Amino Acids (2024) 56:62
https://doi.org/10.1007/s00726-024-03422-5

ORIGINAL ARTICLE

LLM4THP: a computing tool to identify tumor homing peptides by molecular and sequence representation of large language model based on two-layer ensemble model strategy

Sen Yang[1,2] · **Piao Xu**[3]

Received: 18 June 2024 / Accepted: 4 October 2024

Abstract
Tumor homing peptides (THPs) have a distinctive capacity to specifically attach to tumor cells, providing a promising approach for targeted cancer treatment and detection. Although THPs have the potential for significant impact, their detection by conventional methods is both time-consuming and expensive. To tackle this issue, we provide LLM4THP, an innovative computational approach that utilizes large language models (LLMs) to quickly and effectively detect THPs. LLM4THP utilizes two protein LLMs, ESM2 and Prot_T5_XL_UniRef50, to encode peptide sequences. This allows for the capture of complex patterns and relationships within the peptide data. In addition, we utilize inherent sequence characteristics such as Amino Acid Composition (AAC), Pseudo Amino Acid Composition (PAAC), Amphiphilic Pseudo Amino Acid Composition (APAAC), and Composition, Transition, and Distribution (CTD) to improve the representation of peptides. The RDKitDescriptors feature representation approach transforms peptide sequences into molecular objects and computes chemical characteristics, resulting in enhanced THP identification. The LLM4THP ensemble strategy incorporates various features into a two-layer learning architecture. The first layer consists of LightGBM, XGBoost, Random Forest, and Extremely Randomized Trees, which generate a set of meta results. The second layer utilizes Logistic Regression to further refine the identification of sequences as either THP or non-THP. LLM4THP exhibits exceptional performance compared to the most advanced methods, showcasing enhancements in accuracy, Matthew's correlation coefficient, F1 score, area under the curve, and average precision. The source code and dataset can be accessed at the following URL: https://github.com/abcair/LLM4THP.

Keywords Tumor homing peptides · Computational method · Large Language models · Peptide sequence encoding · Ensemble strategy

Introduction

Tumor homing peptides (THPs) (Kondo et al. 2021) are compact, sequence-specific compounds that possess the extraordinary capacity to specifically adhere to cancerous cells or tissues (Wu et al. 2022a). The exceptional characteristic of THP renders them important instruments in diverse domains of cancer research and therapy (Lu et al. 2017). THP can be employed for the targeted delivery of therapeutic medicines to tumor cells, thereby limiting harm to healthy cells and diminishing adverse effects (Lempens et al. 2011). Through the process of conjugating THPs to anticancer medications, nanoparticles, or other therapeutic payloads, these agents can be precisely directed towards tumors, thereby improving their effectiveness and minimizing harm to the rest of the

Communicated by A. de Brevern.

✉ Piao Xu
xupiao@njfu.edu.cn

[1] School of Computer Science and Artificial Intelligence Aliyun School of Big Data School of Software, Changzhou University, Changzhou 213164, China

[2] The Affiliated Changzhou No. 2 People's Hospital of Nanjing Medical University, Changzhou 213164, China

[3] College of Economics and Management, Nanjing Forestry University, Nanjing 210037, China

Published online: 15 October 2024

body (Zhang et al. 2021). THPs can also function as imaging agents for the visualization of cancers (Wu et al. 2022b). Researchers can non-invasively visualize the location, size, and changes over time of tumors by including fluorescent or radioactive markers into THPs. This information is crucial for the precise diagnosis of a medical disease, the evaluation of its severity, and the assessment of the effectiveness of the treatment being provided (Li and Cho 2012). Simply said, tumor homing peptides provide a potent means for precise cancer treatment and detection. Their capacity to specifically adhere to tumor cells possesses significant potential for enhancing cancer treatment results and comprehension of the mechanism (Karami Fath et al. 2022).

Additionally, the function of binding of THPs is primarily mediated by their capacity to recognize and interact with specific receptors or molecules on the surface of cancer cells. This selective binding is contingent on the amino acid composition and spatial conformation of the peptide sequence. The characteristics of THPs include high affinity, specificity, and the capability to penetrate biological barriers. During the binding process, THPs typically exploit the abnormal glycosylation or protein expression on tumor cell surfaces, which is less common in normal cells, thereby achieving precise tumor cell targeting. Additionally, THPs are characterized by their biocompatibility and low immunogenicity, which endows them with significant potential in applications such as drug delivery, diagnostic imaging, and cancer therapy. By delving into the binding mechanisms and characteristics of THPs, researchers can further refine these peptides to develop more effective strategies for cancer treatment.

Nevertheless, the identification of THPs by conventional experimental methods, such as phage display, might incur significant expenses and consume a considerable amount of time (Melssen et al. 2023). These procedures necessitate thorough laboratory work, which include peptide synthesis, screening, and characterization (Sharma et al. 2013a). Moreover, the process of optimizing peptide sequences to improve tumor targeting can be intricate and involves multiple iterations (Li et al. 2019). Therefore, it is necessary to identify THP in a high-throughput way. Machine learning models can rapidly analyze large datasets and identify patterns and relationships that are not apparent through manual analysis (Lin et al. 2015). In 2013, Centre et al. conducted an analysis on a dataset of experimentally validated THPs and non-THPs to uncover important compositional features and preferences for residues (Sharma et al. 2013b). These attributes are further utilized to construct SVM models (Jiang et al. 2020) employing several representations of the peptides, such as amino acid composition, dipeptide composition, and binary profile patterns. The results indicate that the model based on binary profile patterns obtains the maximum level of accuracy, especially when focusing on the N- and C-terminal residues of the peptides (Huttunen-Hennelly 2010). In addition, TumorHPD is a user-friendly online server that utilizes these SVM models and offers tools for creating new THPs with enhanced tumor homing capabilities. In 2019, Shoombuatong et al. introduced a new computational model called THPep (Shoombuatong et al. 2019). This model uses a random forest classifier and various peptide features, such as amino acid composition, dipeptide composition, and pseudo amino acid composition, to predict tumor homing peptides (THPs). The model outperforms existing methods and its interpretability makes it a potentially valuable tool for researchers in the field of cancer therapeutics. Furthermore, to assist researchers in conducting experiments, the authors have developed a publicly accessible web server for THPep. In 2021, He et al. introduce a new meta-learning model called Mutual Information Maximization Meta-Learning (MIMML) (He et al. 2022). The purpose of this model is to accelerate the process of identifying bioactive peptides, which are short chains of amino acids that have potential medicinal uses. MIMML utilizes the ideas of few-shot learning and meta-learning (Langdon et al. 2022) to effectively adjust to novel tasks with limited training data. MIMML uses a Text Convolution Neural Network (TextCNN) (Soni et al. 2023) to convert peptide sequences into feature vectors, capturing their hidden features. It then uses a prototypical network to generate class prototypes from these embeddings, allowing the model to classify new and unknown peptide sequences. In 2022, Charoenkwan et al. introduced NEPTUNE (Charoenkwan et al. 2022b), an innovative computational method designed to precisely and efficiently identify tumor homing peptides (THPs) from sequence data on a wide scale. NEPTUNE is a stacked ensemble learning technique that uses several feature encoding approaches in conjunction with six well-known machine learning algorithms, including random forest, support vector machine, partial least squares, logistic regression, extremely randomized trees, and k-nearest neighbor. Utilizing the probabilistic data obtained from the most effective baseline models. The integrated prediction is derived from the ultimate meta-predictor. In 2022, Charoenkwan et al. introduced SCMTHP (Charoenkwan et al. 2022a), an innovative method for detecting and analyzing tumor homing peptides (THPs) by utilizing predicted propensity scores of amino acids. SCMTHP utilizes a score card approach (SCM) (Charoenkwan et al. 2020) to enhance the accuracy and comprehensibility of predictions. The SCMTHP algorithm calculates propensity scores for 20 amino acids, which are subsequently utilized to discover physicochemical features (PCPs) that are informative and linked with THP bioactivity. In 2023, Guan et al. presented StackTHPred (Guan et al. 2023), an innovative

computational technique specifically developed for the detection of tumor-homing peptides (THPs). StackTHPred employs a stacking ensemble architecture that incorporates feature selection based on gradient boosting decision trees (GBDT) (Liu et al. 2022). The framework utilizes five broad protein descriptors, namely amino acid composition, pseudo-amino acid composition, physicochemical properties, BLOSUM62, and z-scale, to extract informative characteristics. The GBDT algorithm is used to efficiently choose features, which reduces computational complexity and improves prediction accuracy. In 2024, Arif et al. introduced a new computational framework called THP-DF(Arif et al. 2024), specifically developed for the precise detection of tumor homing peptides (THPs) on a wide scale. THP-DF employs a blend of sequential and deep learning characteristics. Initially, peptide sequences are encoded utilizing a range of sequential characteristics. Afterwards, a BiLSTM (Bidirectional Long Short-Term Memory) (Huang et al. 2022) model is used, together with attention layers, to extract profound characteristics from these sequences. The deep features, as well as the sequential features, are combined in an ensemble framework and used as input for a support vector machine (SVM) classifier to create THP-DF. Although current predictors show fair accuracy in identifying THPs, there is much potential for improvement. An effective approach involves utilizing large language models (LLMs) to encode peptide sequences. LLMs have transformed the field of natural language processing by capturing complex patterns and relationships in textual material. By employing analogous methods to peptide sequences, it is possible to achieve more thorough and refined depictions, which may uncover profound understandings of THP functioning. Hence, this work introduces a novel model named LLM4THP for the purpose of detecting THP. We employ large language models (LLMs) (Thirunavukarasu et al. 2023) to create features by extracting relevant information from peptide sequences, including peptide sequence intrinsic features and molecular information features, to encode peptide sequences. The ensemble technique is utilized to construct LLM4THP, which consists of a two-layer learning architecture, using the embedding vectors as a foundation. The first layer has four meta predictors: LightGBM (LGBM, referred to as M1), XGBoost (XGB, referred to as M2), Random Forest (RF, referred to as M3), and Extremely randomized trees (ERT, referred to as M4). The cross-product of the embedding vectors [V1, V2, V3, V4, …, V7] with the meta predictors [M1, M2, M3, M4] produces a collection of outcomes known as VMs. These outcomes represent the predictive ability of each feature when combined with each model. The collection of predictions is further analyzed using Logistic Regression to enhance the differentiation between THP and non-THP sequences. The result of LLM4THP is a classification that determines whether the input peptide sequence is a THP or a non-THP. This classification is based on the combined predictions from the ensemble model. LLM4THP is assessed using many measures and a user-friendly prediction model is developed for academic research purposes.

Table 1 The training and test dataset

	Training dataset		Independent test dataset	
	Primary	Small	Primary	Small
THP	490	350	161	119
non-THP	490	350	161	119

Data and methods

Data

In the current investigation, the dataset architecture was meticulously designed to encompass four distinct subsets for the purpose of training and evaluating the predictive model. The primary training dataset, denoted as TRP, was compiled with an equal representation of 490 tumor homing peptides (THPs) and 490 non-THP sequences. A smaller training subset, referred to as TRS, was also constructed with a balanced distribution of 350 THPs and 350 non-THPs. To independent validation, two additional datasets were curated. The Primary independent test dataset (ITP) was assembled with 161 THPs and 161 non-THPs, while the smaller independent test dataset (ITS) comprised 119 THPs and an equivalent number of non-THP sequences. These datasets were originally created and characterized in a prior study (Shoombuatong et al. 2019). The THP samples were meticulously selected from the TumorHoPe database (Kapoor et al. 2012), ensuring their experimental verification as tumor-seeking peptides. Conversely, the non-THP counterparts were randomly selected from the SwissProt database (Bairoch 2000), which are not in TumorHoPe database without any overlap with THP dataset, providing a diverse and representative control group. A comprehensive summary of the training and independent test datasets, including the number of samples and their respective sources, is provided in Table 1. This structured approach to dataset curation facilitates a robust evaluation of the model's performance across various conditions, enhancing the reliability and generalizability of the findings. Table 1. shows the training and test dataset for LLM4THP. And the distribution of peptide sequence length in Primary dataset and Small dataset is shown in Supplementary Figure S1.

Methods

Peptide sequence encoding

Sequence encoding features by protein large language model a. Evolutionary scale modeling 2 Evolutionar scale mody Scale Model 2 (Lin et al. 2023), abbreviated ESM2, is another protein large language model that uses a massive corpus of protein sequences to train a deep neural network architecture to capture the subtle evolutionary links and functional features of amino acids. ESM2 can construct highly informative and context-aware embeddings of protein sequences. The various hidden layer outputs in ESM2 produce varying output dimensions. ESM2 has six alternative implementations, including esm2_t48_15B_UR50D (5120), esm2_t36_3B_UR50D (2560), esm2_t33_650M_UR50D (1280), esm2_t30_150M_UR50D (640), esm2_t12_35M_UR50D (480), and esm2_t6_8M_UR50D (320). esm2_t48_15B_UR50D indicates that this model has 15 billion parameters and 48 attention layers trained on the UniRef50 (UR50D) protein dataset with 5120 output dimensions. esm2_t33_650M_UR50D indicates that this model has 650 million parameters and 33 attention layers trained on the UniRef50 (UR50D) protein dataset with 1280 output dimensions. Other ESM2 models use the same formats as the two mentioned above. Different ESM2 implements represent different output dimensions. For example, esm2_t33_650M_UR50D can encode the Seq_p with sequence length L to a matrix with dimension L × 1280. Furthermore, due to the varying lengths of the peptides, we calculate the mean of the matrix along the length direction to obtain a vector with dimensions 1280. As a result, using the encoder methods described above, a peptide sequence can be converted into 1280-dimensional vectors. In this research, we choose esm2_t33_650M_UR50D as an encoder to represent peptide sequence since its output can carry most encoded information in a not-too-high dimension vector.

b. Prot_T5_XL_UniRef50 Prot-T5_XL_UniRef50 (Pratyush et al. 2024) are the second pre-trained large language models for protein sequence representation. These T5-based models, trained on enormous protein sequence databases, can create informative representations for a variety of biological applications. The UniRef50 dataset (Suzek et al. 2015) contains over 90 million protein sequences. Both models use extra-large architecture, which has deeper and wider neural networks than the original T5 model, allowing Prot-T5_XL_UniRef50 to capture more complicated patterns and correlations within protein sequences. Prot-T5_XL_UniRef50's powerful pre-trained representations can be fine-tuned for a variety of downstream applications, including protein function prediction, binding site prediction and protein-protein interaction prediction.

In this paper, Prot-T5_XL_UniRef50 is introduced to encode peptide sequences. For peptide sequences Seq_p with length L, Prot-T5_XL_UniRef50 can encode the Seq_p into a matrix with dimension $L \times 1024$. Additionally, because of the different lengths of peptides, we get the mean of the above matrix along the length direction to get a vector with 1024 dimensions. Therefore, following the above encoder methods, we can get a 1024-dimension vectors.

Molecular information encoding feature by SMILES This research uses chemical attribute properties of peptides to encode peptide sequences. RDKitDescriptors (Katubi et al. 2023), a chemical attribute calculation library, is used, which provides a complete array of chemical descriptors tailored to peptides. These descriptors effectively capture peptides' chemical and structural properties, making it easier to represent peptide sequences. RDKitDescriptors is a useful tool in computational biology for encoding peptide sequences. RDKitDescriptors can identify crucial molecular attributes such as molecular weight, hydrophobicity, hydrogen bonding potential, and topological indices. RDKitDescriptors also store molecular structural information that determines their biological activity. Furthermore, RDKitDescriptors provide a standardized and well-documented set of descriptors for encoding THP sequences in a high-throughput manner.

Therefore, in this study, RDKitDescriptors are introduced to determine peptides' chemical information. The technique entails transforming a peptide sequence into a molecular object using the RDKit (Bento et al. 2020) Python library's Chem.MolFromFASTA(seq) function, where seq represents the peptide sequence. Chem.MolToSmiles(mol) converts the molecular object into SMILES format (O'Boyle 2012). RDKitDescriptors is then used to calculate the chemical attribute of the peptide using the SMILES format input, returning a dictionary data structure containing 210 chemical attribute features. As a result, a peptide sequence of length L can be transformed into a 210-dimensional vector.

Peptide intrinsic sequence features In this paper, we use Amino Acid Composition (AAC) (Bartas et al. 2021), Pseudo Amino Acid Composition (PAAC) (Naseer et al. 2022), Amphiphilic Pseudo Amino Acid Composition (APAAC) (Wang et al. 2020)and Composition, Transition

and Distribution (CTD) (Meher et al. 2018) as intrinsic sequence features to encode peptide sequence.

a. Amino acid composition AAC calculated the frequency of each amino acid type in a protein or peptide sequence. The frequency calculation for all 20 natural amino acids (ACDEFGH-IKLMNPQRSTVWY) was shown in following Formula:

$$f(t) = \frac{N(t)}{N}, \mathrm{t} \in \{\ \mathrm{A,C,D,\ldots,Y}\}$$

Where $N(t)$ was the count of amino acids in class t, and N was the length of the protein or peptide sequence. In this study, we used AAC to encode an amino acid sequence to a 20-dimensional vector.

b. Pseudo-amino acid composition PAAC feature encoding method considered the frequency of each amino acid and the influence of sequence order on the amino acid sequence. The calculation method was shown in following Formula:

$$\begin{cases} \theta_i = \sum_{i=1}^{N-d} \frac{(P_{\mathrm{i}} - P_{i+d})^2}{N_p} \\ X_{c(i)} = \frac{N_i}{1 + \omega \times \sum_{i=1}^{30} \theta_i} \\ X_{Clambda_i} = \frac{\omega \times \theta_i}{1 + \omega \times \sum_{i=1}^{30} \theta_i} \end{cases}$$

Here, θ_i represented the number of factors related to sequence order, P_{i} was the property value of the *i-th* amino acid, and N_p was the number of attributes, and N_i was the appearance of the *i-th* amino acid and ω was a parameter set to 0.05. In this study, we used PAAC to encode an amino acid sequence to a 23-dimensional vector.

c. Amphiphilic pseudo amino acid composition APAAC, unlike prior methods, incorporates the physicochemical properties of amino acids, specifically hydrophilicity (PHI) and hydrophobicity (PH) (Li et al. 2016). The APAAC approach was utilized to enhance the acquisition of additional peptide sequence information. The 20 naturally occurring amino acids exhibit PHI and PH values that show their interaction with water molecules. The pH values ranged from 0 (indicating the lowest level of hydrophobicity) to 1 (indicating the highest level of hydrophobicity), whereas the PHI values varied from −1 (representing the lowest level of hydrophilicity) to 1 (representing the highest level of hydrophilicity).

The APAAC begins by determining the amino acid composition of the protein sequence and classifying the amino acids into hydrophilic and hydrophobic categories. APAAC computed the PHI and PH scores for each amino acid in the protein sequence. The construction of APAAC involved the consideration of the PHI and PH values of neighboring amino acids within a predetermined window, which was focused around a given place in the peptide sequence. The APAAC feature vector was constructed by amalgamating the amino acid composition, PHI, and PH scores in the peptide sequence. The APAAC encoding approach can be used to extract the amino acid composition and location-specific features of peptide sequences. For this investigation, we employed APAAC to convert a peptide sequence into a vector of 26 dimensions.

d. Composition, transition and distribution The Composition-Transition-Distribution (CTD) feature has been extensively utilized in numerous studies focused on predicting proteins. The CTD is physical-chemical features which can capture the distribution of amino acids based on their physicochemical properties, such as polarity, hydrophobicity, and charge and this encoding scheme is particularly effective for THPs because it reflects the sequence's propensity to interact with the tumor microenvironment, which is crucial for their homing ability. CTD considers the uneven distribution of amino acids along the peptide sequence, which can affect their binding affinity to tumor-specific receptors or other molecular targets. For instance, a higher presence of hydrophobic amino acids may enhance the peptide's ability to penetrate the lipid bilayer of cancer cells, while an increased number of hydrophilic residues might facilitate interactions with hydrophilic regions on the cell surface. Moreover, the transition aspect of the CTD descriptor is important for understanding how amino acids change from one type to another along the sequence, which can impact the peptide's conformation and its overall interaction with the tumor tissue. The distribution of these transitions can provide insights into the structural flexibility or rigidity of the peptide, which may be essential for its targeting specificity. Therefore, CTD is introduced to encode peptide sequence to identify THP. The classification of 20 types of amino acids into seven groups is based on the levels of Dipole and volume scale. For example, $G_1 = \{A, G, V\}$, $G_2 = \{I, L, F, P\}$, $G_3 = \{Y, M, T, S\}$, $G_4 = \{H, N, Q, W\}$, $G_5 = \{R, K\}$, $G_6 = \{D, E\}$, $G_7 = \{C\}$. Therefore, it is feasible to employ a binary space (V, F) to represent a protein sequence. Where V is the vector space of the sequence features, and each feature

V_i represents a sort of triad type. F is the frequency vector corresponding to V, and the value of the i-th dimension of $F(f_i)$ is the frequency of type V_i appearing in the protein sequence. For the amino acids that have been catalogued into seven classes, the size of V should be 343 $(7 \times 7 \times 7)$; However, the value of f_i correlates to the length (number of amino acids) of protein. In general, a long protein would have a large value of f_i, which complicates the comparison between two heterogeneous proteins. To solve this problem, we defined a new parameter, di, by normalizing f_i with $d_i = \frac{(f_i - \min(f_1, f_2, \ldots, f_{343}))}{\max(f_1, f_2, \ldots, f_{343})}$.

Therefore, we get AAC for 20-dimensional vector, PAAC for 23-dimensional vector, APAAC for 26-dimensional vector and CTD for 343-dimensional vector for a peptide sequence. And we merge AAC, PAAC, APAAC and CTD as peptide intrinsic sequence features with a fused 412-dimensional vector.

In brief, using the peptide sequence encoding method, we obtain the encoding vectors for, ESM2 (1280 dimensions), Prot-T5_XL_UniRef50 (1024 dimensions) and peptide intrinsic sequence features including AAC (20 dimensions), PAAC (23 dimensions), APAAC (26 dimensions) and CTD (343 dimensions). After the encode by SMILES, a peptide sequence is encoded by RDKitDescriptors to 210 dimensions vector. Based on the encoding features, we will build LLM4THP to distinguish THP and non-THP.

Construction of LLM4THP

In this study, we introduce LLM4THP (Fig. 1), a new computational model for identifying THP.

In the development of LLM4THP, an ensemble strategy is adopted to learn the predictive power of multiple models. For the construction of features, we utilize large language models (LLMs) capable of extracting valuable information from peptide sequences and molecular SMILES

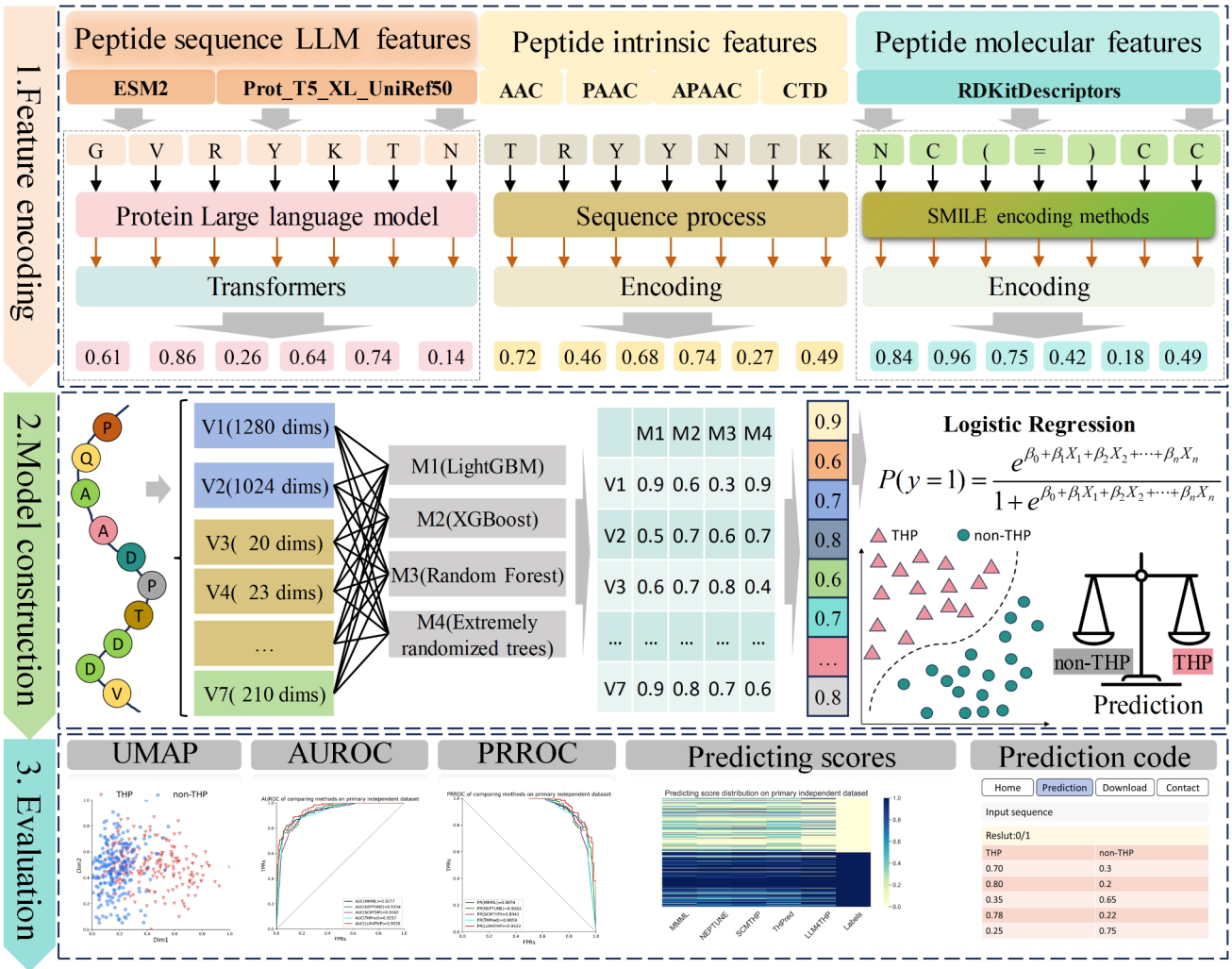

Fig. 1 The integrated pipeline of LLM4THP

representations. This approach yields two distinct types of information: the first derived from the peptide's internal sequence, and the second from the molecular information represented by SMILES. The peptide sequence is transformed into vector representations using ESM2 and Prot_T5_XL_UniRef50, resulting in two vectors V1 and V2 with dimensions of 1280 and 1024, respectively. Additionally, the peptide inner sequence features are also represented by AAC, PAAC, APAAC and CTD called V3, V4, V5 and V6 respectively. Besides, the peptide sequence is converted into a SMILES string format through the RDKit software, which is then encoded into a vectors V7 with 210 dimensions.

These seven embedding vectors collectively encode the peptide sequence, serving as the foundation for the feature encoding phase. Building upon these vectors, the ensemble strategy is implemented to create LLM4THP, which comprises a two-layer learning architecture. The initial layer consists of four meta predictors: LightGBM (LGBM, designated as M1), XGBoost (XGB, designated as M2), Random Forest (RF, designated as M3), and Extremely randomized trees (ERT, designated as M4). The cross-product of the embedding vectors [V1, V2, V3, V4, V5, V6, V7] with the meta predictors [M1, M2, M3, M4] yields a set of results termed VMs, which reflects the predictive capabilities of each feature in conjunction with each model. This ensemble of predictions is then processed by Logistic Regression to refine the distinction between THP and non-THP sequences. The ultimate output of LLM4THP is a classification that determines whether the input peptide sequence is a THP or a non-THP, based on the aggregated predictions from the ensemble model. Finally, LLM4THP is evaluated by multiple metrics and a user-friendly prediction is implemented for academic research. Figure 2 shows the workflow of LLM4THP.

Evaluation metrics

In this study, we adopted a suite of eight established performance metrics to comprehensively evaluate the efficacy of the LLM4THP predictive model. These metrics include Precision, Sensitivity (SE), Specificity (SP), accuracy (ACC), F1-score (F1), Matthew's correlation coefficient (MCC). Additionally, Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision-Recall Curve (APROC) are also used to evaluate LLM4THP performance. The area enclosed by the AUROC and the coordinate axes is called AUC and the area enclosed by the PRROC and the coordinate axes is called AP. The calculations for these metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{SP} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FP})(\text{TN} + \text{FN})}}$$

In the context of binary classification, true positives (TPs) and true negatives (TNs) denote the instances where THPs and non-THPs are correctly identified, respectively. Conversely, false positives (FPs) and false negatives (FNs) represent misclassifications.

Results

Showing the performance of each encoding features

In this study, we introduce LLM4THP, a novel two-layer stack model, which incorporates two distinct large language models (LLMs) encoding features (ESM2 and Prot_T5_XL_UniRef50) alongside molecular descriptors derived from RDKit and a suite of sequence inner encoding features including AAC, PAAC, APAAC, and CTD. These features collectively facilitate the transformation of peptide sequences into hybrid vectors. Subsequently, to assess the efficacy of our feature combination, we implement Uniform Manifold Approximation and Projection (UMAP) (Armstrong et al. 2021) to project the LLM4THP encoding features into a three-dimensional space, visually depicting the distribution of THP and non-THP within both training and test datasets. As illustrated in Fig. 2, the UMAP visualization reveals discernible boundaries between THP and non-THP instances, suggesting a robust discriminative capacity of the integrated features. The distinct distributions of LLM encoding features, molecular information features, and sequence inner encoding features provide empirical evidence supporting the utility and informativeness of our chosen encoding strategy.

Additionally, we concatenate all encoding features and apply UMAP for dimensionality reduction, projecting the aggregate feature set into a three-dimensional space for both the training and test datasets. This visualization strategy is employed to manifest the discriminative power of the concatenated feature set. Figure 3 elucidates the resultant

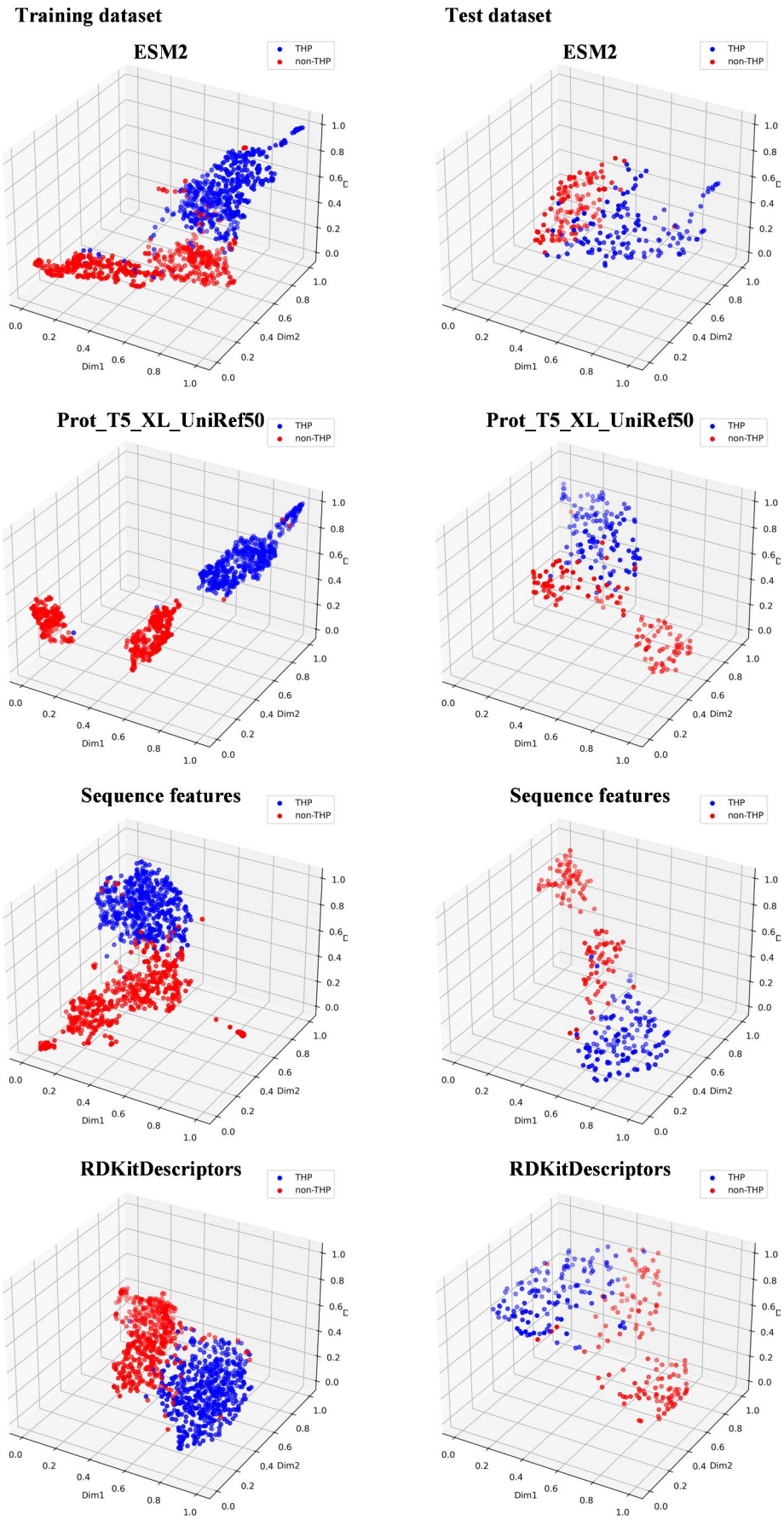

Fig. 2 The distribution of each encoding feature on training and test dataset

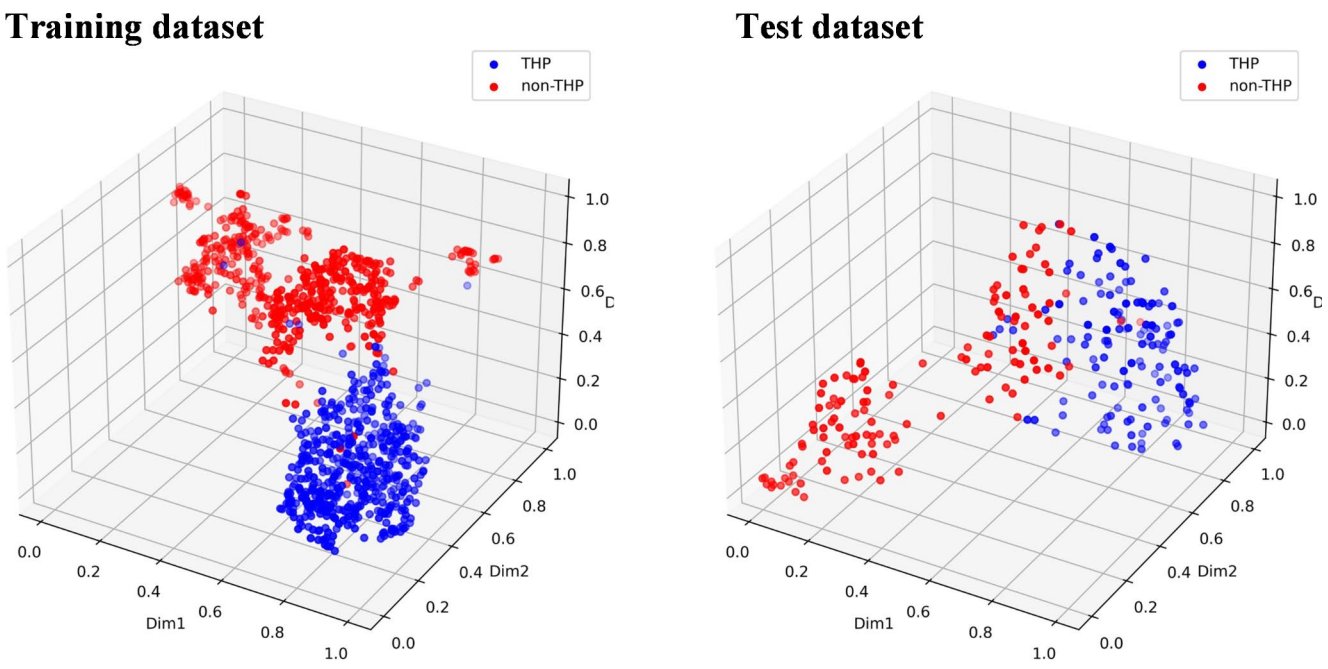

Fig. 3 The distribution of combined features on training and test dataset

Table 2 The results of feature cooperation effect on primary test dataset

Features	Precise	SP	SE	ACC	MCC	F1	AUC	AP
E	0.8423	0.7846	0.9000	0.8423	0.6892	0.8509	0.8610	0.8048
P	0.8239	0.8076	0.9000	0.8538	0.7107	0.8602	0.9030	0.8647
S	0.8108	0.7846	**0.9230**	0.8538	0.7145	0.8633	0.8920	0.8318
R	0.8270	0.8230	0.8461	0.8346	0.6694	0.8365	0.8806	0.8358
EP	0.8428	0.8305	0.9076	0.8692	0.7406	0.8740	0.9023	0.8647
ES	0.8163	0.7923	0.9230	0.8576	0.7215	0.8664	0.9313	0.8986
ER	0.8057	0.7923	0.8615	0.8269	0.6554	0.8327	0.9014	0.8676
PS	0.8108	0.7846	0.9225	0.8538	0.7145	0.8633	0.9384	0.9085
PR	0.8260	0.8153	0.8769	0.8461	0.6936	0.8507	0.9216	0.8999
EPS	0.8321	0.8153	0.9151	0.8653	0.7344	0.8717	0.9341	0.9052
EPR	0.8156	0.8000	0.8846	0.8423	0.6870	0.8487	0.9191	0.8983
PSR	0.8417	0.8307	0.9000	0.8653	0.7325	0.8698	0.9418	0.9220
EPSR	**0.8439**	**0.8307**	0.9153	**0.8730**	**0.7488**	**0.8782**	**0.9528**	**0.9532**

Note ESM2 (E), Prot_T5_XL_UniRef50 (P), Sequence features (S), RDKitDescriptors (R)

mappings, where it is evident that the chosen features are capable of effectively delineating between THP and non-THP instances, underscoring the robustness of our feature selection approach in identifying THP.

The cooperation effect of encoding features to identify THP

Table 2. presents a comprehensive comparison of various feature sets for identifying THP on primary dataset, measured by multiple performance metrics. The features include ESM2 embeddings (E), Prot_T5_XL_UniRef50 embeddings (P), sequence features (S), and RDKit descriptors (R). Additionally, combinations of these features are evaluated, such as E + S, E + P, and E + P + S, among others. First, E + P + S + R emerges as the top-performing feature set, achieving the highest scores across most metrics. Notably, it attains the highest MCC (0.7488), F1 (0.8782), AUC (0.9528) and AP (0.9532), demonstrating its superior ability to distinguish between THP and non-THP. This suggests that integrating diverse information sources, including embeddings, sequence features, and chemical descriptors can enhance the model's performance significantly. Additionally, ESM2 embeddings (E) and Prot_T5_XL_UniRef50 embeddings (P) perform comparably, with E slightly outperforming P in most metrics. This indicates that both embedding methods effectively capture relevant information from the sequences. However, the significant improvement

Table 3 The results of feature cooperation effect on small test dataset

Features	Precise	SP	SE	ACC	MCC	F1	AUC	AP
E	0.8314	0.8404	0.7872	0.8138	0.6285	0.8087	0.8372	0.7959
P	0.8235	0.8404	0.7446	0.7925	0.5878	0.7821	0.8479	0.8079
S	0.8409	0.8510	0.7872	0.8191	0.6396	0.8131	0.8749	0.8487
R	0.8720	0.8829	0.7978	0.8404	0.6833	0.8333	0.8619	0.8408
EP	0.8255	0.8404	0.7553	0.7978	0.5979	0.7888	0.8563	0.8184
ES	0.8539	0.8617	**0.8085**	0.8351	0.6711	0.8306	0.8843	0.8642
ER	**0.8720**	**0.8829**	0.7978	0.8404	0.6833	0.8333	0.8787	0.8615
PS	0.8505	0.8617	0.7872	0.8244	0.6507	0.8176	0.8847	0.8723
PR	0.8452	0.8617	0.7553	0.8085	0.6405	0.7977	0.8760	0.8590
EPS	0.8426	0.8510	0.7978	0.8244	0.6798	0.8196	0.8942	0.8858
EPR	0.8390	0.8510	0.7765	0.8238	0.6894	0.8266	0.8747	0.8556
PSR	0.8470	0.8617	0.7659	0.8238	0.6805	0.8144	0.8914	0.8848
EPSR	0.8470	0.8617	0.7985	0.8391	**0.7056**	**0.8372**	**0.9081**	**0.9119**

Note ESM2 (E), Prot_T5_XL_UniRef50 (P), Sequence features (S), RDKitDescriptors (R)

observed with the combined E + P feature set suggests that they complement each other, providing a more comprehensive representation of the data. For Sequence features (S) and RDKit descriptors (R) individually exhibit moderate performance, with S performing slightly better than R. However, their combination (S + R) significantly enhances performance, indicating a synergistic effect. Besides, the results demonstrate the importance of feature combination in achieving optimal performance. Adding sequence features (S) to E or P improves performance, highlighting the value of incorporating biological information. Further, the addition of RDKitdescriptors (R) to E + P + S yields the best results, suggesting that integrating chemical properties is crucial for this task. Therefore, E + P + S + R emerges as the most effective feature set, demonstrating the importance of integrating embeddings, sequence features, and chemical descriptors for comprehensive representation and improved predictive ability.

From Table 3, we can find The ESM2 feature (E) demonstrates a commendable balance between precision and sensitivity, with respective values of 0.8314 and 0.7872, culminating in a robust F1-score of 0.8087, which suggests that ESM2 is adept at both identifying true positives and minimizing false negatives, a critical trait for models where the cost of missing a positive instance is high. The Prot_T5_XL_UniRef50 feature (P), while slightly less precise than ESM2, exhibits a similar sensitivity, with an F1-score of 0.7821. This indicates that while Prot_T5_XL_UniRef50 may not be as adept at identifying true positives as ESM2, it maintains a respectable balance between precision and recall. For Sequence features (S) and RDKitDescriptors (R) both show high specificity (SP), with values of 0.8510 and 0.8829, respectively. This high specificity indicates that these features are effective at correctly identifying negative instances, which is particularly valuable in contexts where false positives are undesirable. When considering the overall accuracy (ACC), RDKitDescriptors (R) stands out with a score of 0.8404, suggesting that it is the most effective at classifying instances correctly across both positive and negative classes. This is further supported by its high AUC and AP scores, indicating that RDKitDescriptors is particularly adept at distinguishing between positive and negative instances. Additionally, The Matthew's correlation coefficient (MCC) provides a balanced measure of the model's performance, considering both true positives and true negatives. Here, the RDKitDescriptors feature (R) again emerges as a strong contender, with an MCC of 0.6833, which is higher than the other features, reflecting its overall effectiveness to identify THP on small dataset. Besides, all features exhibit strengths in various aspects of predictive performance, the RDKitDescriptors feature (R) stands out as the most informative and useful for the small test dataset, offering high specificity, accuracy, and discriminative power.

Threshold for identification THP

Furthermore, to select the appropriate threshold for identifying THP, we examine the various thresholds to see which one is best for predicting THP. Figure 4 depicts the MCC and F1-score distributions across various thresholds on the primary test dataset and small test dataset. Figure 4 shows that when the threshold is set to 0.5, the MCC and F1-score increase, indicating that the 0.5 threshold is appropriate for discriminating between THP and non-THP. As a result, in this paper, if the prediction probability is more than 0.5, the peptide sequence is considered THP, else it is not.

Comparison with other ensemble strategies

Table 4 presents a comprehensive comparison of three popular ensemble strategies including voting, averaging, and

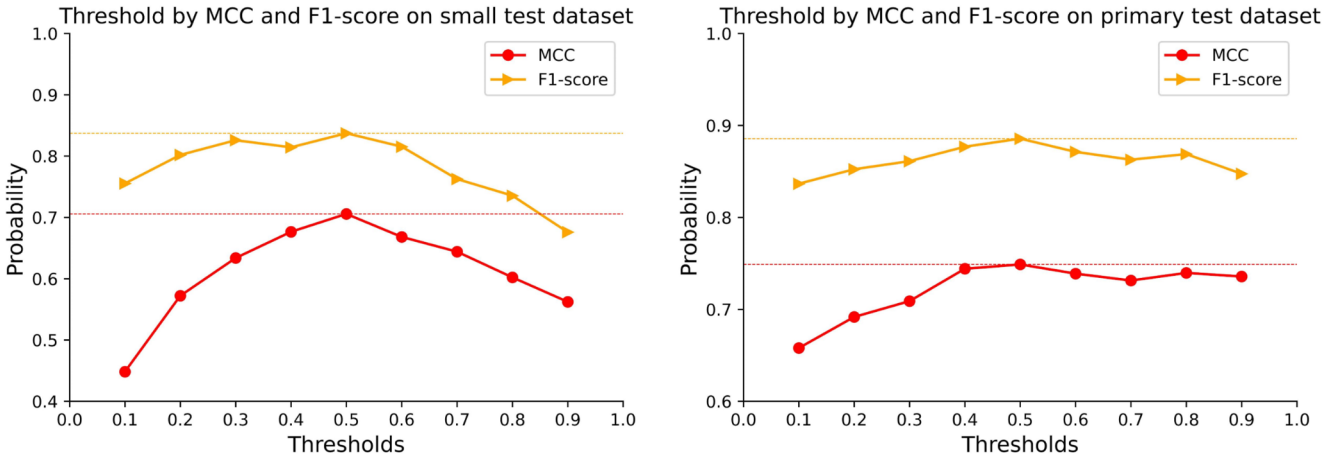

Fig. 4 The threshold distribution to identify THP

Table 4 The comparation with different ensemble strategy

Dataset	Strategy	Precise	SP	SE	ACC	MCC	F1	AUC	AP
Primary	Voting	0.8024	0.8312	0.8742	0.8345	0.7012	0.8445	0.9112	0.8945
	Averaging	0.8212	0.8104	0.8905	0.8512	0.7214	0.8546	0.9145	0.9156
	Stacking	**0.8439**	**0.8307**	**0.9153**	**0.8730**	**0.7488**	**0.8782**	**0.9528**	**0.9532**
Small	Voting	0.8142	0.8214	0.7416	0.7914	0.6512	0.8029	0.8578	0.8471
	Averaging	0.8089	0.8179	0.7342	0.8045	0.6812	0.8012	0.8446	0.8064
	Stacking	**0.8470**	**0.8617**	**0.7985**	**0.8391**	**0.7056**	**0.8372**	**0.9081**	**0.9119**

Table 5 The results of compared with state-of-the-art methods on primary test dataset

Methods	Precise	SP	SE	ACC	MCC	F1	AUC	AP
StackTHPred	0.7852	0.7538	0.9000	0.8269	0.6609	0.8387	0.9257	0.9059
SCMTHP	0.7986	0.7692	**0.9153**	0.8423	0.6920	0.8530	0.9182	0.8941
MIMML	0.8370	**0.8307**	0.8692	0.8500	0.7005	0.8528	0.9277	0.9074
NEPTUNE	0.8226	0.8076	0.8923	0.8500	0.7025	0.8560	0.9334	0.9262
LLM4THP	**0.8439**	**0.8307**	**0.9153**	**0.8730**	**0.7488**	**0.8782**	**0.9528**	**0.9532**

stacking for the task of distinguishing between THP and non-THP. The results, measured across a range of performance metrics including precision, specificity, sensitivity, accuracy, MCC, F1-score, AUC, and AP, reveal more differences between the strategies depending on the dataset size. In the case of the primary dataset, which is presumably larger and more diverse, stacking emerges as the superior strategy across all metrics, demonstrating the potential of this approach to use the complementary strengths of individual models to achieve higher performance. Voting and averaging also exhibit competitive results, with averaging slightly outperforming voting in terms of AUC and AP. Besides, on the small dataset, stacking maintains its dominance, suggesting that the strategy is particularly effective in extracting valuable information from limited data. In contrast, averaging and voting exhibit more comparable performance on the smaller dataset, with voting slightly outperforming averaging in terms of precision and specificity. Therefore, in this paper we utilize stacking strategy to build LLM4THP.

Compared with state-of-the-art methods

Table 5 presents a comprehensive comparison of the proposed LLM4THP model with five state-of-the-art methods (StackTHPred, SCMTHP, MIMML, NEPTUNE) on a primary test dataset. The evaluation metrics used include Precision, Specificity, Sensitivity, Accuracy, MCC, F1 Score, AUC, and AP. The results indicate that LLM4THP achieves superior performance across most metrics compared to the other methods. Specifically, LLM4THP demonstrates the highest values for Precision, Specificity, F1 Score, AUC, and AP, highlighting the effectiveness in accurately identifying target homologs, which suggests that the proposed model's utilization of two-layer stacking ensemble learning, along with the integration of multiple features, contributes to improve predictive capability. In terms of ACC, MCC, F1, AUC and AP, LLM4THP shows improvement by 2.3–4.61%, 4.63–8.79%, 2.22–3.95%, 1.94% to 3.46 and 2.7–5.91%. Furthermore, LLM4THP exhibits a competitive performance in Sensitivity, Accuracy, and MCC,

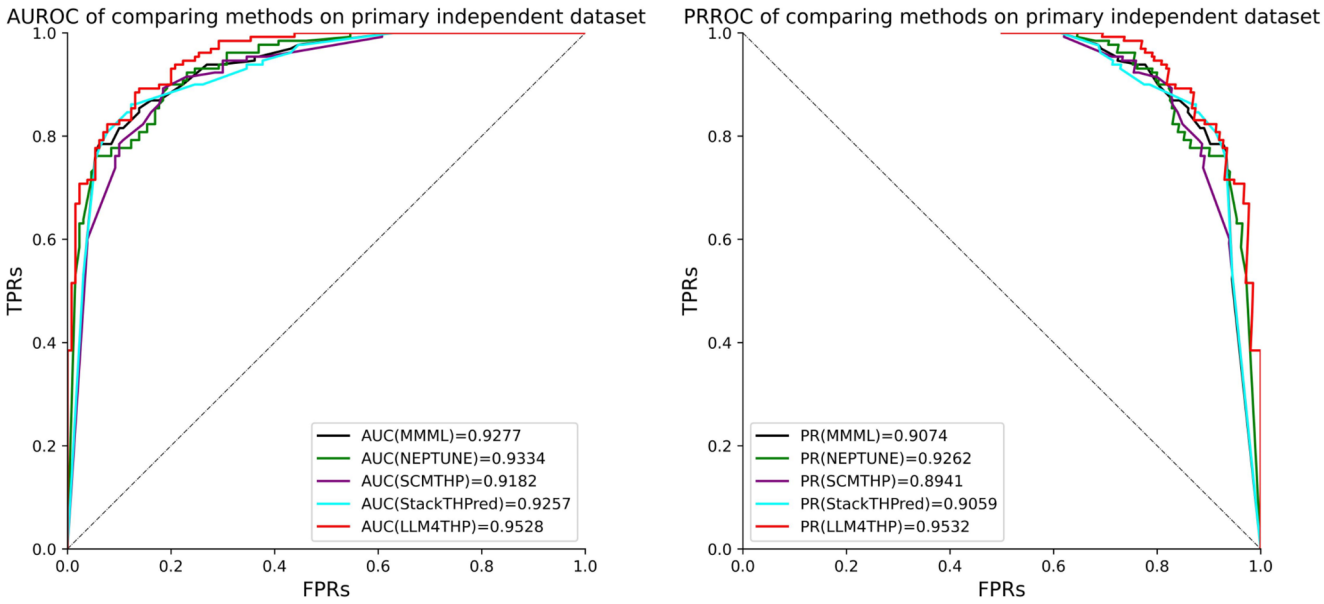

Fig. 5 The AUC curve and PR curve of each comparing method on primary test dataset

Table 6 The results of compared with state-of-the-art methods on small test dataset

Methods	Precise	SP	SE	ACC	MCC	F1	AUC	AP
StackTHPred	0.8390	0.8510	0.7765	0.8138	0.6294	0.8066	0.8825	0.8809
SCMTHP	0.8089	0.8191	0.7659	0.7925	0.5859	0.7868	0.8681	0.8670
MIMML	0.8314	0.8404	0.7872	0.8138	0.6285	0.8087	0.8798	0.8807
NEPTUNE	**0.8522**	**0.8617**	0.7978	0.8297	0.6609	0.8241	0.8785	0.8833
LLM4THP	0.8470	**0.8617**	**0.7985**	**0.8391**	**0.7056**	**0.8372**	**0.9081**	**0.9119**

demonstrating the ability to balance the detection of true positives and true negatives. Overall, the comparison results in Table 5 provide compelling evidence of the effectiveness and robustness of the LLM4THP model for predicting THP on primary test dataset.

Furthermore, to highlight the differences between each comparison approach, Fig. 5 shows the AUC and PR curves. Figure 5 shows both the AUC curve and the PR curve of LLM4THP when compared to other approaches, indicating that LLM4THP has a greater ability to distinguish between LLM4THP and non- LLM4THP.

Besides, we also compare LLM4THP with other state-of-the-art methods on small test dataset listed in Table 6. Among the compared methods, StackTHPred, SCMTHP, MIMML, NEPTUNE, and our proposed method, LLM4THP, each demonstrates unique strengths and areas of improvement. StackTHPred and SCMTHP show competitive performance with precision scores of 0.8390 and 0.8089, respectively, and AUC values of 0.8825 and 0.8681, respectively. MIMML, with a precision of 0.8314 and an AUC of 0.8798, also holds a strong position in terms of predictive accuracy. However, the method that stands out with the most promising results is our proposed LLM4THP. LLM4THP achieves the highest scores across several metrics, with a precision of 0.8470, specificity of 0.8617, sensitivity of 0.7985, accuracy of 0.8391, and an MCC of 0.7056. Notably, LLM4THP excels with an F1-score of 0.8372, AUC of 0.9081, and an AP of 0.9119, indicating a well-balanced model that effectively minimizes both false positives and false negatives. For improvement, in terms of SE, ACC, MCC, F1, AUC and AP, LLM4THP shows improvement by 0.07–3.26%, 0.94–4.66%, 4.47–11.97%, 1.31–5.04%, 2.56–4.0% and 2.86–4.49%. The superior performance of LLM4THP can be attributed to its sophisticated ensemble strategy, which harnesses the complementary strengths of individual models to enhance overall predictive accuracy. This is particularly evident in the AUC and AP scores, where LLM4THP significantly outperforms the other methods, suggesting its exceptional ability to discriminate between THP and non-THP. The comparative analysis of the small test dataset underscores the superiority of our proposed LLM4THP method over existing state-of-the-art approaches. The outstanding performance of LLM4THP across multiple metrics highlights the potential as a leading model to identify THP, offering a robust and reliable solution.

To further elucidate the distinctions between the compared approaches on a smaller test dataset, Fig. 6 depicts the Receiver Operating Characteristic (ROC) and

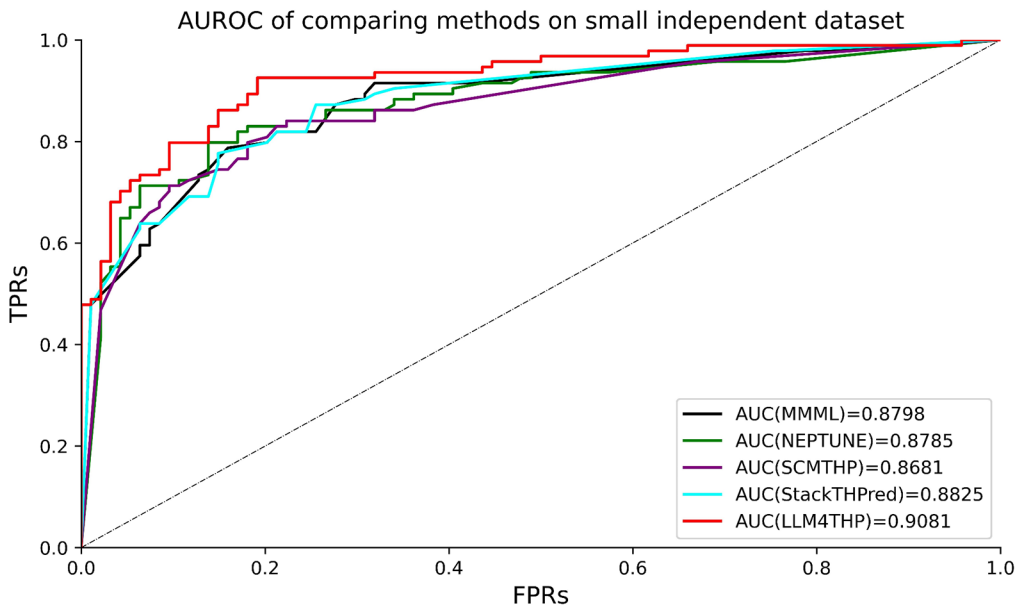

Fig. 6 The AUC curve and PR curve of each comparing method on small test dataset

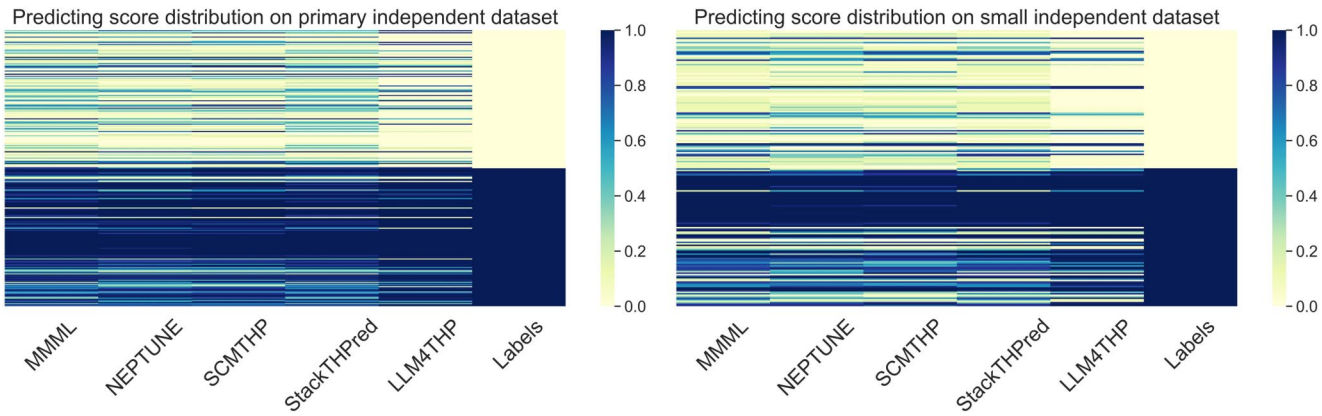

Fig. 7 The predicting probability distribution of each compared method on primary test dataset and small test dataset

Precision-Recall (PR) curves. This visualization reveals the discriminatory power of each method in distinguishing between true positives and false positives. As observed in Fig. 6, LLM4THP consistently demonstrates the highest area under the ROC curve (AUC) and PR curve, surpassing the performance of the other approaches. This indicates that LLM4THP possesses a greater ability to accurately classify instances as either THP or non-THP.

Furthermore, Fig. 7 depicts the forecasting probability of each compared approach. The lighter color, the lower the risk of THP, and the darker the color, the higher the probability. Figure 7 shows that LLMTMP is darker in the THP area and lighter in the non-THP part than the other approaches. The visualization of the prediction distribution demonstrates that LLM4THP can predict THP more accurately.

The comparation between LLM4THP and other methods

Based on the above results, we summary the differences and advantages of LLM4THP for other compared methods, and the results are listed in Table 7. We can find LLM4THP use more comprehensive features and more powerful classification and get better performance. Therefore, the main contribution of LLM4THP is following:

(1) The combined encoding capabilities of a large language model, intrinsic features of peptide sequences, and molecular information all contribute to the identification of THP.

(2) The two-layer ensemble technique demonstrates beneficial to accuracy and robustness in distinguishing between THP and non-THP.

Table 7 The differences and advantages of compared methods

Names	Features	Methods	Advantages	Limitations
StackTH-Pred	AAC, PAAC, PHYC, BLOSUM62, Z-Scale.	GBDT, Stacking Ensemble Architecture, ET, RF, GBDT	Interpretability	Overreliance on Features,
SCMTHP	AAC, Propensity Scores	SCM, GA, PCPs	Simplicity	Without higher Generalization
MIMML	Embedding Technique, Convolution Kernel and Mutual Information.	Meta-Learning Paradigm, Joint Optimization	Mutual Information Maximization	Overfitting Risk
NEPTUNE	AAC, DPC, AAIndex, APAAC, CTD, PAAC, PCP and RS.	Stacking Ensemble Learning, SVM, GA, SAR	Feature Optimization	Complexity of model
LLM4THP	ESM2, AAC, PAAC, APAAC and CTD, RDKit-Descriptors, Prot_T5_XL_UniRef50,	Stacking Ensemble Learning, LightGBM, XGBoost, RF, ET	Protein large language embedding model, SMILES property, Stacking Ensemble Learning	More computation

Note Amino Acid Composition (AAC), Pseudo-Amino Acid Composition (PAAC), Physicochemical Properties (PHYC), Dipeptide Composition (DPC), Amino Acid Index (AAIndex), Amphiphilic Pseudo-Amino Acid Composition (APAAC), Composition Transition and Distribution (CTD), Physicochemical Properties (PCP), Reduced Protein Sequences (RS), Gradient Boosting Decision Tree (GBDT), Extremely Randomized Trees (ET), Random Forest (RF), Scoring Card Method (SCM), Genetic Algorithm (GA), Physicochemical Properties (PCPs), Support Vector Machine (SVM), Self-Assessment-Report (SAR)

(3) The experimental results reveal that LLM4THP outperforms the other approaches that were compared.

Discussion

Tumor homing peptides (THPs) are small, sequence-specific molecules that have the remarkable ability to selectively bind to tumor cells or tissues. This unique property makes THP invaluable tools in various areas of cancer research and treatment. THPs can be used to deliver therapeutic agents directly to tumor cells, minimizing damage to healthy cells and reducing side effects. By conjugating THPs to anticancer drugs, nanoparticles, or other therapeutic payloads, these agents can be specifically targeted to tumors, enhancing their efficacy and reducing systemic toxicity. THPs also can be used as imaging agents to visualize tumors. By attaching fluorescent or radioactive labels to THPs, researchers and clinicians can non-invasively visualize tumor location, size, and progression. This information is crucial for diagnosis, staging and monitoring treatment response. In brief, tumor homing peptides offer a powerful tool for targeted cancer therapy and diagnosis. Their ability to selectively bind to tumor cells holds immense potential for improving cancer treatment outcomes and understanding of the mechanism. However, discovering THPs through traditional experimental approaches, such as phage display, can be costly and time-consuming. These methods require extensive laboratory work, including peptide synthesis, screening, and characterization. Additionally, the optimization of peptide sequences to enhance tumor targeting can be a complex and iterative process. Therefore, it is necessary to identify THP in a high-throughput way. Machine learning models can rapidly analyze large datasets and identify patterns and relationships that are not apparent through manual analysis, which significantly reduces the time and effort required to discover and validate potential THPs.

Therefore, in this paper, we proposed a new computational method to identify THP, called LLM4THP. For the construction of features, we utilize large language models (LLMs) capable of extracting valuable information from peptide sequences, peptide sequence intrinsic features and molecular information features to encode peptide sequences. LLMs have revolutionized natural language processing by capturing intricate patterns and relationships within text data. Applying similar techniques to peptide sequences could lead to more comprehensive and nuanced representations. Leveraging large language models (LLMs) to encode peptide sequences, two protein large language model including ESM2 and Prot_T5_XL_UniRef50 is introduced to encode peptide sequence. ESM2 and Prot_T5_XL_UniRef50 can enhance peptide sequence representation and show superior performance to identify THP. Furthermore, pervious study has shown that peptide inner sequence features are also competitive information to identify THP. Therefore, we employ Amino Acid Composition (AAC), Pseudo Amino Acid Composition (PAAC), Amphiphilic Pseudo Amino Acid Composition (APAAC) and Composition, Transition and Distribution (CTD) as intrinsic sequence features to encode peptide sequence. Furthermore, we find a new feature representment method called RDKitDescriptors shows higher performance to identify THP as well. RDKit-Descriptors entails transforming a peptide sequence into a molecular object using the RDKit Python library's Chem.MolFromFASTA(seq) function, where seq represents the peptide sequence. Chem.MolToSmiles(mol) converts the molecular object into SMILES format. RDKitDescriptors is then used to calculate the chemical attribute of the peptide

using the SMILES format input, returning a dictionary data structure containing 210 chemical attribute features to encode peptide sequence. In all, we from three views to construct feature space to represent THP and these features get a contributive effect to identify THP. Building upon these embedding vectors, the ensemble strategy is implemented to create LLM4THP, which comprises a two-layer learning architecture. The initial layer consists of four meta predictors: LightGBM (LGBM, designated as M1), XGBoost (XGB, designated as M2), Random Forest (RF, designated as M3), and Extremely randomized trees (ERT, designated as M4). The cross-product of the embedding vectors [V1, V2, V3, V4, …, V7] with the meta predictors [M1, M2, M3, M4] yields a set of results termed VMs, which reflects the predictive capabilities of each feature in conjunction with each model. This ensemble of predictions is then processed by Logistic Regression to refine the distinction between THP and non-THP sequences. The ultimate output of LLM4THP is a classification that determines whether the input peptide sequence is a THP or a non-THP, based on the aggregated predictions from the ensemble model. Finally, LLM4THP is evaluated by multiple metrics and a user-friendly prediction is implemented for academic research. Additionally, LLM4THP is compared with other state-of-the-art methods and show better performance. LLM4THP outperformance other compared methods in terms of ACC, MCC, F1, AUC and AP with improvement by 2.3–4.61%, 4.63–8.79%, 2.22–3.95%, 1.94% to 3.46 and 2.7–5.91% on primary test dataset and 0.07–3.26%, 0.94–4.66%, 4.47–11.97%, 1.31–5.04%, 2.56–4.0% and 2.86–4.49% on small test dataset. Therefore, the main contribution of LLM4THP is following: (1) The integrated encoding features from large language model, peptide sequence intrinsic features and molecular information contribute to identify THP; (2) The two-layers ensemble strategy show high accuracy and robust to distinguish THP and non-THP; (3) Experiment result indicate LLM4THP get better performance than other compared methods. Additionally, the source code and dataset are available at https://github.com/abcair/LLM4THP.

Conclusion

In conclusion, the paper presents a novel computational approach, LLM4THP, designed to identify Tumor Homing Peptides (THPs) with high accuracy and efficiency. THPs, due to their unique ability to selectively bind to tumor cells, are significant in cancer research and treatment, offering targeted delivery of therapeutic agents and aiding in non-invasive tumor imaging. The traditional discovery methods for THPs are laborious and costly, necessitating a high-throughput alternative. LLM4THP leverages the power of large language models (LLMs) to extract meaningful information from peptide sequences, employing ESM2 and Prot_T5_XL_UniRef50 to enhance peptide sequence representation. Additionally, the method incorporates intrinsic sequence features such as Amino Acid Composition (AAC), Pseudo Amino Acid Composition (PAAC), Amphiphilic Pseudo Amino Acid Composition (APAAC), and Composition, Transition, and Distribution (CTD). A new feature representation method, RDKitDescriptors, is introduced, demonstrating superior performance in identifying THPs by transforming peptide sequences into molecular objects and calculating chemical attributes. The ensemble strategy of LLM4THP, with its two-layer learning architecture, combines the predictive capabilities of four meta predictors including LightGBM, XGBoost, Random Forest, and Extremely randomized trees with logistic regression to refine the classification of THPs. This approach has been rigorously evaluated and demonstrated a significant improvement over state-of-the-art methods in various metrics, including Accuracy, Matthew's correlation coefficient, F1 score, Area Under the Curve, and Average Precision. In all, LLM4THP represents a significant advancement in the field of computational biology, particularly in the discovery of THPs, and has the potential to greatly enhance the development of targeted cancer therapies and diagnostic tools.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00726-024-03422-5.

Author contributions S. Y. is the project administrator and P. X develops the software. All authors reviewed the manuscript.

Funding This research was funded by Natural Science Foundation of Jiangsu Province of China (Grant No. BK20230626), partly supported by the open funds of the State Key Laboratory of Plant Environmental Resilience (Grant No. SKLPERKF2401), supported by the Open project of State Key Laboratory of Animal Biotech Breeding (Grant No. 2024SKLAB6-1), the Fourth Batch of Leading Innovative Talents Introduction and Training Projects under the Longcheng Talent Plan in Changzhou City (Basic Research and Innovation) (Grant No. CQ20230086), Changzhou Science and Technology Plan (Basic Research Program) 2024 (Grant No. CJ20241083).

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethical approval This article does not contain any studies with animals performed by any of the authors.

Competing interests The authors declare no competing interests.

References

Arif R, Kanwal S, Ahmed S, Kabir M (2024) A computational predictor for Accurate Identification of Tumor homing peptides by integrating sequential and deep BiLSTM features. Interdiscip Sci Comput Life Sci. https://doi.org/10.1007/s12539-024-00628-9

Armstrong G, Martino C, Rahman G et al (2021) Uniform Manifold Approximation and Projection (UMAP) reveals composite patterns and resolves visualization artifacts in Microbiome Data. mSystems 6:e00691–e00621. https://doi.org/10.1128/mSystems.00691-21

Bairoch A (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48. https://doi.org/10.1093/nar/28.1.45

Bartas M, Červeň J, Guziurová S et al (2021) Amino acid composition in various types of nucleic acid-binding proteins. IJMS 22:922. https://doi.org/10.3390/ijms22020922

Bento AP, Hersey A, Félix E et al (2020) An open source chemical structure curation pipeline using RDKit. J Cheminform 12:51. https://doi.org/10.1186/s13321-020-00456-1

Charoenkwan P, Yana J, Nantasenamat C et al (2020) iUmami-SCM: a Novel sequence-based predictor for prediction and analysis of umami peptides using a Scoring Card Method with Propensity scores of Dipeptides. J Chem Inf Model 60:6666–6678. https://doi.org/10.1021/acs.jcim.0c00707

Charoenkwan P, Chiangjong W, Nantasenamat C et al (2022a) SCMTHP: a New Approach for identifying and characterizing of tumor-homing peptides using estimated propensity scores of amino acids. Pharmaceutics 14:122. https://doi.org/10.3390/pharmaceutics14010122

Charoenkwan P, Schaduangrat N, Lio' P et al (2022b) NEPTUNE: a novel computational approach for accurate and large-scale identification of tumor homing peptides. Comput Biol Med 148:105700. https://doi.org/10.1016/j.compbiomed.2022.105700

Guan J, Yao L, Chung C-R et al (2023) StackTHPred: identifying tumor-homing peptides through GBDT-Based feature selection with stacking Ensemble Architecture. IJMS 24:10348. https://doi.org/10.3390/ijms241210348

He W, Jiang Y, Jin J et al (2022) Accelerating bioactive peptide discovery via mutual information-based meta-learning. Brief Bioinform 23:bbab499. https://doi.org/10.1093/bib/bbab499

Huang F, Li X, Yuan C et al (2022) Attention-emotion-enhanced convolutional LSTM for sentiment analysis. IEEE Trans Neural Netw Learn Syst 33:4332–4345. https://doi.org/10.1109/TNNLS.2021.3056664

Huttunen-Hennelly HEK (2010) An investigation into the N- and C-capping effects of glycine in cavitand-based four-helix bundle proteins. Bioorg Chem 38:98–107. https://doi.org/10.1016/j.bioorg.2010.01.004

Jiang H, Zou B, Xu C et al (2020) SVM-Boosting based on Markov resampling: theory and algorithm. Neural Netw 131:276–290. https://doi.org/10.1016/j.neunet.2020.07.036

Kapoor P, Singh H, Gautam A et al (2012) TumorHoPe: a database of tumor homing peptides. PLoS ONE 7:e35187. https://doi.org/10.1371/journal.pone.0035187

Karami Fath M, Babakhaniyan K, Zokaei M et al (2022) Anti-cancer peptide-based therapeutic strategies in solid tumors. Cell Mol Biol Lett 27:33. https://doi.org/10.1186/s11658-022-00332-w

Katubi KM, Saqib M, Mubashir T et al (2023) Predicting the multiple parameters of organic acceptors through machine learning using RDkit descriptors: an easy and fast pipeline. Int J Quantum Chem 123:e27230. https://doi.org/10.1002/qua.27230

Kondo E, Iioka H, Saito K (2021) Tumor-homing peptide and its utility for advanced cancer medicine. Cancer Sci 112:2118–2125. https://doi.org/10.1111/cas.14909

Langdon A, Botvinick M, Nakahara H et al (2022) Meta-learning, social cognition and consciousness in brains and machines. Neural Netw 145:80–89. https://doi.org/10.1016/j.neunet.2021.10.004

Lempens EHM, Merkx M, Tirrell M, Meijer EW (2011) Dendrimer Display of Tumor-Homing peptides. Bioconjug Chem 22:397–405. https://doi.org/10.1021/bc100403e

Li ZJ, Cho CH (2012) Peptides as targeting probes against tumor vasculature for diagnosis and drug delivery. J Transl Med 10:S1. https://doi.org/10.1186/1479-5876-10-S1-S1

Li J, Wang S, Zhang D et al (2016) Amino acids functionalized graphene oxide for enhanced hydrophilicity and antifouling property of poly(vinylidene fluoride) membranes. Chin J Polym Sci 34:805–819. https://doi.org/10.1007/s10118-016-1808-2

Li L, Lu Y, Lin Z et al (2019) Ultralong tumor retention of theranostic nanoparticles with short peptide-enabled active tumor homing. Mater Horiz 6:1845–1853. https://doi.org/10.1039/C9MH00014C

Lin Y, Lim YF, Russo E et al (2015) Multidimensional Design of Anticancer peptides. Angew Chem Int Ed 54:10370–10374. https://doi.org/10.1002/anie.201504018

Lin Z, Akin H, Rao R et al (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123–1130. https://doi.org/10.1126/science.ade2574

Liu W, Fan H, Xia M (2022) Credit scoring based on tree-enhanced gradient boosting decision trees. Expert Syst Appl 189:116034. https://doi.org/10.1016/j.eswa.2021.116034

Lu L, Qi H, Zhu J et al (2017) Vascular-homing peptides for cancer therapy. Biomed Pharmacother 92:187–195. https://doi.org/10.1016/j.biopha.2017.05.054

Meher PK, Sahu TK, Mohanty J et al (2018) nifPred: proteome-wide identification and categorization of Nitrogen-fixation proteins of Diaztrophs based on composition-transition-distribution features using support Vector Machine. Front Microbiol 9:1100. https://doi.org/10.3389/fmicb.2018.01100

Melssen MM, Sheybani ND, Leick KM, Slingluff CL (2023) Barriers to immune cell infiltration in tumors. J Immunother Cancer 11:e006401. https://doi.org/10.1136/jitc-2022-006401

Naseer S, Ali RF, Khan YD, Dominic PDD (2022) iGluK-Deep: computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. J Biomol Struct Dynamics 40:11691–11704. https://doi.org/10.1080/07391102.2021.1962738

O'Boyle NM (2012) Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J Cheminform 4:22. https://doi.org/10.1186/1758-2946-4-22

Pratyush P, Bahmani S, Pokharel S et al (2024) LMCrot: an enhanced protein crotonylation site predictor by leveraging an interpretable window-level embedding from a transformer-based protein language model. Bioinformatics 40:btae290. https://doi.org/10.1093/bioinformatics/btae290

Sharma A, Kapoor P, Gautam A et al (2013a) Computational approach for designing tumor homing peptides. Sci Rep 3:1607. https://doi.org/10.1038/srep01607

Sharma A, Kapoor P, Gautam A et al (2013b) Computational approach for designing tumor homing peptides. Sci Rep 3:1607. https://doi.org/10.1038/srep01607

Shoombuatong W, Schaduangrat N, Pratiwi R, Nantasenamat C (2019) THPep: a machine learning-based approach for predicting tumor homing peptides. Comput Biol Chem 80:441–451. https://doi.org/10.1016/j.compbiolchem.2019.05.008

Soni S, Chouhan SS, Rathore SS (2023) TextConvoNet: a convolutional neural network based architecture for text classification. Appl Intell 53:14249–14268. https://doi.org/10.1007/s10489-022-04221-9

Suzek BE, Wang Y, Huang H et al (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932. https://doi.org/10.1093/bioinformatics/btu739

Thirunavukarasu AJ, Ting DSJ, Elangovan K et al (2023) Large language models in medicine. Nat Med 29:1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Wang C, Wang W, Lu K et al (2020) Predicting Drug-Target interactions with Electrotopological State Fingerprints and Amphiphilic Pseudo amino acid composition. IJMS 21(5694). https://doi.org/10.3390/ijms21165694

Wu C, Zhang Y, Wei X et al (2022a) Tumor homing-penetrating and nanoenzyme-augmented 2D phototheranostics against hypoxic solid tumors. Acta Biomater 150:391–401. https://doi.org/10.1016/j.actbio.2022.07.044

Wu C, Zhang Y, Wei X et al (2022b) Tumor homing-penetrating and nanoenzyme-augmented 2D phototheranostics against hypoxic solid tumors. Acta Biomater 150:391–401. https://doi.org/10.1016/j.actbio.2022.07.044

Zhang J, Chen C, Li A et al (2021) Immunostimulant hydrogel for the inhibition of malignant glioma relapse post-resection. Nat Nanotechnol 16:538–548. https://doi.org/10.1038/s41565-020-00843-7