nature chemistry

ARTICLES

https://doi.org/10.1038/s41557-021-00766-3

Deep learning to design nuclear-targeting abiotic miniproteins

Carly K. Schissel[1,9], Somesh Mohapatra[2,9], Justin M. Wolfe[1,7], Colin M. Fadzen[1,8], Kamela Bellovoda[3], Chia-Ling Wu[3], Jenna A. Wood[3], Annika B. Malmberg[3], Andrei Loas[1], Rafael Gómez-Bombarelli[2]✉ and Bradley L. Pentelute[1,4,5,6]✉

There are more amino acid permutations within a 40-residue sequence than atoms on Earth. This vast chemical search space hinders the use of human learning to design functional polymers. Here we show how machine learning enables the de novo design of abiotic nuclear-targeting miniproteins to traffic antisense oligomers to the nucleus of cells. We combined high-throughput experimentation with a directed evolution-inspired deep-learning approach in which the molecular structures of natural and unnatural residues are represented as topological fingerprints. The model is able to predict activities beyond the training dataset, and simultaneously deciphers and visualizes sequence–activity predictions. The predicted miniproteins, termed 'Mach', reach an average mass of 10 kDa, are more effective than any previously known variant in cells and can also deliver proteins into the cytosol. The Mach miniproteins are non-toxic and efficiently deliver antisense cargo in mice. These results demonstrate that deep learning can decipher design principles to generate highly active biomolecules that are unlikely to be discovered by empirical approaches.

The vast chemical search space hinders the design of functional macromolecules by empirical approaches alone[1]. It is hypothesized that machine learning can enable interpolation in high-dimensional search spaces by bridging the gaps between experimental training data points[2,3]. Recent works have shown promise using a variety of input representations and quantitative activity predictions for the design of new antimicrobial peptides and antibody CDR3 loops[4–6]. For cell-penetrating peptides (CPPs), similar strategies that involve binary classifiers have been used to optimize activity[7–10]. We sought to further address this challenge by using a large standardized dataset and an advanced input representation combined with deep learning to simultaneously design new functional miniproteins and quantitatively predict their activity.

Successful design of functional polymers can have considerable implications for medicine. For example, anticancer miniproteins have been shown to access intracellular targets[11,12]. Similarly, CPPs are short (5–20 residue) sequences that can enhance the intracellular delivery of biomolecules, such as oligonucleotides and proteins, that otherwise cannot efficiently cross the cell membrane[13–17]. Although promising, variation in experimental design has resulted in inconsistent and sometimes contradictory datasets. For example, penetratin has different efficacies as a CPP, which depend on the assay and the cargo[18]. These inconsistent results preclude the development of sequence–activity relationships and complicate the use of machine-learning models to design analogues de novo[19–21].

We overcame these challenges by a de novo design of abiotic miniproteins that deliver an active cargo, antisense phosphorodiamidate morpholino oligomer (PMO), to the nucleus of cells. The miniproteins described here are distinct in that they have a defined function (PMO delivery) and are substantially longer (30–80 residues) than CPPs (5–20 residues). Although PMO has recently been approved for the treatment of Duchenne muscular dystrophy, a major challenge remains with the poor cellular permeability[13–17,22,23]. High doses of PMO of up to 50 mg kg^{-1} are required for in vivo efficacy[24]. It has been shown that nuclear delivery can be improved by attaching PMO to CPPs, and the clinical success of this strategy was demonstrated in 2020[25,26]. Development of advanced, novel sequences for antisense delivery would rapidly accelerate the development of these gene therapies.

Here we report a deep-learning-based design strategy with predictive power fuelled by robust input data that contains unnatural residues and structures. Our framework includes the generation of starting sequences, a predictor to predict the activity of a sequence and an optimizer to improve the activity of the sequence. A library that contained 600 unique antisense–miniprotein conjugates was constructed using linear combinations of three peptides, or 'modules' (Fig. 1a). A quantitative activity readout was achieved using an in vitro assay in which the nuclear delivery of PMO results in enhanced green fluorescent protein (EGFP) fluorescence (Fig. 1b,c). Residues were encoded as fingerprints to provide chemical structure information, labelled with the corresponding activity data and used to train a predictor neural network (Fig. 1d). A 'CPP thesaurus' dataset was used to train a generator neural network to produce novel sequences that are 'CPP-like' and to be used as seeds for optimization. These novel sequences were then optimized in the predictor–optimizer loop to increase the predicted activity, but also to minimize similarity to the library and to minimize length and Arg content to mitigate toxicity[27]. The output was hundreds of

[1]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Sarepta Therapeutics, Cambridge, MA, USA. [4]The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Present address: Ra Pharmaceuticals, Cambridge, MA, USA. [8]Present address: Harvard Medical School, Boston, MA, USA. [9]These authors contributed equally: Carly K. Schissel, Somesh Mohapatra. ✉e-mail: rafagb@mit.edu; blp@mit.edu

de novo designed sequences with a broad spectrum of predicted activity (Fig. 1e).

The model is also interpretable: we can visualize the decision-making process and identify structure–activity relationships that are consistent with empirical observations. From these predictions, we discovered best-in-class abiotic 'Mach' (machine learning) nuclear-targeting miniproteins that improve PMO delivery 50-fold and are effective in animals. Mach miniproteins are non-toxic and non-inflammatory, and are able to deliver macromolecules other than PMO to the cytosol. Our approach has the potential to be extended to the design of peptides with other functions, although further work is required in these directions.

Results

Assembly of a standardized dataset. Recently, we demonstrated that linear combinations of known CPP sequences into chimeric miniproteins can synergistically improve the delivery of PMO compared that of each CPP alone[28]. We hypothesized that expanding this approach to a larger, more diverse library of linear combinations of CPPs would access a wide range of sequences and activities. We designed a synthetic method to assemble this library via the bioconjugation of peptide 'modules' into hundreds of novel PMO–miniproteins. Our rationale is that such a library would enable a broad sequence diversity and spectrum of activities and would be ideal to train machine-learning models (Fig. 1).

Our synthesis strategy employs four modules: one for PMO and three for distinct pools of peptide sequences that contain diverse structure and function, which includes nuclear-targeting peptides and peptides containing unnatural residues and Cys-linked macrocycles (Supplementary Table 12)[29]. The constructs were synthesized in a series of bioconjugation reactions that are chemoselective and irreversible to yield products of sufficient crude purity for direct testing in vitro (Supplementary Fig. 18). The resulting library contained 600 miniproteins, composed of combinations of 57 peptides in total.

The resulting dataset was broad in terms of both peptide sequences and range of activity, quantified by a high-throughput nuclear-targeting assay[19]. The activity-based assay used to acquire the training data provides a direct, quantitative readout of the activity characteristic we want to enhance—specifically, nuclear delivery. In this assay, HeLa cells stably transfected with an EGFP gene interrupted by a mutated intron of β-globin (IVS2-654) produce a non-fluorescent EGFP protein. A successful delivery of PMO IVS2-654 to the nucleus results in corrective splicing and EGFP synthesis. The amount of PMO delivered to the nucleus therefore correlates with EGFP fluorescence, quantified by flow cytometry. Activity is reported as the mean fluorescence intensity (MFI) relative to PMO alone (Fig. 1c,e and Supplementary Information). The most active construct improved PMO delivery by nearly 20-fold, whereas the median activity was 3-fold.

Developing the deep-learning model. Inspired by directed evolution, we leveraged fingerprint sequence representations to develop a machine-learning-based generator–predictor–optimizer triad. In this framework, the generator produces novel cell-penetrating sequences, the predictor quantitatively estimates the activity for a given sequence and the optimizer evolves towards the most optimal miniprotein sequence.

The standardized dataset of activity-labelled sequences from the modular library allowed for the development and training of a quantitative regressor algorithm. This approach enabled us to overcome the limitations of other efforts in the computational CPP literature, which employed binary classifiers of active versus inactive sequences[2,3,8,30–32]. These previous predictors were mostly trained using physicochemical descriptors, with datasets obtained from non-standardized experiments and containing only natural residues[7,10,33,34]. The inclusion of chemically diverse unnatural moieties using this strategy is challenging because such physicochemical descriptors may not be readily available. The ability to predict the effect of unnatural residues expands the chemical search space, and may lead to an enhanced macromolecule delivery[35]. One-hot residue encodings can be extended to represent unnatural residues in the training data. We were interested, however, in encoding the molecular structure of each residue. Therefore, to predict activity of de novo-designed nuclear-targeting abiotic miniproteins, we evaluated a topological representation based on stacking traditional cheminformatics fingerprints for each residue along the sequence[36]. This representation extends the approach of using one-hot encodings for quantitative structure–activity relationship predictors in the peptide literature[4,5], provides chemical structure information for unnatural residues and may leverage weight sharing across structurally similar residues. Combined with quantitative experimental readouts, this polymer representation allows us to access the diverse pool of unnatural residues and structures and quantitatively predict delivery activity.

Peptide sequences are represented as matrices that comprise residue fingerprints in the columns, padded with zeros until each sequence matrix is the same length. Individual residue fingerprints are bit vectors based on the molecular graph of the whole monomer, including backbone and side chain. We used 2,048-bit ECFP6 fingerprints generated by RDKit (Fig. 2a and Supplementary Appendix 2), but other structural descriptors may be used[37]. For analysis and visualization of the fingerprints, we removed all the indices that were inactive across all the residues, which resulted in a condensed 191-bit fingerprint (Supplementary Appendix 2). Each bit in the vector corresponded to a substructure, and was active or inactive depending on the presence or absence of the particular substructure. Representing residues as chemical structures, rather than discrete choices, eases the use of both natural and unnatural residues and leverages chemical similarity between the residues. The fingerprints were then compiled into a row matrix to encode the amide backbone of the peptide sequence (Fig. 2b). This representation method is also more effective than typical one-hot encodings at using the inherent chemistry to predict novel sequences and is able to predict activities of sequences that contain a new residue not in the training set (Supplementary Table 8 and Supplementary Section 2.5).

The predictor neural network quantitatively estimated the normalized MFI for a given sequence. Pairs of sequence representations and corresponding experimental activities were used to train a convolutional neural network (CNN). The training dataset consisted of PMO–miniproteins from the modular library as well as other conjugates previously tested in the same assay[7]. A randomly selected 20% of the dataset was saved for validation of the predictive accuracy of the algorithm. The root mean squared error (RMSE) on the validation set was 0.4 of the standard deviation (s.d.) of the training data. The prediction relative error was found to be 11% as long as the predicted activity fell within the range of the training values (normalized activity of 0.32–19.5) (Fig. 2c). Tests were conducted against other model architectures using both fingerprint and one-hot encodings in both regression and classification tasks (Supplementary Tables 1–5 and Supplementary Figs. 2 and 3). We also explicitly tested whether the reported models (hosted as webservers) were able to predict the activity of the Mach miniproteins accurately (Supplementary Table 6). We observed that most of these models were limited by the range of the training data, and that only the CNN-fingerprint (CNN-FP) model was able to extrapolate in the codomain and generate predicted activity values (validated by experimental activity values) that were greater than any in the training set. This ability to extrapolate, however, came at a cost in the average accuracy because of the increased statistical noise of the extrapolated predictions. Models based on the topological representations

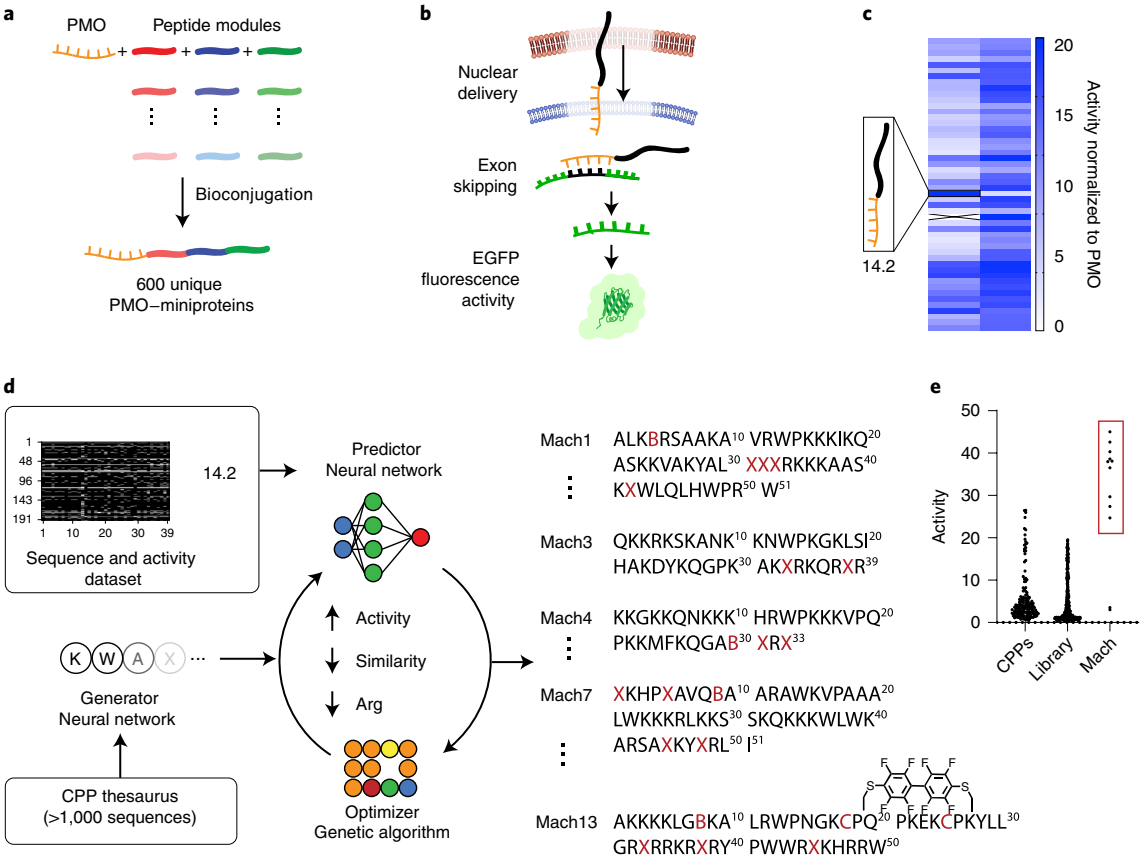

Fig. 1 | Machine-learning model based on directed evolution predicts highly active abiotic miniproteins for macromolecule delivery. a, A 600-membered modular library of PMO–miniprotein conjugates was synthesized using linear combinations of abiotic peptide modules. **b**, Standardized in vitro quantitative activity assay tests for nuclear delivery using a quantitative fluorescence readout. **c**, Members of the modular library exhibit a broad spectrum of activities. Each bit corresponds to a PMO–peptide in the library and its corresponding activity. **d**, Sequences are encoded into a fingerprint matrix, labelled with experimental activity, and used to train a machine-learning model. The model designs novel sequences in a loop based on directed evolution. Unnatural residues include Ahx (X), β-alanine (B), and Cys (C) macrocycles linked through decafluorobiphenyl. **e**, Normalized activity from peptides designed in this work (Mach) compared with that of the peptides in the modular library and known CPPs tested using the same assay.

added only a minimal increase in performance over that of one-hot encodings on the validation dataset, and performed similarly or worse on the Mach dataset, due to outliers with extreme predicted activity values (Supplementary Table 2). However, a CNN model using one-hot encodings, despite its lowest overall average error, was not able to extrapolate in the codomain space, unlike fingerprint representations. To investigate the role of outliers that impact the model performance, we used model ensembling and found that the ensembled CNN one-hot model is superior for the validation dataset, whereas the ensembled CNN-FP model is superior for the Mach dataset, probably due to its ability to extrapolate in the codomain (Supplementary Table 3). Further efforts should be focused on how to accurately predict activity values that reach beyond that of the training set. We investigated the CNN model's ability to extrapolate from the training dataset and found that experimental activity above a threshold of ~8 is necessary to accurately predict peptides with activity beyond that of the training dataset (Supplementary Fig. 4 and Supplementary Table 7). Finally, inclusion of unnatural amino acids was required for high-activity predictions, as predicting the canonical sequences using the same model resulted in a notable drop in predicted activity (Supplementary Fig. 6).

We developed a generator based on a recurrent neural network that captured the ontology of the CPPs and generated 'CPP-like' starter sequences. We trained the generator using a nested long short-term memory (LSTM) neural network architecture, which is better able to capture long-range correlations in sequence data[38]. We trained the algorithm using a 'CPP thesaurus', a collection of sequences from both our modular library and the literature[39]. As the model is learning sequence grammar and has no role in activity predictions, no quantitative labels are necessary and we can use a large dataset of available sequences. Other strategies to generate seed sequences also resulted in predicted peptides with a high predicted activity. For example, the top 50 performers from the PMO–CPP library resulted in the highest predicted activity values. However, we confirmed that the generator approach led to predicted sequences that better met our three criteria simultaneously (high predicted activity, low similarity and low Arg content) than other methods of generating seed sequences (Supplementary Fig. 5 and Supplementary Table 9). It is possible for the other methods of seed selection and optimization to also produce optimal peptide sequences, but experimental validation is required to adequately compare these methods.

The optimizer completed the loop based on directed evolution. Sequences from the generator were randomly mutated and evaluated against an objective function, which maximized activity as predicted by the CNN model and minimized length, Arg content and similarity to the library while retaining water solubility, estimated with the net charge of the sequence (Supplementary Table 10).

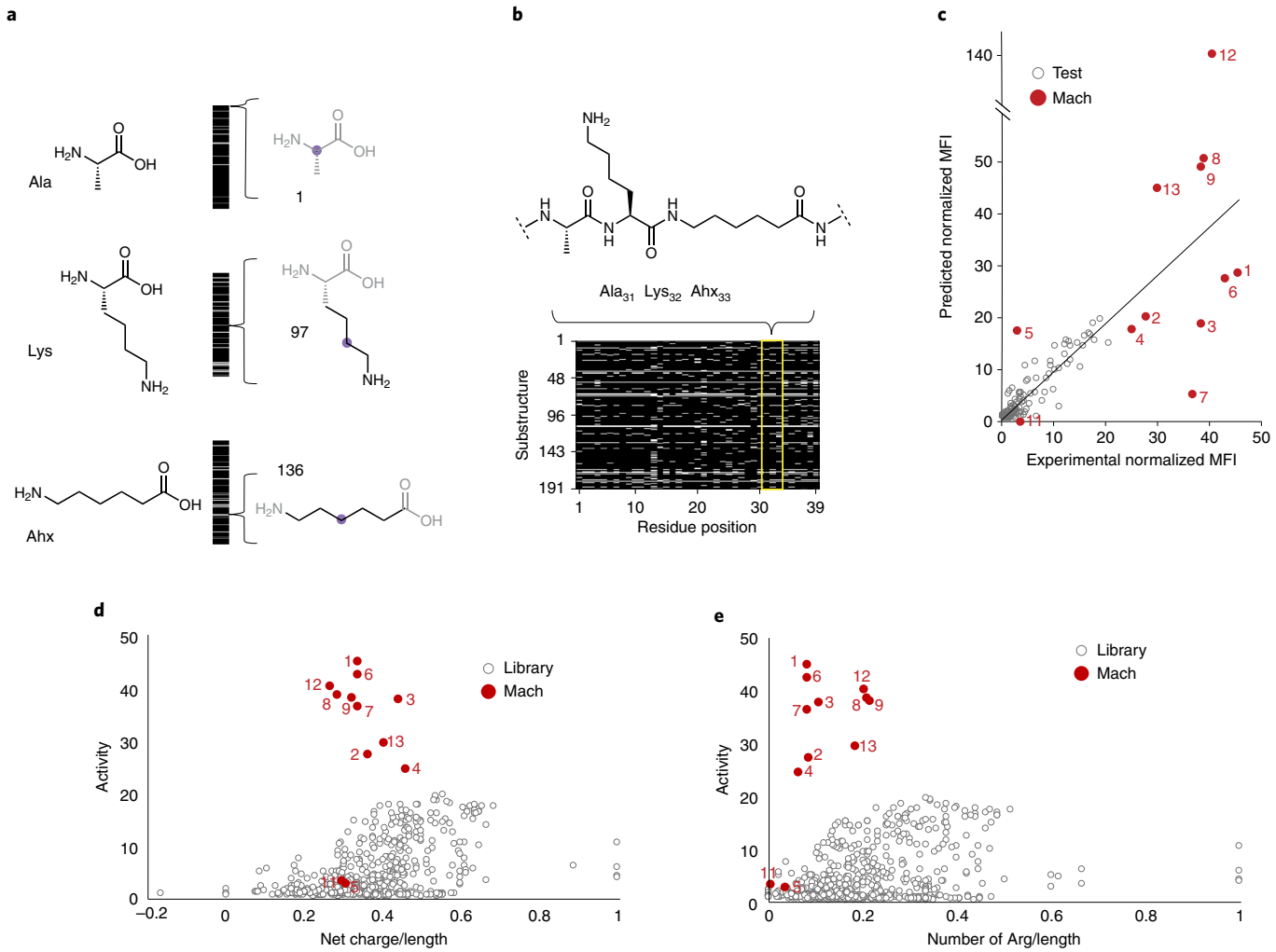

Fig. 2 | Machine-learning-based generator–predictor–optimizer loop predicts nuclear-targeting abiotic miniproteins. a, Each amino acid residue is represented as a unique fingerprint, constructed as a bit vector that encodes for the presence or absence of 191 possible substructures in the residue. **b**, Sequences are represented as residue fingerprints stacked in a row matrix. **c**, Comparison of the predicted and experimental activity values for the holdout test set and novel Mach sequences shows the performance of the machine-learning model. **d**,**e**, Of the predicted Mach peptides, 12 were synthesized and tested in the same activity assay and compared with the modular library in relation to relative charge (**d**) and Arg content (**e**).

After 1,000 iterations over each sequence, the model delivered hundreds of unique sequences with a wide range of predicted activity values. Along with highly active sequences, we predicted inactive sequences as negative controls. By directing the evolution of the optimizer in the opposite direction, that is, minimizing MFI, but keeping other constraints the same, we were able to generate an inactive sequence (Mach11) that appeared similar in amino acid composition to the active predictions. After synthesis, the Mach11 conjugate displayed a low experimental activity, which demonstrated the robustness of the model in predicting the activity of a unique sequence (Fig. 2c).

Interpreting the predictor model. We interpreted the predictor CNN by visualizing the residue substructures that are important in its decision-making process. This type of visualization was a long-standing attribution challenge that was recently addressed for image classification and more recently for small-molecule design[40–42]. We developed an analogous tool to correlate the input sequence representation with the predicted activity. This process generated bit-wise positive and negative activation values for each chemical substructure in the sequence. Bits with a higher activation indicated the features that most strongly influence the final activity prediction.

As an example, for the predicted Mach3 sequence the two C-terminal aminohexanoic acid (Ahx) residues were the most positively activated (Fig. 3a), followed by Arg. The alkyl backbone in Ahx was the most activated substructure (Fig. 3b). A similar trend was observed for active sequences and substructures in the training dataset (Supplementary Figs. 7 and 8).

We used this visualization approach to better understand how the trained model designed sequences. We chose five random sequences of different lengths, seeded them in the predictor–optimizer loop to maximize the activity contingent on other design constraints and visualized the activations for the best predictions. Again, a higher activation can be seen for C-terminal residues (Fig. 3c), most probably due to the attachment of PMO to the N terminus. We also observed that the general composition of the charged and hydrophobic residues remained unchanged across different sequence lengths (Fig. 3d). Particular residue fingerprints were activated irrespective of the sequence length, such as the side chains of Lys, Ser and Asp (Fig. 3e,f). Consistent with previous observations, a strong preference for polar and charged side chains as well as for Ahx was evident. We investigated whether the attribution feature is useful towards post hoc mutations to Mach miniproteins, and found a substantial boost in activity when mutating Ahx (6-carbon

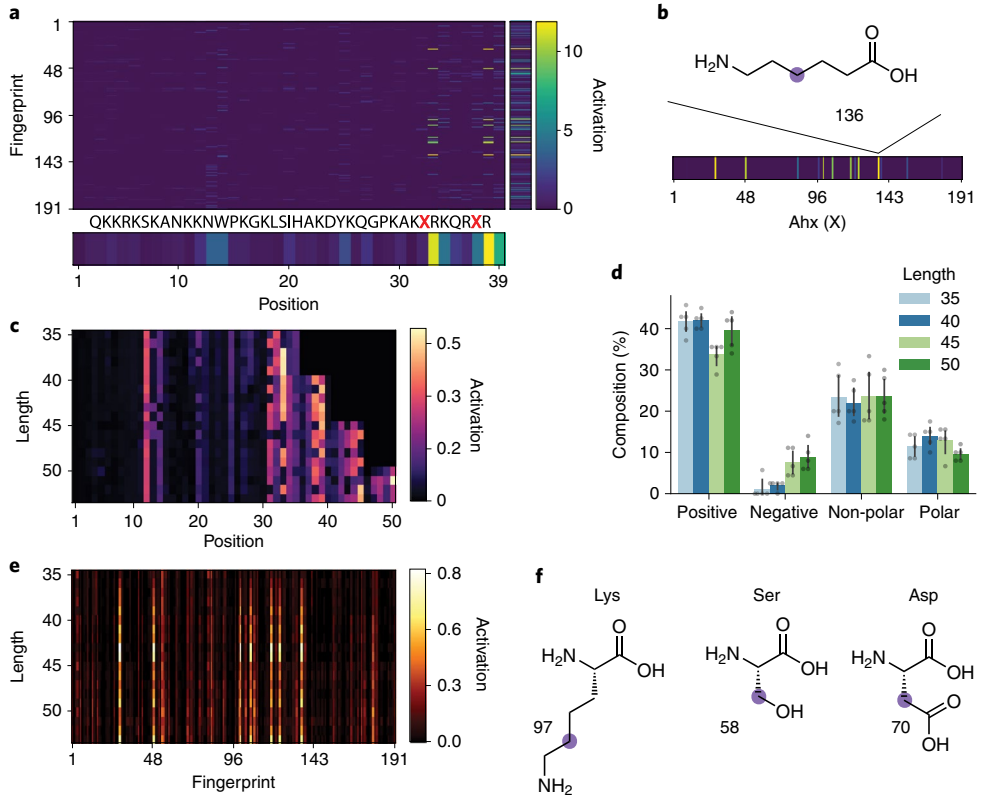

Fig. 3 | Interpretation of predictor CNN unveils activated substructures. a, The CNN positive activation gradient map was calculated for the input sequence representation of Mach3. The averaged activation values over fingerprint indices and residue positions are shown. The fingerprint index represents a corresponding substructure. **b**, The activation gradient map of Ahx in Mach3 indicates the activated substructures of this residue. The alkyl backbone substructure (136) is shown. **c**, Gradient maps of predicted sequences with lengths 35, 40, 45 and 50 are shown relative to the residue position. **d**, The percentage compositions of each type of residue (positive, negative, non-polar and polar) relative to the predicted sequences with lengths 35, 40, 45 and 50 are shown. Each bar represents the group mean ± s.d., $n = 5$. **e**, Gradient maps of predicted sequences with lengths 35, 40, 45 and 50 are shown relative to the substructure fingerprints. **f**, Several residues and substructures that are consistently activated across all the sequence lengths are shown, which include the amine side chain of Lys, the polar side chain of Ser and the carboxylic acid side chain of Asp.

chain) to aminoundecanoic acid (11-carbon chain) in Mach3 (Supplementary Fig. 16).

Mach miniproteins enhance PMO delivery. We synthesized and characterized 12 candidates from hundreds of miniproteins predicted by the model, selecting diverse sequences and predicted activities. Mach1, 2 and 6 were selected because they had a high predicted activity among 50-mer sequences. Mach3 was selected as a mid-length peptide (39 residues), Mach4 was selected as a shorter sequence (33 residues) with only two Arg residues and Mach5 was selected because it was predicted to have moderate activity along with the lowest net charge (10.5). Mach7 was initially designed to be a negative control—in which the sequence of Mach1 was rearranged until the model predicted the lowest activity. Mach8 and 9 were selected from a list of much longer miniproteins (around 80 residues) and Mach12 and 13 were selected from sequences that contained Cys-linked macrocycles. Finally, Mach11 was selected from a list of sequences for which the activity was optimized in the negative direction, to show that the algorithm could predict peptides of similar length, charge and amino acid composition, but with no PMO delivery activity. Each candidate was synthesized using automated fast-flow solid-phase peptide synthesis and, when applicable, the two Cys residues were connected with decafluorobiphenyl, as previously reported (Supplementary Fig. 1)[19,43]. The conjugation of azido-Mach to PMO IVS2-654 was achieved in the same manner as in the library. The final PMO–Mach constructs are described in Supplementary Table 13.

Nearly all the sequences predicted to have an activity greater than 20-fold did, indeed, surpass the highest-performing modular library construct, with the exception of Mach5. As the model was extrapolating outside the range of the training data, the predicted and experimental activity of PMO–Mach constructs shows a greater percentage error than that of the test dataset (Fig. 2c). The PMO–Mach constructs were first tested for PMO delivery in the HeLa 654 assay, as was done with the library (Supplementary Fig. 20).

The physicochemical properties of the validated predictions showed little correlation with PMO activity. We compared the activities of Mach constructs to those in the training library in relation to various physicochemical properties (Fig. 2d,e). Although library constructs clearly show an increase in activity with an increase in Arg content relative to length, and of net charge relative to length, there is no obvious correlation between the activity of Mach constructs and these same properties. In addition, truncated versions of PMO–Mach constructs do not retain the activity of the parent sequences (Supplementary Fig. 17). These observations suggest that the model is taking advantage of sequence–activity relationships that go beyond sequence length and charge.

Several PMO–Mach constructs have a greater potency than previously characterized PMO–CPPs, and also remain non-toxic. This type of macromolecular delivery is a historic challenge as it

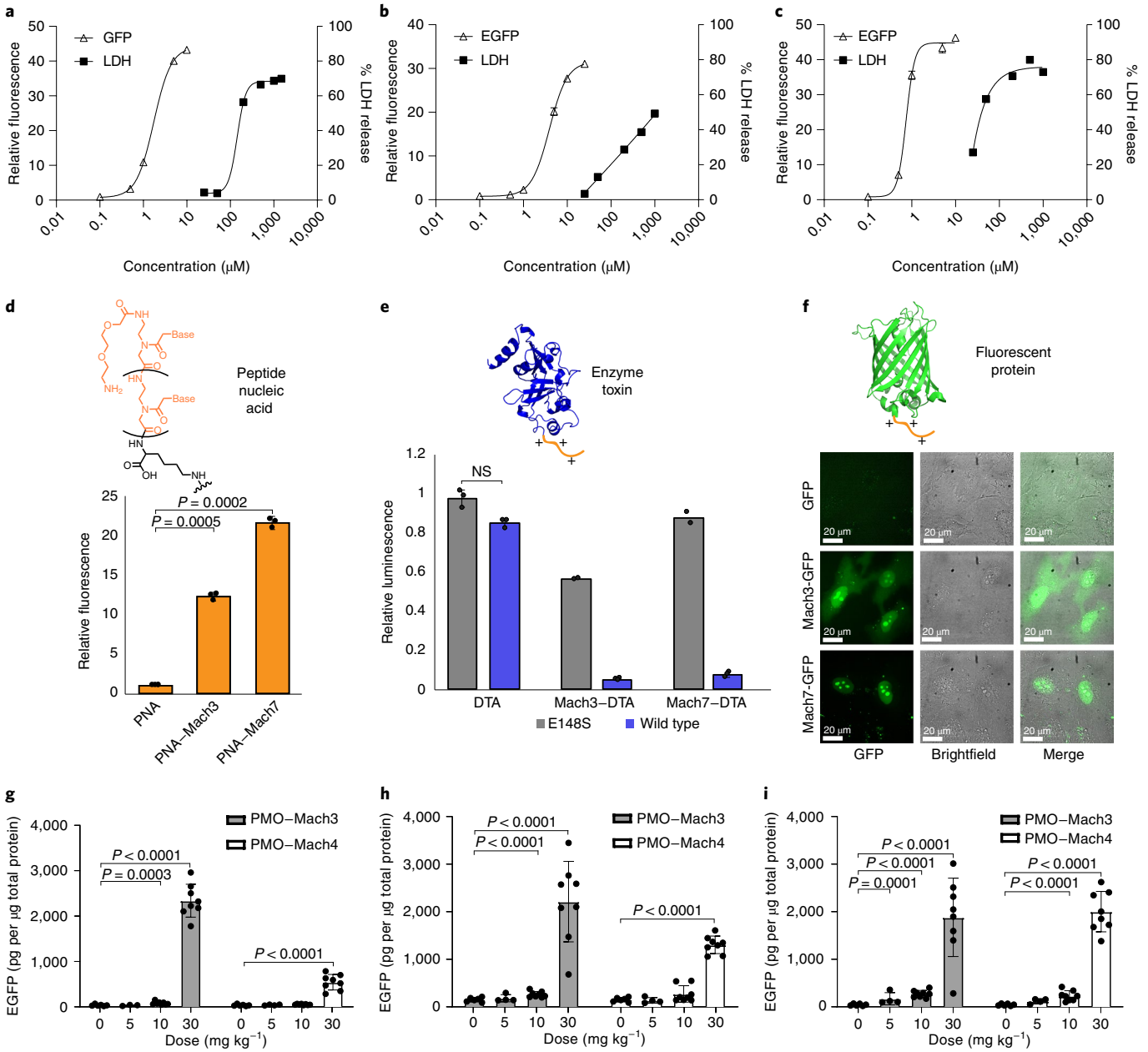

Fig. 4 | Mach miniproteins are highly active in vitro and in vivo and deliver other biomacromolecules into the cytosol. a–c, Shown are dose–response curves corresponding to activity in the EGFP assay and toxicity in the LDH assay for PMO–Mach3 (**a**), 4 (**b**) and 7 (**c**). Activity is shown as the fluorescence intensity relative to that of the unconjugated PMO at 5 μM, and toxicity is shown as the LDH release relative to a lysis control. The symbols show the mean ± s.d. for the EGFP assay, $n = 3$ distinct samples, and the mean for the LDH assay, $n = 2$ distinct samples. EGFP assay experiments were repeated at different concentration ranges with similar results, reported in Supplementary Fig. 21. **d**, The relative fluorescences of Mach3 and 7 conjugated to PNA 654 are compared with that of PNA alone, as determined by EGFP assay. Each bar represents the group mean ± s.d., $n = 3$ distinct samples. *P* values calculated from a two-tailed Student's *t*-test (PNA–Mach3, $P = 0.0005$; PNA–Mach7, $P = 0.0002$). **e**, Comparison of the toxicity of wild-type and inactive mutant DTA and DTA(E148S) alone or conjugated to Mach3 or 7. Delivery of the active toxin to the cytosol results in toxicity as measured by luminescence. Each bar represents the group mean ± s.d., $n = 3$ distinct samples, except for Mach3–DTA(E148S) and Mach7–DTA(E148S), which show the mean, $n = 2$ distinct samples. Full concentration curves are reported in Supplementary Fig. 25. **f**, Confocal micrographs displaying green fluorescence produced by EGFP, Mach3-EGFP or Mach7-EGFP in HeLa cells after 3 h of incubation at 10 μM. This experiment was conducted twice independently with similar results. **g–i**, EGFP synthesis in EGFP transgene mice after treatment with PMO–Mach: dose–response EGFP protein levels in quadriceps (**g**), diaphragm (**h**) and heart (**i**). Saline ($n = 6$ mice), Mach3 and Mach4 at 5 mg kg^{-1} ($n = 4$) and for all the others, $n = 8$ mice. Each bar represents group mean ± s.d. *P* values calculated from the two-tailed Mann–Whitney *U* test. Quadriceps: PMO–Mach3 10 mg kg^{-1}, $P = 0.0003$; PMO–Mach3, 30 mg kg^{-1}, $P < 0.0001$; PMO–Mach4 30 mg kg^{-1}, $P < 0.0001$. Diaphragm: PMO–Mach3 10 mg kg^{-1}, $P < 0.0001$; PMO–Mach3 30 mg kg^{-1}, $P < 0.0001$; PMO–Mach4 30 mg kg^{-1}, $P < 0.0001$. Heart: PMO–Mach3 5 mg kg^{-1}, $P = 0.0001$; PMO–Mach3 10 mg kg^{-1}, $P < 0.0001$; PMO–Mach3 30 mg kg^{-1}, $P < 0.0001$; PMO–Mach4 10 mg kg^{-1}, $P < 0.0001$; PMO–Mach4 30 mg kg^{-1}, $P < 0.0001$. GFP, green fluorescent protein; NS, not significant.

often suffered from either membrane toxicity or endosomal entrapment. We first verified that PMO–Mach constructs enter cells via an energy-dependent uptake using a panel of chemical endocytosis inhibitors and the HeLa 654 assay (Supplementary Fig. 13). We then performed dose–response experiments to characterize the activity in an EGFP assay and the toxicity in a lactate dehydrogenase (LDH) release assay. PMO–Mach2, 3, 4 and 7 each had an EC_{50} (half-maximum effective dose) value near 1 μM and were non-toxic at the concentrations tested, as determined by viability staining with propidium iodide and an LDH release assay (Fig. 4a–c and Supplementary Figs. 21 and 22). Extending toxicity tests to higher concentrations in renal cells showed that no toxicity was observed at the highest concentration needed for the maximum PMO activity in HeLa 654 cells (Supplementary Fig. 23). We compared these results with those of a previously well-performing CPP for PMO delivery, Bpep-Bpep[28]. This peptide has a similar activity, but is composed of mostly Arg residues and exhibits cytotoxicity above 10 μM (Supplementary Fig. 22). This contrast between Mach peptides and Bpep-Bpep indicates that there is no apparent direct connection between the toxicity and cargo delivery efficacy. PMO–Mach constructs have high activity, low Arg content and a wide therapeutic window, which highlights their suitability for cytosolic and nuclear delivery.

Mach miniproteins deliver other biomacromolecules. Mach miniproteins are versatile in that they can deliver other large biomolecules to the cytosol. Peptide nucleic acid (PNA) is a class of synthetic antisense oligonucleotides that has the same mechanism of action as PMO, but also has a highly flexible backbone structure[44]. We tested for the delivery of a PNA variant of PMO 654 that is compatible with the EGFP assay. Each of the four Mach miniproteins tested was able to significantly enhance PNA delivery (Fig. 4d and Supplementary Fig. 24).

In addition to antisense oligonucleotides, Mach peptides can also deliver charged proteins, such as diphtheria toxin A (DTA). DTA is a 21 kDa anionic protein segment that contains the catalytic domain of the toxin but lacks the portions that endow cell entry[45]. Delivery of this enzyme can be monitored using a cell proliferation assay as it inhibits protein synthesis in the cytosol. We found that Mach–DTA constructs were delivered into the cell cytosol significantly more efficiently than protein alone, and that covalent linkage was required for delivery (Fig. 4e and Supplementary Fig. 25). Furthermore, we confirmed that toxicity is due to the cytosolic delivery of the active DTA by comparing the wild-type constructs with those that contained DTA(E148S), a mutant with a 300-fold lower activity that of the wild type[46]. As expected, the mutant DTA conjugates led to a substantially reduced toxicity.

Conjugation to Mach miniproteins also improved the delivery of EGFP, a fluorescent protein commonly used as a reporter. Confocal micrographs of HeLa cells displayed diffuse green fluorescence in the cytosol and intense fluorescence in the nucleus after incubation with Mach–EGFP (Fig. 4f). This observation is in contrast with the EGFP alone condition, in which no diffuse fluorescence was observed in either location, which indicates a reduced uptake.

PMO–Mach restore protein synthesis in mice. After verifying the Mach miniproteins' propensity for in vitro macromolecule delivery, we looked towards in vivo antisense applications. In vitro tests with human macrophages suggested that the constructs are not inflammatory and therefore may be safe to evaluate in animals (Supplementary Fig. 15). Existing predictive models also suggest that Mach sequences would not be T-cell epitopes (Supplementary Fig. 12).

Lastly, we demonstrated that PMO–Mach constructs safely correct protein synthesis in animals. Transgenic mice that contained the same EGFP IVS2-654 gene as used in cell assays were given a single intravenous injection of varying doses of PMO–Mach3 or PMO–Mach4 and evaluated after seven days. Both constructs exhibited a dose-dependent increase in EGFP expression in the quadriceps, diaphragm and heart (Fig. 4g–i). PMO delivery to the heart is a critical but challenging objective. Here we observed similar levels of protein synthesis in both skeletal and cardiac tissue. In addition, there were no significant changes in the level of renal function biomarkers seven-days post-treatment (Supplementary Fig. 26). These findings indicate that Mach miniproteins may be safe delivery materials for PMO to muscle tissues.

Discussion

We demonstrated a method to efficiently sample the vast chemical search space of functional peptides using machine learning and standardized experimentation. Our model was applied to the design of abiotic miniproteins that can deliver an antisense PMO to the nucleus with a very high efficiency for a polypeptide-based variant. Importantly, the new constructs are effective in animals and are non-toxic up to a dose of 30 mg kg^{-1}. These miniproteins are versatile intracellular carriers and can deliver other classes of biomolecules, such as antisense PNA, fluorescent protein and enzymes, to the nucleus and cytosol. The core strengths of our model lie in: (1) standardized quantitative activity data, (2) the model's ability to extrapolate beyond the training set and (3) a visual attribution tool to interpret the decision-making process of the model.

A critical factor in building a robust machine-learning model is the training dataset; the 600-member library was synthesized by combining peptide modules and tested in a standardized assay that provides quantitative activity information. Synthesis and testing of the modular PMO–CPP library produced a broad spectrum of sequence and activity data with which we trained the model. By representing peptide sequences as topological fingerprints rather than categorical choices or descriptors, such as molecular weight, charge and hydrophobicity, the model has access to inherent structural information and can be used on monomers not encountered in the training data. The standardized activity values allowed us to use a quantitative regressor, rather than an active/inactive classifier, and thus design sequences with a broad spectrum of activity predictions. Although we previously tested CPPs designed by other machine-learning methods, we found that they were not able to deliver PMO[7]. The CNN model that uses fingerprints was able to extrapolate predicted activity beyond that of the training dataset, whereas models that used other frameworks and representations were not. Although the other models and methods to generate seed sequences may be able to produce sequences with a high experimental activity, the ability to predict that high activity is critical for the informed selection of predicted sequences to validate. As our goal is to discover unique peptides with a very high activity, a model able to predict values outside the range of the training data is required, which thus necessitates the use of CNN with fingerprint representations.

The interpretability of the model is an additional advantage. By overlaying the output of the predictor with the sequence matrix of a given peptide, we can visualize the activated residues and substructures important for the decision-making process. Several observations from the interpretations match our current understanding of CPP motifs, such as the benefit of cationic residues. The model also identified Ahx as an important residue, one which has only been investigated in the context of endosomal escape in Arg-rich sequences[47]. This tool allowed for post hoc analysis to validate empirical hypotheses and enhance the activity of Mach3 by mutating Ahx to aminoundecanoic acid, an amino acid not present in the training dataset.

In addition to PMO, Mach peptides deliver other antisense oligonucleotides as well as functional proteins into the cell cytosol. Delivery of EGFP reveals diffuse green fluorescence in the cytosol and a clear accumulation of EGFP to the nucleus. We believe that

Mach peptides may contain nuclear localization sequences (NLS), which have been described previously and are typically Lys rich[48]. However, nuclear localization is not solely due to the cationic charge, as shown by a previous study[49]. The model likely selected for such nuclear localization sequences because the activity used in the training was acquired from a nuclear delivery-based assay.

PMO–Mach conjugates effected a dose-dependent increase in protein synthesis in all three examined mouse muscle tissues, which included the heart after a single intravenous injection. The mouse model used contains the same transgene as that in the in vitro assay, and the Mach sequences recapitulated the in vitro results in vivo, which indicates that the model implemented here could be applied to data obtained from animal experiments. If a sequence–activity training set were generated from data obtained in animals, then this model may be applicable further downstream in the drug-design pipeline. A greater challenge remains towards the in vivo delivery to target tissues. In Duchenne muscular dystrophy, PMO must access the nucleus of muscle cells to have a therapeutic effect. Targeting to cardiac tissue is a primary concern given that the leading cause of death from this disease is heart failure. Our animal model confirmed localization of the PMO–Mach constructs to the heart, which suggests a potential solution to the tissue-targeting challenge.

In conclusion, this strategy illustrates how deep learning can be applied to the de novo design of functional abiotic miniproteins. The Mach miniproteins are the most effective PMO delivery constructs developed to date and are effective in animals. Our machine-learning framework could potentially be repurposed to discover sequence-optimized peptides with other desired activities, solely requiring a standardized high-quality input dataset. We envision that this strategy will enable the rapid future design of de novo functional peptides with impacts on chemical, biological and material sciences.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-021-00766-3.

Received: 26 June 2020; Accepted: 5 July 2021;
Published online: 09 August 2021

References

1. Lemonick, S. Exploring chemical space: can AI take us where no human has gone before? *Chemical & Engineering News* (6 April 2020); https://cen.acs.org/physical-chemistry/computational-chemistry/Exploring-chemical-space-AI-take/98/i13
2. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
3. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).
4. Spänig, S. & Heider, D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Min.* **12**, 7 (2019).
5. Witten, J. & Witten, Z. Deep learning regression model for antimicrobial peptide design. Preprint at *bioRxiv* https://doi.org/10.1101/692681 (2019).
6. Liu, G. et al. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126–2133 (2020).
7. Wolfe, J. M. et al. Machine learning to predict cell-penetrating peptides for antisense delivery. *ACS Cent. Sci.* **4**, 512–520 (2018).
8. Su, R., Hu, J., Zou, Q., Manavalan, B. & Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief Bioinform.* **21**, 408–420 (2020).
9. Sanders, W. S., Johnston, C. I., Bridges, S. M., Burgess, S. C. & Willeford, K. O. Prediction of cell penetrating peptides by support vector machines. *PLoS Comput. Biol.* **7**, e1002101 (2011).
10. Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O. & Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* **17**, 2715–2726 (2018).
11. Crook, Z. R., Nairn, N. W. & Olson, J. M. Miniproteins as a powerful modality in drug development. *Trends Biochem. Sci.* **45**, 332–346 (2020).
12. Beaulieu, M.-E. et al. Intrinsic cell-penetrating activity propels Omomyc from proof of concept to viable anti-MYC therapy. *Sci. Transl. Med.* **11**, eaar5012 (2019).
13. Juliano, R. L. The delivery of therapeutic oligonucleotides. *Nucleic Acids Res.* **44**, 6518–6548 (2016).
14. Slastnikova, T. A., Ulasov, A. V., Rosenkranz, A. A. & Sobolev, A. S. Targeted intracellular delivery of antibodies: the state of the art. *Front. Pharmacol.* **9**, 1208 (2018).
15. Miersch, S. & Sidhu, S. S. Intracellular targeting with engineered proteins. *F1000Research* **5**, 1947 (2016).
16. Trenevska, I., Li, D. & Banham, A. H. Therapeutic antibodies against intracellular tumor antigens. *Front. Immunol.* **8**, 1001 (2017).
17. Fu, A., Tang, R., Hardie, J., Farkas, M. E. & Rotello, V. M. Promises and pitfalls of intracellular delivery of proteins. *Bioconjug. Chem.* **25**, 1602–1608 (2014).
18. Illien, F. et al. Quantitative fluorescence spectroscopy and flow cytometry analyses of cell-penetrating peptides internalization pathways: optimization, pitfalls, comparison with mass spectrometry quantification. *Sci. Rep.* **6**, 36938 (2016).
19. Wolfe, J. M. et al. Perfluoroaryl bicyclic cell-penetrating peptides for delivery of antisense oligonucleotides. *Angew. Chem.* **130**, 4846–4849 (2018).
20. Betts, C. et al. Pip6-PMO, a new generation of peptide–oligonucleotide conjugates with improved cardiac exon skipping activity for DMD treatment. *Mol. Ther. Nucleic Acids* **1**, e38 (2012).
21. Boisguérin, P. et al. Delivery of therapeutic oligonucleotides with cell penetrating peptides. *Adv. Drug Deliv. Rev.* **87**, 52–67 (2015).
22. Chery, J. RNA therapeutics: RNAi and antisense mechanisms and clinical applications. *Postdoc J.* **4**, 35–50 (2016).
23. Mendell, J. R. et al. Eteplirsen for the treatment of Duchenne muscular dystrophy. *Ann. Neurol.* **74**, 637–647 (2013).
24. Moulton, J. & Jiang, S. Gene knockdowns in adult animals: PPMOs and vivo-morpholinos. *Molecules* **14**, 1304–1323 (2009).
25. McClorey, G. & Banerjee, S. Cell-penetrating peptides to enhance delivery of oligonucleotide-based therapeutics. *Biomedicines* **6**, 51 (2018).
26. Sarepta Therapeutics announces positive clinical results from MOMENTUM, a Phase 2 clinical trial of SRP-5051 in patients with Duchenne muscular dystrophy amenable to skipping exon 51. *GlobeNewswire News Room* http://www.globenewswire.com/news-release/2020/12/07/2140613/0/en/Sarepta-Therapeutics-Announces-Positive-Clinical-Results-from-MOMENTUM-a-Phase-2-Clinical-Trial-of-SRP-5051-in-Patients-with-Duchenne-Muscular-Dystrophy-Amenable-to-Skipping-Exon-5.html (2020)
27. Cardozo, A. K. et al. Cell-permeable peptides induce dose- and length-dependent cytotoxic effects. *Biochim. Biophys. Acta* **1768**, 2222–2234 (2007).
28. Fadzen, C. M. et al. Chimeras of cell-penetrating peptides demonstrate synergistic improvement in antisense efficacy. *Biochemistry* **58**, 3980–3989 (2019).
29. Wolfe, J. *Peptide Conjugation to Enhance Oligonucleotide Delivery* PhD thesis (MIT, 2018).
30. Wei, L., Tang, J. & Zou, Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics* **18**, 742 (2017).
31. Pandey, P., Patel, V., George, N. V. & Mallajosyula, S. S. KELM-CPPpred: kernel extreme learning machine based prediction model for cell-penetrating peptides. *J. Proteome Res.* **17**, 3214–3222 (2018).
32. Chen, B. et al. Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* **37**, 1332–1343 (2019).
33. Lee, E. Y., Wong, G. C. L. & Ferguson, A. L. Machine learning-enabled discovery and design of membrane-active peptides. *Bioorg. Med. Chem.* **26**, 2708–2718 (2018).
34. Dobchev, D. A. et al. Prediction of cell-penetrating peptides using artificial neural networks. *Curr. Comput. Aided Drug Des.* **6**, 79–89 (2010).
35. Jearawiriyapaisarn, N. et al. Sustained dystrophin expression induced by peptide-conjugated morpholino oligomers in the muscles of mdx mice. *Mol. Ther.* **16**, 1624–1629 (2008).
36. Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113 (1965).
37. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
38. Moniz, J. R. A. & Krueger, D. Nested LSTMs. *Proc. Mach. Learn. Res.* **77**, 530–544 (2017).
39. Agrawal, P. et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res.* **44**, D1098–D1103 (2015).
40. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Proc. IEEE Int. Conf. Comput. Vis.* 618–626 (IEEE, 2017); https://doi.org/10.1109/ICCV.2017.74
41. McCloskey, K., Taly, A., Monti, F., Brenner, M. P. & Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl Acad. Sci. USA* **116**, 11624–11629 (2019).

42. Sanchez-Lengeling, B. et al. Machine learning for scent: learning generalizable perceptual representations of small molecules. Preprint at https://arxiv.org/abs/1910.10685 (2019).
43. Hartrampf, N. et al. Synthesis of proteins by automated flow chemistry. *Science* **368**, 980–987 (2020).
44. Hanvey, J. C. et al. Antisense and antigene properties of peptide nucleic acids. *Science* **258**, 1481–1485 (1992).
45. Choe, S. et al. The crystal structure of diphtheria toxin. *Nature* **357**, 216–222 (1992).
46. Wilson, B. A., Reich, K. A., Weinstein, B. R. & Collier, R. J. Active-site mutations of diphtheria toxin: effects of replacing glutamic acid-148 with aspartic acid, glutamine, or serine. *Biochemistry* **29**, 8643–8651 (1990).
47. Abes, S. et al. Vectorization of morpholino oligomers by the (R–Ahx–R)4 peptide allows efficient splicing correction in the absence of endosomolytic agents. *J. Control. Release* **116**, 304–313 (2006).
48. Cerrato, C. P., Künnapuu, K. & Langel, Ü. Cell-penetrating peptides with intracellular organelle targeting. *Expert Opin. Drug Deliv.* **14**, 245–255 (2017).
49. Nischan, N. et al. Covalent attachment of cyclic TAT peptides to GFP results in protein delivery into live cells with immediate bioavailability. *Angew. Chem. Int. Ed.* **54**, 1950–1953 (2015).

Methods

Peptide synthesis. All the peptides and miniproteins were synthesized by automated solid-phase peptide synthesis, as previously described[43,50]. If the predicted sequence contained a Cys macrocycle, we subsequently utilized S_NAr (nucleophilic aromatic substitution) chemistry to link the two Cys residues with decafluorobiphenyl before additional purification. PMO was acquired from Sarepta and functionalized with a DBCO acid handle. A complete description of the synthesis, purification and characterization protocols can be found in the Supplementary Information.

Mach miniproteins were conjugated to PMO via strain-promoted azide–alkyne cycloaddition. PMO–DBCO (5 mM in water) was stoichiometrically combined with the azide–peptide (5 mM in water) and incubated at room temperature until the reaction completed (between 2 and 12 h), monitored by liquid chromatography–mass spectrometry (LC–MS). The reaction was purified using reversed-phase HPLC (Agilent Zorbax SB C3 column, 21.2 × 100 mm, 5 µm) and a linear gradient from 2 to 60% B (solvent A, 100 mM ammonium acetate in water, pH 7.2; solvent B, acetonitrile) over 58 min (1% B min^{-1}). Pure fractions were pooled as determined by LCMS and lyophilized.

PNA 654 (O-GCTATTACCTTAACCCAG-Lys(DBCO)) (50 nmol) was purchased from PNABio. PNA–DBCO (1 mM in water) was stoichiometrically combined with the azide–peptide (1 mM in water) and incubated at 4 °C for 12 h. The product was then used in cell assays without purification. Conversion was checked by LCMS.

Library synthesis. The library was synthesized in a combinatorial fashion and analysed by LC–MS[51]. The 600-member library was synthesized using 50 peptide members in module 4 (Supplementary Table 12).

Reaction 1. PMO–DBCO was dissolved in water to a 10 mM concentration (determined by ultraviolet–visible spectroscopy). The two peptides of the module were dissolved in water that contained 0.1% trifluoroacetic acid at 10 mM concentration (determined gravimetrically; the molecular mass was calculated to include 0.5 trifluoroacetate counter ions per Lys, Arg and His residue). In a microcentrifuge tube, 50 µl each of the PMO–DBCO solution and module two-peptide solution were mixed and incubated for 1 h. The product was analysed by LC–MS and dried by lyophilization. Lastly, the product was resuspended in 100 µl of dimethylsulfoxide (DMSO) to provide a 5 mM solution and stored at −20 °C.

Reaction 2. Stock solutions were prepared by dissolving module 3 peptides and module 4 peptides in water at 10 mM concentration (determined gravimetrically). For each reaction, 4 µl of the module 3 peptide was mixed with 4 µl of the module 4 peptide in a PCR tube. Separately, the copper bromide solution was prepared by mixing 1 ml of degassed DMSO with 2.8 mg of copper(I) bromide under N_2 to afford a 20 mM solution. Under ambient conditions, 4 µl of the CuBr solution was added to the mixture of module peptides 3 and 4. The reaction was capped and allowed to proceed for 2 h; the small amount of O_2 present during reaction set-up does not substantially impede the reaction progress. After 2 h, 2 µl of a 100 mM solution of Na_2HPO_4 was added. The PCR tube was then sonicated, vortexed and centrifuged. To remove the solvent, the PCR tube was centrifuged under vacuum using a Savant SPD121P Speed-Vac set at 35 °C for 2 h. Last, the product was resuspended in 16 µl of DMSO to provide a 5 mM solution and stored at −80 °C. The product was analysed by LC–MS.

Reaction 3. The final modular construct was synthesized through the combination of module 1–2 and module 3–4. First, 1.6 µl of the reaction 2 solution was added to a 384-well plate. Separately, 30 µl of the reaction 1 solution was mixed with 15 µl of TCEP (tris(2-carboxyethyl)phosphine) solution (100 mM TCEP·HCl in 50/50 water/DMSO that contained 400 mM NaOH) and 75 µl of DMSO. Then, 1.6 µl of the reaction 1 solution was added to the reaction 2 solution in the 384-well plate. Each individual reaction ultimately contained 0.4 µl of the reaction 1 solution (at 5 mM in DMSO), 1.6 µl of the reaction 2 solution (at 5 mM in DMSO), 0.2 µl of the TCEP solution (at 100 mM in water/DMSO) and 1 µl DMSO. Excess reaction 2 solution was used to force the reaction to go to completion; the presence of copper hinders the efficiency of this conjugation. The reaction 1 solution was used as a limiting reagent to avoid excess PMO, which is the active component for the cell-culture assays. The reaction was allowed to proceed for 2 h, and then the plate was stored at −80 °C. The reaction was analysed by LC–MS.

EGFP assay. HeLa 654 cells obtained from the University of North Carolina Tissue Culture Core facility were maintained in MEM supplemented with 10% (v/v) fetal bovine serum (FBS) and 1% (v/v) penicillin–streptomycin at 37 °C and 5% CO_2. The cells were plated at a density of 5,000 cells per well in a 96-well plate in MEM supplemented with 10% FBS and 1% penicillin–streptomycin 18 h prior to treatment.

To test of the library on the day of the experiment, the 384-well plate that contained the crude reaction mixtures in DMSO was diluted to 100 µM by the addition of 16.8 µl of PBS to 3.2 µl of the reaction mixture. Then, each construct was diluted to 5 µM in MEM supplemented with 10% FBS and 1% penicillin–streptomycin. For individual peptide testing, PMO–peptides were dissolved in PBS without Ca^{2+} or Mg^{2+} at a concentration of 1 mM (determined by ultraviolet spectroscopy) before being diluted in MEM. Cells were incubated at the designated concentrations for 22 h at 37 °C and 5% CO_2. Next, the treatment media was removed, and the cells were washed once before being incubated with 0.25% trypsin–EDTA for 15 min at 37 °C and 5% CO_2. Lifted cells were transferred to a V-bottomed 96-well plate and washed once with PBS, before being resuspended in PBS that containing 2% FBS and 2 µg ml^{-1} propidium iodide. Flow cytometry analysis was carried out on a BD LSRII flow cytometer at the Koch Institute. Gates were applied to the data to ensure that cells that were positive for propidium iodide or had forward/side scatter readings that were sufficiently different from those of the main cell population were excluded. Each sample was capped at 5,000 gated events (Supplementary Appendix 3).

Analysis was conducted using Graphpad Prism 7 and FlowJo. For each sample, the MFI and the number of gated cells was measured. To report activity, triplicate MFI values were averaged and normalized to the PMO-alone condition.

Inverse design model. Generator recurrent neural network. The generator is a data-driven tool to generate new peptide sequences that follow the ontology of cell-penetrating peptides to seed the optimization from likely starting points, and is based on a recurrent neural network–nested LSTM architecture[38]. It was trained using one-hot encoding representations of the amino acids to predict the next amino acid in the sequence from the preceding sequence. The inputs were size 5 to 50 amino acids, left-padded with zeros and represented termination with a unique token. The training dataset comprised 1,150 sequences and a total of 19,800 sequence-next character pairs, which included the non-modular sequences used in the creation of the library and sequences from CPPSite2.0[39]. The training was performed using 80% of this dataset, and validated using the remaining 20%. A validation accuracy of 76% was obtained in the training. For the model, multiple combinations of LSTM and nested LSTM layers were tried with different cell sizes[38]. The final model was chosen after the optimization of hyperparameters. All the hyperparameters were optimized using SigOpt (https://sigopt.com/).

Predictor–convolutional neural network. The predictor, based on the CNN, estimates the normalized fluorescence intensity from PMO delivery by a given peptide sequence, as measured in the HeLa 654 assay. The model was trained on a row matrix of residue fingerprints. The row matrix of 2,048-bit vectors (vector of 0s and 1s) represents the arrangement of the residues along the backbone of the peptide chain. This representation is analogous to a 1D image with 2,048 colour channels. Fingerprints have a radius of 3 atoms from the node atom and were generated using RDKit (http://www.rdkit.org/). By combining the CPP library from this work as well as the collection of CPPs from previous work, we compiled 640 PMO–peptide sequences for training[7].

We used fingerprints and one-hot encodings to train non-CNN models, such as those based on support vector regression, Gaussian process regression, kernel ridge regression, k-nearest neighbours regression and XGBoost regression.

Optimizer. The optimization was done using a genetic algorithm, in which single residue mutations involved insertion, deletion and swapping, and multiresidue mutation was done using hybridization. For hybridization, the sequence length and position to be hybridized, and the hybridized sequence (from the list of all the CPPs) were all chosen randomly. In the case of hybridization mutation, the selection and replacement of motifs was done at random without conservation of the sequence length. For the case of mutations with Cys macrocycles, explicit conditions were built in to keep the number and position of Cys residues separate in the case of a single through-space covalent bond or bicycle. A constrained hybridization condition that conserved the sequence length was also set up for specific optimization tasks. In the case of cysteine macrocycles, different fingerprints were used to denote the residues. The genetic algorithm used the following objective function (where R_{count} denotes the number of arginine residues in the sequence), starting from LSTM-generated sequences and taking 1,000 evolution steps:

$$\text{GA score} = \frac{1}{2}\text{intensity} - \frac{1}{2}\left(\frac{1}{2}R_{\text{count}} + \frac{1}{5}\text{length} - \frac{1}{10}\text{net charge} + \text{similarity}\right)$$

Set-up of generator–predictor–optimizer loop. The generator was primed with a 5-long random sequence from the training dataset and sampled until a termination character was produced. The randomly sampled sequences were then set up for optimization. The directed evolution of the generated sequences was carried using the predictor–optimizer feedback loop. Each sequence was mutated by the optimizer. Post mutations, the normalized fluorescence values for the new sequence were predicted by the predictor and the optimization parameters (similarity, % Arg, length and net charge) were calculated. The objective function (the equation with optimization parameters) was evaluated for both the old and mutated sequences. If the value for the mutated sequence was higher for the mutated than the older, then the old sequence was replaced by the mutated sequence. For each sequence, 1,000 such optimization rounds were conducted. The output was hundreds of sequences with varying predicted activity.

Toxicity assays. Cytotoxicity assays were performed in both HeLa 654 cells and human RPTEC (renal proximal tubule epithelial cells, TH-1, ECH001 and Kerafast). RPTEC were maintained in high-glucose DMEM supplemented with 10% (v/v) FBS and 1% (v/v) penicillin–streptomycin at 37 °C and 5% CO_2. Treatment of RPTEC was performed as with the HeLa 654 cells. After treatment, the supernatant was transferred to a new 96-well plate. To each well of the 96-well plate that contained supernatant, described above, was added CytoTox 96 Reagent (Promega). The plate was shielded from light and incubated at room temperature for 30 min. An equal volume of Stop Solution was added to each well, mixed and the absorbance of each well was measured at 490 nm. The blank measurement was subtracted from each measurement, and the % LDH release was calculated as % cytotoxicity = 100 × experimental LDH release (OD490/maximum LDH release (OD490).

Synthesis and testing of Mach–DTA. Mach–LPSTGG peptides were synthesized and purified by a standard protocol as described. G_5-DTA (50 μM) was incubated with either Mach3–LPSTGG (250 μM) or Mach7–LPSTGG (750 μM) and SrtA* (2.5 μM) for 90 min at 4 °C in SrtA buffer (10 mM $CaCl_2$, 50 mM Tris, 150 mM NaCl, pH 7.5). The reaction was monitored by LC–MS and gel electrophoresis. After 90 min, the Mach–DTA conjugate was isolated using a HiLoad 26/600 Superdex 200 prep grade size exclusion chromatography column (GE Healthcare) in a 20 mM Tris, 150 mM NaCl, pH 7.5 buffer. Fractions that contained the pure product as determined by LC–MS and gel electrophoresis were concentrated using a centrifugal filter unit (10K, Millipore).

To test for DTA delivery to the cytosol, HeLa cells were plated at 5,000 cells per well in a 96-well plate the day before the experiment. Wild-type and mutant constructs of G_5–DTA, Mach3–DTA, and Mach7–DTA, as well as Mach3–LPSTGG and Mach7–LPSTGG, were prepared at varying concentrations in complete media and transferred to the plate. Cell proliferation was measured after 48 h using the CellTiter-Glo assay.

Synthesis and testing of Mach–EGFP. G_5–EGFP (60 μM) was incubated with either Mach3–LPSTGG (1,000 μM) or Mach7–LPSTGG (1,000 μM) and SrtA* (5 μM) in SrtA buffer (10 mM $CaCl_2$, 50 mM Tris, 150 mM NaCl, pH 7.5) for 90 min at room temperature under the exclusion of light. The reaction was monitored by LC–MS and gel electrophoresis. After 90 min, the Mach–EGFP conjugate was isolated by cation exchange chromatography using a HiTrap SP HP cation exchange chromatography column (GE Healthcare) with 0–100% B over 20 column volumes, in which A is 50 mM NaCl, 20 mM Tris, pH 7.5 buffer and B is 1 M NaCl, 20 mM Tris, pH 12 buffer. Fractions that contained the pure product as determined by LC–MS and gel electrophoresis were immediately desalted and concentrated using a centrifugal filter unit (10K, Millipore).

To visualize the delivery of EGFP into cells, HeLa cells were plated at 5,000 cells per well in a coverslip glass-bottomed 96-well plate the day before the experiment. Mach3–EGFP, Mach7–EGFP or EGFP were added to each well at 10 μM and incubated at 37 °C and 5% CO_2 for 3 h. Treatment media was replaced with fresh media 1 h before being imaged in the W.M. Keck microscopy facility on an RPI spinning disk confocal microscope on a brightfield and green fluorescent protein setting (488 nm, 150 mW OPSL excitation laser, 525/50 nm emission).

In vivo studies. EGFP-654 transgenic mice (FVB/NJ mice transformed with CX-EGFP-654 plasmid) obtained from R. Kole's lab[52] ubiquitously express the EGFP-654 transgene throughout the body under chicken β-actin promoter. Identical to the HeLa 654 cell line, a mutated nucleotide 654 at intron 2 of the human β-globin gene interrupts the EGFP-654 coding sequence and prevents a proper translation of the EGFP protein. The antisense activity of PMO blocks aberrant splicing and results in EGFP expression, the same as in the HeLa 654 assay. In this study, 6- to 8-week-old male EGFP-654 mice bred at Charles River Laboratory were shipped to the vivarium at Sarepta Therapeutics. These mice were group housed with ad libitum access to food and water. All the animal protocols were approved by and conducted in accordance with the Institutional Animal Care and Use Committee of Sarepta Therapeutics.

After 3 days of acclimation, the mice were randomized into groups to receive a single intravenous tail vein injection of either saline or PMO–peptide (PMO–Mach3 or PMO–Mach4) at the indicated doses—5, 10 and 30 mg kg^{-1}. The mice were euthanized 7 days after the injection for the serum and tissue sample collection. The quadriceps, diaphragm and heart were rapidly dissected, snap-frozen in liquid nitrogen and stored at −80 °C until analysis.

Serums from all the groups were collected 7 days post-injection and tested for kidney injury markers using a Vet Axcel Clinical Chemistry System (Alfa Wassermann Diagnostic Technologies, LLC.) Specifically, serum BUN, creatinine and cystatin C levels were measured using ACE creatinine reagent (Alfa Wassermann, catalogue no. SA1012), ACE blood urea nitrogen reagent (Alfa Wassermann, catalogue no. SA2024) and Diazyme cystatin C immunoassay (Diazyme Laboratories, catalogue no. DX133C-K), respectively, as per the manufacturer's recommendations.

Mouse tissue (20–25 mg) was homogenized in RIPA buffer (Thermo Fisher, catalogue no. 89900) with protease inhibitor cocktail (Roche, 04693124001) using a Fast Prep 24-5G instrument (MP Biomedical). Homogenates were centrifuged at 12,000*g* for 10 min at 4 °C. The resultant supernatant lysates were quantified by a Pierce BCA Protein Assay Kit (Thermo Fisher, catalogue no. 23225) and saved for the EGFP expression measurement. Specifically, 80 μg of lysates were aliquoted in each well in a black-wall clear-bottom 96-well microplate (Corning). The EGFP fluorescent intensity of each sample was measured in duplicate using a SpectraMAx i3x microplate reader (Molecular Devices) by default setting. The average EGFP fluorescent intensity of each sample was then plotted against a standard curve constructed by recombinant EGFP protein (Origen, catalogue no. TP790050) to quantify the EGFP protein level per μg of protein lysate.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The main data supporting the findings of the current study are available within the paper and its Supplementary Information, which provides additional methods information, supplementary figures and data. Supplementary Table 1 includes sequences and activity of the modular library. Data used for training of the model is available at https://github.com/learningmatter-mit/peptimizer, and archived in the Zenodo repository[53]. Source data are provided with this paper.

Code availability

All the code used for model training and analysis is available at https://github.com/learningmatter-mit/peptimizer, and archived in Zenodo repository at https://zenodo.org/record/4815385#.YK_VCjZKhhE. Tutorial Jupyter notebooks are also in the repository, and demo Google Colab notebooks can be found at github.com/pikulsomesh/tutorials.

References

50. Mijalis, A. J. et al. A fully automated flow-based approach for accelerated peptide synthesis. *Nat. Chem. Biol.* **13**, 464–466 (2017).
51. Wolfe, J. M. *Peptide Conjugation to Enhance Oligonucleotide Delivery* (Massachusetts Institute of Technology, 2018).
52. Sazani, P. et al. Systemically delivered antisense oligomers upregulate gene expression in mouse tissues. *Nat. Biotechnol.* **20**, 1228–1233 (2002).
53. Mohapatra, S. learningmatter-mit/peptimizer: initial release. *Zenodo* https://doi.org/10.5281/zenodo.4815385 (2021).

Acknowledgements

We thank A. R. Loftis and J. Rodriguez for assistance with recombinant protein expression, C. Backlund for assistance with immunoassays, W. C. Salmon at the W. M. Keck Microscopy Facility at the Whitehead Institute for help with imaging, the Swanson Biotechnology Center Flow Cytometry Facility at the Koch Institute for the use of their flow cytometers and B. Mastis and S. Foley for help with the in vivo studies. We also thank Z.-N. Choo for igniting our interest in machine learning. This research was funded by Sarepta Therapeutics, by the MIT-SenseTime Alliance on Artificial Intelligence and by an award from the Abdul Latif Jameel Clinic for Machine Learning in Health (J-Clinic). C.K.S. (NSF Award no. 4000057398) acknowledges the National Science Foundation Graduate Research Fellowship (NSF grant no. 1122374) for research support.

Author contributions

C.K.S., S.M., J.M.W., B.L.P. and R.G.-B. conceptualized the research. J.M.W. and C.M.F. synthesized and tested the modular library. S.M. and R.G.B. developed the machine learning model with input from C.K.S. and B.L.P. C.K.S. synthesized the Mach peptides and constructs, performed experiments and analysed the results. K.B., C.-L.W. and J.A.W. performed the in vivo study with input from A.B.M. C.K.S., S.M., A.L., B.L.P. and R.G.-B. wrote the manuscript with input from all the authors.

Competing interests

B.L.P. is a co-founder of Amide Technologies and of Resolute Bio. Both companies focus on the development of protein and peptide therapeutics. The following authors are inventors on patents and patent applications related to the technology described: J.M.W., C.M.F. and B.L.P are co-inventors on patents WO 2020028254A1 (6 February 2020), WO2019178479A1 (19 September 2019), WO2019079386A1 (25 April 2019) and WO2019079367A1 (24 April 2019), which describe trimeric peptides for antisense delivery, chimeric peptides for antisense delivery, CPPs for antisense delivery and bicyclic peptide oligonucleotide conjugates, respectively. A.B.M., K.B., C.-L.W. and J.A.W. are employees of Sarepta Therapeutics, and Sarepta Therapeutics provided a portion of the funding for the work.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-021-00766-3.

Correspondence and requests for materials should be addressed to R.G.-B. or B.L.P.

Peer review information *Nature Chemistry* thanks Dominik Heider, Ülo Langel, Zigang Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Bradley L Pentelute, Rafael Gomez-Bombarelli

Last updated by author(s): May 28, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Agilent MassHunter B.06.1 and B10.1, BD FACSDiva v8.0, Peptimizer-v0.1
Data analysis	Agilent MassHunter B.06.0 and B10.0, FlowJo v10.7, GraphPad Prism v8.4.3, BioLegend LegendPlex v8.0, Peptimizer-v0.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data used for training of the model has been made available at https://github.com/learningmatter-mit/peptimizer, and archived in Zenodo repository (https://zenodo.org/record/4815385#.YK_VCjZKhhE). All other data has been provided in the main text or supplementary information.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Mouse studies: n = 8, 6, or 4. In vitro assays varied in sample size from n = 2 to 3 and is indicated in each figure caption, with the exception of the library activity assay in which n=1.
Data exclusions	The modular peptide library members that resulted in low cell count (high cell death) as determined by propidium iodide staining were excluded from the Predictor training dataset as they would bias the sequence-activity predictions. A single mouse treated at 60 mg/kg Mach4 was not responsive to stimuli several hours after injection, so this dosage group was not continued and the data from the single animal was not included.
Replication	The library peptides were tested in a single biological replicate. In vitro assays varied in sample size n=2 to 3 as described in figure captions. For delivery assays, independent replication experiments at varying concentration ranges were carried out and included in supplementary information, with similar results. The mouse study was conducted as a single experiment with multiple animals in each group. All replication attempts were successful.
Randomization	Randomization is not relevant in this study as all animals were identical male transgenic mice.
Blinding	Blinding was not relevant to our study, because biased interpretation is not likely due to either automated data collection or quantitative consistent measurements.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa 654 cells obtained from the University of North Carolina Tissue Culture Core facility. Human Renal Proximal Tubule Epithelial cells, THP-1, ECH001, Kerafast. THP-1 cells (ATCC TIB-202) . HeLa cells (ATCC CCL-2).
Authentication	Cell lines were used as purchased/received and the authors did not perform authentication.
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	EGFP-IVS2 654 transgenic mice (FVB/NJ mice transformed with CX-EGFP-654 plasmid) were obtained from Dr. Ryszard Kole's lab and

Laboratory animals	bred at Charles River Laboratory. 6- to 8- week-old male EGFP-654 mice were shipped to the vivarium at Sarepta Therapeutics (Cambridge, MA).
Wild animals	None.
Field-collected samples	None.
Ethics oversight	Institutional Animal Care and Use Committee (IACUC) of Sarepta Therapeutics.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	HeLa 654 cells obtained from the University of North Carolina Tissue Culture Core facility were maintained in MEM supplemented with 10% (v/v) fetal bovine serum (FBS) and 1% (v/v) penicillin-streptomycin at 37 C and 5% CO2. 18 h prior to treatment, the cells were plated at a density of 5,000 cells per well in a 96-well plate in MEM supplemented with 10% FBS and 1% penicillin-streptomycin. For individual peptide testing, PMO-peptides were dissolved in PBS without Ca2+ or Mg2+ at a concentration of 1 mM (determined by UV) before being diluted in MEM. Cells were incubated at the designated concentrations for 22 h at 37 C and 5% CO2. Next, the treatment media was removed, and the cells were washed once before being incubated with 0.25 % Trypsin-EDTA for 15 min at 37 C and 5% CO2. Lifted cells were transferred to a V-bottom 96-well plate and washed once with PBS, before being resuspended in PBS containing 2% FBS and 2 µg/mL propidium iodide (PI).
Instrument	BD LSRII HTS flow cytometer
Software	Data was collected using BD FACSDiva v8.0 and analyzed using Graphpad Prism 8.0 and FlowJo v.10.7
Cell population abundance	Each sample included at least 5,000 gated events. Only HeLa 654 cells were analyzed, and the main populations of concern were PI-positive and PI-negative cells. For most samples, PI-positive cells were outliers and not a main population. The cells gated as healthy (uniform and PI-negative) were then analyzed for EGFP fluorescence.
Gating strategy	Gates were applied to the data to ensure that cells that were positive for propidium iodide or had forward/side scatter readings that were sufficiently different from the main cell population were excluded.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.