Expert Opinion on Drug Discovery

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iedc20

Approaches for the discovery of new cell-penetrating peptides

Ly Porosk , Ilja Gaidutšik & Ülo Langel

To cite this article: Ly Porosk , Ilja Gaidutšik & Ülo Langel (2020): Approaches for the discovery of new cell-penetrating peptides, Expert Opinion on Drug Discovery, DOI: 10.1080/17460441.2021.1851187

To link to this article: https://doi.org/10.1080/17460441.2021.1851187

Published online: 01 Dec 2020.

EXPERT OPINION ON DRUG DISCOVERY
https://doi.org/10.1080/17460441.2021.1851187

Taylor & Francis
Taylor & Francis Group

REVIEW

Approaches for the discovery of new cell-penetrating peptides

Ly Porosk[a], Ilja Gaidutšik[a] and Ülo Langel [b]

[a]Institute of Technology, University of Tartu, Tartu, Estonia; [b]Department Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

ABSTRACT

Introduction: The capability of cell-penetrating peptides (CPP), also known as protein transduction domains (PTD), to enter into cells possibly with an attached cargo, makes their application as delivery vectors or as direct therapeutics compelling. They are generally biocompatible, nontoxic, and easy to synthesize and modify. Three decades after the discovery of the first CPPs, ~2,000 CPP sequences have been identified, and many more predicted. Nevertheless, the field has a strong commitment to authenticate new, more efficient, and specific CPPs.

Areas covered: Although a scattering of CPPs have been found by chance, various systematic approaches have been developed and refined over the years to directly aid the identification and depiction of new peptide-based delivery vectors or therapeutics. Here, the authors give an overview of CPPs, and review various approaches of discovering new ones. An emphasis is placed on *in silico* methods, as these have advanced rapidly in recent years.

Expert opinion: Although there are many known CPPs, there is a need to find more efficient and specific CPPs. Several approaches are used to identify such sequences. The success of these approaches depends on the advancement of others and the successful prediction of CPP sequences relies on experimental data.

ARTICLE HISTORY
Received 27 July 2020
Accepted 11 November 2020

KEYWORDS
CPP; HTS; PTD; delivery vectors; CPP prediction

1. Introduction

The cell membrane is a major limiting barrier for therapeutics with intracellular targets. Surmounting this barrier could enable new treatment possibilities for various human pathologies that are currently unaddressed. Besides therapeutic applications, efficient and specific delivery vectors open up an immense range of scientific and biotechnological applications. Both development of new, more efficient, and specific delivery systems and improving the ones already available, commit to increase the bioavailability and reduce the side effects associated with therapeutics today.

Cell-penetrating peptides are generally 4–40 amino acid (AA) long peptides that can enter into cells, and enhance the cellular uptake of various molecular cargoes; that otherwise are not able to cross the cell membrane [1]. They are considered nontoxic and do not permanently damage the cell membranes while entering, or at least this should be the aim. Although we presently know of several efficient CPPs, they are unable to cover a wide range of possible applications. The discovery of new and efficient CPPs for research, therapeutic, and diagnostic applications remains a challenge. Over the years, several approaches aimed at finding new CPP have been developed, and adopted from other fields.

The remarkable notion of protein transduction was based on observations that some proteins that could shuttle within the cell and from one cell to another. In 1988 the cellular uptake of the tat protein of HIV-1 into cell was described [2,3]. A few years later, in 1991, it was demonstrated that the 60 AA homeodomain of Antennapedia (a *Drosophila* homeodomain) could enter into cells, and later it was shown that a short pAntp$_{43\text{-}58}$ peptide derived from this protein was sufficient for translocation [4,5]. In 1997, the peptide sequence Tat$_{48\text{-}60}$ derived from HIV tat protein, was identified as being required for cell entry [6], and the delivery of a non-covalently formed complex between a nucleic acid and CPP MPG was achieved [7]. In 2000, Wender *et al* demonstrated the entry of synthetic polyarginine peptides [8]. Later, Futaki *et al* demonstrated the beneficial effect of attaching stearic acid to R8 to achieve improved delivery [9]. The first *in vivo* use of a CPP-cargo conjugate came from Langel´s group in 1998 [10]. Subsequently, many other CPPs have been discovered, designed, and tested. In 2003, the first CPP mediated therapeutic agent, PsorBan®, a cyclosporine-poly-arginine conjugate, entered clinical trial phase II, which opened the opportunity for other CPPs. According to the CPPSite 2.0 database, there are almost 2,000 known CPPs, with 1,699 unique sequences [11], some of these have entered into clinical trials, such as XG-102, KAI-9803, R-002, AM-111 [12].

CPPs can be used as independent delivery vectors, but with the emerge of new polymer-based delivery approaches, polypeptide and peptide-polymer hybrids have been developed to overcome the limitations of peptide-based delivery vectors and shortcomings of polymer-based approaches. The incorporation or conjugation of CPPs into polymers by physical or chemical modifications produces multifunctional vectors with improved transfection efficacies, prolonged blood circulation times, enhanced accumulation at tumors sites, and targeting. CPPs can be used to functionalize polymers such as polysaccharides, proteins, lipids, micelles, and nanoparticles [13–16].

CONTACT Ly Porosk ly.porosk@ut.ee Institute of Technology, University of Tartu, Tartu 50411, Estonia

Article highlights

- *In silico* prediction from both protein and peptide sequences significantly increases the probability of finding new CPPs and reduce the need for extensive *in vitro* experiments.
- Although new prediction algorithms and approaches are developed, educated guess and trial-and-error approaches retain their importance in discovering new CPPs.
- Predictions rely on known experimental data and there is a need for experimental HTS approaches that increase prediction accuracy and versatility.
- There is no clear-cut strategy to design new CPPs.
- CPPs must be tested and modified for specific applications.

This box summarizes key points contained in the article.

Among other examples polymers accessorized with CPP have demonstrated inhibition of tumor growth [17], and enhanced intranasal delivery of siRNA [18,19].

CPPs comprise an eminently diverse group of peptides, with moderate similarities between their sequences, secondary structures, internalization mechanisms, and efficacies. There are resemblances between CPPs such as AA compositions, origins, biochemical and physico-chemical properties, etc. These can be used to classify the vast number of CPPs. Classification itself is sometimes complicated, as one peptide may fall into more than one category, due to its overlapping properties. Different suggestions for classification systems have been introduced over the years, for example, based on their origin, CPPs can be classified into designed and protein-derived peptides, or based on their physico-chemical or structural properties, between predicted or random CPPs, nonspecific, or targeted, linear or modified, etc [1]. The diverse, nonetheless finite, number of possibilities of classification speaks to the versatility of CPPs. The various possibilities are indicative of the difficulties that may arise when trying to apply uniform rules to the diverse range of CPPs.

2. Discovery of new CPPs

The discovery of the first CPPs was a fortunate occurrence, and even today some new CPPs are discovered by serendipitous coincidence. Before attempting to find new, unique CPP candidates, one should recognize the incoherence between different CPPs, and lack of specific guidelines. The approaches require different prerequisites, such as skills, knowledge, and available infrastructure (Table 1). Therefore, combining different approaches and willingness to test more peptide candidates, would be beneficial. Often, the discovery is not straightforward and requires several rounds of predictions, optimizations, and modifications (Figure 1).

Various inputs can be used as a starting point to find new potential CPP sequences. Without a specific protein or a peptide sequence in mind, a broader screen in different databases can be done. There are several databases available such as Uniprot [21], Human Protein Atlas available from http://www.proteinatlas.org, AMP databases such as DBAASP [22], DRAMP [23], or signal peptide databases such as SignalP 5.0 [24], to name a few.

Typically protein sequences are used [5; 6; 10; 24–28]. Proteins have specific motifs, which encode functions generally characterized as binding, posttranslational modifications, and trafficking [25]. Proteins with specific localization, functions, or amino acid patterns may include sequences for prospective CPP sequences, and including these patterns in the peptide sequences may help to fine-tune the interactions between cargo DNA and peptide, or increase specificity by targeting organelles. Proteins, their sequences, and their known functions may give a base upon which new targeting or biologically active peptide (e.g. inhibitor or effector) sequences with cell internalization properties can be derived. For example, a C-terminal H/KDEL sequences provide a signal for retrieval from the Golgi complex [26]. Often, cell-penetrating properties are found in viral protein-derived peptides [27–31] or sequences containing nuclear localization sequences (NLS) [32,33]. NLS is an amino acid motif found in protein sequences that directs protein transport or shuffling between the cytoplasm and the nucleus. The length and features of NLS sequences vary substantially; however, NLS sequences are usually abundant in positively charged residues, with the consensus sequence K-K/R-X-K/R [34]. Monopartite and bipartite NLS motifs have been thoroughly described. They are characterized by a cluster (monopartite) or two clusters (bipartite) of basic residues preceded by a helix-breaking residue or separated by 9–12 residues. There are several variations to this, such as longer linker regions, tri-partide NLS motifs, proline-tyrosine NLS with the motif R/K/H-$X_{(2-5)}$-P-Y, etc [35]. Overlapping with NLS motifs are molecular sleds, which facilitate the sliding of peptide/protein on DNA [36] and could potentially help to fine tune the interactions between CPPs and nucleic acid.

Another class of peptides, antimicrobial peptides (AMP) can be used as a starting point because they have several characteristics similar to CPPs. From AMP, with slight modifications, new CPPs can also be designed, such as the synthetic peptide CIGB-552 [37]. Several reports indicate that the biological function of highly cationic peptides could be switched between antimicrobial and cell-penetrating peptides [38,39]. Peptides with both CPP and AMP function are, for example, Buforin II and SynB. Some AMPs, for example, hipposin, a histone-derived antimicrobial peptide isolated from Atlantic halibut, naturally contains a CPP sequence, that does not have antimicrobial activity itself [40]. Another CPP, sC18 was derived from the C-terminal domain of the cationic AMP CAP18 [41]. In addition to AMPs, inhibitor or effector peptides (derived from proteins or independent peptides) can be modified to add CPP properties.

Besides trying to find CPP sequences from proteins or making biologically active peptides into CPPs, various building blocks composed of specific AAs, motifs, peptides, or other modifications (e.g. fatty acid) can be used to design new CPPs. The first designed CPPs included individual known CPPs fused with sequences containing specific functions of interest [42]. Addition of a known CPP, mostly Tat and polyarginine, to other peptides has been used to provide them with the ability to penetrate cells [43–46]. Several other CPPs have also been harnessed, such as elastin-like polypeptide added to SynB1, Tat, or Bac CPPs [47] or PEGA added to pVec [48]. This approach is useful when the cargo (e.g. peptide sequences with bioactivity) should be modified as little as possible to retain its activity. When adding known CPPs to desired sequences or cargoes, the CPP activity may be decreased or the cargo sequence may lose its activity. For

Table 1. Approaches for the discovery of new CPPs, their requisite inputs, outputs, and requirement level for use.

Approaches		Inputs	Outputs	Requirements
Experimental approaches	Trial-and-error	Protein sequences; AMP	Experimental data	***
	Rational design	Peptide/cargo sequences with known bioactivity; parts of protein sequences with selected motifs; viral proteins; CPP sequences with low efficacy; AMP, and knowledge of needed traits and/or modification possibilities	Experimental data; structure-activity relationship	****
	Phage display	a gene encoding a protein/peptide inserted into a phage coat protein gene, and selection target	Targeting sequences (intracellular or membrane); internalized peptide candidates	***
	mRNA display	mRNA-DNA-puromycin library and immobilized selection target	Internalized peptide candidates; amplified and isolated sequences	***
	Experimental HTS approaches			
In silico approaches	*In silico* prediction	Peptide or protein sequences, depending on the prediction size limitations	Peptide sequences with predicted scores	*
	In silico modelling/ simulation of CPP-membrane interactions	Peptide sequences, lipid bilayer system data	Possible interactions with membranes, mechanism of entry, that has to be validated by experimentally.	***
Combined approaches	Synthetic molecular evolution	iterative (synthesized) library of peptides, and several rounds of screening	Experimental data; gain-of-function sequences	**
	Protein mimicry	Protein-protein interaction sequences from databases, or peptides with inhibitory/activation activity	Peptide sequences with predicted scores and possible interactions, must be tested experimentally	***

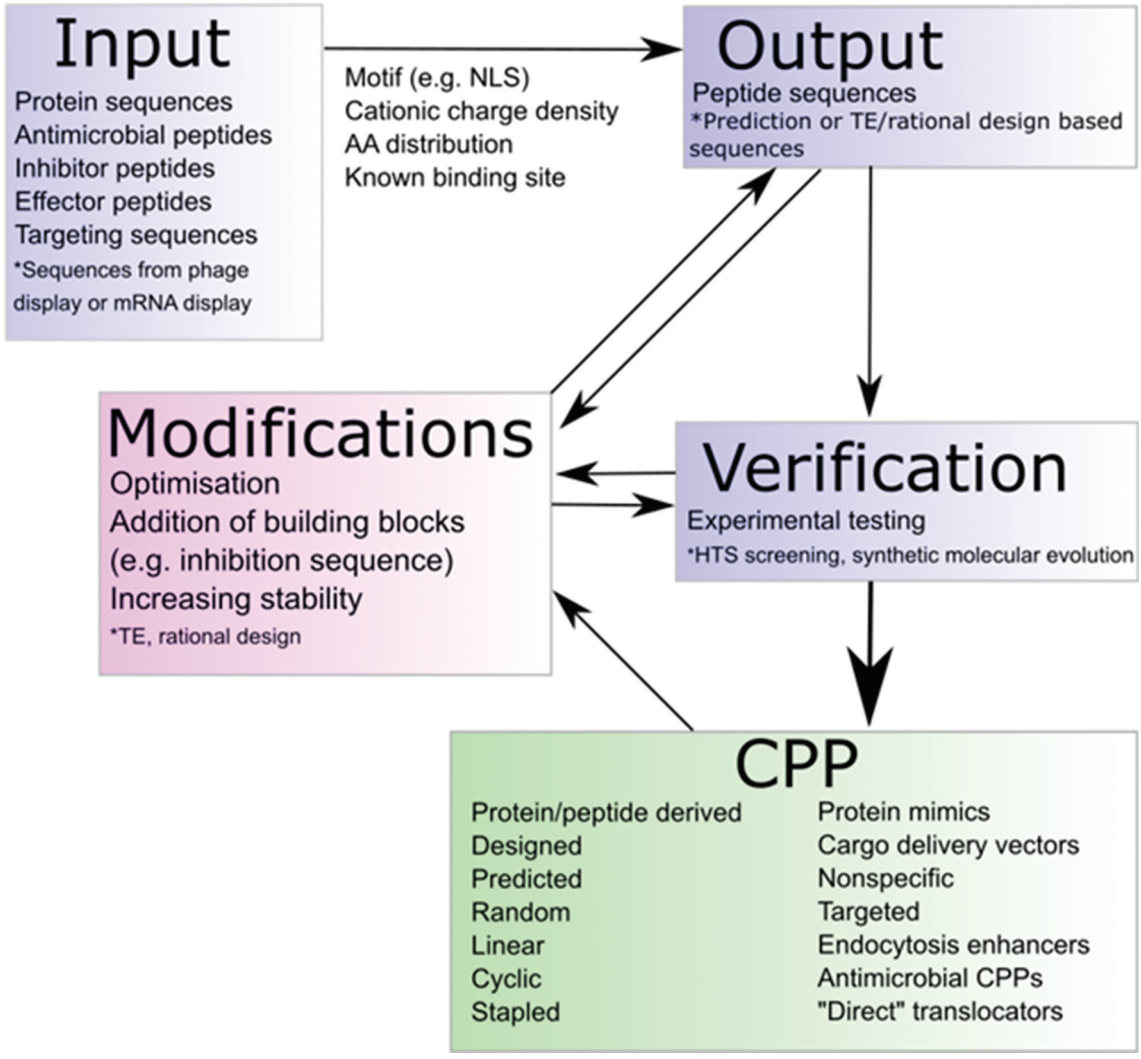

Figure 1. Workflow of CPP discovery, and CPP classes.

separating different functional parts of the peptide, additional AA, such as spans of glycines, or linkers, such as aminohexanoic acid, can be added. Nevertheless, testing several analogs is advised, especially when the cargo interactions should remain unaffected. Therefore, each modification should be considered from both a bioactivity and a cell-penetration point of view.

3. Approaches for the discovery of new CPPs

The approaches used to discover new CPPs can be broadly divided into experimental, *in silico*, and combined approaches. Each of which have their own advantages.

3.1. Experimental approaches

Experimental approaches allow one to collect results obtained with synthesized peptides. The CPP efficacy may be significantly affected by experimental conditions, therefore, ideally all the peptide candidates would be screened under the same conditions. An important contribution has been made by Remaker *et al*, who screened and ranked 474 sequence motifs under the same experimental conditions [49]. Another *E.coli* based screening of 55 peptides was performed by Oikawa *et al*, using fluorescence spectroscopy and confocal laser scanning microscopy [50]. Nevertheless, the peptides should also be tested for specific applications which may require suboptimal experimental conditions.

3.1.1. Trial-and-error

Trial-and-error (TE) is a fundamental approach, characterized by repeated, varied attempts to solve a problem. TE was frequently exploited in the early years of CPP research because there were no apparent rules to attach, and little structural resemblance was found between different CPPs. Although TE is expensive and laborious, it permits collecting experimental indications and step-by-step follow up of peptide efficacy and physiochemical characteristics. This information can be used to further refine predictions or modifications of CPPs. TE was used to discover and optimize several CPPs, such as CyLoP-1 derived from crotamine [51], and the chimeric CPP transportan [10].

The structure–activity relationship between point-mutations and protein function is widely used. One amino acid change may help to increase the efficacy of a peptide under specific conditions, such as *in vivo* [52], or to choose peptides with highest remaining bioactivity and penetration

capacity. In order to determine which AA and which position of the peptide sequence can and should be changed, a systematic approach generally helps. For example, the efficacy of CPP sC18 attached to silica nanoparticles was investigated by an alanine scan, by substituting each of the 16 amino acids systematically with an alanine residue [53]. Deletions and additions can also be introduced to the sequence, with the same objective. For example, deletion analogs of transportan were designed, synthesized, and tested [54], leading to the discovery of the most efficient analog.

Peptides in their native forms are rapidly degraded by peptidases under biological conditions. By introducing different modifications, the stability, internalization, and/or specificity can be improved. Because cells contain exopeptidases, cell-penetrating peptides are often capped to increase their stability, for example amidation of the C-terminus [55] and/or acetylation of the N-terminus [56]. Other modifications such as myristoylation [57], substitutions to non-coded AA, N-terminal addition of fatty acid, cyclization [58], stapling [59], and the use of D-AA have been used [60]. Usually, these modifications help with increasing the efficacy/specificity/stability of a peptide, but the penetration ability may be lost or decreased.

Although the TE approach is most often applied, rational design is instituting itself in the CPP field. The data gained from previous structure–activity relationships instructs new designs. Still, each peptide and especially peptides with cargo must be evaluated exclusively.

3.1.2. Rational design

The first designs of CPPs were mainly based on the properties of the AA, the primary structure of the peptide, and its tendency to form secondary structures. However, in complicated systems, it is difficult to predict the potency of CPPs based only on these criteria. Several chemical and physio-chemical properties, such as the charge, chirality, aromatic and hydrophobic content and also their co-action drive the internalization of CPPs and should all be considered [42]. Although several groups have tried to develop rational approaches, they frequently combine it with TE. The rational design of new CPPs falls on one´s thorough knowledge of CPPs and possible modifications. Essentially, the properties of a peptide should be known in advance, and secondly, the modifications introduced should help to achieve the desired outcome.

Most CPPs are classified as synthetic [11], and a common approach for developing new CPPs is to add new sequences to known CPPs thereby creating chimeric/synthetic peptides. In the first synthetic CPP sequences, both high cationic charge in polyarginine [8,9], and amphipathicity in a model amphipathic peptide [61] were taken as a basis for rational design. Further modifications in the polyarginine resulted in a tryptophan containing synthetic peptide and demonstrated how the spacing of the arginine residues influence uptake [62]. Rational design can also be applied to other types of CPPs, with the aim of increasing their specificity or efficacy. In the PepFect and NickFect CPP families several modifications have been introduced over the years. Although some of the modifications are one-two amino acid substitutions or additions of different length fatty acids, the transfection efficacies differ substantially [52,63,64]. For example, reducing and redistributing the net charge within the peptide sequence enhanced the plasmid delivery capability *in vivo* [52]. For siRNA delivery, a pH-sensitive increase in net charge and fatty acid modifications are advantageous [65] and new siRNA delivery vectors have been designed based on these parameters [63]. In mitochondrial-penetrating peptides (MMPs), two main parameters found to be important: positive charge and lipophilic character. Further fine-tuning of these improved the internalization [66] and new MPPs have been designed by the same group [63]. As an example of a more complex rational design that has been tested, a chimeric trifunctional peptide was created as a fusion of NLS, CPP, and an interfering peptide [67].

3.1.3. Phage display

Phage display methods harness bacteriophages to display foreign peptides on their surfaces by fusing the library or peptide sequence into the virus´s capsid protein [68]. The resulting heterogenous phages are then presented to immobilized targets such as proteins, peptides, or DNA sequences. Only the displayed peptides or proteins that are interacting with target molecules are detected. It is a potent technology for screening and isolating target-specific peptides. Several tissue or cell line-specific peptide sequences have been identified, such as cardiac targeting peptides [69], fibroblast growth factor receptor binding sequences [70], HUVEC cell line-specific sequences [71]. The phage display method can also be used to find CPPs [72] that are not tissue-specific, such as a CPP from M13 phage library [73]. Phage display is a suitable method for selection of peptides that can be used as targeting sequences or cell-penetration peptides. One limitation of phage display is the peptide/protein length, which should not interfere with the phage assembly. As a result, peptides that are too long, may be left out due to selection bias. The main advantage of phage display is that it allows one to screen through large numbers of possible candidates and select peptides based on their interactions and/or their capability for internalization.

Isolation of peptides that are capable of targeting specific receptors on specific cells would benefit the development of targeted gene therapy or therapeutics. CPPs can be modified with these specific ligands to improve their efficacy and specificity. Peptide sequences that are able to target in specific was are termed 'homing peptides.' There are several examples of chimeric and synthetic (cell-penetrating) peptides with targeting sequences, such as the glioma targeted drug delivery vector gHoPe [74], and brain-specific phage-derived peptide carrier [75]. Peptides that enable tissue specificity can be modified to become CPPs, or CPP sequences can be attached to targeting sequences. For example, adding an iRGD sequence increased the tumor specificity of the attached CPP [76].

Phage display can screen two or more required traits simultaneously and was used to detect both the targeting and membrane crossing ability of the cardiac targeting peptide CTP [69] from the M13 phage display library. A BirA-based (Biotin ligase) CPP discovery screen was introduced in 2018 [77]. Inspired by phylomer peptides [78] and virus-derived CPPs [79] the authors used a phage-based screening platform

to identify 'Phylomer' CPPs, derived from bacterial and viral genomes. Only peptides with intracellular uptake and cytosolic delivery are biotinylated inside cells that stably express BirA and these were chosen as CPPs. Thirteen unique CPPs derived from diverse organisms were identified using this approach.

3.1.4. mRNA display

mRNA display is also a high-throughput screening method [80] that can be used to both select functional peptides and evolve their properties. mRNA display is performed entirely *in vitro*. Although other in vitro translation approaches have been used [81], mRNA display has some advantages over other methods. It can scan through large libraries of peptide variants that are orders of magnitude larger than libraries that can be screened by other display technologies. In translation, due to the puromycin insertion at the 3´ end of mRNA sequence, the translated peptides stay associated with their mRNA progenitors. These mRNA representing peptides can afterward be used for cell experiments, from where internalized and interacting peptides can be isolated and amplified using reverse transcription and PCR [82]. An advantage over other display approaches is that only the peptide-nucleotide conjugates, that have translocated into the cells, can be enriched and amplified from the lysed cells. In addition, that ability to screen large libraries allows one to discover very rare sequences and screen through a more diverse range of candidate sequences. Two peptides with the ability to penetrate cell membranes were found using mRNA display [83].

3.1.5. Other high-throughput screening (HTS) approaches

The accurate prediction of CPPs relies on data gained from experimental studies of CPPs. The HTS methods, at least in their first stages, requires testing of many peptide sequences and their properties, such as the capability of penetrating into cells. In the long term, the information gained from these laborious studies form the basis upon which new guidelines can be based on. HTS methods provide major advantages when screening vast numbers; however, the parameters must be chosen carefully and the limitations of each screening methods must be considered.

The Kodadek group uses a different HTS approach. They monitor the relative cell permeability of large numbers of compounds by tagging every molecule in the library with a dexamethasone derivative [84,85]. Only peptides that have entered the cells will be counted as hits.

The one-bead-one-compound (OBOC) combinatorial method can be used to screen membrane-active peptides. Although each tested peptide must be synthesized, the output values can be ranked by intensity and enabling one to choose the most efficient peptide candidates. Libraries screened with this method can contain peptides with D-amino acids, and β-amino acids and the method can be used to test various internalization environments (e.g. pH, lipid composition). A large number of possible hits allow one to classify motifs within sequences that can be adjusted depending on the application (e.g. lower pH to mimic endosomal conditions and the capability of endosomal escape) [86].

3.2. *In silico* approaches

The vast majority of known cell-penetrating peptides are derived from known protein sequences. In the modern era of genomic sequencing researchers are provided with an enormous amount of protein-encoding sequences that are impossible to process manually, using trial and error wet-lab approaches. In contrast, *in silico* approaches are faster, cheaper, and less laborious. They allow one to conduct large-scale screenings and search for new CPPs that fulfill the needs of fundamental science and applied research fields such as biomedicine and pharmacology. As a recent example, an *in silico* method was used to screen the whole proteome of severe acute respiratory syndrome coronavirus 2 for CPPs [87].

3.2.1. In silico prediction of CPPs

Here we focus on *in silico* approaches used to predict CPP sequences. The CPP predictor construction usually contains three main stages.

1. The collection of datasets of proven CPPs and non-CPPs from the literature and/or databases. The data is usually split into two parts: a) the training set for algorithm learning, and b) testing set (aka independent set), which was not involved in the training process and is used to check the performance of the model.
2. Generation representative features that reflect different characteristics of selected peptides. These features may include AA composition, dipeptide composition, physicochemical properties of AAs, peptide structure, or combinations of different features. The generated features are often optimized to select the most influential feature(s) and discard irrelevant ones.
3. Model development by providing this representation of selected features to a machine learning (ML) algorithm.

To evaluate the performance of the constructed model, and to compare between different models, four metrics are frequently used. These are Sensitivity (SN), Specificity (SP), Accuracy (ACC) and the Mathew correlation coefficient (MCC), as expressed in the following equations:

$$SN = 100 * \frac{TP}{(TP + FN)}$$

$$SP = 100 * \frac{TN}{(TN + FP)}$$

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

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

Footnote to equations: TP, TN, FP, FN denote the true positives, true negatives, false positives, and false negatives. SN

and SP are the rate of true positives and true negatives. ACC is the ability of the model to differentiate between true positive and true negative and MCC is a correlation coefficient between observed and predicted. The value range of MCC holds between −1 and 1. In case of perfect prediction, the coefficient is 1, when coefficient is 0 the prediction is not better than random guess and −1 is total disagreement between observed and predicted.

The creation of a CPP prediction model is a complicated, multistep iterative process with several pitfalls that must be avoided. Firstly, the creation of datasets. Usually, they are split into two uneven parts: the first is used for training of the classifier or algorithm, and second – an independent (validation) dataset – is used to test the performance of resulting model. In general, the datasets should contain a sufficient number of diverse peptides, while avoiding redundant peptides. This works to limit overfitting of the model to the training dataset. In addition, the training set should be balanced and contain an even number of positive (CPPs) and negative (nCPPs) sample. Using unbalanced datasets decreases the accuracy of the resulting model. Initially, validated CPPs were scrupulously picked from literature; however, since 2012, a publicly available database of experimentally verified CPPs – CPPsite, can be used as the source of CPP information. It was upgraded to CPPsite 2.0 over time [11] and it currently contains more than 1800 entries.

The information stored in the CPP sequence is of key importance. The feature representation method reveals the stored information to a machine learning algorithm. In the beginning, the first models used single type of features that were based on AA physico-chemical properties [87;88] and biochemical properties [88,89]. Further works started using different features and their combinations. The most widely used features, in addition to the aforementioned, are the composition of the AA and dipeptide, pseudo AA composition (pseAAC) and motif-based features [89–93]; their mergers and application to machine learning algorithm in various combinations (hybrid features) has been shown to be the most efficient strategy. The efficiency of this hybrid approach is due to its overwhelming nature: the combination of the most influential prediction feature types. This approach will be the standard for CPP prediction models, until the understanding of the peptide uptake mechanism becomes clearer.

Machine learning algorithms are widely used in the field of CPP prediction. The selected features are presented to an ML algorithm that generates a predictive model. The most widely used methods for this purpose are support vector machine (SVM), random forest (RF), and the use of artificial neural networks (ANN). The SVM method creates a high dimensional space and finds a separation hyperplane that maximizes the distance between two classes of features [94]. The trained algorithms demonstrate accuracies between 81%-91% [89,90,95]. ANN-based algorithms are algorithmic models that simulate the structure of the brain to process information. The two examples of NN mentioned in this work [89; 97] perform with similar accuracies (83%). The most widely applied algorithm in CPP prediction is RF which generates the number of decision trees trained on the subset of information and the final result is determined by a combined total score of all decision trees [96]. The accuracies of RF-based methods are around 90% [91,92,97–99]. During the last several years, in pursuit of increased predictive power, CPP prediction models based on other algorithms have appeared. Manavalan *et al* [100,101] developed a two-layer method that first used extremely randomized trees (ERT) method [102] to predict the CPP and then a second-layer RF-based prediction is used for uptake efficiency prediction [100]. Pandey *et al* developed a model based on a kernel version of extreme learning machine (ELM) algorithm [102,103] and more recently, a method [89] based on gradient boost decision trees [104]. All three methods demonstrate comparable accuracies with state-of-art models (89.6%, 86.2%, 88.5%).

Because there are several approaches to developing computational methods for predicting CPP sequences, it is important to have some background on each in order to choose the one most suitable for the application. The first CPP prediction method was an algorithm based on z-scales [105]. The z-scales were initially calculated by Hellberg *et al* [106] for 20 coding AA. This idea was expanded and developed further, which led to the generation of z-scales for 87 AA (including non-coding AAs) based on their different physicochemical properties [107]. The bulk property values (ZΣ/n) of every peptide in the set of 24 published CPPs and 17 non-CPPs (nCPP) were used for training the algorithm [108]. Further [109], z-descriptors were subjected to partial least squares and principal component analysis (PCA). The amount of CPPs and nCPPs in the training set increased to 85. These models allow one to predict potential CPP sequences from whole protein sequences.

Several new machine learning-based CPP prediction algorithms have been developed in the last decade. In 2010, an ANN for CPP prediction was developed [88]. For this study, 101 CPPs were selected and >250 molecular features were generated for every peptide. Out of these, the six most influential were chosen as input data for ANN. Part of the dataset (30) was used to validate the performance of the ANN, and showed 83% accuracy. Further manipulations with dataset and improvements of the model resulted in almost 100% prediction accuracy.

Another machine learning related approach, SVM, was introduced for CPP prediction [89]. The work included 111 CPPs and 34 nCPPs. Basic biochemical properties (number of AAs, peptide length, net charge *etc*.) were generated for each peptide, and various combination of the properties were applied to train the SVM. The problems presented by this unbalanced dataset (CPP≫nCPP) was overcome by increasing the negative set with randomly generated nCPPs. The prediction accuracy of the algorithm increased to 91.7%. Furthermore, SVM models based on the AA position in the CPP sequence and AA motifs, have been developed and used together with much larger data sets (708 CPPs) [90]. The best SVM model outperformed all previously reported models on their datasets; however, the maximum accuracy of the most productive model when applied to an independent dataset was 81.3%. The algorithms developed in this study can be found as a web tool under the name CellPPD.

In 2013, another example of using ANN to predict CPPs was presented [110]. For the training of N-to-1 network, they used the datasets from [89]. All redundant peptides were eliminated to avoid overfitting and improved the predictive power of the

model. The advantage of N-to-1 NN over frequency-based methods is that it takes into account the entire motif and relative position of the AA in the sequence. The prediction accuracy of this model was 83%.

The repertoire of machine learning methods used for CPP prediction was enlarged by Chen *et al* [99]. The peptides from various previously published sources were encoded by pseAAC [111] representing many AA features. The most influential features were selected by a minimum redundancy maximum relevance (mRMR) method [112], the optimal prediction method and the set of optimal combination of features were found using an incremental feature selection (IFS) method and random forest (RF) classifier. The resulting prediction accuracy using this method was 83.5%.

The predictive power of algorithms has improved and further training sets have been added to predict multifunctional peptides. The first two-layer prediction algorithm [91] was created by training the RF algorithm on CPP and nCPP sets to predict whether any given sequence is a CPP (ACC 91%) and also how efficient its uptake is (ACC 66%). This trained algorithm was used to develop a new approach to design multifunctional CPPs with compatible activities. Furthermore, information about the residue order was presented to the algorithm in the form of dipeptides instead of just the AA composition [95]. The most efficient dipeptide features were selected by an analysis of variance (ANOVA) based technique and used to train an SVM algorithm. The resultant method was examined using ten-fold cross-validation (CV) and provided a maximum ACC of 83.6%.

The next aspects that were focused on to improve the performance of machine learning was the improvement of feature representation and selection methods [92]. For this, four feature descriptors were employed that enabled one to generate an enormous amount of features, and more importantly, rank them by using mRMR method. The optimal feature subset from this ranking was selected by Sequential Forward Search (SFS) and used to train an RF algorithm. The highest accuracy achieved by this method for CPP prediction was 91.6%. It is important to note that this method, available online as CPPred-RF is a two–layer method that also predicts the efficiency CPP uptake with 77% accuracy. The same authors published another publicly available prediction model (SkipCPP-Pred; accuracy 90.6%) [97] where the peptide sequence is processed by a k-skip-n-gram model which provides its resulting vector to an RF algorithm.

Another two-layer prediction framework that has been developed is termed MLCPP [100]. Four different types of features were extracted from the peptides and analyzed by an SVMQA method [113]. The most effective set of features was generated and presented to four different machine learning algorithms. Out of these, two showed the best scores: RF and extremely randomized trees (ERT, a variation of RF). The ERT outperformed other methods as the first layer predictor (CPP vs nCPP) and RF was selected as the second layer predictor (low vs high uptake). When tested on an independent dataset, MLCPP outperformed all the state-of-art methods of both algorithms with accuracies of 89.6% and 72.5%.

The majority of known CPPs contain only encoded AAs, however, various modifications to these CPPs are commonly used in their design and development. Kumar *et al* [98] used different structural information acquired from tertiary structures of known peptides to develop a method capable of predicting CPPs from modified peptides. RF displayed the best performance compared with other machine learning algorithms providing an ACC of 92.3%. The CellPPDMod server that utilizes this model could assist researchers in the design of CPPs that have various modifications and help to assess the influence of these modifications on the uptake of the peptide. Not only does optimizing the selection procedure for finding the most representative features help to improve the CPP prediction model efficacy, but also, as shown by the authors of KELM-CPPpred web server [102], utilization of a kernel version of extreme learning machine (ELM) algorithm [103]. In total, six types of features were used to train the algorithm. The resultant average accuracy obtained by 10-fold CV was 86.2% and 83.1% when tested on an independent dataset.

Currently, TargetCPP is the latest CPP prediction framework published [93]. The method utilizes a hybrid feature set composed of four representative groups that are selected by an mRMR selection algorithm and presented to a Gradient boost decision tree (GBDT) algorithm. The model obtained an accuracy of 88.3% and MCC 0.675 when tested on an independent data set. Summary of prediction programs and links (if applicable) are shown in Table 2.

3.2.2. Models and simulations of CPP-membrane interactions

Membranes consist of hundreds of different lipids and are crowded with proteins, creating distinct areas over the membranes that lead to processes such as membrane fusion, protein trafficking, signal transduction, and entry of therapeutics. Various computational methods have dramatically reduced the time and cost of drug discovery [114]. Molecular dynamics (MD) is a technique of computer simulations capable of describing the interactions between all the components in the system at atomic resolution, acting like a 'computational microscope.' The first MD simulations of surfactants and lipids appeared in the 1980s, and today they have a growing range of applications, including simulations of pore formation by AMPs, interactions with membrane-active peptides, and CPPs [115–117]. MD simulations have been used to describe the entry of arginine-rich CPPs [118] and HIV-1 Tat peptide [119], AMPs [120], and Spontaneous Membrane Translocating Peptides [116]. It has great potential as a large-scale computational screening of peptides and elucidation of their entry mechanism.

3.3. Combined approaches

In silico methods and experimental methods can be combined to find new CPPs by using the best of both approaches. Experimental data can confirm the predicted CPP activity and *in silico* predictions or data analysis can help to reduce both manual work and bias.

3.3.1. Synthetic molecular evolution

Synthetic evolution uses, by definition, modern molecular and synthetic biology approaches to iterate diversity and select

Table 2. Prediction methods are presented in this work and their features.

Predictor name	Feature representation/selection/classifier	ACC	MCC	Uptake (ACC)	Modifications	Reference	Accessible
N.A.	Z-scales of the physicochemical descriptors, PCA for descriptor selection	77%	N.A.	N.A.	Yes	[105]	No
N.A.	Z-scales of the physicochemical descriptors, PCA for descriptor selection, PLS	68%	N.A.	N.A.	Yes	[109]	No
N.A.	Molecular features, PCA for descriptor selection. ANN conjoined with QSAR	83%	N.A.	N.A.	N.A.	[88]	No
N.A.	Biochemical and physicochemical properties. SVM algorithm, 10-fold CV	91.7%	N.A.	N.A.	N.A.	[89]	No
CellPPD	AA composition, physicochemical properties, pattern profiles and motifs, SVM algorithm	81.3%	0.63	N.A.	N.A.	[90]	yes
CPPpred	Motif information, N-to-1 neural network	83%	0.69	N.A.	N.A.	[110]	Currently unavailable
N.A.	Presentation of CPP using pseAAC, analysis of features and models by mRMR and IFS. RF algorithm.	83.5%	0.49	N.A.	N.A.	[99]	no
N.A.	AA frequencies and physicochemical properties, RF algorithm	91%	N.A.	66%	N.A.	[91]	no
C2Pred	Dipeptides, Analysis of variance based technique (ANOVA), SVM algorithm, 5-fold CV.	83.6%	N.A.	N.A.	N.A.	[95]	yes
CPPred-RF	Four feature descriptors, feature selection mRMR and SFS, RF algorithm. Jackknife test validation.	91.6%	0.83	71.1%	N.A.	[92]	yes
SkipCPP-Pred	Adoptive k-skip-n-gram feature, RF algorithm, jackknife CV	90.6%	0.81	N.A.	N.A.	[97]	yes
MLCPP	Four feature descriptors, feature selection by SVMQA, 10-fold CV. 1st layer – ERT and 2nd layer – RF algorithm	89.6%	0.79	72.5%	N.A.	[100]	yes
CellPPDMod	Features from peptide structures and AA composition, RF algorithm.	92.3%	0.85	N.A.	Yes	[98]	yes
KELM-CPPpred	Six feature descriptors, Kernel-ELM algorithm, 10-fold and jackknife CV	83.1%	0.67	N.A.	N.A.	[102]	yes
TargetCPP	Four feature descriptors, selection by mRMR, GBDT algorithm	88.5%	0.68	N.A.	N.A.	[93]	Currently unavailable

Predictor name – Name of the predictor given by authors.
Feature representation/selection/classifier – Features, approaches and algorithms used by authors to develop the models.
ACC – model prediction accuracies. Accuracies in the table are taken from the original articles as the authors declared and if possible from the results obtained from independent (validation) datasets were chosen. Note that comparisons between the performance of some methods not utilized in the original studies are sometimes biased due to the reasons discussed in [20].
The MCC – Matthews correlation coefficient represents the correlation between observed and predicted. When the prediction is perfect it is value is 1, if the coefficient is 0 the prediction is not better than random guess and −1 means total disagreement between observed and predicted.
Uptake (ACC) – Some models allow one to predict the uptake efficiency of predicted peptides. The numerical values represent the accuracy of this prediction.
Modifications. – Some models allow one to work with chemically modified peptides that contain non-coded peptides or other modifications in their sequences
Accessible – whether the model is accessible online to the general public

desired functions of phenotypes [121]. With slight modifications, it can be applied to find new CPPs, by starting with known sequences, and screening libraries orthogonally for members that have a set of desired properties. After each set of gain-of-function peptides is selected for the next round of screening, so the next generation of sequences will have a more refined set of properties that one is seeking [122]. It has been used to select of new pore-forming peptides [122], hybrid CPPs [123], and AMPs [124]. The main advantage of this approach is that after each round peptides are experimentally tested and the most efficient peptides are taken as input for next library screening round.

3.3.2. Protein mimicry

Protein mimicry is an approach where peptides with the function of the original protein are used to influence protein interactions by inhibition or activation. Proteins and their interactions are essential components of normal cellular processes, and aberrations in their interactions or functions may cause disease states. Modulation of cellular processes through influencing these protein–protein interactions has the potential to restore normal cellular functions [125], or to help investigate these processes in the cells. Designing new protein mimicry CPPs, is an attractive approach, because it addresses both internalization and bioactivity. Various tools can be used to aid in selecting possible protein–protein interaction peptides from different databases, such as BioGRID [126], and PinaColada [127], or PepCrawler [128]. Additional approaches are also used in CPP discovery, such as phage display, HTS, and rational design [129]. The peptide sequences can be designed into CPPs or be additionally screened for their ability to penetrate cell membranes. The term 'bioportide' was introduced to distinguish between CPPs and bioactive CPPs. For example, the peptides camptide, and nosangiotide have both CPP activity and bioactivity [130].

4. Conclusions

The discovery of new and efficient CPPs for research, therapeutic, and diagnostic applications remains a challenge. Over the years, several approaches have been developed and adopted from other fields to find new CPPs. According to the level of experience, input information, and desired output, one can choose between experimental, *in silico*, or combined approaches.

Experimental approaches require a less previous background in the CPP field, and lead to more success in its refined form. However, experimental approaches are laborious and require the synthesis and testing of each candidate. Nevertheless, direct experimentation enables one to collect new data that can help improve both rational design and *in silico* approaches. In addition, experimentation encourages CPP findings that are not limited to rules that apply only to known CPPs. New and efficient HTS methods would significantly contribute to the development of other approaches.

In silico approaches rely both on the previous knowledge and databases or libraries of CPPs. While several accessible prediction sites exist, they nevertheless fall short when more complicated modifications are included or several traits are simultaneously investigated. Display approaches encourage one to find new CPPs with higher specificity, but this is strongly depended on the quality of display.

Although several approaches are available, the discovery and design of new CPPs is complicated and requires integrating previous knowledge with novel propositions. Currently, no single approach can guarantee success, yet new CPPs are found each year. With an increase in the number of known CPPs, and the development of new approaches, it is likely that new guidelines for discovering novel CPP sequences will appear.

5. Expert opinion

Currently, there are no strict guidelines for the successful selection of sequences that can cross cell membranes. Nevertheless, substantial progress has been made in developing *in silico* approaches that predict potential sequences and these have been used in the search for new CPPs. Although the prediction algorithms are trained on known CPP sequences, and this may create a bias in selection, it enables one to scan through large libraries that would otherwise be unsurmountable for traditional wet-lab approaches. Today, the main approaches display varying degrees of success for CPP discovery. These include a) educated guess integrated with trial and error and rational design, b) prediction of new CPPs from protein sequences or peptides, c) peptide sequences found from HTS approaches and display methods.

Trial and error, although primitive in its nature, have still maintained its position for finding new CPP sequences. It is laborious and costly, and requires synthesis and experimental testing of each candidate; however, it enables one to collect background data that can be used to improve HTS approaches. It is possible to outsource peptide synthesis; therefore, this approach does not require specific equipment or facilities. Cell-penetration of the peptide may be registered as an anomaly in one experiment, and further investigation leads to the discovery of a new CPP. In other cases, based on its sequence, structure, or protein function, a part of a protein is synthesized and tested for internalization. Going one step further, in an attempt to find the underlying rules that define efficacy, rational design is applied. In rational design, the success rate strongly relies on the knowledge of the researcher. It requires a heightened awareness of how CPPs work, what components are necessary to add to the sequence, and how and where to introduce the modification. In rational design, the number of possible candidates is reduced, based on an educated guess or other indications. Often rational design is accompanied with trial and error, as the efficacy of the CPP relies not only on the primary structure, but also several chemical and physio-chemical properties. The charge, chirality, aromatic and hydrophobic content and their co-action are often unpredictable. *In silico* approaches help to reduce the number of unsuccessful candidates and screen through more peptide sequences before synthesis and testing. There are several prediction platforms accessible online and upon request for screening from the authors. The predictions use different algorithms, which also influences their accuracy. Predictions are trained on known CPP sequences, which consequently may introduce selection bias. The success of *in silico* methods is strongly connected with the experimental data that has been collected and the other content within the CPP databases. The prediction itself does not require thorough knowledge from the user, and the output is quite easy to comprehend. Display methods allow one to screen possible candidates and select the ones with desired traits; however, this approach has several limitations, such as size, and require experience in the display method itself. Phage display is especially advantageous when one needs to scout tissue specificity in addition to cell internalization. In addition to display methods, other HTS approaches have been developed. Nevertheless, they are not widely used, due to their unavailability, elaborate setup, or other special requirements. There are several approaches for discovering and designing new CPPs, however, there is a need to devise higher throughput approaches that are more user-friendly, versatile, and can account for the inclusion of modifications.

Although there are several *in silico* methods that have been developed, discovering new, fully unique CPPs often begins with educated guess accompanied with serendipitous observations from typically unrelated experiments. Wet lab testing of known CPPs, investigating their physiochemical properties and interactions is the basis upon which rational design is built. This, hopefully, will lead to new approaches that enable screening of peptides or protein sequences for specific and efficient CPPs for each unique application. Development of new, more efficient, and accurate HTS methods could significantly increase the versatility and applicability of CPPs both as delivery vectors and as direct therapeutics. Data collected by HTS approaches and improved databases could enable even more accurate and diverse predictions, with more refined screening. Today, more and more researchers use available prediction programs to screen through their potential sequences before testing. This is encouraged, but again, may restrain further discovery of fully unique CPP sequences that do not fall into the same categories as know CPPs. Today, finding sequences with internalization properties only, is not enough. The sequence itself should be able to target specific tissues (e.g. targeting) or modulate processes (e.g. protein mimicry) within the cells. The search for multifunctional CPPs even more strongly necessitates the development of diverse HTS approaches that can account for modifications.

Funding

The authors are supported by EU project [2014-2020.4.01.15-0013] from the European Regional Development Fund through the project Center of Excellence in Molecular Cell Engineering.

Declaration of interest

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

ORCID

Ülo Langel http://orcid.org/0000-0001-6107-0844

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Langel Ü, CPP. Cell-penetrating peptides. Singapore (Singapore): Springer Nature Singapore Pte Ltd; 2019.
2. Frankel AD, Pabo CO. Cellular uptake of the tat protein from human immunodeficiency virus. Cell. 1988 Dec;55(6):1189–1193.
3. Green M, Loewenstein PM. Autonomous functional domains of chemically synthesized human immunodeficiency virus tat trans-activator protein. Cell. 1988;55(6):1179–1188.
4. Joliot A, Pernelle C, Deagostini-Bazin H, et al. Antennapedia homeobox peptide regulates neural morphogenesis. Proc Natl Acad Sci U S A. 1991 Mar;88(5):1864–1868.
5. Derossi D, Joliot AH, Chassaing G, et al. The third helix of the Antennapedia homeodomain translocates through biological membranes. J Biol Chem. 1994 Apr;269(14):10444–10450.
6. Vivès E, Brodin P, Lebleu B. A truncated HIV-1 Tat protein basic domain rapidly translocates through the plasma membrane and accumulates in the cell nucleus. J Biol Chem. 1997 Jun;272(25):16010–16017.
7. Morris MC, Vidal P, Chaloin L, et al. A new peptide vector for efficient delivery of oligonucleotides into mammalian cells. Nucleic Acids Res. 1997 Jul;;25(14):2730–2736.
8. Wender PA, Mitchell DJ, Pattabiraman K, et al. The design, synthesis, and evaluation of molecules that enable or enhance cellular uptake: peptoid molecular transporters. Proc Natl Acad Sci U S A. 2000 Nov;;97(24):13003–13008.
9. Futaki S, Ohashi W, Suzuki T, et al. Stearylated arginine-rich peptides: a new class of transfection systems. Bioconjug Chem. 2001 Nov-Dec;12(6):1005–1011.
10. Pooga M, Soomets U, Hällbrink M, et al. Cell penetrating PNA constructs regulate galanin receptor levels and modify pain transmission in vivo. Nat Biotechnol. 1998 Sep;;16(9):857–861.
11. Agrawal P, Bhalla S, Usmani SS, et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res. 2016 Jan;44(D1): D1098–103.

•• **Database of known CPPs44D1D1098-D1103.**

12. Habault J, Poyet JL. Recent advances in cell penetrating peptide-based anticancer therapies. Molecules. 2019 Mar;24(5):5.
13. Olson SE, Jiang T, Aguilera TA, et al. Activatable cell penetrating peptides linked to nanoparticles as dual probes for in vivo fluorescence and MR imaging of proteases. Proc Nat Acad Sci Mar. 2010;107(9):4311–4316.
14. Böhmová E, Pola R, Pechar M, et al. Etrych T polymer cancerostatics containing cell-penetrating peptides: internalization efficacy depends on peptide type and spacer length. Pharmaceutics. 2020;12(1):59.
15. Golan M, Feinshtein V, David A. Conjugates of HA2 with octaarginine-grafted HPMA copolymer offer effective siRNA delivery and gene silencing in cancer cells. Eur J Pharm Biopharm. 2016; (109):103–112,
16. Bartlett RL, Sharma S, Panitch A. Cell-penetrating peptides released from thermosensitive nanoparticles suppress pro-inflammatory cytokine response by specifically targeting inflamed cartilage explants. Nanomedicine. 2013;9(3):419–427.
17. Zhu Y, Jiang Y, Meng F, et al. Highly efficacious and specific anti-glioma chemotherapy by tandem nanomicelles co-functionalized with brain tumor-targeting and cell-penetrating peptides. J Control Release. 2018;278(18–8):0168–3659.
18. Kanazawa T, Akiyama F, Kakizaki S, et al. Delivery of siRNA to the brain using a combination of nose-to-brain delivery and cell-penetrating peptide-modified nano-micelles. Biomaterials. 2013;34(36):9220–9226.
19. Kanazawa T, Morisaki K, Suzuki S, et al. Prolongation of life in rats with malignant glioma by intranasal siRNA/drug codelivery to the brain with cell-penetrating peptide-modified micelles. Mol Pharm. 2014;11(5):1471–1478.
20. Su R, Hu J, Zou Q, et al. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform. 2020 Mar;21(2):408–420.
21. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. NAR. 2019;47:D506–515.
22. Pirtskhalava M, Gabrielian A, Cruz P, et al. 2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. Nucl Acids Res. 2016;44(D1):D1104–D1112.
23. Kang X, Dong F, Shi C, et al. DRAMP 2.0, an updated data repository of antimicrobial peptides. Sci Data. 2019;6(148). DOI:10.1038/s41597-019-0154-y
24. Armenteros JJA, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37(420–423). DOI:10.1038/s41587-019-0036-z
25. Sharma S, Schiller MR. The carboxy-terminus, a key regulator of protein function. Crit Rev Biochem Mol Biol. 2019 04;54(2):85–102.
26. Barlowe C, Helenius A. Cargo capture and bulk flow in the early secretory pathway. Annu Rev Cell Dev Biol. 2016 10;32(1):197–222.
27. Zhang P, Moreno R, Lambert PF, et al. Cell-penetrating peptide inhibits retromer-mediated human papillomavirus trafficking during virus entry. Proc Natl Acad Sci U S A. 2020 03;117 (11):6121–6128.
28. Montrose K, Yang Y, Krissansen GW. X-pep, a novel cell-penetrating peptide motif derived from the hepatitis B virus. Biochem Biophys Res Commun. 2014 Oct;;453(1):64–68.
29. Ohgita T, Takechi-Haraya Y, Nadai R, et al. A novel amphipathic cell-penetrating peptide based on the N-terminal glycosaminoglycan binding region of human apolipoprotein E. Biochim Biophys Acta Biomembr. 2019 03;1861(3):541–549. .
30. Fazil MHUT, Chalasani MLS, Choong YK, et al. C-terminal peptide of TFPI-1 facilitates cytosolic delivery of nucleic acid cargo into mammalian cells. Biochim Biophys Acta Biomembr. 2020 Feb;1862 (2):183093.
31. Trenner A, Godau J, Sartori AA, et al. BRCA2-derived cell-penetrating peptide targets RAD51 function and confers hypersensitivity toward PARP inhibition. Mol Cancer Ther. 2018 07;17(7):1392–1404.
32. Zhang P, Monteiro da Silva G, Deatherage C, et al. Cell-penetrating peptide mediates intracellular membrane passage of human papillomavirus L2 protein to trigger retrograde trafficking. Cell. 2018 Sept;174(6):1465–1476. . e13.
33. Yu W, Zhan Y, Xue B, et al. Highly efficient cellular uptake of a cell-penetrating peptide (CPP) derived from the capsid protein of porcine circovirus type 2. J Biol Chem. 2018 09;293 (39):15221–15232.
34. Bonifaci N, Moroianu J, Radu A, et al. Karyopherin beta2 mediates nuclear import of a mRNA binding protein. Proc Natl Acad Sci U S A. 1997 May;94(10):5055–5060.
35. Bernhofer M, Goldberg T, Wolf S, et al. NLSdb-major update for database of nuclear localization signals and nuclear export signals. Nucleic Acids Res. 2018 *Database of NLS 01;46(D1):D503–D08.
36. Xiong K, Blainey PC. Molecular sled sequences are common in mammalian proteins. Nucleic Acids Res. 2016 Mar;;44(5):2266–2273.
37. Astrada S, Fernández Massó JR, Vallespí MG, et al. Cell penetrating capacity and internalization mechanisms used by the synthetic peptide CIGB-552 and its relationship with tumor cell line sensitivity. Molecules. 2018 Mar;23(4):4.
38. Henriques ST, Melo MN, Castanho MA. Cell-penetrating peptides and antimicrobial peptides: how different are they? Biochem J. 2006 Oct;399(1):1–7.

39. Bahnsen JS, Franzyk H, Sayers EJ, et al. Cell-penetrating antimicrobial peptides – prospectives for targeting intracellular infections. Pharm Res. 2015 May;;32(5):1546–1556.
40. Bustillo ME, Fischer AL, LaBouyer MA, et al. Modular analysis of hipposin, a histone-derived antimicrobial peptide consisting of membrane translocating and membrane permeabilizing fragments. Biochim Biophys Acta. 2014 Sep;;1838(9):2228–2233.
41. Neundorf I, Rennert R, Hoyer J, et al. Fusion of a short HA2-derived peptide sequence to cell-penetrating peptides improves cytosolic uptake, but enhances cytotoxic activity. Pharmaceuticals (Basel). 2009 Sep;2(2):49–65. .
42. Eiríksdóttir E, Konate K, Langel Ü, et al. Secondary structure of cell-penetrating peptides controls membrane interaction and insertion. Biochim Biophys Acta. 2010 Jun;;1798(6):1119–1128.
43. Wadia JS, Stan RV, Dowdy SF. Transducible TAT-HA fusogenic peptide enhances escape of TAT-fusion proteins after lipid raft macropinocytosis. Nat Med. 2004 Mar;;10(3):310–315.
44. Fang SL, Fan TC, Fu HW, et al. A novel cell-penetrating peptide derived from human eosinophil cationic protein. PLoS One. 2013;8(3):e57318. .
45. Chen JX, Xu XD, Yang S, et al. Self-assembled BolA-like amphiphilic peptides as viral-mimetic gene vectors for cancer cell targeted gene delivery. Macromol Biosci. 2013 Jan;;13(1):84–92.
46. Pastushok L, Fu Y, Lin L, et al. A novel cell-penetrating antibody fragment inhibits the DNA repair protein RAD51. Sci Rep. 2019 08;9 (1):11227.
47. Walker L, Perkins E, Kratz F, et al. Cell penetrating peptides fused to a thermally targeted biopolymer drug carrier improve the delivery and antitumor efficacy of an acid-sensitive doxorubicin derivative. Int J Pharm. 2012 Oct;;436(1–2):825–832.
48. Myrberg H, Zhang L, Mäe M, et al. Design of a tumor-homing cell-penetrating peptide. Bioconjug Chem. 2008 Jan;;19(1):70–75.
49. Ramaker K, Henkel M, Krause T, et al. Cell penetrating peptides: a comparative transport analysis for 474 sequence motifs. Drug Deliv. 2018;25(1):928–937.
50. Oikawa K, Islam MM, Horii Y, et al. Screening of a cell-penetrating peptide library in Escherichia coli: relationship between cell penetration efficiency and cytotoxicity. ACS Omega. 2018;3(12):16489–16499.
51. Jha D, Mishra R, Gottschalk S, et al. CyLoP-1: a novel cysteine-rich cell-penetrating peptide for cytosolic delivery of cargoes. Bioconjug Chem. 2011 Mar;22(3):319–328. .
52. Freimann K, Arukuusk P, Kurrikoff K, et al. Optimization of in vivo DNA delivery with NickFect peptide vectors. J Control Release. 2016 11;241:135–143.
53. Gessner I, Neundorf I. Nanoparticles modified with cell-penetrating peptides: conjugation mechanisms, physicochemical properties, and application in cancer diagnosis and therapy. Int J Mol Sci. 2020 Apr;21(7):7.
54. Soomets U, Lindgren M, Gallet X, et al. Deletion analogues of transportan. Biochim Biophys Acta. 2000 Jul;;1467(1):165–176. .
55. Soleymani-Goloujeh M, Nokhodchi A, Niazi M, et al. Effects of N-terminal and C-terminal modification on cytotoxicity and cellular uptake of amphiphilic cell penetrating peptides. Artif Cells Nanomed Biotechnol. 2018;46(sup1):91–103. .
56. Nguyen LT, Chau JK, Perry NA, et al. Serum stabilities of short tryptophan- and arginine-rich antimicrobial peptide analogs. PLoS One. 2010 Sep;5:9.
57. Park S, Kim M, Hong Y, et al. Myristoylated TMEM39AS41, a cell-permeable peptide, causes lung cancer cell death. Toxicol Res. 2020 Apr;36(2):123–130. .
58. Insua I, Montenegro J. 1D to 2D self assembly of cyclic peptides. J Am Chem Soc. 2020 01;142(1):300–307.
59. Khatri A, Mishra A, Chauhan VS. Characterization of DNA condensation by conformationally restricted dipeptides and gene delivery. J Biomed Nanotechnol. 2017 Jan;;13(1):35–53.
60. Zhang C, Ren W, Liu Q, et al. Transportan-derived cell-penetrating peptide delivers siRNA to inhibit replication of influenza virus in vivo. Drug Des Devel Ther. 2019;13:1059–1068.
61. Robbins PF, Kantor JA, Salgaller M, et al. Transduction and expression of the human carcinoembryonic antigen gene in a murine colon carcinoma cell line. Cancer Res. 1991 Jul;51(14):3657–3662.
62. Rydberg HA, Matson M, Amand HL, et al. Effects of tryptophan content and backbone spacing on the uptake efficiency of cell-penetrating peptides. Biochemistry. 2012 Jul;51(27):5531–5539.
63. Porosk L, Arukuusk P, Põhako K, et al. Enhancement of siRNA transfection by the optimization of fatty acid length and histidine content in the CPP. Biomater Sci. 2019 Sep;7(10):4363–4374.
64. Kurrikoff K, Veiman KL, Künnapuu K, et al. Effective in vivo gene delivery with reduced toxicity, achieved by charge and fatty acid - modified cell penetrating peptide. Sci Rep. 2017 12;7(1):17056. .
65. Pärnaste L, Arukuusk P, Langel K, et al. The formation of nanoparticles between small interfering RNA and amphipathic cell-penetrating peptides. Mol Ther Nucleic Acids. 2017 Jun;7:1–10.
66. Horton KL, Pereira MP, Stewart KM, et al. Tuning the activity of mitochondria-penetrating peptides for delivery or disruption. Chembiochem. 2012 Feb;;13(3):476–485.
67. Dominguez-Berrocal L, Cirri E, Zhang AL, et al. New therapeutic approach for targeting hippo signalling pathway. Sci Rep. 2019;9 (1):4771.
68. Saw PE, Song EW. Phage display screening of therapeutic peptide for cancer targeting and therapy. Protein Cell. 2019 11;10(11):787–807.
69. Zahid M, Feldman KS, Garcia-Borrero G, et al. Cardiac targeting peptide, a novel cardiac vector: studies in bio-distribution, imaging application, and mechanism of transduction. Biomolecules. 2018 11;8(4):4. .
70. McConnell SJ, Thon VJ, Spinella DG. Isolation of fibroblast growth factor receptor binding sequences using evolved phage display libraries. Comb Chem High Throughput Screen. 1999 Jun;2(3):155–163.
71. Nicklin SA, White SJ, Watkins SJ, et al. Selective targeting of gene transfer to vascular endothelial cells by use of peptides isolated by phage display. Circulation. 2000 Jul;102(2):231–237.
72. Mukai Y, Sugita T, Yamato T, et al. Creation of novel Protein Transduction Domain (PTD) mutants by a phage display-based high-throughput screening system. Biol Pharm Bull. 2006 Aug;29 (8):1570–1574. .
73. Mi Z, Mai J, Lu X, et al. Characterization of a class of cationic peptides able to facilitate efficient protein transduction in vitro and in vivo. Mol Ther. 2000 Oct;2(4):339–347.
74. Eriste E, Kurrikoff K, Suhorutšenko J, et al. Peptide-based glioma-targeted drug delivery vector gHoPe2. Bioconjug Chem. 2013 Mar;24(3):305–313.
75. Wu LP, Ahmadvand D, Su J, et al. Crossing the blood-brain-barrier with nanoligand drug carriers self-assembled from a phage display peptide. Nat Commun. 2019 Oct;10(1):4635.
76. Sugahara KN, Teesalu T, Karmali PP, et al. Tissue-penetrating delivery of compounds and nanoparticles into tumors. Cancer Cell. 2009 Dec;;16(6):510–520.
77. Hoffmann K, Milech N, Juraja SM, et al. A platform for discovery of functional cell-penetrating peptides for efficient multi-cargo intracellular delivery. Sci Rep. 2018 08;8(1):12538.
78. Watt PM, Milech N, Stone SR. Structure-diverse Phylomer libraries as a rich source of bioactive hits from phenotypic and target directed screens against intracellular proteins. Curr Opin Chem Biol. 2017 Jun;38:127–133.
79. Durzyńska J, Przysiecka Ł, Nawrot R, et al. Viral and other cell-penetrating peptides as vectors of therapeutic agents in medicine. J Pharmacol Exp Ther. 2015 Jul;354(1):32–42.
80. Blanco C, Verbanic S, Seelig B, et al. High throughput sequencing of in vitro selections of mRNA-displayed peptides: data analysis and applications. Phys Chem Chem Phys. 2020 Mar;22(12):6492–6506.
81. Kamide K, Nakakubo H, Uno S, et al. Isolation of novel cell-penetrating peptides from a random peptide library using in vitro virus and their modifications. Int J Mol Med. 2010 Jan;25(1):41–51.
82. Liu R, Barrick JE, Szostak JW, et al. Optimized synthesis of RNA-protein fusions for in vitro protein selection. Methods Enzymol. 2000;318:268–293.
83. Lee JH, Song HS, Park TH, et al. Screening of cell-penetrating peptides using mRNA display. Biotechnol J. 2012 Mar;7(3):387–396.
84. Kwon YU, Kodadek T. Quantitative evaluation of the relative cell permeability of peptoids and peptides. J Am Chem Soc. 2007 Feb;129(6):1508–1509.

85. Yu P, Liu B, Kodadek T. A convenient, high-throughput assay for measuring the relative cell permeability of synthetic compounds. Nat Protoc. 2007;2(1):23–30.
86. Carney RP, Thillier Y, Kiss Z, et al. Combinatorial library screening with liposomes for discovery of membrane active peptides. ACS Comb Sci. 2017 May;19(5):299–307.
87. Hemmati S, Behzadipour Y, Haddad M. Decoding the proteome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) for cell-penetrating peptides involved in pathogenesis or applicable as drug delivery vectors. Infect Genet Evol 2019. 2020;104474. DOI:10.1016/j.meegid.
88. Dobchev DA, Mager I, Tulp I, et al. Prediction of cell-penetrating peptides using artificial neural networks. Curr Comput Aided Drug Des. 2010;6(2):79–89.
89. Sanders WS, Johnston CI, Bridges SM, et al. Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol. 2011 Jul;;7(7):e1002101.
90. Gautam A, Chaudhary K, Kumar R, et al. In silico approaches for designing highly effective cell penetrating peptides. J Transl Med. 2013;11:74.
91. Diener C, Garza Ramos Martínez G, Moreno Blas D, et al. Effective design of multifunctional peptides by combining compatible functions. PLoS Comput Biol. 2016 Apr;12(4):e1004786.
92. Wei L, Xing P, Su R, et al. CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. J Proteome Res. 2017 05;16(5):2044–2053.
•• **Advanced prediction approach.**
93. Arif M, Ahmad S, Ali F, et al. TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree. J Comput Aided Mol Des. 2020 Aug;34(8):841–856.
•• **Advanced prediction of CPPs, currently unavailable**
94. Noble WS. What is a support vector machine? Nat Biotechnol. 2006 Dec;24(12):1565–1567.
95. Tang H, Su ZD, Wei HH, et al. Prediction of cell-penetrating peptides with feature selection techniques. Biochem Biophys Res Commun. 2016 Aug;477(1):150–154.
96. Breiman L. Random Forests. Mach Learn. 2001 Oct;45:5–32.
97. Wei L, Tang J, Zou Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genomics. 2017 Oct;18(Suppl 7):742.
98. Kumar V, Agrawal P, Kumar R, et al. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. Front Microbiol. 2018;9:725.
•• **Prediction approach containing chemically modified residues**
99. Chen L, Chu L, Huang T, et al. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. Amino Acids. 2015 Jul;47(7):1485–1493.
100. Manavalan B, Subramaniyam S, Shin TH, et al. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. J Proteome Res. 2018 08;17 (8):2715–2726.
•• **Advanced prediction approach.**
101. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006 Mar;63:3–42.
102. Pandey P, Patel V, George NV, et al. KELM-CPPpred: kernel extreme learning machine based prediction model for cell-penetrating peptides. J Proteome Res. 2018 17;9:3214–3222.
103. Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing. 2006 Dec;;70(1–3):489–501.
104. Chen T, Guestrin C. XGBoost: a scalable tree boosting system: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. August 2016.
105. Hällbrink M, Kilk K, Elmquist A, et al. Prediction of cell-penetrating peptides. Int J Pept Res Ther. 2005;11:249–259.
106. Hellberg S, Sjöström M, Skagerberg B, et al. Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem. 1987 Jul;;30(7):1126–1135.
107. Sandberg M, Eriksson L, Jonsson J, et al. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem. 1998 Jul;;41(14):2481–2491.
108. Langel Ü. Handbook of cell-penetrating peptides, 2nd ed. CRC Press Inc. 2006.
109. Hansen M, Kilk K, Langel Ü. Predicting cell-penetrating peptides. Adv Drug Deliv Rev. 2008 Mar;60(4–5):572–579.
110. Holton TA, Pollastri G, Shields DC, et al. CPPpred: prediction of cell penetrating peptides. Bioinformatics. 2013 Dec;29(23):3094–3096.
111. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins. 2001 May;43(3):246–255.
112. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005 Aug;27 (8):1226–1238.
113. Manavalan B, Lee J. SVMQA: support–vector-machine-based protein single-model quality assessment. Bioinformatics. 2017 Aug;33 (16):2496–2503.
114. Lin X, Li X. Lin XA review on applications of computational methods in drug screening and design. Molecules. 2020;25 (6):1375.
115. Marrink SJ, Corradi V, Souza PCT, et al. Computational modeling of realistic cell membranes. Chem Rev. 2019;119(9):6184–6226.
116. Cao Z, Liu L, Hu G, et al. Interplay of hydrophobic and hydrophilic interactions in sequence-dependent cell penetration of spontaneous membrane-translocating peptides revealed by bias-exchange metadynamics simulations. Biochimica Et Biophysica Acta (BBA) – Biomembranes. 2020;1862(10): 0005–2736. 183402.
117. Ulmschneider JP, Ulmschneider MB. Molecular Dynamics Simulations Are Redefining Our View Of Peptides Interacting With Biological Membranes. Acc Chem Res. 2018;51(5):1106–1116.
118. Allolio C, Magarkar A, Jurkiewicz P, et al. Arginine-rich cell-penetrating peptides induce membrane multilamellarity and subsequently enter via formation of a fusion pore. Proc Nat Acad Sci Nov. 2018;115(47):11923–11928.
119. Herce HD, Garcia AE. Molecular dynamics simulations suggest a mechanism for translocation of the HIV-1 TAT peptide across lipid membranes. Proc Nat Acad Sci Dec. 2007;104 (52):20805–20810.
120. Ma R, Wong SW, Ge L, et al. In vitro and MD simulation study to explore physicochemical parameters for antibacterial peptide to become potent anticancer peptide. Mol Ther Oncolyt. 2020;16:7–19.
121. Simon AJ, d'Oelsnitz S, Ellington AD. Synthetic evolution. Nat Biotechnol. 2019 Jul;37(7):730–743.
122. Li S, Kim SY, Pittman AE, et al. Potent macromolecule-sized poration of lipid bilayers by the macrolittins, a synthetically evolved family of pore-forming peptides. J Am Chem Soc. 2018 May;140 (20):6441–6447.
123. Kauffman WB, Guha S, Wimley WC. Synthetic molecular evolution of hybrid cell penetrating peptides. Nat Commun. 2018 Jul;9(1):2568.
124. Wimley WC. Application of synthetic molecular evolution to the discovery of antimicrobial peptides. Adv Exp Med Biol. 2019;1117:241–255.
125. Johansson HJ, Andaloussi SE, Langel Ü. Mimicry of protein function with cell-penetrating peptides. Methods Mol Biol. 2011;683:233-247.
126. Oughtred R, Stark C, Breitkreutz BJ, et al. The BioGRID interaction database: 2019 update. NAR. 2019, Jan;47(D1):D529–D541.
127. Zaidman D, Wolfson HJ. PinaColada: peptide-inhibitor ant colony ad-hoc design algorithm. Bioinformatics. 2016 Aug 1;32 (15):2289–2296.
128. Donsky E, Wolfson HJ. PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. Bioinformatics. 2011 Oct 15;27(20):2836–2842.
129. Cunninham AD, Qvit N, Mochly-Rosen D. Peptides and peptidomimetics as regulators of protein–protein interactions. Curr Opin Struct Biol. 2017 Jun;44:59–66.
130. Howl J, Matou-Nasri S, West DC, et al. Bioportide: an emergent concept of bioactive cell-penetrating peptides. Cell Mol Life Sci. 2012;69:2951–2966.