Machine Learning Guides Peptide Nucleic Acid Flow Synthesis and Sequence Design

Chengxi Li, [a] † Genwei Zhang, [a] † Somesh Mohapatra, [b] Alex J. Callahan, [a] Andrei Loas, [a] Rafael Gómez-Bombarelli, [b] and Bradley L. Pentelute [a,c,d,e]*

Abstract: Peptide nucleic acids (PNAs) are potential antisense therapies for genetic, acquired, and viral diseases. Efficiently selecting candidate PNA sequences for synthesis and evaluation from a genome containing hundreds to thousands of options can be challenging. To facilitate this process, we leverage here machine learning (ML) algorithms and automated synthesis technology to predict PNA synthesis efficiency and guide rational PNA sequence design. The training data was collected from individual fluorenylmethyloxycarbonyl (Fmoc) deprotection reactions performed on a fully automated PNA synthesizer. Our optimized ML model allows for 93% prediction accuracy and 0.97 Pearson's *r*. The predicted synthesis scores were validated to be correlated with the experimental HPLC crude purities (correlation coefficient $R^2 = 0.95$). Furthermore, we demonstrated a general applicability of ML through designing synthetically accessible antisense PNA sequences from 102,315 predicted candidates targeting exon 44 of the human dystrophin gene, SARS-CoV-2, HIV, as well as selected genes associated with cardiovascular diseases, type II diabetes, and various cancers. Collectively, ML provides an accurate prediction of PNA synthesis quality and serves as a useful computational tool for rational PNA sequence design.

Introduction

In the past five years, antisense oligonucleotide (ASO)-based drug development resulted in five Food and Drug Administration (FDA)-approved drugs, *i.e.* Eteplirsen,[1] Golodirsen,[2] Casimersen,[3] Viltepso[4] (based on phosphorodiamidate morpholino oligomers, PMOs) and Spinraza[5] (based on 2′-*O*-methoxyethyl-phosphorothioate). Backbone modifications increase the therapeutic potential of ASO-based drugs due to improved pharmacokinetic and pharmacodynamic profiles. By assembling a charge-neutral ASO, peptide nucleic acid (PNA)-based chemistry is also gaining popularity for developing gene-specific therapies.[6] The amide-based backbone of PNAs offers unique physicochemical properties including enhanced chemical, thermal, and enzymatic stability, as well as high hybridization affinity and specificity with DNA and RNA.[7]

To evaluate biologically active PNA sequences for a given indication, the existing approach is to screen a small PNA library that typically contains up to dozens of candidates, each with a length of about 20 bases. There typically are hundreds to thousands of sequence design options available when targeting a specific gene or genome. For example, the genome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) contains nearly 30,000 bases,[8] raising a selection challenge when designing anti-SARS-CoV-2 sequences. Therefore, it is crucial to select "high value" PNA sequences from the multitude of available options to minimize costs and workload in the development process. In addition, sequence-dependent coupling efficiency should also be considered for each variant produced. The availability of routine computational algorithms, such as those enabled by machine learning (ML) to predict the efficiency of PNA synthesis, would represent a major step forward in improving overall PNA sequence design. To achieve this goal, a large high quality data set and reliable training methods are essential.

In chemical synthesis, access to high quality, interpretable, and standardized collections of data suitable for machine learning remains limited.[9] The data from published literature are usually collected using different reaction conditions and setups, and the reported results often exist in different formats.[9a] Furthermore, it is difficult to ascertain the irreproducible literature data.[10] Each of these aspects can contribute to an unsatisfactory ML model performance. The automated experimental platforms, on the other hand, can generate reproducible and highly consistent data, which could improve the model performance, but the data set size is usually limited. We recently demonstrated the advantages of automated fast-flow antisense PMO and PNA synthesis over traditional batch techniques in terms of higher synthetic fidelity, improved purity, and significantly decreased synthesis time.[11] The high-throughput reproducible flow synthesis data can provide a foundation for building robust machine learning models to predict and improve synthesis quality.

[a] Prof. Dr. B. L. Pentelute, Dr. C. Li, Dr. G. Zhang, A. J. Callahan, Dr. A. Loas, Department of Chemistry, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

[b] Prof. Dr. R. G. Bombarelli, S. Mohapatra, Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

[c] Prof. Dr. B. L. Pentelute, The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02142, USA.

[d] Prof. Dr. B. L. Pentelute, Center for Environmental Health Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

[e] Prof. Dr. B. L. Pentelute, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA.

* Correspondence to blp@mit.edu.

† These authors contribute equally to this work

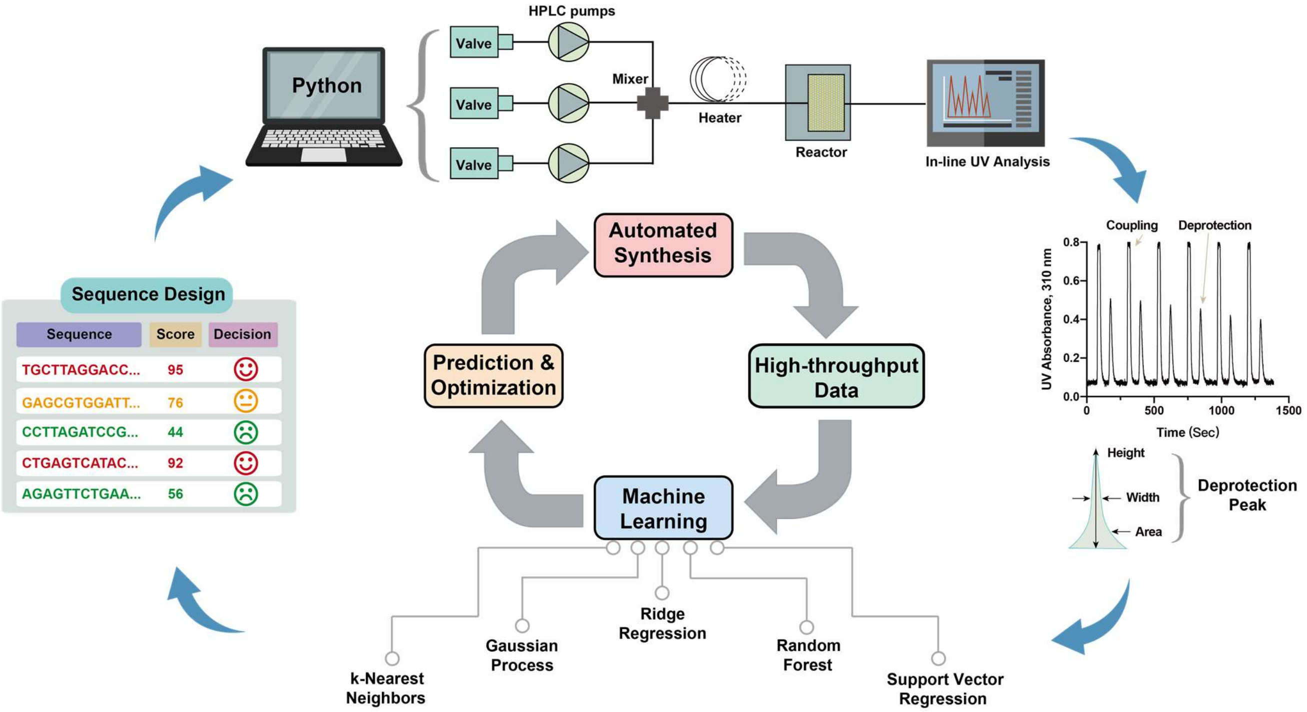


Figure 1. **Combining machine learning with automated synthesis technology delivers a design-build-test-learn cycle for rational PNA sequence design**. A Python program-controlled automated oligonucleotide synthesizer is used to synthesize PNAs, with a real-time UV-vis trace monitoring all coupling and deprotection reactions. Machine learning was applied over the integral peak areas calculated from the deprotection steps in the experimental data. A trained and optimized ML model makes prediction on the synthesis efficiency for any arbitrary PNA sequences, and therefore, enables a rational sequence design.

ML algorithm advancement can aid in uncovering nonobvious complex relationships. In biological transformations, ML has been previously applied to identification of drug-resistant cell phenotypes,[12] analysis of singe-cell metabolomics data,[13] and prediction of antibody toxicity.[14] Furthermore, the combination of state-of-the-art ML with automated chemical synthesis platforms can facilitate drug lead design and therapeutic development. In this regard, ML methods have been recently used to assist organic synthesis design,[15] and predict efficient organic synthetic pathways.[16] In addition, ML has also found applications in facilitating biopolymer productions, for example, optimizing fast-flow peptide synthesis using deep learning approach,[9b] discovering effective antimicrobial peptides through evolutionary algorithms,[17] and designing nuclear-targeting abiotic miniproteins.[18] Overall, using ML algorithms to mine the complex data set can unveil hidden patterns through performing data clustering, model regression, and trend prediction.

Here, we demonstrate that the in-line collected synthesis UV data can be utilized to train effective ML models to predict the synthesis yield of PNA sequences (Figure 1). After training and optimizing 10 different modern ML methods using 239 individual PNA coupling reactions, we developed a predictive ML model that allows for 93% prediction accuracy of the PNA synthesis. The predicted synthesis scores (deprotection peak area of the last coupling after normalization) were found to be highly correlated with the experimental HPLC crude purity, with a correlation coefficient $R^2 = 0.95$.

To further demonstrate the applicability of our optimized ML model towards efficient antisense PNA sequence design, we predicted all possible 18-mer antisense PNA candidates targeting human dystrophin gene exon 44, which contributes to 8% of Duchenne muscular dystrophy (DMD) patients but currently lacks treatments.[19] Three antisense PNA sequences were selected to represent easy, neutral, and difficult sequences for synthesis, and the purified product yields validated the model predictions. To benefit DMD antisense therapy development, the top 100 synthetically facile antisense PNA sequences targeting the exon 44 were reported. Similarly, top antisense PNA sequences were designed as potential candidates to target therapeutic-relevant genes that associated with SARS-CoV-2, HIV-1, as well as cardiovascular-related diseases, type II diabetes, and solid tumors. Taken together, nominating candidates that are synthetically easy to obtain can accelerate the overall process of producing bioactive PNAs. As a small step forward, in this study, we show that optimized ML model can guide efficient PNA sequence design and potentially accelerate the process of antisense drug development.

Results and Discussions

Training Data was Compiled from a Fully Automated PNA Synthesizer

Recently, our laboratory developed a Python program-controlled fully automated PNA synthesizer,[11] which enables rapid formation of each amide bond in approximately 10 seconds, a process significantly more rapid than either commercial peptide synthesizers or routine batch protocols.[11b] On our platform, the deprotection of fluorenylmethyloxycarbonyl (Fmoc) groups during PNA synthesis can be monitored using an in-line UV-vis detector (at 310 nm). Under optimized reaction conditions,[11b] the

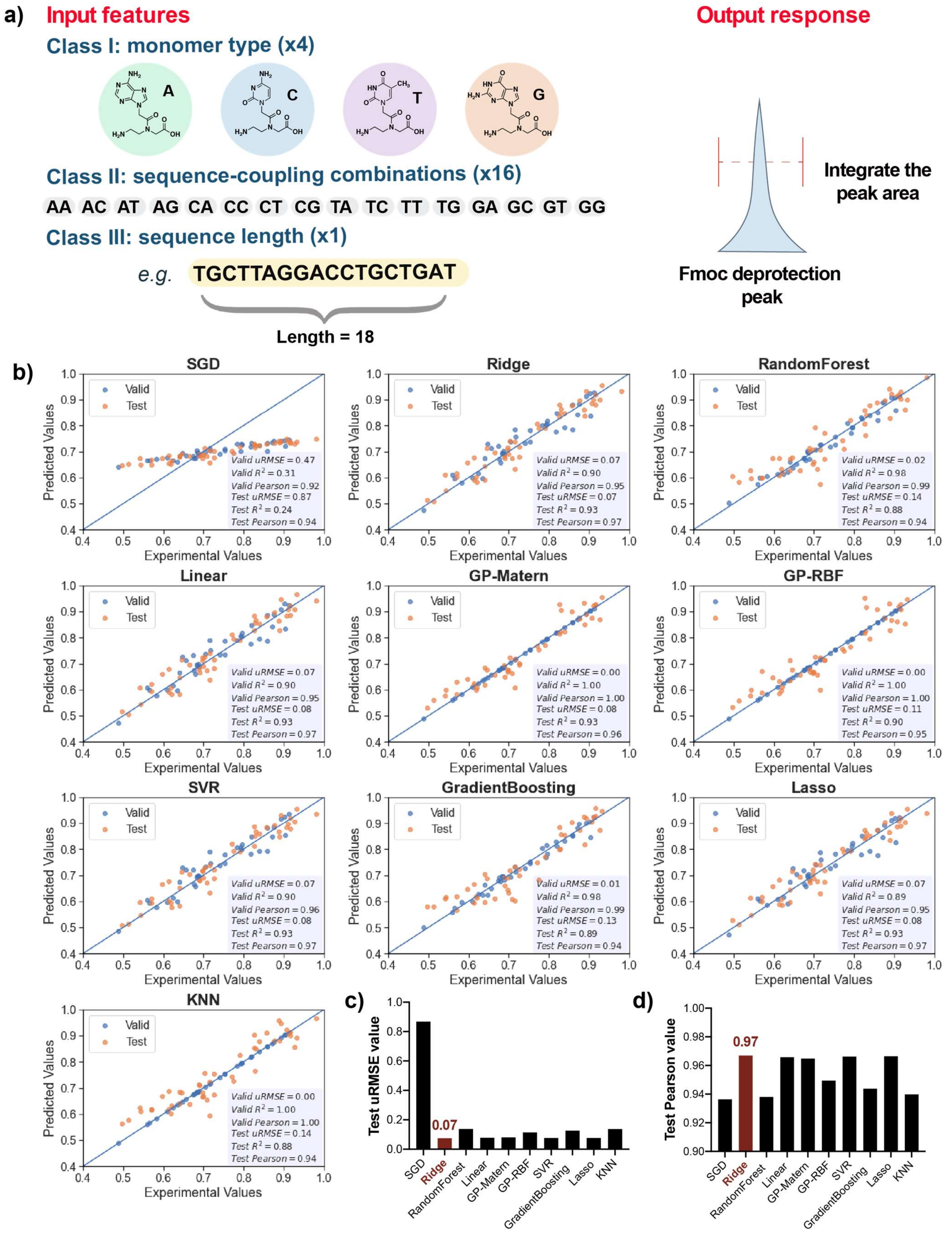


Figure 2. Benchmark 10 ML model architectures for accurate PNA synthesis prediction. (a) The input features include 4 PNA monomers, 16 sequence-coupling combinations, and sequence length. The integration of the Fmoc deprotection peak area is the output response. (b) Performance of 10 different ML model architectures on validation and testing datasets, visualized using parity plots. Individual scatter plots have points in blue for sequences in the validation dataset, and points in orange for sequences in the held-out testing dataset.

Metrics for model performance, unitless or relative root-mean-squared-error (uRMSE), R^2, and Pearson's correlation, have been noted for validation and testing datasets in the inset textboxes. Titles of the subplots refer to the specific model architectures. (c) Test uRMSE values of 10 ML models of which Ridge model presents the lowest value: 0.07. (d) Test Pearson values of 10 ML models of which Ridge model presents the highest score: 0.97. For more model performance details, see supporting information Table S1 and Table S2. Abbreviations: SGD – Stochastic gradient descent, GP – Gaussian process, SVR – Support vector regression, RF – Random forest, GB – gradient boosting, kNN – k-nearest neighbors, uRMSE – unitless/relative root-mean-squared error.

Fmoc deprotection UV trace can be used as an indicator of the synthesis quality. To fully utilize this information, we attempted to quantitatively investigate the relationship between the deprotection UV traces and the overall PNA synthesis efficiency via ML. To our knowledge, such a standardized UV-vis data set on PNA synthesis is not previously accessible with conventional PNA synthesis protocols.

To prepare the training data for our ML algorithm, we installed a 3-mer lysine linker on the C-terminus of each PNA sequences for data normalization. The peak area of every deprotection peak was then computed in Python environment.[11] The PNA sequence information was used to prepare training features and the deprotection peak areas were used as the response. Due to the peak variations caused by the resin amount loaded onto the synthesizer, the integral of deprotection peaks were normalized to the average peak area of the first three lysine residues. The final data set obtained contains 239 unique PNA pre-chain and nucleotide combinations.

ML Provides a Robust Tool for Accurate PNA Synthesis Prediction

Establishing a reliable training approach is key to achieving accurate model prediction. Many modern ML methods can be found to implement complex biological and chemical data analysis previously.[18, 20] To find the best ML approach, we benchmarked the performance of ten machine learning model architectures, i.e, Linear, Ridge, Lasso, stochastic gradient descent (SGD), gaussian process (GP) with two different kernel functions ('Matern' and 'RBF'), support vector regression (SVR), random forest (RF), gradient boosting (GB), and k-nearest neighbors (kNN). Three-fold cross-validation was used for a random split of 60% training, 20% validation, and 20% held-out testing data sets.[18] The input features consist of 21 different parameters including the PNA sequence length, 4 PNA monomers, and 16 possible sequence-coupling combinations within the sequence, while the integrated Fmoc deprotection peak area is treated as the output response (Figure 2a). The last-step synthesis efficiency was used as the final prediction score for each input PNA sequence.

The compiled data set collected on our automated PNA synthesizer was used to train and build all aforementioned ML architectures. After parameter optimization using a grid search approach, each optimized ML model was validated using the same validation data set and their prediction accuracy was tested and compared using the same testing data set. The model performances of 10 ML models were listed in Figure 2b. Except SGD, all models were able to achieve a near perfect validation R^2 and *Pearson's* correlation coefficient, indicating robust model fitting. On the held-out testing data set, Ridge, Linear, Lasso and SVR yielded the same *Pearson's* correlation coefficient, but Ridge regression outperformed all other model architectures by achieving a unitless/relative root-mean-squared error (uRMSE) of 0.07. Thus, we selected the optimized ML model based on Ridge regression for subsequent experimental validations and predictions.

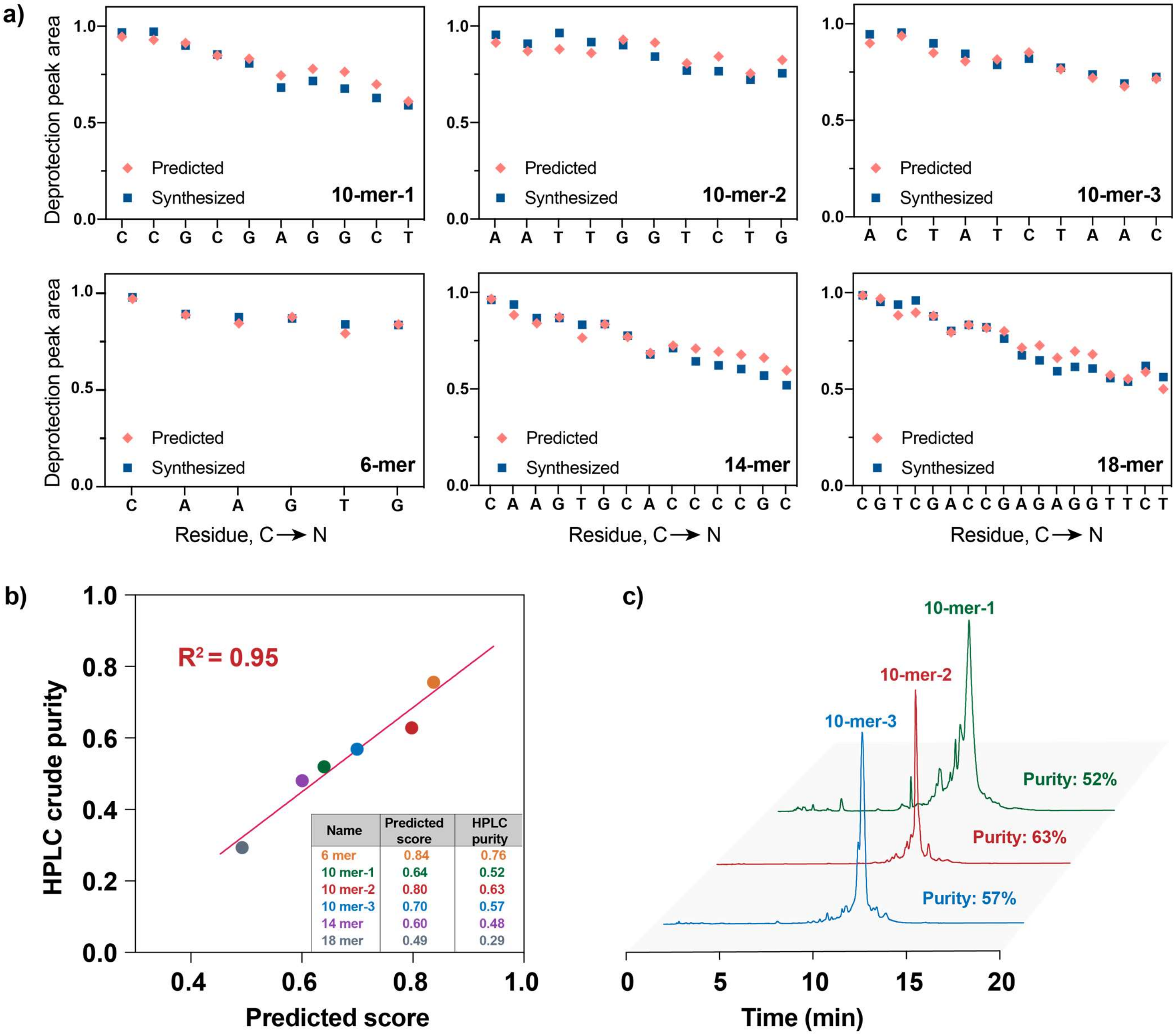
ML Informs the Feature Importance for Model Performance

Data mining over the training data set informs the feature importance for model performance. The relative feature importance contributing to the model prediction was summarized using *n-grams* representation approach and Ridge ML algorithm respectively (Figures S4 and S5). In line with the common intuition, the PNA chain length was ranked as a top important feature in both cases. In addition, besides the sequence length, we observed that four PNA monomers, *i.e.,* guanine (G), thymine (T), cytosine (C), and adenine (A), contribute significantly to the model performance. Overall, chain length and four monomers play a more important role than any of the 16 possible dimer permutations with respect to our model performance.

ML Predictions Agree with Experimental Data

To experimentally validate the prediction accuracy of optimized ML model, we randomly generated six PNA sequences for re-synthesis, including three 10-mers, one 6-mer, one 14-mer, and one 18-mer. Synthesis efficiency, denoted as the deprotection peak area at each coupling step, was predicted using the optimized ML model (Figure 3a). The six randomly generated sequences were individually synthesized on the automated PNA synthesizer, and the in-line deprotection data were collected and integrated. Notably, the experimental synthesis data were found highly consistent with the predicted traces (Figure 3a), indicating that our model enables an accurate prediction on the PNA synthesis quality based off sequences.

Side-reactions such as monomer deletion, rearrangement, and isomerization can occur during PNA synthesis,[11b] which cause lower reaction yield than predicted scores or potentially inconsistent results, and this information is difficult to track using UV-vis surveillance. To validate the correlation between the ML-predicted synthesis scores and the actual yield of the synthetic materials, all six synthesized PNAs were cleaved off the resin and their crude sample purities were measured. After a high-performance liquid chromatography (HPLC) analysis, the crude product yield was calculated via integration over the main product peaks, which were characterized with liquid chromatography-mass spectrometry (LC-MS, in supporting information Section 3). As shown in Figure 3b and 3c, the HPLC crude purities of the six randomly generated PNAs show strong correlation ($R^2 = 0.95$) with ML-predicted synthesis scores, suggesting that ML-predicted synthesis scores can further indicate the crude product yield.



Name	Predicted score	HPLC purity
6 mer	0.84	0.76
10 mer-1	0.64	0.52
10 mer-2	0.80	0.63
10 mer-3	0.70	0.57
14 mer	0.60	0.48
18 mer	0.49	0.29

Figure 3. **Predicted PNA synthesis scores agree with experimental validation.** (a) Six PNA sequences were randomly generated, including three 10-mers, one 6-mer, one 14-mer, and one 18-mer. ML predicts the synthesis efficiency, denoted as deprotection peak area of each step, and the trace were found consistent with the experimentally recorded UV data. (b) The HPLC crude purities of the six randomly generated PNAs show strong correlation (R^2 = 0.95) with ML-predicted synthesis scores. (c) The crude HPLC traces of three same-length PNAs were compared to demonstrate the distinguishing capability of the ML model. Integration was applied over the main product peaks, as indicted by the LC-MS data (supporting information Section 3).

ML Designs Antisense PNA Sequences Targeting Various Diseases and Cancers

To further demonstrate the practical application of our optimized ML model, we predicted all the potential antisense PNA sequences (14,854 18-mers in total, Fig. 4a) targeting the exon 44 of human dystrophin gene, which contributes to ~8% of all DMD patients and for which, at present, no drug treatment is available.[19] To validate the prediction accuracy experimentally, we selected one easy sequence (sequence I, predicted score: 0.71), one medium sequence (sequence II, predicted score: 0.59), and one difficult sequence (sequence III, predicted score: 0.32), and re-synthesized them on the automated flow instrument. As the mass spectrum shows in Figure 4a, only trace amounts of the desired product were found for sequence III, indicating an unsatisfactory synthesis. In contrast, for both sequences I and II, the major peaks were identified as the desired products with an observation that sequence I presented a cleaner mass spectrum ion trace than sequence II. Moreover, all the three PNA samples were purified with mass-directed reversed-phase high performance liquid chromatography (RP-HPLC). After purification, 1.2 mg and 0.7 mg of pure products were obtained for easy PNA (sequence I) and medium PNA (sequence II), respectively (Figure 4b and 4c). Unfortunately, we failed to obtain measurable pure product for the difficult PNA (sequence III) due to the low crude quality. Taken together, we confirmed that the PNA synthesis and purification outcomes are correlated with ML predictions. To potentially accelerate DMD antisense therapy development, we reported the top 100 easy antisense PNA sequences for targeting

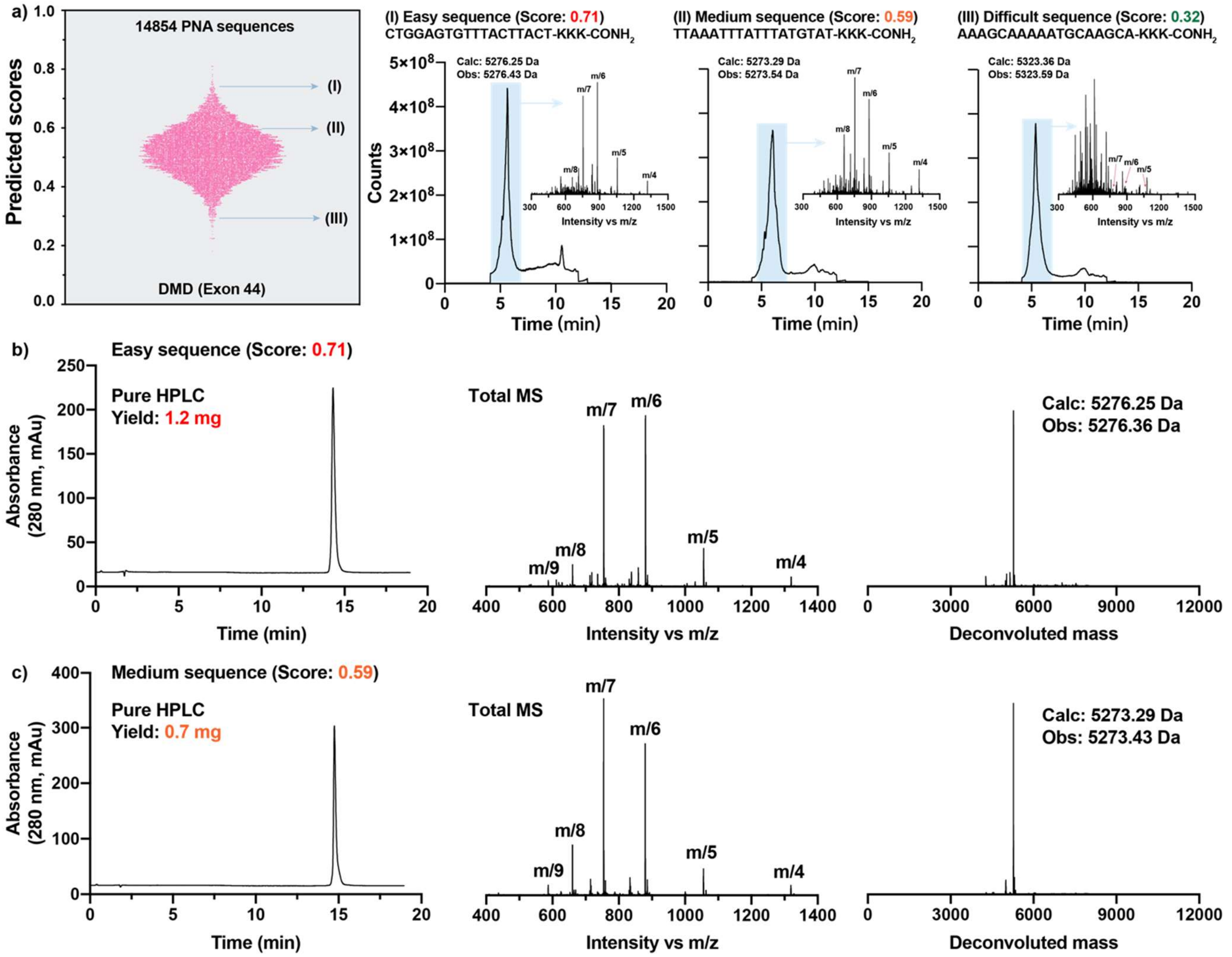


Figure 4. **ML predicts "high value" antisense PNA sequences for DMD disease.** (a) Left, predicted scores for 14,854 18-mer PNA sequences targeting exon 44 of human dystrophin gene; right, the crude total ion current (TIC) chromatogram and whole mass spectrum of three representative PNA sequences after re-synthesis. (b) Yield, HPLC trace, total mass spectrum, and deconvoluted mass of purified easy sequence I. (c) Yield, HPLC trace, total mass spectrum, and deconvoluted mass of purified medium sequence II. Failed to obtain pure product of difficult PNA sequence III after purification.

exon 44 of human dystrophin gene (sequences and predicted scores are available in the supporting information Section 9).

In addition, to show a broad applicability of our ML model, we attempted to design antisense PNA sequences for viral diseases, cardiovascular-related diseases, and various cancer types. Based on literature precedence, we selected two viral diseases, the ongoing pandemic-maker SARS-CoV-2[8, 11a] and incurable HIV-1,[21] as well as six protein targets (ANGPTL3, ANGPTL4, APOB, APOC3, LPA, and PCSK9) for cardiovascular-related diseases,[22] two protein targets (GCGR and SGLT2) for type 2 diabetes,[23] and seven protein targets (BRAF, EGFR, HER2, KRAS, MDM2, PD-L1, and VEGF) for various cancers[24] in consideration of their pharmaceutical potentials of developing antisense therapies (Figure 5). After predicting all possible antisense PNAs targeting the corresponding mRNA coding regions of the aforementioned protein targets, the top 100 most synthetically facile sequences were also reported (sequences and predicted scores can be found in the supporting information Section 9). In principle, the ML model can be used to guide antisense PNA sequence design for targeting any pharmaceutically relevant oligonucleotide sequences.

Collectively, we believe our ML prediction results are encouraging because the ability to design high-yielding PNA sequences from a vast candidate pool can save tremendous amounts of lab effort and reduce the overall costs of the synthesis process. The presented data processing and ML workflow can be used in principle for any flow chemistry reaction setup with the capability of in-line analysis. Towards accelerating the antisense drug development, we envision our strategy, combining automated synthesis technology with ML algorithms, can also be applied to guide other oligonucleotide sequence design, e.g. PMO,[11a] locked nucleic acid (LNA),[25] or DNA with already demonstrated potentials for therapeutic development.

Conclusion

In this study, a large training set was generated on an automated PNA synthesizer, providing suitable input data for the development of a robust machine learning (ML) algorithm. We then applied the optimized ML model to predict the efficiency of sequence-dependent solid-phase synthesis events. This model

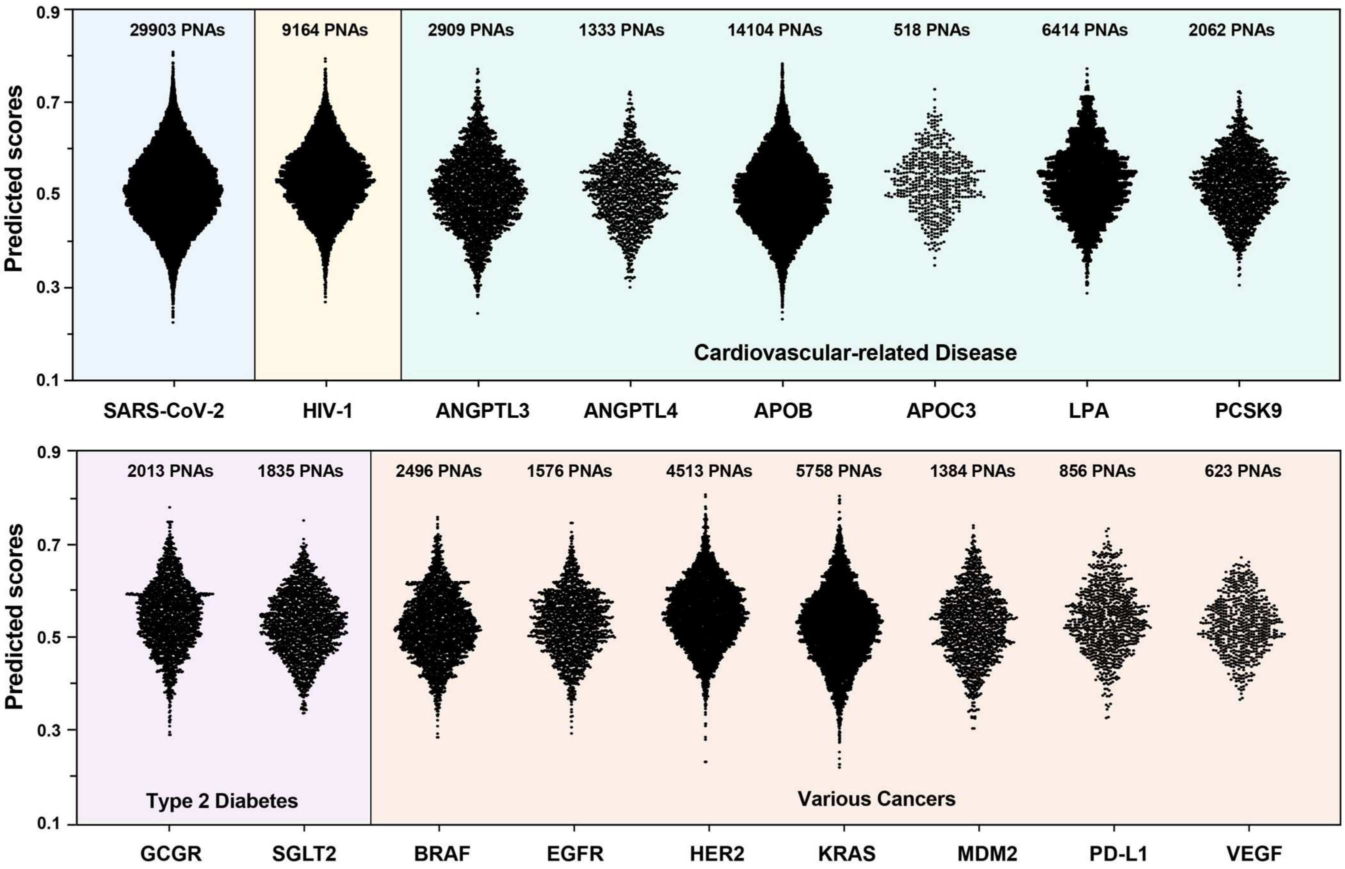


Figure 5. **ML predicts synthetically accessible antisense PNA sequences for various diseases and cancer targets.** Predicted scores for all possible 18-mer PNA sequences targeting the whole genome of SARS-CoV-2 and HIV-1, or mRNA sequences of ANGPTL3, ANGPTL4, APOB, APOC3, LPA, PCSK9, GCGR, SGLT2, BRAF, EGFR, HER2, KRAS, MDM2, PD-L1, and VEGF. Top 100 antisense PNA sequences for each target can be found in the supplemental information Section 9.

allows for accurate prediction of PNA synthesis efficiency and can serve as a useful tool to guide rational PNA sequence design.

Ten state-of-the-art ML algorithms were compared in our study. Ridge stands out as a robust approach among tested ML methods after hyper parameter tuning and optimization, allowing for 93% prediction accuracy of the synthesis using PNA sequences as the only input. Moreover, the predicted synthesis scores were validated to have a strong correlation with the experimental HPLC crude purities.

As a broad application of our ML model, we showed that it can design antisense PNA sequences for genetic and viral diseases, as well as cardiovascular disorders and cancers. Several representative protein targets and two viral genomes were selected as showcases in consideration of their pharmaceutical potentials to develop antisense therapies, and top antisense PNA sequences were reported. To conclude, the ML model we developed here are effective to design synthetically accessible PNA sequences, with the potential to accelerate antisense oligonucleotide drug development.

Acknowledgements

This research was supported by Novo Nordisk A/S. S. M. acknowledges the MIT-Takeda Fellowship program for research support. A.J.C. is supported by the Koch Institute MIT School of Science Fellowship in Cancer Research.

Conflict of interest

B.L.P. is a co-founder of Amide Technologies and Resolute Bio. Both companies focus on developing protein and peptide therapeutics.

Notes

All the data generated or analyzed during this study are included in the main text and supporting information. Predicted top 100 PNA sequences for various diseases are included in the supporting information.

The Python code for automated operation of the flow synthesis instrument is available at: https://github.com/L-Chengxi/MechWolf_Pull. All code used for training and optimization of the model is available at: https://github.com/genweizhang/Tiny_Tide.

Keywords: Machine learning · Automated flow synthesis · Peptide nucleic acid · Synthesis optimization · Sequence design

[1] Y. Y. Syed, *Drugs* **2016**, *76*, 1699-1704.
[2] Y. A. Heo, *Drugs* **2020**, *80*, 329-333.
[3] M. Shirley, *Drugs* **2021**, *81*, 875-879.
[4] S. Dhillon, *Drugs* **2020**, *80*, 1027-1031.

[5] V. Prakash, *Gene Ther.* **2017**, *24*, 497-497.

[6] a) S. Montazersaheb, M. S. Hejazi, H. Nozad Charoudeh, *Adv. Pharm. Bull.* **2018**, *8*, 551-563; b) C. Sharma, S. K. Awasthi, *Chem. Biol. Drug Des.* **2017**, *89*, 16-37.

[7] V. V. Demidov, V. N. Potaman, M. D. Frank-Kamenetskil, M. Egholm, O. Buchard, S. H. Sönnichsen, P. E. Nlelsen, *Biochem. Pharmacol.* **1994**, *48*, 1310-1313.

[8] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, *Nature* **2020**, *579*, 265-269.

[9] a) C. W. Coley, N. S. Eyke, K. F. Jensen, *Angew. Chem. Int. Ed.* **2020**, *59*, 23414-23436; b) S. Mohapatra, N. Hartrampf, M. Poskus, A. Loas, R. Gómez-Bombarelli, B. L. Pentelute, *ACS Cent. Sci.* **2020**, *6*, 2277-2286.

[10] M. Baker, *Nature* **2016**, *533*, 452-454.

[11] a) C. Li, A. J. Callahan, M. D. Simon, K. A. Totaro, A. J. Mijalis, K.-S. Phadke, G. Zhang, N. Hartrampf, C. K. Schissel, M. Zhou, H. Zong, G. J. Hanson, A. Loas, N. L. B. Pohl, D. E. Verhoeven, B. L. Pentelute, *Nat. Commun.* **2021**, *12*, 4396; b) C. Li, Callahan Alex J., Phadke Kruttika-Suhas, Bellaire Bryan, Farquhar Charlotte E., Zhang Genwei, Schissel Carly K., Mijalis Alexander J., Hartrampf Nin, Loas Andrei, Verhoeven David E., P. B. L.. *ChemRxiv* **2021**. 10.26434/chemrxiv.14099042.v1.

[12] R. Liu, G. Zhang, Z. Yang, *Chem. Commun.* **2019**, *55*, 616-619.

[13] X. Tian, G. Zhang, Y. Shao, Z. Yang, *Anal. Chim. Acta.* **2018**, *1037*, 211-219.

[14] M. Garofalo, L. Piccoli, M. Romeo, M. M. Barzago, S. Ravasio, M. Foglierini, M. Matkovic, J. Sgrignani, R. De Gasparo, M. Prunotto, L. Varani, L. Diomede, O. Michielin, A. Lanzavecchia, A. Cavalli, *Nat. Commun.* **2021**, *12*.

[15] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434-443.

[16] X. Wang, Y. Qian, H. Gao, Connor W. Coley, Y. Mo, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2020**, *11*, 10959-10972.

[17] M. Yoshida, T. Hinkley, S. Tsuda, Y. M. Abul-Haija, R. T. McBurney, V. Kulikov, J. S. Mathieson, S. G. Reyes, M. D. Castro, L. Cronin, *Chem* **2018**, *4*, 533-543.

[18] C. K. Schissel, Somesh Mohapatra, Justin M. Wolfe, Colin M. Fadzen, Kamela Bellovoda, Chia-Ling Wu, Jenna A. Wood, Annika B. Malmberg, Andrei Loas, Rafael Gómez-Bombarelli, B. L. Pentelute, *Nat. Chem.* **2021**, *13*, 992-1000.

[19] R. T. Wang, F. Barthelemy, A. S. Martin, E. D. Douine, A. Eskin, A. Lucas, J. Lavigne, H. Peay, N. Khanlou, L. Sweeney, R. M. Cantor, M. C. Miceli, S. F. Nelson, *Hum. Mutat.* **2018**, *39*, 1193-1202.

[20] a) R. Burbidge, M. Trotter, B. Buxton, S. Holden, *Comput. Chem.* **2001**, *26*, 5-14; b) S. Mahadevan, S. L. Shah, T. J. Marrie, C. M. Slupsky, *Anal. Chem.* **2008**, *80*, 7562-7570; c) J. Schmidt, J. M. Shi, P. Borlido, L. M. Chen, S. Botti, M. A. L. Marques, *Chem. Mater.* **2017**, *29*, 5090-5103; dN. E. Helwig, *Quant. Meth. Psychol.* **2017**, *13*, 1-19.

[21] a) S. A. Prasad, A. Wany, B. Mazumdar, *Res. J. Biotechnol.* **2008**, 168-170; b) B. Dropulic, L. Humeau, G. K. Binder, X. B. Lu, V. Slepushkin, R. Merling, P. Echeagaray, M. Pereira, T. Slepushkina, S. Barnett, L. K. Dropulic, R. Carroll, B. Levine, R. R. MacGregor, C. H. Junes, *Mol. Ther.* **2004**, *9*, S384-S384.

[22] A. Laina, A. Gatsiou, G. Georgiopoulos, K. Stamatelopoulos, K. Stellos, *Front. Physiol.* **2018**, *9*.

[23] S. X. Chen, N. Sbuh, R. N. Veedu, *Nucleic Acid Ther.* **2021**, *31*, 39-57.

[24] a) P. Khan, J. A. Siddiqui, I. Lakshmanan, A. K. Ganti, R. Salgia, M. Jain, S. K. Batra, M. W. Nasser, *Mol. Cancer* **2021**, *20*; b) Y. H. Wu, W. Y. Gu, J. Li, C. Chen, Z. P. Xu, *Nanomedicine* **2019**, *14*, 955-968; cB. T. Le, P. Raguraman, T. R. Kosbar, S. Fletcher, S. D. Wilton, R. N. Veedu, *Mol. Ther. Nucleic Acids* **2019**, *14*, 142-157.

[25] D. A. Braasch, D. R. Corey, *Chem. Biol.* **2001**, *8*, 1-7.