Deep Learning Enables Discovery of a Short Nuclear Targeting Peptide for Efficient Delivery of Antisense Oligomers

Eva M. López-Vidal, Carly K. Schissel, Somesh Mohapatra, Kamela Bellovoda, Chia-Ling Wu, Jenna A. Wood, Annika B. Malmberg, Andrei Loas, Rafael Gómez-Bombarelli,* and Bradley L. Pentelute*

Cite This: JACS Au 2021, 1, 2009–2020

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Therapeutic macromolecules such as proteins and oligonucleotides can be highly efficacious but are often limited to extracellular targets due to the cell’s impermeable membrane. Cell-penetrating peptides (CPPs) are able to deliver such macromolecules into cells, but limited structure–activity relationships and inconsistent literature reports make it difficult to design effective CPPs for a given cargo. For example, polyarginine motifs are common in CPPs, promoting cell uptake at the expense of systemic toxicity. Machine learning may be able to address this challenge by bridging gaps between experimental data in order to discern sequence–activity relationships that evade our intuition. Our earlier data set and deep learning model led to the design of miniproteins (>40 amino acids) for antisense delivery. Here, we leveraged and expanded our model with data augmentation in the short CPP sequence space of the data set to extrapolate and discover short, low-arginine-content CPPs that would be easier to synthesize and amenable to rapid conjugation to desired cargo, and with minimal in vivo toxicity. The lead predicted peptide, termed P6, is as active as a polyarginine CPP for the delivery of an antisense oligomer, while having only one arginine side chain and 18 total residues. We determined the pentalysine motif and the C-terminal cysteine of P6 to be the main drivers of activity. The antisense conjugate was able to enhance corrective splicing in an animal model to produce functional eGFP in heart tissue in vivo while remaining nontoxic up to a dose of 60 mg/kg. In addition, P6 was able to deliver an enzyme to the cytosol of cells. Our findings suggest that, given a data set of long CPPs, we can discover by extrapolation short, active sequences that deliver antisense oligomers.

KEYWORDS: cell-penetrating peptides, machine learning, antisense, oligonucleotides, delivery

INTRODUCTION

The delivery of therapeutic macromolecules to intracellular targets is a major challenge. For example, phosphorodiamidate morpholino oligomers (PMOs, >6500 Da) are synthetic antisense oligonucleotides that can bind strongly to nucleic acids. The first approved PMO therapy Exondys 51 (etplirsen) was a breakthrough for treatment of Duchenne muscular dystrophy (DMD) in 2016 and acts via corrective splicing of mRNA.¹ To access its genetic target and elicit its therapeutic effect, PMO must reach the nucleus of cells in target tissues. Although a few antisense oligonucleotides have been approved by the Food and Drug Administration (FDA), a major obstacle for clinical advancement of these synthetic biopolymers is their poor cell permeability.² As a consequence, the clinical use of PMOs without delivery vehicles requires large administered doses.³ Traditional approaches to deliver PMO using liposomes and nanoparticles rarely advance to clinical work, often suffering from poor endosomal escape or significant toxicity.⁴ Other methods of macromolecule delivery
exist with varying efficiencies. However, delivery of PMO by covalent attachment to cell-penetrating peptides (CPPs) has been studied broadly and has recently shown promise in clinical trials. SRP-5051, a cell-penetrating peptide attached to eteplirsen, led to higher tissue exposure, exon-skipping, and dystrophin production in patients taking a monthly dose compared to patients taking weekly doses of Exondys 51.

The use of CPPs is a promising mode of macromolecule drug delivery. CPPs are short peptides of cationic, amphipathic, or hydrophobic nature that facilitate intracellular delivery of cargoes that are otherwise non-cell-penetrant, such as large and charged hydrophilic biomolecules. CPPs can enhance delivery through covalent linkages or through the formation of noncovalent nanoparticle complexes in the case of negatively charged oligonucleotides. The known sequence—activity motifs for CPPs largely rely on the guanidinium group of arginine (Arg). While polyarginine peptides have been shown to promote cell uptake with greater efficiency than other cationic residues due to membrane affinity and a difference in effective protonation, these sequences often remain trapped in the endosomes. A promising polyarginine peptide, Bpep, triggers endosomal escape by interspacing arginine residues with long alkyl 6-aminohexanoic acid and β-alanine residues. Still, one of the key challenges to clinical translation of polyarginine PMO–CPPs is their demonstrated in vivo toxicity caused by the peptide portion. While a linear relationship has been shown between the number of arginine residues and LD50 in mice, there is a stark difference.

Figure 1. A deep learning model trained on long (>20 residues) sequences can be repurposed to predict high-activity short (<20 residues) sequences. (A) Scheme of the machine learning-based CPP design approach. A deep learning triad trained with a combinatorial library of PMO–CPPs was first developed to generate unique miniproteins. In this work, the model is adapted to predict short CPPs that retain high PMO delivery activity. (B) Structure of PMO–CPP conjugates studied. PMO anti-IVS2-654 was attached to the N-terminus of peptides via strained alkyne–azide cycloaddition. (C) Schematic of the activity-based in vitro assay used to quantify PMO delivery activity. In order to produce fluorescence output, the PMO–peptide must enter the cell, escape the endosome, and localize to the nucleus where the PMO achieves exon-skipping activity.
between observed toxicity in vitro and in vivo.\textsuperscript{19,24} It is thought that the systemic toxicity induced by nonaarginine at 5 $\mu$mol/kg doses may be due to mast cell degranulation caused by the positive charges.\textsuperscript{22,23} Even so, recent clinical trial results show promise for the future of PMO–CPPs in patients.\textsuperscript{8} Although CPPs have been extensively studied, structure–activity relationships remain obscure due to inconsistent findings, making the rational design of unique highly active CPPs challenging.\textsuperscript{11} These experimental inconsistencies are due in part to the dependence of CPP performance on cargo, cell type, and treatment conditions, meaning that there is no universally applicable CPP. For example, studies in our laboratory have demonstrated that the cell-penetrating ability of common CPPs differs when bound to a cyanine dye versus a PMO drug, with no discernible trend.\textsuperscript{7} Therefore, to design unique CPPs that can deliver a particular macromolecular cargo and that do not rely on arginine, that cargo must be included.

One promising method for connecting such gaps in experimental space is the use of machine learning. Recent studies using a variety of strategies have been proven useful in the quantitative activity prediction for the design of antimicrobial peptides.\textsuperscript{24,25} The design of CPPs by machine learning has typically been limited to the use of binary classifiers because the training data are of inadequate size and from nonstandardized experiments.\textsuperscript{9,26–28} Our group has recently developed a machine learning strategy that combined a large standardized data set with deep learning to simultaneously design nuclear-targeting miniproteins and quantitatively predict their activity.\textsuperscript{12} This approach took advantage of the availability of a collection of hundreds of chimeric PMO–CPPs for the deep learning model training and used fingerprint representations to include unnatural amino acids aiding in cytosolic delivery,\textsuperscript{18} producing highly active miniproteins (40–80 residues). Most importantly, this machine learning framework was able to extrapolate activity data beyond the range of the training data set, suggesting that it could be leveraged to design shorter CPPs that distill sequences down to their critical motifs (Figure 1A). Short peptides would provide advantages over longer sequences, with lower molecular weight, and are of particular interest to the broad community to enhance the internalization of macro-molecules with minimal modification.\textsuperscript{29}

Here, we report the use of our deep learning model repurposed for the discovery of an 18-mer CPP that is able to deliver PMO with the same efficiency as the widely used polyarginine peptide, Bpep.\textsuperscript{18} Our peptide, P6, contains only one arginine residue and improves the delivery of PMO by 25-fold compared to PMO alone. Sequence–activity studies show the reliance on an N-terminal pentalysine motif, the single arginine residue, and the C-terminal cysteine residue. Peptide P6 has a wide therapeutic index in vivo and was able to deliver an anionic enzyme to the cytosol. Finally, P6 was able to restore protein synthesis in vivo after a single intravenously administered dose without evidence of kidney toxicity. These results demonstrate that we can use a machine learning model trained with long PMO–CPPs to design optimized, short CPPs (<20 residues) with minimal arginine content that are able to deliver PMO to the nucleus of cells with high efficiency and minimal toxicity in animals.

### RESULTS

#### Machine Learning-Based Design

To generate our sequences, we repurposed the machine learning model previously developed for the design of nuclear-targeting abiotic miniproteins in order to optimize short CPPs. In brief, this strategy includes the generation of starting sequences, a predictor of activity, and an optimizer to improve the activity of the sequence (Figure 1A). The model was trained using a chimeric library composed of 600 unique PMO–peptide conjugates and tested in an activity-based assay where the delivery of PMO to the nucleus of cells results in fluorescence, measured by flow cytometry (Figure 1B,C).\textsuperscript{30} The model optimizes sequences by increasing predicted activity while minimizing length and arginine content to mitigate potential cell toxicity.\textsuperscript{31} The output is hundreds of de novo designed sequences containing unnatural amino acids with a broad spectrum of predicted activity. Our previous report resulted in high-activity sequences reaching 40–80 residues in length. In this work, we demonstrate that the machine learning strategy can be used to predict high-activity, short sequences (<20 residues) for the delivery of PMO.

#### Data Augmentation Improves Extrapolation for Shorter CPPs

Because the original model was trained using a data set composed mostly of long (>40 residues) sequences, we used data augmentation to circumvent the sequence length imbalance and extrapolate predicted activities in the short (<20 residues) CPP sequence space. The data set obtained

---

Figure 2. Data augmentation improves extrapolation of activity for sequences with 20 residues or less. (A) A majority of sequences in the training data set, 601 out of 653, are longer than 20 residues (scatter plot). The desired design space of sequences with 20 or less residues is denoted in orange. (B) The distribution of lengths of sequences as-is and with 10x augmentation of sequences with 20 residues or less shows the change from a nearly normal distribution with a mean of 41 residues and a small peak at 23 residues, to a bimodal distribution. (C) The 10x CNN model ensemble predicts mean fluorescence intensity (MFI) for sequences in the validation (blue), test (orange), and experiment (green) data sets with reasonable accuracy. The points indicate the mean value of the predicted MFI from all of the models in the ensemble, and the error bars denote the standard deviation of prediction.
from our previous experimental studies had a mean sequence length of 41 residues, with 92% of sequences being longer than 20 residues. In contrast, here, we intended to design sequences with 20 residues or less (Figure 2A). To achieve this goal, in the training set, we replicated sequences with 20 residues or less 9 times, thereby showing the model 10x sequences in the desired design space (Figure 2B). This augmentation ensured that the model is trained over a similar sequence space as desired, which contains a shorter length from the augmented data set and a higher activity from the longer sequences already present in the data set. The caveat with this approach is that the diversity of sequences in the shorter search space has not changed, although the length domain has been shifted down toward shorter residues.

With less than 10% of sequences in the training data set in the desired design space and most with low experimental mean fluorescence intensity (MFI), data augmentation slightly improved extrapolation of predicted activities for sequences with fewer than 20 residues. The predicted peptides were validated experimentally, as described below, and their experimental activities were compared to the model’s predicted activity to judge the prediction accuracy (Figure 2C). In retrospect, we noted that the convolutional neural network (CNN) model performed the best among all tested models on the data set of sequences that were experimentally tested with a 0.148 R² score and 0.512 Pearson’s correlation (Tables S1−S3). For the unbiased root-mean-square error (uRMSE) on the experiment data set, we noted that simpler models had slightly better values than CNN models, but the ensemble variance denoted by the error bars in the parity plots were significantly large, thereby rendering the mean uRMSE values unreliable. A significantly larger training data set, containing a greater diversity of sequences and activities among short (<20 residue) peptides, would aid the effort of enhancing predictive models to extrapolate activity prediction, as observed from a retrospective active learning study (Figure S13). A complete summary of the model architectures tested with and without data augmentation is discussed in the Supporting Information, Section 10.

Figure 3. Predicted PMO−peptide conjugates enhance PMO delivery in a length- and sequence-dependent manner. (A) Delivery activities of predicted PMO−peptide conjugates PMO−P1 to PMO−P7 compared to PMO−Bpep, measured by the EGFP assay and calculated as fluorescence relative to PMO alone. (B) List of predicted peptides. Sequences selected represent those with the highest predicted activity of each length. Peptides selected for experimental validation are named P1−P7, were attached to PMO anti-IVS2-654, and tested in the EGFP assay. Bpep was used as a positive control for a comparison. Predicted and experimental activities are shown as relative to PMO alone. (C) Experimental PMO delivery activity of the library peptides (gray) and validated predicted peptides (red) versus length and arginine content relative to length. Bars in part A represent group mean ± SD, n = 3. Each sample was measured at a concentration of 5 μM. Shown here is one representative biological replicate, which was repeated twice with similar results, as shown in the Supporting Information. (*p < 0.01, **p < 0.001, ***p < 0.0001, ****p < 0.000 01 compared to PMO. n.s. = not significant as determined by Student’s unpaired two-tailed t test.)
Validation of Predictions

We had the model generate the highest predicted activity sequences with increasing lengths (from 5 to 20 residues), from which we selected 13 sequences to analyze and seven to validate experimentally (Figure 3). These sequences contain 0 or 1 arginine residue and have predicted activities ranging from a 4- to 11-fold increase with respect to unconjugated PMO. Of note, the sequences were cationic and contained several lysine residues as well as the extended alkyl backbone of unnatural residues β-alanine (B) and 6-aminohexanoic acid (X). A clear trend of the predicted activity increasing with length can be seen. Notably, the predicted sequences contained similar common motifs, indicating that the model consistently identifies these motifs as potential drivers of high predicted activity. We then selected seven of these sequences to validate experimentally, by ranking them by length and selecting every other sequence (P1−P7, Figure 3B). These sequences were synthesized via semiautomated fast-flow solid-phase peptide synthesis (SPPS) and coupled with 5-azidopentanoic acid to the N-terminus before cleavage and purification as previously reported. Separately, dibenzocyclooctyne (DBCO) was coupled to the 3′ end of PMO anti-IVS2-654 (Figure S2) before the two constructs were combined into the PMO−CPP conjugate via strain-promoted azide−alkyne cycloaddition (Figure S3).

PMO−CPP constructs were then validated with the HeLa 654 enhanced green fluorescent protein (EGFP) assay as previously described (Figure 1C). HeLa 654 are cells stably transfected with an EGFP coding sequence interrupted by an intron from the human β-globin gene (IVS2-654) containing a mutation that alters the normal pre-mRNA splice site. The change in splicing leads to retention of an unnatural mRNA fragment in the spliced EGFP mRNA and the translation of a nonfluorescent form of EGFP. The PMO anti-IVS2-654 hybridizes to the mutant β-globin exon in the stably transfected HeLa cells, altering gene splicing and leading to full-length EGFP expression. The amount of PMO delivered is therefore correlated to the amount of functional EGFP expressed, quantified via flow cytometry. This assay directly informs on the amount of PMO that is internalized into the nucleus of these cells. As a positive control, we used PMO conjugated to Bpep, a high-arginine-content peptide with the sequence RXRRBRRXRRBR studied extensively for PMO delivery.

The selected constructs exhibited a length-dependent increase in activity up to the 18-mer, P6. Analysis of the exon skipping activity measured in the EGFP assay performed with three different biological replicates (Figure 3A, Figures S4−S6) indicated that PMO−P6 is the conjugate that displays the highest activity, with a 23-fold increase relative to PMO alone. Bars represent group mean ± SD (n = 3, *p < 0.01, **p < 0.001, compared to PMO−P6; n.s. = not significant; determined using an unpaired two-tailed Student’s t test).

Figure 4. The Cys-containing aminohexanoic acid C-terminus has a mild effect on the activity of PMO−P6. (A) Construct P6, highlighting the KXXC C-terminal motif. B = beta-alanine, X = 6-aminohexanoic acid. (B) Activities of various mutated sequences as determined by EGFP assay and reported as fluorescence relative to PMO alone. Shown is a representative experiment, with biological replicates in the Supporting Information. (C) Table of Ala-mutated sequences corresponding to activities shown in part B as well as net charge of each sequence and activity relative to PMO alone. Bars represent group mean ± SD (n = 3, *p < 0.01, **p < 0.001, compared to PMO−P6; n.s. = not significant; determined using an unpaired two-tailed Student’s t test).
unconjugated PMO. This activity is comparable to the activity of the standard conjugate PMO−Bpep (24-fold increase) while containing only 1 arginine residue. When compared to the library data set, it is clear that P6 stands out as being shorter with fewer arginine residues, while having greater activity than any sequence in the library data set (Figure 3C). While having a minimal number of arginine residues, the predicted sequences are still polycationic, and all contain an oligo-lysine N-terminus. The second and third best performers among the conjugates of PMO with these 7 predicted peptides were PMO−P4 and PMO−P5 with an 8-fold and 9-fold increase in activity relative to unconjugated PMO, respectively. The rest of the conjugates showed a 6-fold increase or lower.

Interestingly, we observed a similar trend to the predicted activity, of increasing activity with increasing length, with the exception of the 20-mer PMO−P7 which had a moderate 6-fold activity. This observation suggests that while machine learning models have the potential to predict high-activity sequences, random errors leading to false positives are common.

**Alanine Mutations Reveal Sequence−Activity Relationships**

We employed alanine scanning to reveal that the C-terminal cysteine residue of P6 is partially responsible for the peptide’s high experimental activity. The sequences of P6 and P5, the best and second-best performers, respectively, are identical except for a few amino acids located toward the C-terminus; P6 contains a KXXC motif while P5 contains a MG motif. Despite this similarity, P6 has a nearly 3-fold higher activity than P5. We hypothesized that the presence of the KXXC motif, or specific residues within this motif, could be responsible for the activity of P6. To test this hypothesis, we substituted each one of the residues in the KXXC motif with alanine residues, and we deleted the KXXC motif completely (P8−P12 in Figure 4). After the conjugation of each P6 analogue to PMO, we carried out the HeLa 654 EGFP assay for activity profiling of PMO delivery. Activities of PMO−P8 to PMO−P11 demonstrated that, of the alanine mutations of the different positions in the KXXC motif, only the mutation of cysteine at the final position led to a significant, but not complete, decrease in activity. This observation suggests that, while not solely responsible for the high activity observed for PMO−P6, the C-terminal cysteine contributes to uptake. It is possible that the disulfide dimer form could be the active form of this conjugate, as a synthetic dimer exhibited a similar activity (SI, Section 11). Still curious about the presence of the KXXC motif, which is not found in the lower-performing PMO−P5, we made a truncated version of P6 lacking the KXXC C-terminus (PMO−P12, Figure 4). We observed a decrease in activity of the truncated sequence similar to that of PMO−P8 and only slightly higher than that of PMO−P5.
indicating again that the cysteine residue is the most critical C-terminal residue.

We then hypothesized that the cationic motifs may be responsible for the enhanced activity observed for PMO−P6, including the pentalysine chain present at the N-terminus, or the single arginine residue. Therefore, we substituted one, two, or three lysine for alanine residues, as well as the single arginine residue. Together, these observations suggest that while no single residue is responsible for the high activity of P6, the cationic residues may contribute the most to the activity. Nonetheless, the dramatic difference in activity between P5 and P6 despite their cationic sequence similarity suggests that the high activity of P6 is dependent on its unique sequence.

Concentration-Response of Activity and Toxicity

PMO−P6 and its truncated derivative, PMO−P12, have concentration-dependent activities while remaining nontoxic in vitro. We investigated these two sequences further by measuring the concentration−response for both activity and toxicity in vitro (Figure 6A,B). The half maximal effective concentration (EC50) was calculated by measuring the EGFP fluorescence in HeLa 654 cells of each conjugate along a range of concentrations (between 0.1 and 100 μM). The EC50 value of PMO−P6 was 4 μM, and activity ultimately reached 40-fold over PMO at 25 μM. The EC50 of PMO−P12 was 7 μM, and activity ultimately reached 30-fold at 50 μM. Therefore, PMO−P6 remained slightly more active than its truncated derivative PMO−P12.

We then measured the concentration−response of toxicity for PMO−P6 and PMO−P12. Membrane toxicity was assessed using the lactate dehydrogenase (LDH) release assay, where cytosolic LDH detected in the cell supernatant indicates membrane permeabilization. Toxicity is shown as a percentage relative to cells fully lysed with the detergent sodium dodecyl sulfate (SDS). We first assayed the cell supernatant of the HeLa 654 cells used in the concentration−response activity assay and found no toxicity at the concentrations tested, indicating that the observed delivery activity was not due to membrane permeabilization (Figure S9). We then analyzed a wider concentration range with the LDH assay in renal epithelial cells (TH1 RPTEC), since polyamionic sequences sometimes result in renal toxicity in vivo.38 We tested concentration ranges between 1 and 200 μM of PMO−P6 and between 5 and 400 μM of PMO−P12 (Figure 6A,B). The median lethal concentration (LC50) for PMO−P6 was observed at 100 μM, and at concentrations below 100 μM, the toxicity is negligible. LC50 for PMO−P12 was 220 μM, and at concentrations below 100 μM, the toxicity is negligible. These data demonstrate a large concentration window between activity and toxicity of these constructs in vitro, with PMO−P12 having a slightly larger in vitro window of 30, compared to 25 for PMO−P6. Elimination of the KXXC motif at the C-terminus of PMO−P6 slightly decreases activity, indicating that while PMO−P12 is less toxic, the machine-learning-designed PMO−P6 is more effective.

Mechanism of Uptake

We then investigated the mechanism of cell entry for PMO−P6 and PMO−P12 using chemical endocytosis inhibitors. For the uptake of PMO−CPPs at physiologically relevant concentrations, the primary mechanism of internalization is energy-dependent, and multiple endocytic mechanisms can take place.10,12 However, there is no main mechanism for the cell entry of peptides or PMO−peptide conjugates, and activity is highly dependent on the treatment concentrations, cell line, and cargo attached. Therefore, we investigated the mechanism of internalization of PMO−P6 and PMO−P12 in HeLa 654 cells using a panel of chemical endocytosis inhibitors: chlorpromazine, cytochalasin D, wortmannin, EIPA, and Dynasore. Chlorpromazine inhibits clathrin-mediated endocytosis,39,40 Dynasore inhibits clathrin-mediated endocytosis by preventing the assembly and disassembly of clathrin lattices on cell surfaces and on endosomes. Cytochalasin D inhibits phagocytosis and macropinocytosis, and wortmannin alters various endocytosis pathways by inhibiting phosphatidylinositol kinases; 5′-(N-ethyl-N-isopropyl) amiloride (EIPA) inhibits macropinocytosis, and Dynasore, a dynamin inhibitor, also inhibits clathrin-mediated endocytosis.

The mechanism of cell uptake of PMO−P6 and PMO−P12 is possibly endocytosis. The experiment was conducted in a pulse-chase format in which HeLa 654 cells were preincubated with the inhibitors. After 30 min, PMO−P6 or PMO−P12 was added to each well at a concentration of 5 μM. After incubation at 37 °C for 3 h, the treatment media was replaced with fresh media, and the cells were allowed to grow for

Figure 6. PMO−P6 and PMO−P12 induce a concentration-dependent activity and have low toxicity in vitro. Activity was determined using the EGFP assay at varying concentrations in HeLa 654 cells and is represented as fluorescence relative to PMO alone. Toxicity was determined using the LDH release assay at varying concentrations in TH1 RPTEC cells and is represented as % LDH release relative to cells fully lysed with SDS. (A) For PMO−P6, EC50 = 4 μM, and LC50 = 100 μM. (B) For PMO−P12, EC50 = 7 μM, and LC50 = 220 μM. Each point represents group mean ± SD, n = 6.
another 22 h at 37 °C. Sample preparation and flow cytometry were then performed as usual. The activities of PMO−P6 and PMO−P12 were reduced by treatment with chlorpromazine, cytochalasin D, and wortmannin. Chlorpromazine reduced activity in a concentration-dependent manner, suggesting that clathrin-mediated endocytosis may be the dominant uptake activity in a concentration-dependent manner, suggesting that cytochalasin D, and wortmannin. Chlorpromazine reduced receptor-mediated endocytosis.

**PMO Can Deliver Other Macromolecules to the Cytosol of Cells**

In addition to delivery of antisense oligonucleotides, we found that P6 can also deliver an anionic, active enzyme. Selected as a model protein, diphtheria toxin A (DTA) is a 21 kDa anionic protein segment of diphtheria toxin containing the catalytic domain of the toxin but lacking the portions for cell entry.37 Once inside the cytosol, this toxin inhibits protein synthesis and kills the cell. Therefore, its delivery can be measured with a cell proliferation inhibition assay using CellTiter-Glo, which quantifies adenosine triphosphate (ATP). Each point represents group mean ± SD, n = 3 distinct samples.

**In Vivo Studies**

We then found that PMO−P6 was able to correct aberrant splicing in vivo. We chose for our studies transgenic mice containing the same gene target as the HeLa 654 cells, enabling a direct transition of the PMO−peptide from in vitro to in vivo studies. Mice were injected with a single dose of PMO−P6 at three different concentrations (10, 30, and 60 mg/kg). Seven days later, serum and tissues including the quadriceps, diaphragm, and heart were harvested. We observed a dose-dependent increase in EGFP fluorescence in each muscle tissue examined, including the heart (Figure 9A–C). In the case of Duchenne muscular dystrophy, therapies that access cardiac tissue are critical for treatment because the primary cause of death in DMD patients is cardiac arrest. Therefore, the activity observed for PMO−P6 in the heart is particularly advantageous. Notably, EGFP levels measured in the heart were quite variable between individual mice treated with 30 or 60 mg/kg PMO−P6, whereas the samples from other tissues did not display a variability that was as high. A possible explanation is the differential oxidative environment in the heart compared to other tissues affecting the local oxidative state of our Cys-containing PMO−peptide, which is known to affect uptake of Cys-containing peptides.38 Nonetheless, despite this observed variability, PMO−P6 represents a short, low-arginine-content peptide able to deliver PMO to the heart and muscles of mice in vivo.

At the same time, we found that PMO−P6 does not induce kidney toxicity in the mice at the tested doses. The in vivo use
of polycationic CPPs can cause kidney toxicity. We analyzed kidney toxicity by measuring kidney biomarkers isolated from post-treatment serum, specifically blood urea nitrogen (BUN), creatinine, and cystatin C. No changes were found in the levels of these three renal function markers 7 days post-treatment which indicates that PMO−P6 is not toxic to the kidney at the doses tested (Figure 9D−F).

A key reason for the design of low-arginine-containing peptides is that high-arginine-content PMO−CPPs have demonstrated in vivo toxicity.19 It was observed that reduction in the number of arginine residues confers improved toxicology outcomes, but at the same time, screening studies in mdx mice showed a general pattern of reduced dystrophin splice correction activity as arginine residues are removed from the CPP.20 We show in this study that in vivo administration of PMO−P6 to mice induces a dose-dependent increase in splice correction and EGFP synthesis in the quadriceps, diaphragm, and heart while having no effect on kidney biomarkers up to a dose of 60 mg/kg. The EGFP mice used in this study are an appropriate model system in that our in vitro cell assay can be directly translated to this animal model to validate activity on the same gene target.

■ CONCLUSION

Designing cell-penetrating peptides to deliver a specific macromolecular cargo is a longstanding challenge in chemical biology. Here, a machine learning model trained on longer sequences (>20 residues) drove the discovery of a short, low-arginine-content peptide (<20 residues) that efficiently delivers PMO to the nucleus of cells. This cationic 18-mer peptide (P6) contains one arginine residue within its sequence and delivers PMO in vivo.

It is important to note that the machine learning model used here was trained with a PMO−CPP library composed of long sequences, with the vast majority being linear combinations of known CPPs. However, because our model is able to extrapolate predictions beyond the training set, we were able to generate unique short, low-arginine peptides that were not previously encountered. The challenge of designing effective CPP sequences became apparent during the experimental validation of the predicted peptides. We observed a length-dependent increase in activity up to 18 residues. However, two predicted peptides that differ only in their C-terminal regions, i.e., P5 (C-terminus: MG) and P6 (C-terminus: KXXC), displayed significantly different experimental activities, with the former about a third as active as the latter. Data augmentation was able to slightly improve the accuracy of activity prediction, but more work is required to more accurately design and predict unique short peptides for antisense delivery.

We found that the C-terminal cysteine residue boosted activity of PMO−P6, but the cationic N-terminal motif was primarily responsible for its high activity. Mutation of the C-terminal cysteine to alanine resulted in a decrease in activity, but mutations of the other residues within the KXXC motif had no effect. Removal of the entire motif (PMO−P12) resulted in a similar activity to the C18A mutation but also led to a slightly larger concentration window between activity and toxicity compared to P6 in vitro. The cysteine residue could be

Figure 9. PMO−P6 leads to dose-dependent increase in EGFP expression without affecting kidney toxicity biomarkers in EGFP-654 transgenic mice. EGFP expression levels in (A) quadriceps, (B) diaphragm, and (C) heart tissue 7 days after a single i.v. injection of various doses of PMO−P6. Concentrations of kidney toxicity biomarkers: (D) blood urea nitrogen (BUN), (E) creatinine, and (F) cystatin C isolated from serum 7 days after a single i.v. injection of various doses of PMO−P6. Each bar represents group mean ± SD; dots represent individual samples. Quadriceps: saline (0 mg/kg) n = 6; 10 mg/kg, n = 6; 30 mg/kg, n = 7; 60 mg/kg, n = 4. Diaphragm: saline, n = 6; 10 mg/kg, n = 6; 30 mg/kg, n = 6; 60 mg/kg, n = 4. Heart: saline, n = 6; 10 mg/kg, n = 7; 30 mg/kg, n = 8; 60 mg/kg, n = 4. Statistical significance was determined relative to the 0 mg/kg dose (*p < 0.01, unpaired two-tailed Student’s t test).
enhancing uptake by partaking in thiol-mediated uptake or by forming a more active “chimera” through disulfide formation, although it is clear that this residue is necessary but not sufficient for the activity of P6. Further structure–activity studies using analogues of P6 containing alanine mutations revealed that the pentalysine chain is the primary driver for its high activity, while the C-terminal cysteine residue only moderately boosts the activity. These observations indicate that a combination of length, composition, sequence, and charge influence the activity of P6. These short peptides demonstrate how minor variations in side chain composition can significantly affect activity, making empirical optimization of short peptides an arduous and unpredictable task.

While P6 contains a single arginine residue, it is still a polycationic peptide with nine lysine residues. The reason for this feature may be 2-fold; cationic sequences may be ideal for PMO delivery, and the prediction model may have been biased toward polycationic sequences. In the past, our lab has tested hundreds of CPPs for the delivery of PMO and have found that cationic sequences are in general more effective than other sequences. At the same time, the library that was used to train the machine learning model was composed of cationic sequences, likely biasing the outcomes of prediction. With the constraint of reducing arginine, lysine residues appear to have taken its place. However, we do not expect that the lysine residues have simply “replaced” arginine, as several past studies have shown that polarynine peptides outperform polycylsine peptides.

Here, we have found a low-arginine-containing peptide with comparable activity to a commonly studied oligoarginine CPP.

The design strategy discussed here is promising for the generation of peptide sequences able to deliver macromolecular cargo other than antisense oligonucleotides into cells and may be applicable for the design of other functional peptides with a suitable training data set. An improvement in PMO–CPP library synthesis and testing to access a wider range of sequences and activity may be one strategy to improve the machine learning-guided discovery of highly effective peptide sequences.

**EXPERIMENTAL SECTION**

**Peptide Synthesis, Purification, and Conjugation**

Peptides were synthesized using the Fmoc/tBu strategy via continuous fast-flow peptide synthesis, using either fully automated or semiautomated technology as previously described. S-Azidopentanoic acid was installed manually before cleavage and global deprotection using trifluoroacetic acid (TFA) and scavengers. Crude peptide was purified using reverse-phase chromatography before characterization by liquid-chromatography mass spectrometry (LC-MS). A complete summary of protocols can be found in the SI.

Azide-peptides (5 mM, water) were conjugated to PMO–DBCO (5 mM, water) via strain-promoted azide–alkyne cycloaddition and incubated at room temperature until reaction completion as monitored by LC-MS. The reaction was purified using reverse-phase HPLC with solvent A, 100 mM ammonium acetate in water pH 7.2, and solvent B, acetonitrile. Pure fractions were pooled as determined by LC-MS and lyophilized.

**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% CO2. The night before treatment, the cells were plated at 5000 cells per well in a 96-well flat-bottomed plate. On the day of the experiment, PMO–peptides were dissolved in PBS without cations at a concentration of 1 mM, determined by UV–vis, before being diluted to the specified treatment concentrations in MEM supplemented with 10% (v/v) fetal bovine serum (FBS) and 1% (v/v) penicillin–streptomycin. Cells were incubated for 22 h at 37 °C and 5% CO2, before the treatment media was removed and the cells washed with PBS and lifted using 0.25% trypsin-EDTA for 15 min at 37 °C and 5% CO2. Lifted cells were transferred to a 96-well V-bottomed plate and washed once with PBS before being resuspended in PBS containing 2% FBS and 0.2% propidium iodide (PI). EGFP fluorescence in the cells was analyzed via flow cytometry on a BD LSR II flow cytometer at the Koch Institute. Cells were gated such that cells positive for PI or that had a forward/side scatter sufficiently different from the healthy cell population were excluded. Each sample was capped at 5000 gated events. Analysis was conducted using FlowJo and Graphpad Prism 7. Activity is reported as mean fluorescence intensity (MFI) of experimental samples normalized to the PMO only condition.

**Protein Delivery**

P6-LPSTGG was synthesized as described above, and Gly5-DTA and SrtA were recombinantly expressed in *Escherichia coli* as described in the Supporting Information. Gly5-DTA (50 μM) was incubated with P6-LPSTGG (250 μM) and SrtA (2.5 μM) for 90 min on ice in SrtA buffer (10 mM CaCl2, 50 mM Tris, 150 mM NaCl, pH 7.5). The conjugate was isolated using a HiLoad 26/600 Superdex 200 prep grade size exclusion chromatography column (GE Healthcare, UK) in 20 mM Tris, 150 mM NaCl, pH 7.5 buffer. Fractions containing the pure product, as determined by LC-MS and gel electrophoresis, were concentrated using a centrifugal filter unit (10K, Millipore).

Cytosolic delivery of P6-DTA was tested in HeLa cells. Cells were plated at 5000 cells/well in a 96-well plate the day before the experiment. P6-DTA was prepared at varying concentrations in complete media, transferred to the plate, and incubated at 37 °C and 5% CO2. Cell proliferation was measured after 48 h using the CellTiter-Glo assay, and ATP concentration was reported normalized to the no treatment condition.

**In Vivo Studies**

EGFP-654 transgenic mice obtained from Dr. Ryszard Kole’s lab ubiquitously express EGFP-654 transgene throughout the body, identical to the HeLa 654 cell line. In this study, EGFP-654 mice were randomized into groups to receive a single i.v. tail vein injection of either saline or PMO–P6 at 10, 30, and 60 mg/kg. 7 days after the injection, the mice were euthanized for 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. Sera from all groups were collected 7 days postinjection and tested for kidney injury markers (serum BUN, creatinine, and cystatin C). 20–25 mg of mouse tissue was homogenized, and 80 μg of lysates was analyzed for EGFP fluorescence. Full protocols are described in the SI.

**Statistics**

Statistical analysis and graphing were performed using Prism (Graphpad) or Excel (Microsoft). Concentration–response curves were fitted using Prism using nonlinear regression. The listed replicates for each experiment indicate the number of distinct samples measured for a given assay. Significance for activities between constructs was determined using a Student’s two-sided, unpaired t test.

**ASSOCIATED CONTENT**

**Supporting Information**

The Supporting Information is available free of charge at [https://pubs.acs.org/doi/10.1021/jacsau.1c00327](https://pubs.acs.org/doi/10.1021/jacsau.1c00327). Materials; methods for LC-MS analysis; methods for peptide preparation; PMO functionalization and purification; preparation of PMO–peptide conjugates; EGFP assay and supplementary data; LDH assay; preparation and characterization of crystalline film.
Complete contact information is available at:

Supporting Information. Colab notebooks can be found at https://github.com/pikulsomesh. Jupyter notebooks are also in the repository, and demo Google Colab notebooks can be found at https://github.com/pikulsomesh/tutorials. All other data is available in the main text or the Supporting Information.

## AUTHOR INFORMATION

### Corresponding Authors

Rafael Gómez-Bombarelli — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-9495-8599; Email: rafgb@mit.edu

Bradley L. Pentelute — Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0002-7242-801X; Email: blp@mit.edu

### Authors

Eva M. López-Vidal — Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Carly K. Schissel — Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0003-0773-5168

Somesh Mohapatra — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-9498-3834

Kamela Bellovoda — Sarepta Therapeutics, Cambridge, Massachusetts 02142, United States

Chia-Ling Wu — Sarepta Therapeutics, Cambridge, Massachusetts 02142, United States

Jenna A. Wood — Sarepta Therapeutics, Cambridge, Massachusetts 02142, United States

Annika B. Malmberg — Sarepta Therapeutics, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-5640-1645

Andrei Loas — Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-5640-1645

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.1c00327

### Author Contributions

E.M.L.-V. and C.K.S. contributed equally to this work.

### Notes

The authors declare the following competing financial interest(s): B.L.P. is a co-founder of Amide Technologies and 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: B.L.P. is a co-inventor on patents WO 2020028254A1 (February 6, 2020), WO2019178479A1 (September 19, 2019), WO2019079386A1 (April 25, 2019), and WO2019079367A1 (April 24, 2019), describing trimeric peptides for antisense delivery, chimeric peptides for antisense delivery, cell-penetrating peptides for antisense delivery, and bicyclic peptide oligonucleotide conjugates, respectively. Sarepta Therapeutics has filed a provisional patent application related to the compounds described in this work.

### ACKNOWLEDGMENTS

This research was funded by Sarepta Therapeutics. We thank the Swanson Biotechnology Center Flow Cytometry Facility at the Koch Institute for the use of their flow cytometers, and Bryan Mastis and Sam Foley for help with in vivo studies. C.K.S. (4000057398) acknowledges the National Science Foundation Graduate Research Fellowship (NSF Grant 1122374) for research support. S.M. acknowledges an MIT Jameel Clinic—AI for Health (J-Clinic) fellowship for research support.

### REFERENCES

1. Lim, K. R. Q.; Maruyama, R.; Yokota, T. Eteplirsen in the Treatment of Duchenne Muscular Dystrophy. Drug Des., Dev. Ther. 2017, 11, 533–545.
2. Juliano, R. L. The Delivery of Therapeutic Oligonucleotides. Nucleic Acids Res. 2016, 44 (14), 6518–6548.
3. Moulton, H. M.; Moulton, J. D. Morpholinos and Their Peptide Conjugates: Therapeutic Promise and Challenge for Duchenne Muscular Dystrophy. Biochim. Biophys. Acta, Biomembr. 2010, 1798 (12), 2296–2303.
4. Malam, Y.; Loizidou, M.; Seifalian, A. M. Liposomes and Nanoparticles: Nanosized Vehicles for Drug Delivery in Cancer. Trends Pharmacol. Sci. 2009, 30 (11), 592–599.
5. Angelov, B.; Garamus, V. M.; Drechsler, M.; Angelova, A. Structural Analysis of Nanoparticulate Carriers for Encapsulation of Macromolecular Drugs. J. Mol. Liq. 2017, 235, 83–89.
6. Feger, G.; Angelov, B.; Angelova, A. Prediction of Amphiphilic Cell-Penetrating Peptide Building Blocks from Protein-Derived Amino Acid Sequences for Engineering of Drug Delivery Nanoassemblies. J. Phys. Chem. B 2020, 124 (20), 4069–4078.
7. Fu, A.; Tang, R.; Hardie, J.; Farkas, M. E.; Rotello, V. M. Promises and Pitfalls of Intracellular Delivery of Proteins. Bioconjugate Chem. 2014, 25 (9), 1602–1608.
8. 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. 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-51.html (accessed 2020-12-07).
9. Wolfe, J. M.; Fadzen, C. M.; Choo, Z.-N.; Holden, R. L.; Yao, M.; Hanson, G. J.; Pentelute, B. L. Machine Learning To Predict Cell-Penetrating Peptides for Antisense Delivery. ACS Cent. Sci. 2018, 4 (4), 512–520.
10. Fadzen, C. M.; Holden, R. L.; Wolfe, J. M.; Choo, Z.-N.; Schissel, C. K.; Yao, M.; Hanson, G. J.; Pentelute, B. L. Chimeras of Cell-Penetrating Peptides Demonstrate Synergistic Improvement in Antisense Efficacy. Biochemistry 2019, 58, 3980–3989.
11. Bechara, C.; Sagan, S. Cell-Penetrating Peptides: 20 Years Later, Where Do We Stand? FEBS Lett. 2013, 587 (12), 1693–1702.
12. Schissel, C. K.; Mohapatra, S.; Wolfe, J. M.; Fadzen, C. M.; Bellovoda, K.; Wu, C.-L.; Wood, J. A.; Malmberg, A. B.; Loas, A;
