
- [28] M. Reches, E. Gazit, *Science* **2003**, *300*, 625.
- [29] B. Y. Feng, B. H. Toyama, H. Wille, D. W. Colby, S. R. Collins, B. C. H. May, S. B. Prusiner, J. Weissman, B. K. Shoichet, *Nat. Chem. Biol.* **2008**, *4*, 197.
- [30] M. W. Mullooney, K. R. Duncan, S. S. Elsayed, N. Garg, J. J. van der Hooft, N. I. Martin, D. Meijer, B. R. Terlouw, F. Biermann, K. Blin, J. Durairaj, M. Gorostiola González, E. J. N. Helfrich, F. Huber, S. Leopold-Messer, K. Rajan, T. de Rond, J. A. van Santen, M. Sorokina, M. J. Balunas, M. A. Beniddir, D. A. van Bergeijk, L. M. Carroll, C. M. Clark, D.-A. Clevert, C. A. Dejong, C. Du, S. Ferrinho, F. Grisoni, A. Hofstetter, et al., *Nat. Rev. Drug Discovery* **2023**, *22*, 895.
- [31] G. Turon, J. Hlozek, J. G. Woodland, A. Kumar, K. Chibale, M. Duran-Frigola, *Nat. Commun.* **2023**, *14*, 5736.
- [32] C. Yan, D. J. Pochan, *Chem. Soc. Rev.* **2010**, *39*, 3528.
- [33] K. Li, J. Wang, Y. Song, Y. Wang, *Nat. Commun.* **2023**, *14*, 2789.
- [34] S. S. V. J. N. Law, C. E. Tripp, D. Duplyakin, E. Skordilis, D. Biagioni, R. S. Paton, P. C. S. John, *Nat. Machine Intell.* **2022**, *4*, 720.
- [35] Y. Cheng, Y. Gong, Y. Liu, B. Song, Q. Zou, *Brief Bioinform.* **2021**, *22*, bbab344.
- [36] J. Meyers, B. Fabian, N. Brown, *Drug Discov. Today* **2021**, *26*, 2707.
- [37] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
- [38] A. Gupta, J. Zou, *Nat. Machine Intell.* **2019**, *1*, 105.
- [39] P. Das, T. Sercu, K. Wadhawan, I. Padhi, S. Gehrman, F. Cipcigan, V. Chenthamarakshan, H. Strobelt, C. Dos Santos, P. Y. Chen, Y. Y. Yang, J. P. K. Tan, J. Hedrick, J. Crain, A. Mojsilovic, *Nat. Biomed. Eng.* **2021**, *5*, 613.
- [40] R. Chowdhury, N. Bouatta, S. Biswas, C. Floristean, A. Kharkar, K. Roy, C. Rochereau, G. Ahdriz, J. Zhang, G. M. Church, P. K. Sorger, M. AlQuraishi, *Nat. Biotechnol.* **2022**, *40*, 1617.
- [41] X. Zeng, F. Wang, Y. Luo, S. G. Kang, J. Tang, F. C. Lightstone, E. F. Fang, W. Cornell, R. Nussinov, F. Cheng, *Cell Rep. Med.* **2022**, *3*, 100794.
- [42] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R. R. Shayakhmetov, A. Zhebrak, L. I. Minaeva, B. A. Zagribelnyy, L. H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, *Nat. Biotechnol.* **2019**, *37*, 1038.
- [43] R. Batra, T. D. Loeffler, H. Chan, S. Srinivasan, H. Cui, I. V. Korendovych, V. Nanda, L. C. Palmer, L. A. Solomon, H. C. Fry, S. K. R. S. Sankaranarayanan, *Nat. Chem.* **2022**, *14*, 1427.
- [44] T. Xu, J. Wang, S. Zhao, D. Chen, H. Zhang, Y. Fang, N. Kong, Z. Zhou, W. Li, H. Wang, *Nat. Commun.* **2023**, *14*, 3880.
- [45] J. Wang, Z. Liu, S. Zhao, T. Xu, H. Wang, S. Z. Li, W. Li, *Adv. Sci.* **2023**, *10*, 2301544.
- [46] D. Weininger, *J. Chem. Inform. Comp. Sci.* **1988**, *28*, 31.
- [47] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, *Sci. Technol.* **2020**, *1*, 045024.
- [48] G. Landrum Google Scholar **2006**,
- [49] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, *Nat. Biotechnol.* **2019**, *37*, 38.
- [50] L. McInnes, J. Healy, J. Melville, arXiv preprint arXiv:1802.03426 **2018**,
- [51] F. Li, J. Han, T. Cao, W. Lam, B. Fan, W. Tang, S. Chen, K. L. Fok, L. Li, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 11259.
- [52] Y. Zhang, Y. Kuang, Y. Gao, B. Xu, *Langmuir* **2011**, *27*, 529.
- [53] N. J. Greenfield, *Nat. Protoc.* **2006**, *1*, 2876.
- [54] H. Dong, M. Wang, S. Fan, C. Wu, C. Zhang, X. Wu, B. Xue, Y. Cao, J. Deng, D. Yuan, J. Shi, *Angew. Chem. Int. Ed. Engl.* **2022**, *61*, e202212829.
- [55] K. A. Oberg, J.-M. Ruyschaert, E. Goormaghtigh, *Eur. J. Biochem.* **2004**, *271*, 2937.
- [56] P. I. Haris, D. Chapman, *Biopolymers: Orig. Res. Biomol.* **1995**, *37*, 251.
- [57] J. Lou, D. J. Mooney, *Nat. Rev. Chem.* **2022**, *6*, 726.
- [58] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Felix, M. P. Magarinos, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Maranon, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, *Nucleic Acids Res.* **2019**, *47*, D930.
- [59] P. Agrawal, S. Bhalla, S. S. Usmani, S. Singh, K. Chaudhary, G. P. Raghava, A. Gautam, *Nucleic Acids Res.* **2016**, *44*, D1098.
- [60] J. Shi, J. P. Schneider, *Angew. Chem., Int. Ed.* **2019**, *58*, 13706.
- [61] G. Guidotti, L. Brambilla, D. Rossi, *Trends Pharmacol. Sci.* **2017**, *38*, 406.
- [62] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324.
- [63] J. Shi, X. Du, D. Yuan, R. Haburcak, N. Zhou, B. Xu, *Bioconjugate Chem.* **2015**, *26*, 1879.
- [64] J. Shi, X. Du, Y. Huang, J. Zhou, D. Yuan, D. Wu, Y. Zhang, R. Haburcak, I. R. Epstein, B. Xu, *J. Am. Chem. Soc.* **2015**, *137*, 26.
- [65] X. Zhang, J. Yi, G. Yang, C. Wu, T. Hou, D. Cao, *Brief Bioinform* **2022**,
- [66] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, *Front. Pharmacol.* **2020**, *11*, 565644.
- [67] K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, G. Klambauer, *J. Chem. Inf. Model.* **2018**, *58*, 1736.
- [68] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, *ChemMedChem* **2008**, *3*, 1503.
- [69] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887.
- [70] S. D. Moran, M. T. Zanni, *J. Phys. Chem. Lett.* **2014**, *5*, 1984.
- [71] T. Li, C. Zhu, C. Liang, T. Deng, X. Wu, K. Wen, X. Feng, D. Yuan, B. Xu, J. Shi, *ACS Appl. Nano Mater.* **2023**, *6*, 7785.