European Journal of Medicinal Chemistry 275 (2024) 116628

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: www.elsevier.com/locate/ejmech

Research paper

PepExplainer: An explainable deep learning model for selection-based macrocyclic peptide bioactivity prediction and optimization

Silong Zhai [a,1], Yahong Tan [b,1], Cheng Zhu [a], Chengyun Zhang [a], Yan Gao [c], Qingyi Mao [a], Youming Zhang [b], Hongliang Duan [d,**], Yizhen Yin [b,e,*]

[a] School of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou, 310014, China
[b] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao, 266237, China
[c] Qilu Institute of Technology, Jinan, 250200, China
[d] Faculty of Applied Sciences, Macao Polytechnic University, Macao, 999078, China
[e] Shandong Research Institute of Industrial Technology, Jinan, 250101, China

ARTICLE INFO

Keywords:
Macrocyclic peptide
Machine learning (ML)
Graph neural network (GNN)
Bioactivity prediction
Structure-activity relationship (SAR)
Optimization

ABSTRACT

Macrocyclic peptides possess unique features, making them highly promising as a drug modality. However, evaluating their bioactivity through wet lab experiments is generally resource-intensive and time-consuming. Despite advancements in artificial intelligence (AI) for bioactivity prediction, challenges remain due to limited data availability and the interpretability issues in deep learning models, often leading to less-than-ideal predictions. To address these challenges, we developed PepExplainer, an explainable graph neural network based on substructure mask explanation (SME). This model excels at deciphering amino acid substructures, translating macrocyclic peptides into detailed molecular graphs at the atomic level, and efficiently handling non-canonical amino acids and complex macrocyclic peptide structures. PepExplainer's effectiveness is enhanced by utilizing the correlation between peptide enrichment data from selection-based focused library and bioactivity data, and employing transfer learning to improve bioactivity predictions of macrocyclic peptides against IL-17C/IL-17 RE interaction. Additionally, PepExplainer underwent further validation for bioactivity prediction using an additional set of thirteen newly synthesized macrocyclic peptides. Moreover, it enabled the optimization of the IC_{50} of a macrocyclic peptide, reducing it from 15 nM to 5.6 nM based on the contribution score provided by PepExplainer. This achievement underscores PepExplainer's skill in deciphering complex molecular patterns, highlighting its potential to accelerate the discovery and optimization of macrocyclic peptides.

1. Introduction

Macrocyclic peptides, a unique class of molecules positioned between small molecules and biologics, exhibit distinctive features that make them highly promising as a drug modality [1]. Noteworthy characteristics include their synthetic feasibility, high affinity and specificity, tissue-penetrating capabilities, and low toxicity levels [2–4]. Leveraging these inherent attributes, macrocyclic peptides have exhibited notable potential in targeting both intracellular and extracellular "undruggable" targets, such as protein-protein interactions [5–7] (PPIs), across various therapeutic fields [8–10].

Given the potential of macrocyclic peptides, the evolution of screening techniques [11,12] has played a critical role in facilitating the *de novo* discovery of macrocyclic peptides. The *in vitro* method employing mRNA display [13,14] offers a vast library of over 10^{12} sequences for screening, significantly increasing the possibility of identifying peptides with desired properties [13,15]. Furthermore, the mRNA display can be integrated with the flexible *in vitro* translation (FIT) system [16,17], powered by advanced Flexizymes technology, to give the random non-standard peptide integrated discovery (RaPID) system. The system has emerged as a revolutionary platform, streamlining the screening process for macrocyclic peptides that include non-canonical amino acids (NCAAs) [18]. The incorporation of NCAAs broadens the chemical diversity of macrocyclic peptides and introduces novel

* Corresponding author. State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao, 266237, China.
** Corresponding author.
E-mail addresses: hduan@mpu.edu.mo (H. Duan), yizhenyin.1987@sdu.edu.cn (Y. Yin).
[1] These authors contributed equally: Silong Zhai and Yahong Tan.

https://doi.org/10.1016/j.ejmech.2024.116628
Received 17 April 2024; Received in revised form 21 June 2024; Accepted 24 June 2024
Available online 25 June 2024
0223-5234/© 2024 Elsevier Masson SAS. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

possibilities for identifying molecules with distinctive biological activities.

While screening techniques have indeed facilitated the discovery of novel macrocyclic peptides, the integration of artificial intelligence has the potential to further enhance the discovery efficacy [19–21]. These investigations relied on deep learning algorithms, which are artificial neural networks with multiple processing layers capable of modeling complex nonlinear input-output relationships, performing pattern recognition, and feature extraction from low-level data representations [22]. These deep learning approaches have demonstrated the ability to not only rival but also surpass the effectiveness of traditional machine learning [23–25] and quantitative structure-activity relationship [26,27] (QSAR) approaches in drug discovery [28–30]. However, deep learning often lacks model interpretability in capturing complex nonlinear relationships [31,32], especially in medicinal chemistry [33,34], where interpretability is crucial not only for validating scientific hypotheses but also for providing valuable insights for molecular structure optimization.

Currently, researchers are exploring the use of models trained with graph neural network (GNN) architectures [35] to enhance interpretability [36]. GNNs have been applied in the field of drug discovery, for instance, in predicting molecular properties [37,38] and in generative models for *de novo* drug design [39]. However, traditional GNNs face challenges in analyzing macrocyclic peptides, as they typically offer interpretations at the atomic or bond level, rather than at a more macroscopic amino acid level. This limitation makes it difficult for GNNs to provide SAR that are useful for medicinal chemists in structural optimization and new drug design of macrocyclic peptides. The recent Substructure Masking Explanation (SME) method proposed by Wu et al. [40] has been effective in identifying and interpreting key molecular substructures in graph convolutional networks (GCNs), particularly in tasks like estimated solubility (ESOL), mutagenicity, and blood-brain barrier permeability (BBBP). SME incorporates various well-designed molecular segmentation methods such as BRICS [41] substructures, Murcko [42,43] frameworks, and functional groups, providing coherent interpretations with chemical principles. Despite these advances, to the best of our knowledge, there has not yet been a graph interpretation model specifically developed for macrocyclic peptides.

In this study, we developed PepExplainer, a tool that utilize SME to identify key chemical fragments within macrocyclic peptides. Our research shows that there is a correlation between the enrichment of macrocyclic peptides from the focused library [20] and their bioactivity, with a Pearson correlation coefficient of 0.84. To improve the predictive precision of PepExplainer, we thus integrated a transfer learning strategy [44,45]. PepExplainer initially pretrained on large scale data from selection to learn the relationship between peptide structure and properties. This approach significantly improves the model's efficiency and accuracy in identifying potential macrocyclic peptides, which is reflected in the enhanced R^2 and RMSE metrics. The efficacy of PepExplainer is further showcased in a case study, where it significantly aids in optimizing macrocyclic peptides.

2. Results and discussion

2.1. Using PepExplainer with transfer learning strategy to predict the bioactivity of macrocyclic peptides

Aiming to accurately predict the biological activity and interpret the SAR of macrocyclic peptides, we attempt to develop a robust model with superior performance. While most deep learning models developed for peptide-related task [46–48] are Transformer-based [49], which is well-regarded for its exceptional sequence fitting capability. Its application to linear peptides appears more suitable than to macrocyclic peptides. When dealing with macrocyclic peptides that contain unique structures or NCAAs [50], amino acid sequence representation methods with numerical encodings face challenges [51]. These methods are difficult to capture the intricate macrocyclic structure adequately and often falter in representing NCAAs. In addressing this challenge, GNNs are able to consider macrocyclic peptides as comprehensive molecular graphs. Consequently, it enables the proficient encoding of the intricate molecular structure of macrocyclic peptides, paving the way for in-depth exploration and prediction of their properties.

Next, we also seek to incorporate an interpretability mechanism in our model and dissect the contribution of amino acids to macrocyclic peptide bioactivity. Thus, we employed the SME technique, harnessing the capabilities of Relational Graph Convolutional Networks [52] (RGCNs). This specialized subclass of GNNs is tailored to adeptly manage heterogeneously structured data. As depicted in Fig. 1b, SME employs a masking mechanism to reveal the relationships between substructures within macrocyclic peptides and their impact on bioactivity.

By integrating GNNs for prediction and SME for SAR discussion, our model, designated as PepExplainer, was constructed. To further enhance the reliability and precision of both prediction and explanation, our methodology incorporates a multi-model ensemble. This ensemble comprises ten PepExplainers, each initialized with random seeds, ultimately forming a consensus model (Fig. 1c). The average of predictions from the ten PepExplainers determines the model's output.

To assess the contribution of each amino acid, we calculated its value using the formula derived from the difference between predictions of biological activity without amino acid masking and predictions with amino acid masking (Fig. 1d). This approach allows for a thorough examination of each substructure's specific impact on the bioactivity of macrocyclic peptides, commonly known as SAR discussion. Moreover, by performing attribution analysis on amino acids at various positions (Fig. 1e), we can optimize specific positions based on their contributions, resulting in optimized macrocyclic peptides.

During the development of PepExplainer, predicting bioactivity in drug development posed a challenge due to limited bioactivity data, often leading to overfitting of the model. To overcome this, we utilized a two-step transfer learning strategy as shown in Fig. 1a. In the first step, we trained a regression predictor for log enrichment using extensive data on macrocyclic peptide enrichment from the focused library against IL-17C (selection dataset). This data is related to affinity and provided a pre-training source for PepExplainer to learn the relationship between peptides' structure and activity. This training helped PepExplainer to gain insights into macrocyclic peptides and their interactions with IL-17C. In the second step, we fine-tuned the pre-trained PepExplainer using bioactivity dataset. This refinement step aimed to enhance the model's precision in predicting biological activity and reduce its dependence on extensive biological activity data.

Overall, our approach combines PepExplainer and transfer learning to predict the bioactivity of macrocyclic peptides. PepExplainer functions as an interpretable tool, offering insights for analyzing peptide substructures. This work addresses key challenges in macrocyclic peptide research, including modeling, data scarcity, and model interpretability.

2.2. Comparison of PepExplainer with baseline models

After constructing the model's architecture, we next seek to evaluate the performance of the model. The model's overall performance was quantified using the R^2 and RMSE calculated on the pIC_{50} values. The R^2 metric is essential for evaluating how well the model's predictions correspond to the actual observed values, serving as a crucial benchmark for evaluating the model's predictive power. Meanwhile, the RMSE offers a quantitative measure of the prediction errors, providing insight into the precision of the model's predictions.

In this study, we examined four traditional machine learning algorithms commonly used for SAR prediction: K-Nearest Neighbors [53] (KNN), Random Forest [54] (RF), Gradient Boosting Machine [55] (GBM), and Support Vector Machine [56] (SVM). To broaden the

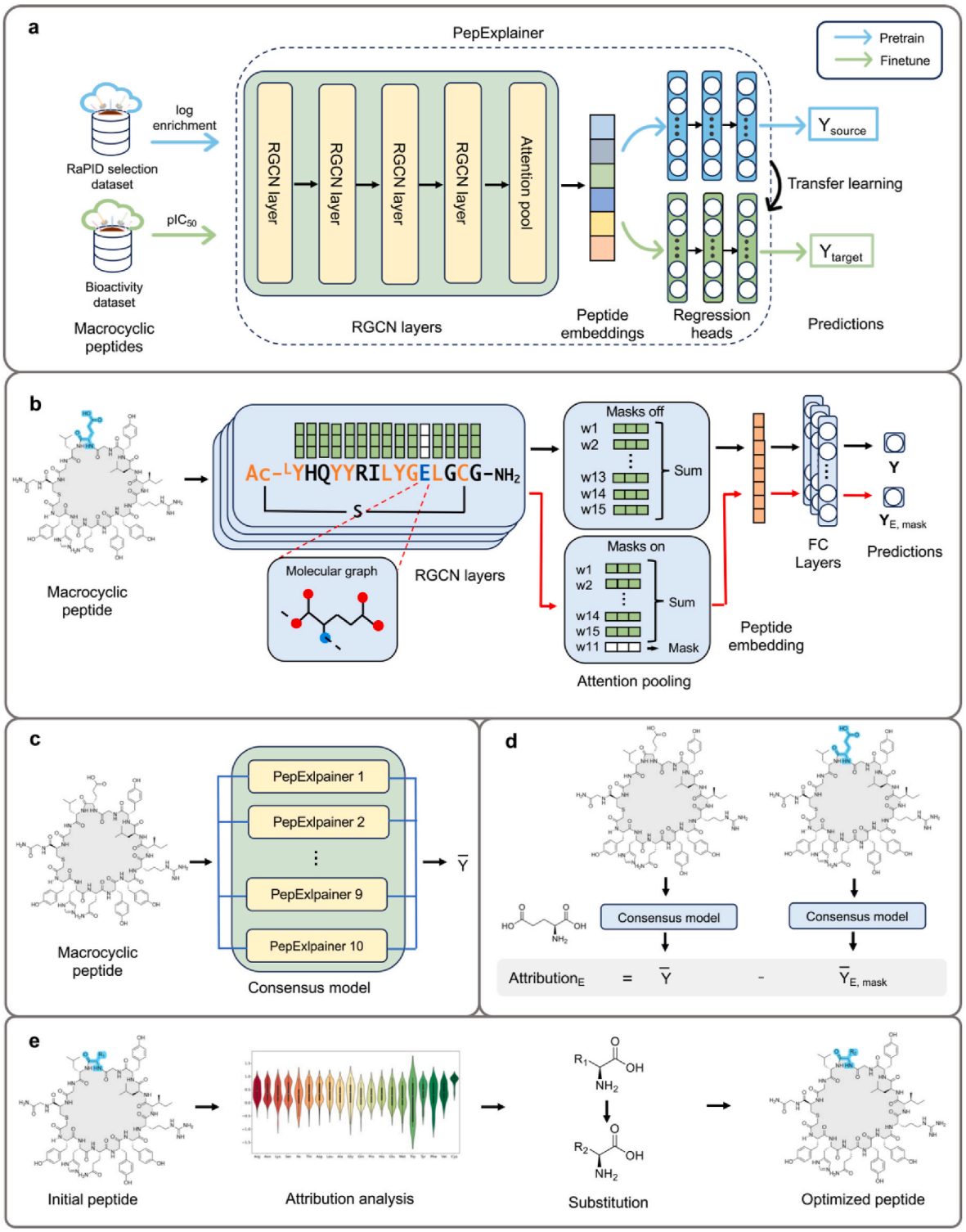

Fig. 1. PepExplainer and substructure mask explanation for macrocyclic peptide bioactivity prediction and optimization. (**a**) Transfer learning strategy for improved prediction of macrocyclic peptide bioactivity. (**b**) PepExplainer's architecture incorporates substructure masking, involving the input of a macrocyclic peptide structure, its conversion into a molecular graph, and subsequent processing through RGCN layers within PepExplainer to generate embeddings for prediction. Macrocyclic peptides are encoded as molecular graphs, with a highlighted amino acid denoted as E, featuring a blue node representing a nitrogen (N) atom and a red node representing an oxygen (O) atom. (**c**) An ensemble of multiple PepExplainers aggregates predictions, ensuring a more robust consensus on peptide bioactivity. (**d**) The consensus model focuses on attribution analysis, revealing how bioactivity changes when an amino acid (e.g., E) is masked. (**e**) The optimization of the target macrocyclic peptide according to amino acid contributions.

baseline models for comparative analysis, we combined these algorithms with four molecular descriptors and two peptide descriptors. These descriptors include Extended Connectivity Fingerprints [57] (ECFP), Molecular ACCess System [58] (MACCS) keys, Weighted Holistic Invariant Molecular [59] (WHIM) descriptors, and Physicochemical Properties (PHYSICHEM), along with two peptide representations of amino acid encoding: One-Hot Encodings (SEQ) and Physicochemical Properties (PHYSICHEM[b]). Additionally, we integrated deep learning models based on the Transformer architecture, like BERT [60] (bi-directional encoder representations from transformers), which have shown excellent performance in peptide-related tasks. By combining different peptide encoding techniques with machine learning algorithms and introducing sequence-based deep learning models, we developed a diverse set of predictive models.

For comprehensive details regarding the construction of the baseline models, readers are directed to refer to the Methods section and Table S1. Compared to baseline models, PepExplainer exhibits outstanding performance, especially when enhanced with transfer learning on enrichment datasets, substantially improving its predictive abilities. Fig. 2a and b illustrate this enhancement. It significantly outperforms PepBERT in key metrics, R^2 and RMSE when incorporating transfer learning. Specifically, it achieved an R^2 of 0.96 and an RMSE of 0.29 on the training set, an R^2 of 0.94 and an RMSE of 0.38 on the test set, as shown in Fig. 2c and d. In bioactivity prediction, a high R^2 and low RMSE indicate that the model accurately and reliably predicts with minimal error. And the consistent metrics scores across training and test sets indicate PepExplainer's effective avoidance of overfitting.

The improvement from transfer learning is evidenced by the

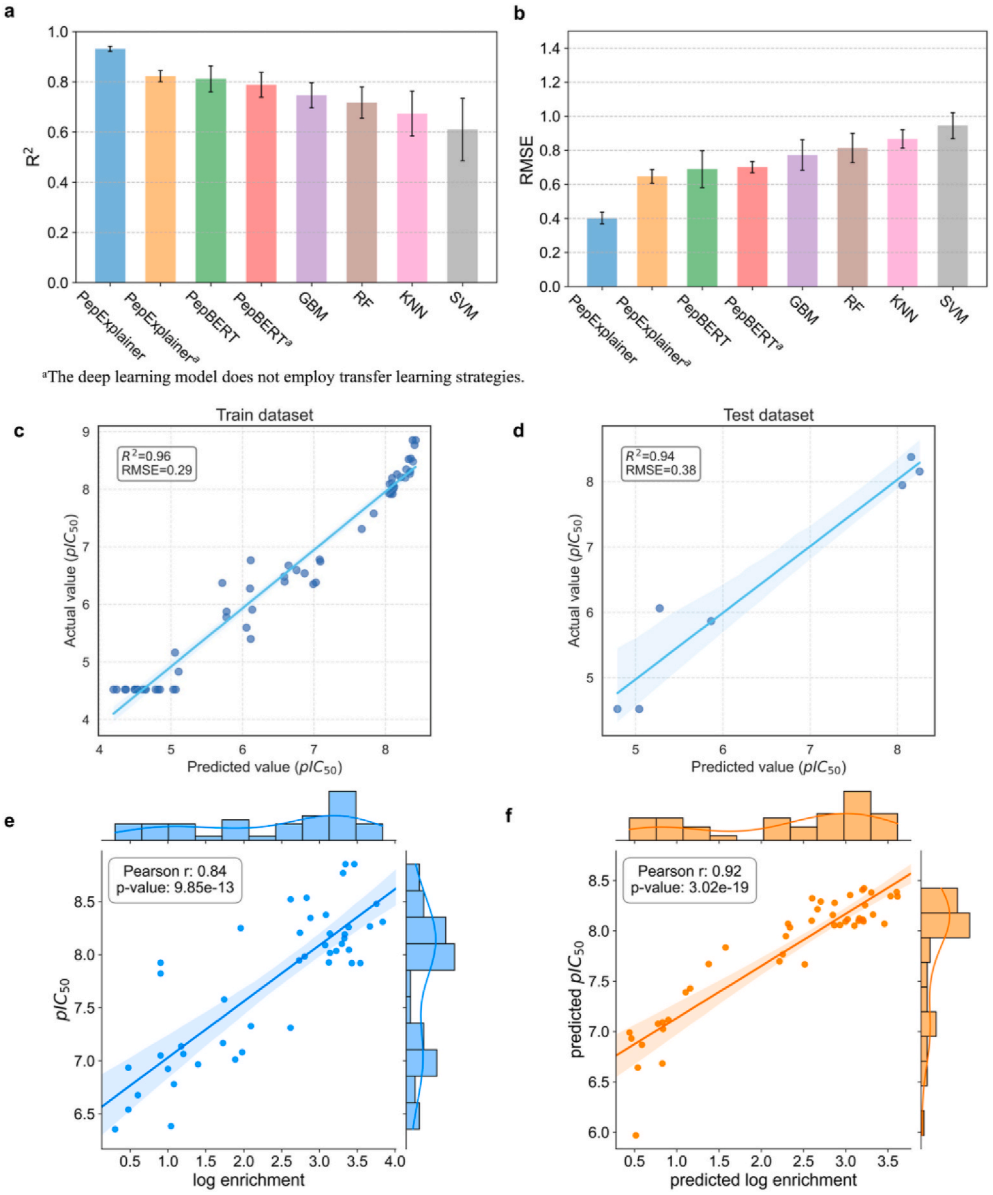

Fig. 2. Comparison of prediction results between PepExplainer and baseline models. It includes a comparison of PepExplainer with baseline models on the test dataset with (**a**) R^2 and (**b**) RMSE as performance metrics. The predicted values provided by PepExplainer in the bioactivity dataset: (**c**) for the training dataset and (**d**) for the testing dataset. It also explores the correlation between enrichment and bioactivity in the focused library of the selection against IL-17C. A logarithmic transformation is applied to both datasets to manage the data effectively, allowing for a clearer understanding of the relationship between variables and reducing the impact of outliers. (**e**) Shows a scatter plot for 45 peptides, revealing a strong correlation with a Pearson coefficient of 0.84 and a p-value of 9.85×10^{-13}. (**f**) Displays a scatter plot for the same peptides but using predicted labels from PepExplainer, showing an enhanced correlation with a Pearson coefficient of 0.92 and a p-value of 3.02×10^{-19}, further solidifying the relationship between enrichment and bioactivity.

correlation between the enrichment and bioactivity of macrocyclic peptides in the focused library, achieving a Pearson R of 0.84 (see Fig. 2e). Using PepExplainer for predictions further enhanced this correlation, with the Pearson R at 0.92 for predicted values (refer to Fig. 2f). This increase demonstrates that PepExplainer had learned the correlation and thus enhanced predictive accuracy after transfer learning. Moreover, the flexibility of transfer learning is proven by its practical application in a different architecture like BERT.

The effectiveness of deep learning models in understanding their inputted data depends on the appropriate representation. PepExplainer distinguishes itself in this aspect, as it uses a molecular graph to represent these peptides, capturing their atomic topological structure with greater precision. As shown in Fig. 3a, we employed the t-SNE [61] (t-distributed Stochastic Neighbor Embedding) algorithm to visualize high-dimensional data. Each point on the plot represents a macrocyclic peptide, with proximity suggesting similarity and color indicating enrichment levels. By incorporating enrichment level labels, PepExplainer effectively groups structurally similar peptides and distinctly separates clusters using color-coded enrichment levels. In contrast, the sequence-based approach of BERT, focusing on amino acid positions and types, falls short in depicting the detailed structure of macrocyclic peptides (Fig. 3b). This limitation means that it only learns the sequence-to-label information, not the specific structures, leading to less effective clustering of points compared to PepExplainer. As a result, BERT is difficult to clearly show the relationship between the structure of peptides and their enrichment levels in the t-SNE mapping. Furthermore, in an unsupervised learning scenario (Fig. S2a), BERT's representation appears scattered and disordered in the t-SNE plot.

Another representation method in bioinformatics is molecular fingerprints (Fig. S1), which turn molecular structures into mathematical vectors for a discrete description, including details about atoms, bonds, and charge distributions. We used ECFP to study macrocyclic peptides' chemical/biological features (Fig. S2b). ECFP and PepExplainer exhibit similar clustering patterns, with numerous clusters forming. This similarity underscores PepExplainer's proficiency in accurately capturing atomic-level physicochemical structural details, demonstrating a performance level equivalent to that of ECFP.

Overall, PepExplainer outperforms traditional machine learning models in accurately predicting macrocyclic peptide activity and excels in molecular structure visualization through t-SNE mapping. It effectively groups and distinguishes peptides, offering high predictive accuracy and enhanced structural insights through visualization.

2.3. The interpretability of PepExplainer for macrocyclic peptides

SAR enables a more profound comprehension of the relationships among molecular structure and properties. Interpreting peptide structures is straightforward for peptide chemists. However, interpreting deep learning models is challenging due to their "black box" nature. Even when predictions align with experimental data, the underlying mechanisms are complex. To address this, we designed PepExplainer to discriminate the contribution levels of positions and amino acid types in macrocyclic peptides to biological activity. By inputting the SMILES or graph representation, the model predicts the activity and assesses the impact of each amino acid, providing intuitive explanations for activity changes. Leveraging PepExplainer, we can ascertain the contribution of each amino acid. In this study, we utilized distinct colors to highlight the impacts of peptide substructures on their activities toward the IL-17C target protein.

Fig. 4 illustrates the visual outcomes for the top three peptides with the highest activity and those with lowest activity. Each amino acid within the macrocyclic peptides was colored to denote its impact on biological activity: gray signifies a minor impact (either positive or negative), light colors (light green for positive impact, yellow for negative impact) indicate a moderate impact, and dark colors (dark green for positive impact, orange for negative impact) signify a significant impact. The amino acid sequence of each macrocyclic peptide is presented below its structure, with amino acid characters colored to correspond with the structure. In Fig. S3, detailed contribution values are provided, displayed on a color spectrum ranging from −1 to 1, offering a continuous numerical representation of the varying impact levels of individual amino acids on enrichment values. The visualization for the whole bioactivity dataset is provided in Fig. S4. The visual representation of color patterns enables an intuitive understanding of how individual amino acids contribute to the activity of a macrocyclic peptide. A predominant presence of positively contributing amino acids is observed in peptides with high activity, whereas peptides with low activity typically exhibit negative values or minimal contributions. For example, in peptides demonstrating high activity, a noteworthy positive contribution is observed in the first half of the sequence, as highlighted in dark green for peptides Lib2-1, Lib2-2, and Lib2-3. Conversely, in peptides with low activity, the negative contribution becomes evident specifically at positions 2, 3, and 5, as highlighted in orange for peptides L20–15 and L20-16. Generally, macrocyclic peptides with low activity values display a color pattern where most amino acids appear in shades of gray. Moreover, when an amino acid at a particular position transitions from one type to another, substituting it with a structurally similar amino acid leads to minimal overall impact on activity, and the visual color patterns in the structure remain largely consistent. For instance, in peptides Lib2-1 and Lib2-2, the amino acids at position 10 differ, with Lib2-1 featuring N and Lib2-2 featuring D. These two amino acids differ by only one atom in their structures, possessing similar physicochemical

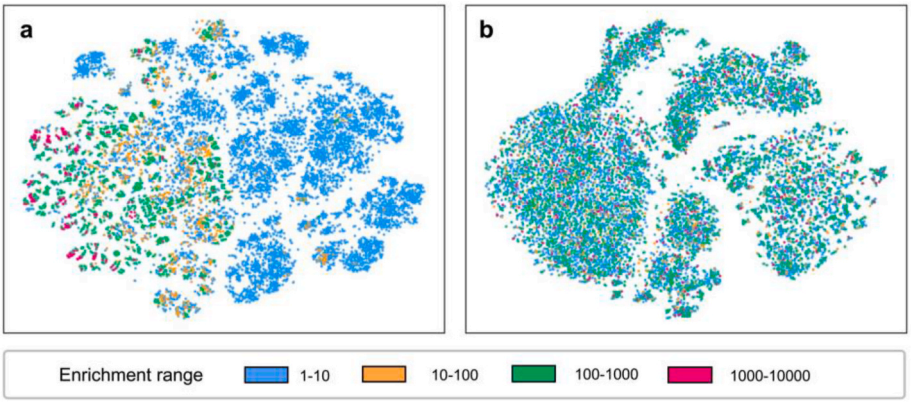

Fig. 3. Visualization of macrocyclic peptide molecular representations in selection dataset using t-SNE mapping. This figure delineates peptide structural representations through (**a**) PepExplainer embeddings and (**b**) PepBERT embeddings. Each point on the plot represents a macrocyclic peptide, with proximity suggesting similarity and color indicating enrichment levels.

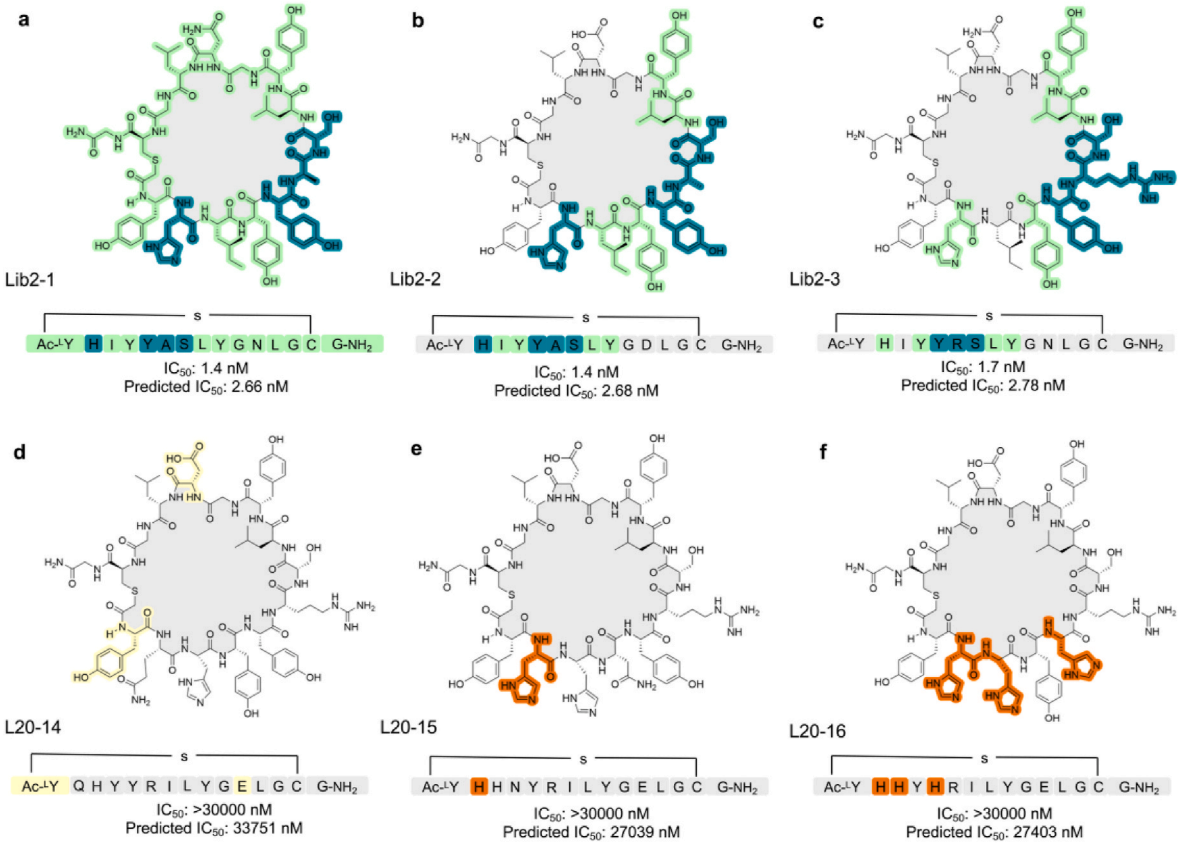

Fig. 4. Comparative analysis of macrocyclic peptides highlighting amino acid contributions to enrichment values. Six peptides are compared: the top three active (**a-c**) and the least active (**d-f**). The amino acids within each peptide are color-coded to reflect their affinity impact: gray indicates minimal impact, light colors (green for positive, yellow for negative) denote moderate influence and dark colors (green for positive, orange for negative) hues for a significant contribution. Beneath peptide's structural depiction is its amino acid sequence, color-matched for easy comparison. The color coding is adjusted based on the contribution values derived from PepExplainer.

properties. Consequently, they can exhibit the same IC_{50} value, 1.4 nM.

Thus, PepExplainer offers a visual analysis method for assessing the activity of individual peptides. It reveals the potential for fine-tuning molecular activity through the modification of particular amino acids or chemical structures.

2.4. Application in bioactivity prediction and optimization

Upon showing superior performance compared to other machine learning models and demonstrating the capability for interpreting the SAR of macrocyclic peptides, PepExplainer was subsequently subjected to additional validation to affirm its capability in predicting bioactivity. We chose an additional set of 13 macrocyclic peptides, initially predicting their bioactivities (Fig. 5). Subsequently, we proceeded to chemically synthesize these peptides, followed by their evaluation using enzyme-linked immunosorbent assay (ELISA). According to the predicted and experimentally assessed activities, we confirmed that the model is reliable, notwithstanding some observed fluctuations. Most macrocyclic peptides demonstrated notable consistency between the predicted IC_{50} values and the experimentally measured values. Take macrocyclic peptide cP1, for instance, where the predicted IC_{50} value of 21.3 nM aligns commendably with the experimentally measured value of 15 nM. Nevertheless, in the case of macrocyclic peptide cP3, the predicted IC_{50} value is 9.2 nM, whereas the actual value is 68 nM. Despite this notable difference, both values still fall within the logarithmic scale range. This is attributed to our choice of using pIC_{50} as the training target for activity, which imposes an upper accuracy limit of a 10-time difference on the logarithmic scale. Another reason is the relative scarcity of leucine (L) at the third position of cP3 from the N-terminus in the selection dataset. This suggests that the model lacks similar structural data for learning, consequently impacting the predictive accuracy for this cyclic peptide. PepExplainer performed 0.7 in R^2 and 0.54 in RMSE on the independent test set. Although these metrics are slightly lower than those observed in the previous training and test sets, the model still shows dependable predictive accuracy, especially when considering the logarithmic scale. Additionally, we evaluated the impact of reducing the training set size using 13 newly tested peptide data points as an independent test set, calculating R^2 and RMSE metrics (Fig. 5a). We systematically reduced the training set size (total is 59) by 10 %, 20 %, 40 %, 60 %, and 80 %, assessing its effect on model performance (Fig. S5). Our findings indicate that a minimum of 30 bioactivity data points in the training set is recommended for more reliable predictions. Despite the importance of bioactivity data, our model shows potential for accurate prediction and optimization even with a smaller dataset. Moreover, under the assistance of PepExplainer, we identified ten macrocyclic peptides with enhanced activity compared to 17C-L20, which showed moderate inhibitory activity in the initial library. In particular, cP1 has an IC_{50} of 15 nM.

To further check the predictive accuracy, we introduced mutations to the four most variable positions (3, 6, 7, 11 from the N-terminus) of the 17C-L20 peptide (IC_{50} = 166 nM), substituting each with 19 different amino acids. This resulted in the generation of 76 mutated macrocyclic peptides (4 × 19). We employed PepExplainer to predict the activities of these mutated macrocyclic peptides. The heatmap in Fig. 5b illustrates the variance between the predicted pIC_{50} values of the mutated macrocyclic peptides and the actual activity of 17C-L20. In the heatmap, green denotes increased activity for the mutated macrocyclic peptides. At the same time blue signifies a decrease, with the color intensity

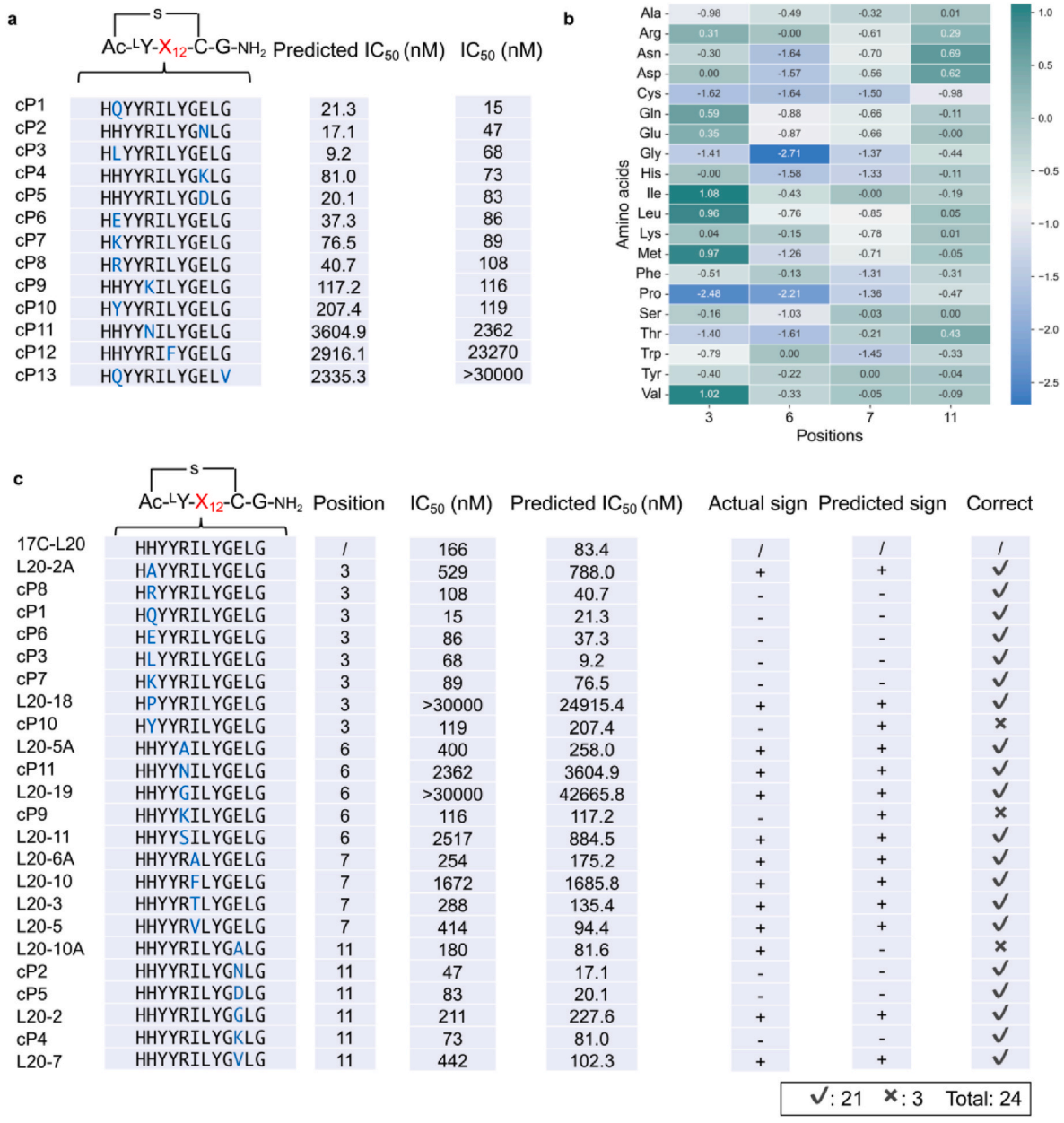

Fig. 5. Peptide position and amino acid type analysis using PepExplainer. (**a**) Newly tested peptides' bioactivity along with their IC_{50} values predicted by PepExplainer. (**b**) Single amino acid mutations of peptide 17C-L20 (YHHYYRILYGELGCG, IC_{50} = 166 nM) at four variable positions (position 3, 6, 7, 11). Mutations resulting in improved predicted IC_{50} values are highlighted in green. Conversely, mutations that negatively impact these values are distinctly marked in blue. In instances where a very high predicted standard deviation occurs, a label of 0 is assigned. (**c**) Detailed prediction outcomes for mutational analysis of the base macrocyclic peptide 17C-L20 at the four most variable positions.

reflecting the magnitude of the activity change. Out of these 76 variants, actual activity data were available for 24 macrocyclic peptides. When comparing the model's predicted ranking with the actual activity ranking, we identified only three macrocyclic peptides with predicted IC_{50} values that deviated from the actual results (Fig. 5c). In summary, PepExplainer exhibited a high accuracy of approximately 87.5 % (21 out of 24) in predicting the activity changes of mutated macrocyclic peptides, underscoring its effective prediction and ranking capability.

To further expand the application scope of PepExplainer, we implemented an average contribution optimization strategy to improve the biological activity of a specific peptide, cP1 (Fig. 6a). This strategy can enhance the biological activity of a macrocyclic peptide by introducing specific mutations at individual amino acid positions. The optimization process selected a single amino acid for modification based on the average impact of various amino acids on the biological activity of the macrocyclic peptide. cP1, with a mutation at position 3, demonstrated a significant enhancement in inhibitory activity compared to 17C-L20. Meanwhile, mutations at positions 6 and 7 appeared to negatively impact activity, as shown in Fig. 5b. Therefore, we opted to focus on mutations at position 11 to further enhance the activity of cP1. In this study, we randomly selected 2000 macrocyclic peptides from selection dataset from the focused library [20] and employed PepExplainer to assess the precise influence of amino acids at position 11 on affinity. In Fig. 6b, the attribution scores are visualized through a scatter plot, where orange denotes positive contributions and blue signifies negative ones. Adjacent to each amino acid, the average contribution is displayed in green (indicating a positive effect) or red (indicating a negative effect), providing insight into the overall impact on the

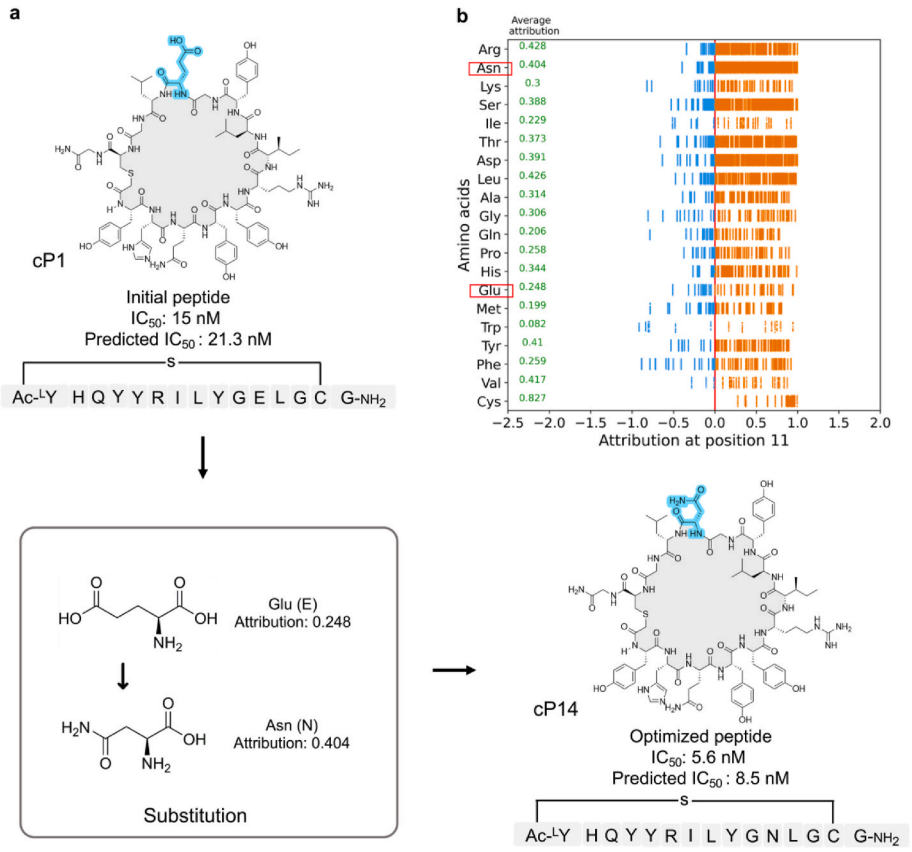

Fig. 6. The optimization of macrocyclic peptide according to the average attribution at position 11. (**a**) cP1 (YHQYYRILYGELGCG, IC_{50} = 15 nM) was optimized by substituting the amino acid from Glu (E) to Asn (N), resulting in cP14 (YHQYYRILYGNLGCG, IC_{50} = 5.6 nM). (**b**) The average attribution for substructures within sampled macrocyclic peptides in selection dataset.

peptide's affinity at that specific position. For a comprehensive analysis of average contributions at other crucial positions, please refer to Figs. S6 and S7. Amino acid E (glutamate) at position 11 contributes with a value of 0.248, whereas N (asparagine) and R (arginine) exhibit higher positive contributions of 0.404 and 0.428, respectively. Considering that introducing R could reduce protease resistance, despite its highest contribution score (Fig. 6b), we opted to mutate E to N instead and proceeded to validate the activity of the macrocyclic peptide cP14 after peptide synthesis. Notably, substituting E at position 11 with N resulted in a remarkable enhancement in the biological activity of the macrocyclic peptide, showing the IC_{50} value at 5.6 nM, as illustrated in Fig. 6a. It should be noted that the attribution value for C (cysteine) is 0.827 (Fig. 6b), indicating that cysteine might significantly impact the bioactivity of the peptide. However, we chose not to conduct further testing cysteine based on the following considerations: 1) During the synthesis of cyclic peptides, if cysteine is present at position 11, it may compete with the cysteine at position 14 to react with the N-terminal chloroacetyl group, forming a thioether cyclic peptide. 2) Despite the high attribution value for cysteine, its frequency is relatively low in our randomly sampled set of 2000 compounds. This suggests that although cysteine theoretically has a significant impact on activity, its representation in the overall dataset is insufficient to justify in-depth individual testing.

To assess the broad applicability of PepExplainer, we tested it with three distinct datasets. These datasets include the nonfouling dataset introduced by Ansari [62], the HLA class I binding dataset proposed by Chu et al. [46], and the thioether-cyclized peptides dataset from Merz et al. [63]: 1) The nonfouling dataset: Comprising 17,185 sequences with peptide lengths ranging from 5 to 20 residues, this dataset allowed us to establish baseline models for binary classification. PepExplainer showed strong performance, outperforming the LSTM [62] baseline but lagging behind PeptideBERT [64]. 2) The HLA class I binding dataset: This well-curated dataset includes peptide-HLA class I binding data for deep learning model training and optimization. It provided a robust platform for explanation analysis. PepExplainer achieved an accuracy of 0.935 an independent dataset of peptides binding to the HLA-A68:01 allele, comparable to TransPHLA [46], and successfully provided an optimization example based on attribution scores. 3) The thioether-cyclized peptides dataset: This dataset involves a combinatorial synthesis and screening approach based on sequential cyclization and one-pot peptide acylation and screening. It is closely aligned with the enrichment data techniques in this work. The dataset comprises 8448 cyclic peptides screened against the disease target thrombin, showcasing PepExplainer's performance (R^2 = 0.469) on macrocyclic peptides. Our results demonstrated that PepExplainer could effectively identify significant contributors within the dataset, confirming its effectiveness in screening complex peptide libraries, including NCAA and cyclic peptides. For further details, please refer to the Supporting Information in section "Extend PepExplainer to other datasets".

Thus, PepExplainer has proven to be a valuable tool for predicting and optimizing the biological activity of macrocyclic peptides. This contribution marks a substantial advancement in library-based screening, showing great potential in refining peptide drug development processes.

3. Conclusion

Macrocyclic peptides, bridging the gap between small molecules and proteins, hold significant promise in therapeutic applications. Their complex structures, however, pose a challenge for AI models in assessing

their biological activity. To tackle this, we developed PepExplainer, a model that utilizes advanced deep learning techniques, integrating GNNs and SME. This model excels in predicting and interpreting the activity of macrocyclic peptides, drawing on datasets from selection-based screening. A key finding is that incorporating enrichment data enhances prediction accuracy, as evidenced by improved R^2 and RMSE metrics. The improvement is attributed to the discovered correlation (Pearson R is 0.84) between bioactivity and enrichment values, addressing the issue of insufficient activity data hampering prediction capabilities. PepExplainer outperforms traditional machine learning methods, offering a novel approach to handle NCAAs and as an interpretable data analysis tool. Its practicality and efficacy in predicting the biological activity of macrocyclic peptides have been validated experimentally. We can decide which amino acids to modify to optimize a peptide through attribution analysis. The advent of PepExplainer could accelerate the discovery and optimization of peptides, representing an advancement over existing methodologies.

4. Methods

4.1. Datasets

Our research utilizes two principal datasets: the selection dataset and a bioactivity dataset (Fig. 1). The selection dataset, sourced from a focused library constructed via the RaPID system (Fig. S8) and the PepScaf [20] framework, underwent rigorous filtering to ensure data quality. Valid macrocyclic peptide sequences were defined as starting with "M" and ending with "CGSGSGSamber". Sequences deviating from this format were excluded, resulting in 163,949 valid macrocyclic peptide data points, each with 15 residues including ClAc-^{L}Y and C as linkage nodes.

For the focused library construction, we fixed amino acids at six key positions (2, 3, 6, 7, 11, and 13) across most macrocyclic peptides. The specific amino acid distribution at these positions is detailed in Fig. S9. In the selection dataset, the minimum enrichment value, equivalent to DNA sequencing reads, was set at 1, indicating negligible binding ability. Peptides with enrichment values below 10 constituted 86.1 % of the dataset.

PepExplainer's representational capabilities effectively capture the SAR of macrocyclic peptides. We demonstrate these capabilities through an interactive macrocyclic peptide atlas, offering visual clustering and similarity searching based on Schwaller's methodology [64] (Fig. S10).

The bioactivity dataset contains 66 peptide samples (59 for training and 7 for test), each with IC_{50} values, divided into training and test sets for comprehensive model evaluation. Notably, these sets include publicly available peptide data from previous studies, complemented by an additional independent test set involving new activity tests. These tests were conducted in this study to assess the model's activity prediction capability under varying distributions.

4.2. Amino acids substructure mask explanation

For effective peptide analysis, particularly concerning amino acid substitutions, we implemented the SME model with consensus methods. By training the model using various random seeds, we generate multiple independent predictions. The average and standard deviation of these predictions are then calculated, using the standard deviation as a measure of prediction reliability. This approach helps us evaluate the consistency and credibility of the predictions. Significant variation in the model's predictions for the same compound suggests substantial random error, reducing reliability. While our model does not directly account for specific experimental errors in each measurement, assessing the predictions' consistency and stability allows us to indirectly identify and reflect the potential impact of experimental errors on the results.

These models use RGCNs for enhanced molecular property prediction, focusing on meaningful amino acid substructures often missed by traditional GNN explanation methods. RGCNs, with edge feature integration, outperform standard physicochemical descriptors and deep learning models in interpretability. Our approach uniquely encodes macrocyclic peptides into molecular graphs, with specific amino acids and their constituent atoms represented as distinct colored nodes. Through RGCN layers, each node updates its state by integrating information from its neighbors, analogous to creating molecular fingerprints. The propagation rule for each node v is calculated via:

$$\mathbf{h}_v^{l+1} = \sigma\left(\sum_{r\in R}\sum_{u\in N_v^r}\mathbf{W}_r^{(l)}\mathbf{h}_u^{(l)} + \mathbf{W}_0^{(l)}\mathbf{h}_v^{(l)}\right)$$

where $h_v^{(l+1)}$ represents the state vector of a target node v after $l+1$ iterations, The term N_v^r refers to the neighbors of node v inked by an edge of type r, where R is the set of all edge types. The neighbors are nodes directly connected to v. We use $\sigma(\cdot)$ as an element-wise activation function, specifically adopting ReLU in our approach. The weight $W_r^{(l)}$ corresponds to the neighbor node u connected to v by an edge with relation $r \in R$, while $W_0{}^{(l)}$ is the weight for the target node v itself. This framework allows for the explicit incorporation of edge information into the RGCN for each relation $r \in R$. The weight $W_r^{(l)}$ is derived from a linear combination of basis transformations.

By leveraging attention pooling to aggregate node information, we can derive molecular embeddings and predict molecular properties via three fully connected (FC) layers. The attention pooling mechanism is defined as follows:

$$\begin{aligned} w_v &= \mathrm{sigmoid}(W\cdot\mathbf{h}_v + \mathrm{bias}) \\ \text{molecular embedding} &= \sum_{v=1}^{N}(\mathbf{w}_v\cdot\mathbf{h}_v) \end{aligned}$$

where W and bias are trainable matrices in the model, utilized for linear transformations during training. N represents the total number of nodes in a molecular graph. The sigmoid function is employed as an activation function, ensuring that the attention weight w_v of each node v remains bounded between 0 and 1. This weight w_v reflects the significance of node v's feature h_v in the molecular graph, enabling a focused and precise molecular property prediction.

The essence of perturbation-based methods lies in using masks to obscure certain atoms, bonds, or fragments in a GNN model. This helps identify substructures whose absence significantly impacts model predictions. However, using non-chemically informed mask units often results in confusing patterns that are hard for chemists to interpret. To address this, our study employs common computational chemistry techniques for splitting compounds, utilizing well-defined substructures as mask units (Fig. S11 illustrates the concept of masking with ChemDraw for cP14). This approach focuses on identifying key substructures, like the masked amino acid Glu shown in Fig. 1b, whose absence significantly impacts a GNN's predictive outcomes. Considering that a GNN determines molecular properties by passing a molecular embedding through a trained fully connected (FC) layer, generating a precise embedding for a molecule with masked elements is vital. The methodology for this process is defined as follows:

$$\begin{aligned} \text{molecular embedding}_{mask} &= \sum_{v=1}^{N}(\mathbf{w}_v\cdot\mathbf{h}_v\cdot\mathrm{mask}_v) \\ \mathrm{mask}_v &= \begin{cases} 0, & \text{if node v is mask} \\ 1, & \text{otherwise} \end{cases} \end{aligned}$$

In this formula, mask v is defined as 0 if node v is masked and 1 otherwise. Here, N represents the total number of nodes, w_v is the attention weight for each node v, and h_v signifies the general feature of node v.

In Fig. 1d, we define attribution as the degree to which a masked substructure affects the overall GNN prediction. To determine this, we

compare GNN predictions before and after applying masks to the molecular graph. The attribution is the difference in these predictions, where sub represents the masked substructure, m the number of RGCN models (10 in this study), and i the specific RGCN model.

$$Y = \sum_{i}^{m} Y_i$$

$$Y_{sub} = \sum_{i}^{m} Y_{i,sub}$$

$$\text{Attribution}_{sub} = Y - Y_{sub}$$

For clearer interpretation, we normalize these attribution scores (Attribution N) to a 0–1 scale. Masking a node impact not just the atom itself but also its chemical surroundings, including adjacent atoms/bonds. This effect diminishes with distance from the central atom. Thus, while masking a node primarily obscures information about that atom, it partially affects nearby atoms/bonds too. However, for simplicity, we illustrate it as masking only the node itself in this article.

4.3. Traditional machine learning strategy

In this study, we explored four traditional machine learning models commonly used for SAR prediction, as illustrated in Fig. S1. The models were constructed using the MoleculeACE [65] platform and encompass the following algorithms:

1. K-Nearest Neighbors (KNN): This method employs a nonparametric strategy, predicting the response of a new molecule based on the average responses of the k most similar molecules in the training set.
2. Random Forest (RF): RF is an ensemble method that consists of multiple distinct decision trees, each trained on varied subsets of the training data generated through bootstrapping. The final prediction for a molecule is made by averaging the outputs of each tree.
3. Gradient Boosting Machine (GBM): Like RF, GBM also utilizes multiple decision trees. However, it refines each subsequent tree to reduce the residuals left by its predecessor, enhancing prediction accuracy.
4. Support Vector Machine (SVM): SVM works by projecting data into higher dimensions using a kernel function (such as a radial basis function in our study) to identify an optimal separating hyperplane that best segregates the training data.

To broaden the baseline models for comparison analysis, we combined these algorithms with four types of molecular descriptors and two types of peptide descriptors. These descriptors, crafted to encapsulate specific chemical characteristics, vary in complexity:

1. Extended Connectivity Fingerprints (ECFP): Represent atom-centered radial substructures in a binary format.
2. Molecular ACCess System (MACCS) keys: Indicate the presence of predefined substructures in binary form.
3. Weighted Holistic Invariant Molecular (WHIM) descriptors: Emphasize three-dimensional aspects like molecular size, shape, symmetry, and atom distribution.
4. Physicochemical Properties (PHYSICHEM): Physicochemical properties relevant for drug-likeness [66], these serve as a baseline.

And the amino acid encoding [67] for peptide representations include:

1. One-Hot Encodings (SEQ): Encoding for the 20 amino acids.
2. Physicochemical Properties (PHYSICHEM[b]): A combination of common QSAR descriptors for peptide sequences, such as BLOSUM [68] (BLOcks SUbstitution Matrix), VHSE [69] (principal components score Vectors of Hydrophobic, Steric, and Electronic properties), etc.

4.4. PepBERT

In this study, we employed PepBERT as the baseline model, which is a modification of the original BERT framework. To tailor it specifically for peptide bioactivity regression tasks, we reduced the complexity of the original BERT model. This reduction was achieved primarily by decreasing the number of layers and minimizing the size of the vocabulary. Further information on the various hyperparameters involved can be accessed through our published code.

In our encoding process, unknown amino acids were substituted with the "[UNK]" token, differing from the standard amino acid alphabet. We employed an embedding matrix to represent amino acids as continuous vectors, a typical method in deep learning. Additionally, our scheme included special tokens from the original BERT model, like "[PAD]", "[CLS]", "[SEP]", and "[MASK]". For macrocyclic peptides, despite their cyclic structure, we encoded them as linear 14-length sequences with special tokens, simplifying their complex nature for computational efficiency.

4.5. Chemical synthesis of the macrocyclic peptides

The synthesis of macrocyclic peptides followed a standard Fmoc solid-phase procedure, utilizing 0.5 g of Rink Amide MBHA resin. The resin underwent swelling in a dichloromethane (DCM) solution with 0.3 mmol hydroxybenzotriazole (HOBT), 0.3 mmol Fmoc-Gly-OH, and 5 % N,N′-Diisopropylcarbodiimide (DIC) at room temperature (RT) for 1 h under nitrogen gas. Subsequent steps included filtration, washing, Fmoc deprotection with 20 % piperidine in dimethylformamide (DMF), and amino acid coupling reactions using 0.9 mmol HOBT, 0.9 mmol Fmoc-AA-OH, and 10 % DIC in 10 mL DMF. The full peptide synthesis involved cycles of deprotection and coupling, and a bromoacetyl group was attached to the N-terminal amide for macrocyclic formation. The peptides were cleaved, precipitated with diethyl ether (Et_2O), redissolved in 10 mL dimethylsulfoxide (DMSO), adjusted to pH 8.0, and incubated for 1 h to enable cyclization. The cyclization reaction was quenched by adjusting the pH to 3 ∼ 4 with trifluoroacetic acid (TFA). Purification was performed by reverse-phase HPLC with a linear gradient from aqueous solution with 1 % TFA to acetonitrile (MeCN) with 1 % TFA. Peptide purity over 95 % was confirmed by LC-2020 (Shimazu), and mass spectra were verified by LCMS-2020 (Shimazu) before lyophilization.

4.6. Evaluation of macrocyclic peptides

To assess the activities of the selected macrocyclic peptides, competitive ELISA were conducted. In this study, 80 μ L of 1 μ g/mL interleukin-17 receptor E (IL-17RE) was added to each well and incubated at 4 °C overnight for immobilization. Following four washes with 150 μ L 1 × PBST buffer, the wells were blocked with 100 μ L of 1 × PBST buffer containing 2 % BSA at RT for 1 h. Subsequently, the wells were washed again with 150 μ L 1 × PBST buffer before being mixed with 100 μ L of biotinylated IL-17C (0.5 nM) and various concentrations of each macrocyclic peptide, followed by incubation at RT for 1 h. After four washes, 150 μ L of Streptavidin-HRP solution (1:1000 dilution) in 1 × PBS was added to each well and incubated at room temperature for 1 h. Following another washing step, 100 μ L of 3,3′,5,5′-tetramethylbenzidine (TMB) solution was added to each well and incubated at RT for 10 min to develop color. After incubation, ELISA stop solution (Absin) was added to the wells, and absorbance was measured at 450 nm using the Tecan Spark multimode reader. The IC_{50} values were calculated using the nonlinear regression method: (inhibitor) vs response–variable slope in GraphPad Prism.

4.7. Data and software availability

Source code for PepExplainer and data analysis, and instructions to

reproduce the work can be found at https://github.com/zhaisilong/PepExplainer. The construction of the model is executed using Python 3.7, supplemented by dgl-cuda11.3 (version 0.7.1) and PyTorch (version 1.12.1). Data processing and metric calculations are also carried out using Python 3.7, utilizing scikit-learn (version 1.0.2), NumPy (version 1.19.5), and Pandas (version 2.1.1).

Our strategy employs 5-fold cross-validation to split the dataset into training and testing sets. We adhere to the optimal hyperparameters suggested in the original model's paper, making only simple and necessary modifications to suit our peptide property prediction parameters. The resulting optimized hyperparameters are then employed to construct RGCN submodels with different random seeds. These ten submodels are subsequently integrated to form a consensus model.

Our model can also handle linear and cyclic peptides with Cys-Cys disulfide bonds as well as the peptides for targeting different targets. However, we first need to screen against the target protein to gather enrichment and partial activity data before training and testing.

CRediT authorship contribution statement

Silong Zhai: Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Yahong Tan:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. **Cheng Zhu:** Writing – review & editing, Data curation. **Chengyun Zhang:** Writing – review & editing, Data curation. **Yan Gao:** Writing – review & editing. **Qingyi Mao:** Writing – review & editing. **Youming Zhang:** Writing – review & editing. **Hongliang Duan:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Yizhen Yin:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (NSFC) (No. 22277065), Natural Science Foundation of Zhejiang Province (LD22H300004), Taishan Scholar Program of Shandong Province (No. tsqn202103148), Qilu Youth Scholar Startup Funding of Shandong University (to Yizhen Yin) and State Key Laboratory of Microbial Technology Open Projects Fund (Project NO. 20220904-17).

Abbreviations

ADMET	absorption, distribution, metabolism, excretion, toxicity
AI	artificial intelligence
BBBP	blood-brain barrier permeability
BERT	bi-directional encoder representations from Transformers
BLOSUM	blocks substitution matrix
DCM	dichloromethane
DIC	N,N′-diisopropylcarbodiimide
DMF	dimethylformamide
DMSO	dimethylsulfoxide
ECFP	extended connectivity fingerprints
ELISA	enzyme-linked immunosorbent assay
ESOL	estimated solubility
Et_2O	diethyl ether
FC	fully connected
FIT	flexible *in vitro* translation
GBM	gradient boosting machine
GCNs	graph convolutional networks
GNN	graph neural networks
HOBT	hydroxybenzotriazole
IC_{50}	50 % inhibitory concentration
IL-17C	interleukin-17C; IL-17RE, interleukin-17 receptor E
KNN	K-nearest neighbors
MACCS	molecular access system keys
MeCN	acetonitrile
NCAAs	non-canonical amino acids
PepBERT	BERT model for peptide regression
PPIs	protein-protein interactions
QSAR	quantitative structure-activity relationships
RaPID	random non-standard peptide integrated discovery
RF	random forest
RGCNs	relational graph convolutional networks
R^2	coefficient of determination
RMSE	root mean square error
RT	room temperature
SAR	structure-activity relationships
SME	substructure mask explanation
SVM	support vector machine
TFA	trifluoroacetic acid
t-SNE	t-distributed stochastic neighbor embedding
TMB	3,3′,5,5′-tetramethylbenzidine
VHSE	principal components score vectors of hydrophobic, steric, and electronic properties
WHIM	weighted holistic invariant molecular descriptors
XAI	explainable artificial intelligence

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmech.2024.116628.

References

[1] K. Fosgerau, T. Hoffmann, Peptide therapeutics: Current status and future directions, Drug Discov. Today 20 (2015) 122–128, https://doi.org/10.1016/j.drudis.2014.10.003.
[2] A.K. Malde, T.A. Hill, A. Iyer, D.P. Fairlie, Crystal structures of protein-bound cyclic peptides, Chem. Rev. 119 (2019) 9861–9914, https://doi.org/10.1021/acs.chemrev.8b00807.
[3] P.G. Dougherty, A. Sahni, D. Pei, Understanding cell penetration of cyclic peptides, Chem. Rev. 119 (2019) 10241–10287, https://doi.org/10.1021/acs.chemrev.9b00008.
[4] J. Damjanovic, J. Miao, H. Huang, Y.-S. Lin, Elucidating solution structures of cyclic peptides using molecular dynamics simulations, Chem. Rev. 121 (2021) 2292–2324, https://doi.org/10.1021/acs.chemrev.0c01087.
[5] M. Gao, K. Cheng, H. Yin, Targeting protein-protein interfaces using macrocyclic peptides, Peptide.Sci. 104 (2015) 310–316, https://doi.org/10.1002/bip.22625.
[6] C.H.M. Rodrigues, D.E.V. Pires, T.L. Blundell, D.B. Ascher, Structural landscapes of PPI interfaces, Briefings Bioinf. 23 (2022), https://doi.org/10.1093/bib/bbac165 bbac165.
[7] J.A. Wells, C.L. McClendon, Reaching for high-hanging fruit in drug discovery at protein–protein interfaces, Nature 450 (2007) 1001–1009, https://doi.org/10.1038/nature06526.
[8] T. Passioura, T. Katoh, Y. Goto, H. Suga, Selection-based discovery of druglike macrocyclic peptides, Annu. Rev. Biochem. 83 (2014) 727–752, https://doi.org/10.1146/annurev-biochem-060713-035456.
[9] A.A. Vinogradov, Y. Yin, H. Suga, Macrocyclic peptides as drug candidates: recent progress and remaining challenges, J. Am. Chem. Soc. 141 (2019) 4167–4181, https://doi.org/10.1021/jacs.8b13178.
[10] L.K. Buckton, M.N. Rahimi, S.R. McAlpine, Cyclic peptides as drugs for intracellular targets: the next frontier in peptide therapeutic development, Chem. Eur J. 27 (2021) 1487–1513, https://doi.org/10.1002/chem.201905385.
[11] A.A. Sadybekov, A.V. Sadybekov, Y. Liu, C. Iliopoulos-Tsoutsouvas, X.-P. Huang, J. Pickett, B. Houser, N. Patel, N.K. Tran, F. Tong, N. Zvonok, M.K. Jain, O. Savych, D.S. Radchenko, S.P. Nikas, N.A. Petasis, Y.S. Moroz, B.L. Roth, A. Makriyannis,

V. Katritch, Synthon-based ligand discovery in virtual libraries of over 11 billion compounds, Nature 601 (2022) 452–459, https://doi.org/10.1038/s41586-021-04220-9.
[12] Z. Li, X. Li, Y.-Y. Huang, Y. Wu, R. Liu, L. Zhou, Y. Lin, D. Wu, L. Zhang, H. Liu, X. Xu, K. Yu, Y. Zhang, J. Cui, C.-G. Zhan, X. Wang, Hai-Bin Luo, Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs, Proc. Natl. Acad. Sci. USA 117 (2020) 27381–27387, https://doi.org/10.1073/pnas.2010470117.
[13] Y. Huang, M.M. Wiedmann, H. Suga, RNA display methods for the discovery of bioactive macrocycles, Chem. Rev. 119 (2019) 10360–10391, https://doi.org/10.1021/acs.chemrev.8b00430.
[14] Y.V. Guillen Schlippe, M.C.T. Hartman, K. Josephson, J.W. Szostak, In vitro selection of highly modified cyclic peptides that act as tight binding inhibitors, J. Am. Chem. Soc. 134 (2012) 10469–10477, https://doi.org/10.1021/ja301017y.
[15] H. Peacock, H. Suga, Discovery of de novo macrocyclic peptides by messenger RNA display, Trends Pharmacol. Sci. 42 (2021) 385–397, https://doi.org/10.1016/j.tips.2021.02.004.
[16] H. Murakami, A. Ohta, H. Ashigai, H. Suga, A highly flexible tRNA acylation method for non-natural polypeptide synthesis, Nat. Methods 3 (2006) 357–359, https://doi.org/10.1038/nmeth877.
[17] Y. Goto, T. Katoh, H. Suga, Flexizymes for genetic code reprogramming, Nat. Protoc. 6 (2011) 779–790, https://doi.org/10.1038/nprot.2011.331.
[18] Y. Goto, H. Suga, The RaPID platform for the discovery of pseudo-natural macrocyclic peptides, Accounts Chem. Res. 54 (2021) 3604–3617, https://doi.org/10.1021/acs.accounts.1c00391.
[19] J.S. Chang, A.A. Vinogradov, Y. Zhang, Y. Goto, H. Suga, Deep learning-driven library design for the de novo discovery of bioactive thiopeptides, ACS Cent. Sci. 9 (2023) 2150–2160, https://doi.org/10.1021/acscentsci.3c00957.
[20] S. Zhai, Y. Tan, C. Zhang, C.J. Hipolito, L. Song, C. Zhu, Y. Zhang, H. Duan, Y. Yin, PepScaf: harnessing machine learning with in vitro selection toward de novo macrocyclic peptides against IL-17C/IL-17RE interaction, J. Med. Chem. 66 (2023) 11187–11200, https://doi.org/10.1021/acs.jmedchem.3c00627.
[21] A.A. Vinogradov, J.S. Chang, H. Onaka, Y. Goto, H. Suga, Accurate models of substrate preferences of post-translational modification enzymes from a combination of mRNA display and deep learning, ACS Cent. Sci. (2022), https://doi.org/10.1021/acscentsci.2c00223.
[22] J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, Nat. Mach. Intell. 2 (2020) 573–584, https://doi.org/10.1038/s42256-020-00236-4.
[23] R.L. Marchese Robinson, A. Palczewska, J. Palczewski, N. Kidley, Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets, J. Chem. Inf. Model. 57 (2017) 1773–1792, https://doi.org/10.1021/acs.jcim.6b00753.
[24] F. Grisoni, V. Consonni, D. Ballabio, Machine learning consensus to predict the binding to the androgen receptor within the CoMPARA Project, J. Chem. Inf. Model. 59 (2019) 1839–1848, https://doi.org/10.1021/acs.jcim.8b00794.
[25] Y. Chen, C. Stork, S. Hirte, J. Kirchmair, NP-scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules, Biomolecules 9 (2019), https://doi.org/10.3390/biom9020043.
[26] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57 (2014) 4977–5010, https://doi.org/10.1021/jm4004285.
[27] M.F. Sanner, K. Zoghebi, S. Hanna, S. Mozaffari, S. Rahighi, R.K. Tiwari, K. Parang, Cyclic peptides as protein kinase inhibitors: structure–activity relationship and molecular modeling, J. Chem. Inf. Model. 61 (2021) 3015–3026, https://doi.org/10.1021/acs.jcim.1c00320.
[28] T. Tian, S. Li, M. Fang, D. Zhao, J. Zeng, MolSHAP: interpreting quantitative structure–activity relationships using shapley values of R-groups, J. Chem. Inf. Model. (2023), https://doi.org/10.1021/acs.jcim.3c00465.
[29] G.B. Goh, C. Siegel, A. Vishnu, N.O. Hodas, N. Baker, Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. https://arxiv.org/abs/1706.06689, 2017.
[30] E.B. Lenselink, N. ten Dijke, B. Bongers, G. Papadatos, H.W.T. van Vlijmen, W. Kowalczyk, A.P. Ijzerman, G.J.P. van Westen, Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set, J. Cheminf. 9 (2017) 45, https://doi.org/10.1186/s13321-017-0232-0.
[31] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.
[32] G.P. Wellawatte, A. Seshadri, A.D. White, Model agnostic generation of counterfactual explanations for molecules, Chem. Sci. 13 (2022) 3697–3705, https://doi.org/10.1039/D1SC05259D.
[33] M. Gupta, H.J. Lee, C.J. Barden, D.F. Weaver, The blood–brain barrier (BBB) score, J. Med. Chem. 62 (2019) 9824–9836, https://doi.org/10.1021/acs.jmedchem.9b01220.
[34] P.D. Leeson, R.J. Young, Molecular property design: does everyone get it? ACS Med. Chem. Lett. 6 (2015) 722–725, https://doi.org/10.1021/acsmedchemlett.5b00157.
[35] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, T. Langer, A compact review of molecular property prediction with graph neural networks, Drug Discov. Today Technol. 37 (2020) 1–12, https://doi.org/10.1016/j.ddtec.2020.11.009.
[36] F. Baldassarre, H. Azizpour, Explainability techniques for graph convolutional networks, ArXiv. abs/1905 (2019) 13686.
[37] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, Chem. Sci. 9 (2018) 513–530, https://doi.org/10.1039/C7SC02664A.
[38] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, Molecular graph convolutions: moving beyond fingerprints, J. Comput. Aided Mol. Des. 30 (2016) 595–608, https://doi.org/10.1007/s10822-016-9938-8.
[39] W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018-07-10/2018-07-15, pp. 2323–2332.
[40] Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh, T. Hou, Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking, Nat. Commun. 14 (2023) 2585, https://doi.org/10.1038/s41467-023-38192-3.
[41] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, On the art of compiling and using 'drug-like' chemical fragment spaces, ChemMedChem 3 (2008) 1503–1507, https://doi.org/10.1002/cmdc.200800178.
[42] G.W. Bemis, M.A. Murcko, The properties of known drugs. 1. Molecular frameworks, J. Med. Chem. 39 (1996) 2887–2893, https://doi.org/10.1021/jm9602928.
[43] Y. Hu, D. Stumpfe, J. Bajorath, Computational exploration of molecular scaffolds in medicinal chemistry, J. Med. Chem. 59 (2016) 4062–4076, https://doi.org/10.1021/acs.jmedchem.5b01746.
[44] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, J. Pei, Transfer learning for drug discovery, J. Med. Chem. 63 (2020) 8683–8694, https://doi.org/10.1021/acs.jmedchem.9b02147.
[45] R.S. Simões, V.G. Maltarollo, P.R. Oliveira, K.M. Honorio, Transfer and multi-task learning in QSAR modeling: advances and challenges, Front. Pharmacol. 9 (2018) 74, https://doi.org/10.3389/fphar.2018.00074.
[46] Y. Chu, Y. Zhang, Q. Wang, L. Zhang, X. Wang, Y. Wang, D.R. Salahub, Q. Xu, J. Wang, X. Jiang, Y. Xiong, D.-Q. Wei, A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design, Nat. Mach. Intell. 4 (2022) 300–311, https://doi.org/10.1038/s42256-022-00459-7.
[47] X.-C. Zhang, C.-K. Wu, Z.-J. Yang, Z.-X. Wu, J.-C. Yi, C.-Y. Hsieh, T.-J. Hou, D.-S. Cao, MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction, Briefings Bioinf. 22 (2021) bbab152, https://doi.org/10.1093/bib/bbab152.
[48] P. Charoenkwan, C. Nantasenamat, M.M. Hasan, B. Manavalan, W. Shoombuatong, BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides, Bioinformatics 37 (2021) 2556–2562, https://doi.org/10.1093/bioinformatics/btab133.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017.
[50] M.-L. Lee, S. Farag, J.S. Del Cid, C. Bashore, K.K. Hallenbeck, A. Gobbi, C. N. Cunningham, Identification of macrocyclic peptide families from combinatorial libraries containing noncanonical amino acids using cheminformatics and bioinformatics inspired clustering, ACS Chem. Biol. 18 (2023) 1425–1434, https://doi.org/10.1021/acschembio.3c00159.
[51] R. Zhang, H. Wu, Y. Xiu, K. Li, N. Chen, Y. Wang, Y. Wang, X. Gao, F. Zhou, PepLand: a large-scale pre-trained peptide representation model for a comprehensive landscape of both canonical and non-canonical amino acids. https://arxiv.org/abs/2311.04419, 2023.
[52] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2018, pp. 593–607.
[53] E. Fix, J.L. Hodges, Discriminatory analysis. Nonparametric discrimination: consistency properties, Int. Stat. Rev./Rev. Int. Stat. 57 (1989) 238–247, https://doi.org/10.2307/1403797.
[54] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, https://doi.org/10.1007/BF00058655.
[55] Jerome H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232, https://doi.org/10.1214/aos/1013203451.
[56] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, 2010.
[57] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754, https://doi.org/10.1021/ci100050t.
[58] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL keys for use in drug discovery, J. Chem. Inf. Comput. Sci. 42 (2002) 1273–1280, https://doi.org/10.1021/ci010132r.
[59] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani, MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids, J. Comput. Aided Mol. Des. 11 (1997) 79–92, https://doi.org/10.1023/A:1008079512289.
[60] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR (2018) 04805 abs/1810, https://arxiv.org/abs/1810.04805.
[61] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[62] M. Ansari, A.D. White, Serverless prediction of peptide properties with recurrent neural networks, J. Chem. Inf. Model. 63 (2023) 2546–2553, https://doi.org/10.1021/acs.jcim.2c01317.

[63] M.L. Merz, S. Habeshian, B. Li, J.-A.G.L. David, A.L. Nielsen, X. Ji, K. Il Khwildy, M. M. Duany Benitez, P. Phothirath, C. Heinis, De novo development of small cyclic peptides that are orally bioavailable, Nat. Chem. Biol. 20 (2024) 624–633, https://doi.org/10.1038/s41589-023-01496-y.

[64] P. Schwaller, D. Probst, A.C. Vaucher, V.H. Nair, D. Kreutter, T. Laino, J.-L. Reymond, Mapping the space of chemical reactions using attention-based neural networks, Nat. Mach. Intell. 3 (2021) 144–152, https://doi.org/10.1038/s42256-020-00284-w.

[65] D. van Tilborg, A. Alenicheva, F. Grisoni, Exposing the limitations of molecular machine learning with activity cliffs, J. Chem. Inf. Model. 62 (2022) 5938–5951, https://doi.org/10.1021/acs.jcim.2c01073.

[66] W.P. Walters, M.A. Murcko, Prediction of "drug-likeness,", Comp. Method.Pred. ADME.Toxi. 54 (2002) 255–271, https://doi.org/10.1016/S0169-409X(02)00003-0.

[67] H. ElAbd, Y. Bromberg, A. Hoarfrost, T. Lenz, A. Franke, M. Wendorff, Amino acid encoding for deep learning applications, BMC Bioinf. 21 (2020) 235, https://doi.org/10.1186/s12859-020-03546-x.

[68] A.G. Georgiev, Interpretable numerical descriptors of amino acid space, J. Comput. Biol. 16 (2009) 703–723, https://doi.org/10.1089/cmb.2008.0173.

[69] H. Mei, Z.H. Liao, Y. Zhou, S.Z. Li, A new set of amino acid descriptors and its application in peptide QSARs, Peptide.Sci. 80 (2005) 775–786, https://doi.org/10.1002/bip.20296.