Check for updates

Characterizing ssRNA and dsRNA electrophoretic behavior: empirical insights with neural network-aided predictions†

Nina Sheng Li, [iD] ‡[a] Adriana Coll De Peña, [iD] ‡[b] Matei Vaduva,[c] Somdatta Goswami[d] and Anubhav Tripathi [iD] *[b]

RNA-based therapeutics are currently at the forefront of the biopharmaceutical industry because of their safety, efficacy, and shortened time from disease discovery to therapy development. Microfluidic electrophoresis provides a great analytical platform to analyze nucleic acids in unprecedented detail. However, while DNA has been studied extensively within microfluidic systems, there is limited data available for RNA, particularly of chemically modified molecules, such as those used in the COVID-19 mRNA vaccines, and for long double-stranded RNA molecules, which may accompany, intentionally or as a by-product, RNA therapeutics. To this end, this study focused on the empirical microfluidic electrophoretic analysis of double- and single-stranded RNA, non-modified and pseudouridine-modified, at varying gel concentrations. It then compared the findings to the electrophoretic mobility models in the literature. This work was then complemented with data-driven and physics-informed neural networks that successfully predicted the migration time and length of different RNA molecules with an average error of 12.34% for the data-driven model and 0.77% for the physics-informed model. The low error in the physics-informed neural networks opens the doors to the electrophoretic characterization of molecules, even beyond RNA, without the need for extensive experimental data.

Introduction

The advent of microfluidic electrophoresis as a nucleic acid analytical tool has allowed scientists to study the migration of DNA and RNA at an unprecedented level of detail. Despite the promise this platform brings due to its characteristic short runtime, high resolution, and increased sample throughput,[1,2] the migration patterns of these molecules in microfluidic conditions have not been extensively studied. Furthermore, ideal experimental conditions for nucleic acids of various lengths have not been clearly defined. Studying the biophysical interactions between migratory nucleic acids and their medium is vital for developing nucleic acid purity and integrity analysis assays, such as mRNA vaccine quality assessment.[3,4] It is also

[a]*The Warren Alpert Medical School, Brown University, Providence, RI 02906, USA*
[b]*Center for Biomedical Engineering, School of Engineering, Brown University, 182 Hope Street, Providence, RI, USA*
[c]*Department of Molecular Biology, Cell Biology, and Biochemistry, Division of Biology and Medicine, Brown University, Providence, RI, USA*
[d]*Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, USA*
†Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5an00381d
‡These authors contributed equally.

crucial for developing more complex assays to resolve similar molecules. Such advancements may improve the characterization and quality control assays of exosome genomes[5–8] and samples of nucleic acids mixtures (*e.g.* multivalent nucleic acid-based vaccines), which has proven to be more difficult.[9–11] Understanding the mechanisms behind nucleic acid migration is also crucial for troubleshooting assays, allowing researchers to identify experimental conditions that better suit their parameters and desired outcomes.

The current models describing nucleic acid electrophoretic mobility are differentiated according to the relative sizing between the pore size of the semi-dilute polymeric network and the nucleic acid size, as defined by the radius of gyration (R_g).[12] The main models include the Ogston model, describing DNA molecules with R_g smaller than the pore size, and the Biased Reptation with Fluctuation (BRF), describing DNA molecules with R_g greater than the pore size. The BRF model is further differentiated into two scenarios: reptation without orientation and reptation with orientation. The Ogston model assumes the DNA is a spherical object moving through a sieve driven by the electric field.[12,13] In this model, mobility is proportional to the exponential of the negative concentration of the polymer solution.[12,13] According to the BRF, mobility scales as $1/N$ for short chains and levels off for large sizes and/

or high electric fields.[12,14] Each model has respective limitations and alternative modifications have also been made to better describe nucleic acid movement in capillary electrophoresis.[15–18] While there has been abundant research in the electrophoretic separation of DNA, much less work has focused on the separation and mobility models of RNA, especially long nucleoside-modified mRNA and double-stranded RNA (dsRNA).[19–21]

This study aims to explain the electrophoretic mobility of different RNA molecules in microfluidic systems and understand the underlying physical principles that govern the migratory patterns of differently sized RNA in varying concentrations of semi-dilute polymer solutions. Given their clinical relevance in mRNA vaccines and therapies, a focus is placed on the mobility of both single- and double-stranded RNA and the potential impact of nucleoside modifications.[22–24] To the authors' knowledge, this is the first time the electrokinetic of dsRNA fragments, especially that of longer length and nucleoside-modified RNA, has been studied in microfluidic capillary electrophoresis. With the rise in RNA research, the mobility of RNA with pseudouridine modifications, which enhance RNA stability and decrease their immunogenic response,[23,25] will be important to characterize. Additionally, immunogenic dsRNA, intentional or residual, will be a critical component in future vaccine and alternative therapy research and development.

This paper also aims to provide predictive modeling of the electrophoretic mobility of single- and double-stranded RNA of varying lengths under different conditions using artificial neural networks (ANNs), a class of machine learning tools,[26–29] to provide guidelines for future assay development, diagnostic protocols, quality control platforms, and genetic material differentiation. Both data-driven[30,31] and physics-driven[32,33] ANNs obtaining low margins of error were trained. This work highlights how ANNS, particularly physics-informed neural networks (PINNs), can be used to increase the understanding of physical behavior of biological samples, such as their electrophoretic mobility, and to support the development of analytical methods.

Materials and methods

Samples and sample preparation

The ssRNA (catalog #N0364S) and dsRNA ladders (catalog #N0363S) used in this study were purchased from New England Biolabs (New England Biolabs, Ipswich, MA). The ssRNA ladder contains fragments of 50 nt, 80 nt, 150 nt, 300 nt, 500 nt, and 1000 nt, while the dsRNA ladder contains fragments of 21 bp, 30 bp, 50 bp, 80 bp, 150 bp, 300 bp and 500 bp. The non-chemically modified, 4001 bases long mRNA and dsRNA samples were custom ordered from Genewiz (Genewiz Genomics Headquarters, South Plainfield, NJ). The pseudouridine chemically modified mRNA (818, 1198, 1913, 3406, and 4451 nt) and dsRNA (700 and 1800 bp) fragments, and the 700 and 1800 bp non-chemically modified dsRNA

fragments were purchased from CATUG Biotechnology (CatPure™; Cambridge, USA). Prior to use, the ssRNA and dsRNA ladders were diluted 1 : 5 in 1× TE buffer, and the custom mRNA and dsRNA fragments were diluted to 5 ng μL^{-1} in 1× TE buffer.

Microfluidic measurements and analysis

For the microfluidic electrophoretic characterization of the samples, the LabChip GXII Touch platform (Revvity, Waltham, MA) was used to control the electric fields used to control the migration of the samples for analysis. The platform was combined with a custom RNA microfluidic chip, SYTO 61, poly(N, N-dimethyl acrylamide) (PDMA) solution, and lower marker, all provided by Revvity, as described in our previous study.[4] To analyze the electrophoretic behavior of the samples at different gel concentrations, the stock PDMA solution was diluted with a gel diluent (Revvity) that maintained the conductivity constant while the gel concentration changed. After diluting the gel to the desired concentration, it was mixed with 2.34% v/v of the fluorescent stain and spun down according to the instructions from the provider. The chip was loaded with the gel-dye mixture and the lower marker for electropherogram alignment during analysis.

Once the samples were diluted to the desired concentration, 10–15 µL were loaded onto a 384-well plate, and the well plate and chip were loaded onto the platform. Independent of gel concentration, a script containing the same loading, injecting, and separation voltages, which have been described in our previous study, was used for all experiments.[4] However, the separation time was increased depending on the gel concentration to ensure all peaks were captured. After the script was run, the LabChip Reviewer software (Revvity) was used to visualize the electropherograms.

Unless otherwise specified, all experiments were conducted in 2–3 experimental repeats with 2–3 instrumental repeats, yielding 6–9 data points per condition and sample tested. The statistical analyses were conducted using GraphPad Prism 9.4.1 (681), and the significance was by a Tukey *post hoc* test with a confidence interval of 95%; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$. GraphPad Prism 9.4.1 was also used to generate the non-linear regression fits reported across the study.

Theoretical operating regime

Unlike traditional slab gel electrophoresis, where the cross-linked polymers create fixed pores through which analytes can migrate, capillary electrophoresis in entangled polymers relies on transient pores created by entanglement points that undergo a "random walk". The pore size of various gel concentrations can be approximated by the blob size (ξ_b) and is calculated by:[12]

$$\xi_b = 1.43 R_g \left(\frac{c}{c^*}\right)^{-(1+a)/3a}, \tag{1}$$

where c is the polymer concentration, c^* is the entanglement threshold for the polymer, a is an exponent of the Mark–

Houwink equation, and R_g is the radius of gyration of the polymer calculated by:

$$R_g \approx \sqrt[3]{\frac{[\eta]M_r}{6.2N_A}},\qquad(2)$$

where $[\eta]$ is the intrinsic viscosity of the polymer solution at the concentration, M_r is the polymer molecular mass, and N_A is Avogadro's number. Using the Mark–Houwink coefficients established previously for PDMA[34] and intrinsic viscosity found through linear interpolation of PDMA solutions of various molecular weights,[35] an operating effective pore size of 37 nm, 21 nm, 15 nm, 12 nm, and 10 nm were established for 1%, 2%, 3%, 4%, and 5%, respectively.

When analyzing mobility, it is important to examine the relationship between the pore size of the sieving matrix and the radius of gyration of the nucleic acid. Defined as the average distance squared between different parts of the object and its center of mass, this measurement provides information regarding the average shape of the nucleic acid that end-to-end distance does not, which is essential to characterize as electric fields can induce different molecular conformations.[36] The radius of gyration of nucleic acids can be approximated by:[37]

$$R_g{}^2 = \frac{1}{3}pL\left[1 - \frac{p}{L} + \frac{p}{L}e^{-L/p}\right],\qquad(3)$$

where p is the persistence length, a measurement of polymer chain stiffness, and L is the length of the polymer fragment. Values of 64 nm and 2 nm for the persistence length of dsRNA and ssRNA were used for calculations based on respective averages of current literature reporting.[38–40] Approximating the transition from Ogston-like motion to BRF as the size where the R_g of the nucleic acid is equal to the effective pore size, we predicted that all dsRNA samples would operate under reptation with the possible exception of 80 bp under 1% gel. For ssRNA, this transition point is expected to occur from approximately 54–2150 nucleotides for 1–5% gel. However, we acknowledge the true transitions may vary due to the heavy reliance of the formula on persistence length, which has been reported with large ranges, and the approximations made in formulaic calculations of pore size.

Artificial neural network development

In this study, our goal is to employ ANNs to distinctly solve the following two tasks:

• Task 1: to predict length of the base sequence/number of base pairs (n_b), given the migration time (m_t), gel concentration (g_c), information about the type of RNA (ssRNA or dsRNA) (t_p), and the corresponding persistence length (l_p).

• Task 2: to predict the migration time (m_t), given the length of the base sequence/number of base pairs (n_b), gel concentration (g_c), information about the type of RNA (ssRNA or dsRNA) (t_p), its persistence length (l_p), and the molecular weight of the sequence (M_w).

In these two tasks, all the variables except t_p have multiple discrete values as t_p is used as an identifier with a value of 1

(for ssRNA) or 2 (for dsRNA). The additional parameter of sequence molecular weight (M_w) was included in task 2 to improve model robustness and allow for the model to capture trends that may deviate from theoretical expectations. Additionally, due to varying base composition, modifications, sequence-specific variations, the inclusion of (M_w) may help in capturing additional variability not stated by the other input parameters. M_w is directly tied to the unknown output of task 1 and therefore was not included as an input parameter. To efficiently solve the tasks, we developed a data-driven and physics-informed neural network frameworks, employing deep neural networks. More information regarding deep neural networks as well as basis of both data-driven and physics-driven neural networks can be found in our ESI.†

One primary bottleneck of data-driven neural networks (both deep and shallow) rests in the fact that a considerable amount of training data is required. In this work, we employ data-augmentation schemes to generate additional labeled datasets given the datasets obtained from the experiments. To that end, we plot the curves n_b vs. m_t for every g_c obtained from the experimental data and obtain the equation of the curve using logarithmic regression. For every g_c, we generate an additional 20 sample points considering sample n_b points uniformly distributed between 100 and 4000 bases. The original experimental data as well as the obtained data from the data-augmentation schemes were using for training of the ANNs. We describe the anatomy of the deep neural networks and components of the physics-driven approach below.

Deep neural networks. For task 1, we aim to learn $\mathcal{N}(\cdot;\theta):[g_c, m_t, t_p, l_p] \rightarrow n_b$, where $Y = \mathcal{N}(x;\theta)$, represents a compressed form of a feed-forward neural network (see ESI†). In this scenario, we have one network, taking all the varying parameters as input to the network and outputs the solution, n_b. The network consists of two hidden layers with 64 neurons each. To introduce non-linearity in the network, we have employed a leaky ReLU activation function between the two hidden layers. The connection between the input layer and the first hidden layer as well as the connection between the second hidden layer and the output layer considers linear activation function.

For task 1, we have considered the migration time as an input parameter. However, these details might not be available a priori for an unseen case. Therefore, in task 2, we aim to learn the migration time given the other parameters. Hence, we design the network such that $\mathcal{N}(\cdot;\theta):[g_c, n_b, M_w, t_p, l_p] \rightarrow m_t$. Similar to the previous task, we design a framework with one deep neural network consisting of two hidden layers with 64 neurons each. The network takes as input one of the five quantities in the input space and output a scalar quantity denoting the solution, m_t. A schematic representation of the framework is shown in Fig. 1. In this configuration, a single network takes the varying conditions as inputs and predicts the desired solution field, with the loss function consisting solely of the data loss. Notably, the data-driven architecture does not include the second network that outputs α or incorporates a residual loss.
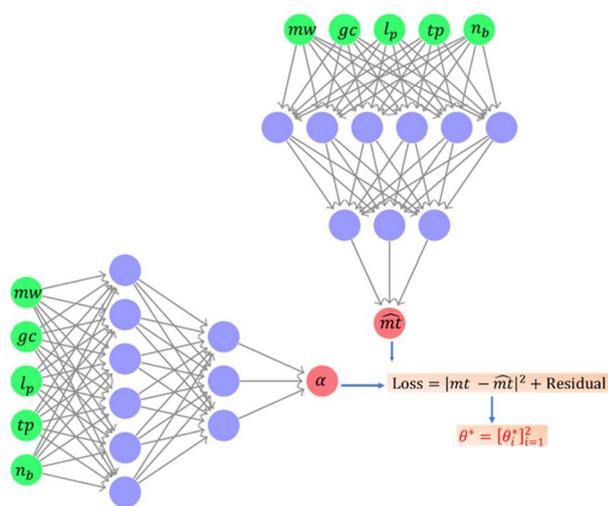
Fig. 1 Schematic representation of the PINNs frameworks, where $\theta*$ denotes the collection of optimized parameters of all the networks. The data-driven framework includes only the neural network that predicts the target quantity and excludes the network responsible for predicting α. Furthermore, the data-driven model is trained solely using data loss, without incorporating a residual loss term.

The network parameters are optimized for both tasks using the – loss function and Adam optimizer. The learning rate for the optimizer is 1×10^{-4}. The primary goal for developing ANN-based surrogate models is to generalize well-known to new and unseen data. However, this is a challenging problem. An under-parametrized model with too little capacity cannot learn the problem. In contrast, an over-parametrized model with too much capacity can learn it too well and overfit the training dataset. For our work, we experience overfitting due to the sparse representation of the input space because of the limited availability of labeled data. One popular approach to improve the generalization of deep neural networks is to use regularization during training that keeps the weights of the model small. These techniques not only reduce overfitting, but they can also lead to faster optimization of the model and better overall performance. In the end, we employed $\mathcal{L}_2$ weights regularization with a regularization coefficient of 9×10^{-3}.

Physics-informed neural networks (PINNs). This section considers the physics of a problem defined by a generic equation (eqn (4)), where b is the pore size, l is the Kuhn length, N_k is the number of Kuhn segments, a is the limiting ratio of free mobility to in-gel mobility for large chains and large fields, and ε_k is the "molecular" reduced field. The insights provided later in this paper demonstrate why eqn (4), describing the biased reptation model, was chosen as the basis of our computation.

$$\frac{\mu}{\mu_0} = \left\{ \left[\frac{\left(\frac{b}{l}\right)^2}{3N_k} \right]^2 + \left[\frac{2\varepsilon_k \left(\frac{b}{l}\right)^2}{5 + 2\alpha\varepsilon_k \left(\frac{b}{l}\right)^2} \right]^2 \right\}^{\frac{1}{2}} \quad (4)$$

For this work, our goal was to obtain the scalar parameter α along with n_b for task 1 and m_t for task 2. For this task, we consider the biased reptation model defined as:

$$\frac{\mu}{\mu_0} = \sqrt{ \frac{\left(\frac{b}{2lp}\right)^4}{8 \frac{n_b^2}{N_l^2}} + \frac{4\varepsilon_k^2 \left(\frac{b}{2lp}\right)^4}{\left[5 + 2\alpha\varepsilon_k \left(\frac{b}{2lp}\right)^2\right]^2} }, \quad (5)$$

where N_l is a constant which is 457 for dsRNA and 14 for ssRNA, μ_0 is the free solution mobility, b is pore size, and p, persistence length, is 64 nm for dsRNA and 2 nm for ssRNA. To that end, we have two deep neural networks consisting of three hidden layers of 8 neurons each and a hyperbolic tan activation function to introduce non-linearity in the PINNs framework (Fig. 1). While both the networks take as inputs all the input space parameters discussed above in the data-driven framework, the first network outputs α, and the second network outputs the task-specific quantity of interest (n_b for task 1 and m_t for task 2). To construct the loss function, we obtain the residual loss from the governing equation and the data loss from the experimental data. The standard Adam optimizer minimizes the loss to obtain the optimized network parameters.

Results and discussion

Electrokinetic behavior of dsRNA and ssRNA

The electrokinetic behavior of migratory nucleic acids is mainly affected by the relative sizing of the molecule and the matrix pore size, the strength of the electric field, and the interactions between the matrix and nucleic acid during electrophoresis. This was demonstrated by the raw migration trends of dsRNA and ssRNA, which were reflective of typical manipulation of gel concentration and fragment size. Fragments of the same size migrated slower as gel concentration increased due to increased viscosity and decreased pore size; however, resolution between longer fragments of nucleic acids significantly decreased at lower gel concentrations (Fig. S1†). Additionally, the migration time of dsRNA and ssRNA of small fragment lengths (<300 bases) were relatively similar and within a 10% difference of each other for respective gel concentrations. As fragment size increased, the faster migration of dsRNA relative to ssRNA became more prominent, especially at higher gel percentages, where there was approximately a 3%, 12%, 51%, 53%, and 71% difference in migration time between the 4001 base fragment size of the two nucleic acids in 1 to 5% gel, respectively. It is also interesting to note that in the observed size ranges, as gel concentration increased from 1% to 5%, the migration difference between the largest and smallest fragment (150 and 4001 fragment length) of dsRNA was less pronounced, with a 28% to 43% difference, than for ssRNA, which had a 28% to 100% difference. This can also be seen visually through how the overall shape of the trend remains constant between the gel percen-

tages as fragment size increases for dsRNA, but the data points span outwards as fragment size increases (Fig. S1a†). In a previous comparison of RNA, ssDNA, and dsDNA, it was also found that single-stranded nucleic acids had better separation and higher resolution over a larger range of sizes compared to dsDNA.[41] This may be due to the greater flexibility of single-stranded molecules and a tendency to form secondary and/or tertiary structures that could cause interactions with the sieving matrix.

The electrophoretic mobility of RNA of varying sizes and type were calculated and demonstrated in Fig. 2. The overall sigmoidal curve demonstrated by the double logarithmic mobility *vs.* size plots agrees with previous DNA and RNA capillary electrophoresis separation studies.[34,42] However, this shape is much less evident for ssRNA, especially at higher gel concentrations (Fig. 2b). Transitions were demarked upon visual inspection with solid gray lines between Ogston-like sieving (regime I), reptation without orientation (regime III), and reptation with orientation (regime IV) with patterns outlined by previous studies.[43,44] Regime II, which was only prominent in dsRNA, marks the regime described by Heller,[34] where may be greater than pore size, but the dsRNA is still too stiff to reptate.

According to the Ogston model,

$$\frac{\mu}{\mu_0} = e^{-K_R \times C} \quad (6)$$

where μ_0 is free solution mobility, K_R represents the retardation coefficient, and c is gel concentration, linear Ferguson plots (logarithm of mobility as a function of gel concentration) would indicate RNA separation within the Ogston sieving regime. RNA is presumed to migrate through the gel matrix in a globular confirmation in a random array of cylindrical obstacles in this region. Interestingly, for dsRNA, none of the fragment lengths demonstrated a strong linear relationship (Fig. 2c). Comparing the straight line fit for 80 bp ($Y = 2.025e^{-0.1024c}$; $R^2 = 0.96$) and 4001 bp ($Y = 1.523e^{-0.1524}$; $R^2 = 0.96$), it appears the Ogston model does not fit the smaller strands any better than the longer fragment, where mobility seems to fail, confirming that dsRNA fragment lengths as small as 80 bp may be too large to sieve through the pores without stretching. For ssRNA, the data points noticeably deviated from a straight line fit of negative slope, with the possible exception of ssRNA of 150 bases (Fig. 2d). These inconsistencies demonstrate that the Ogston mobility model may not be an accurate method of describing the mobility of RNA. It may also indicate that more of the regime I indicated in Fig. 2b may fall under reptation motion. However, definitive delineations cannot be made due to model limitations and without larger magnitude of data points, which is often not possible due to sample and device constraints. Previous studies have also found the Ogston-like sieving mechanism to not accurately describe the mobility of ssRNA, but instead be better for dsDNA and ssDNA.[41] Our study demonstrates that
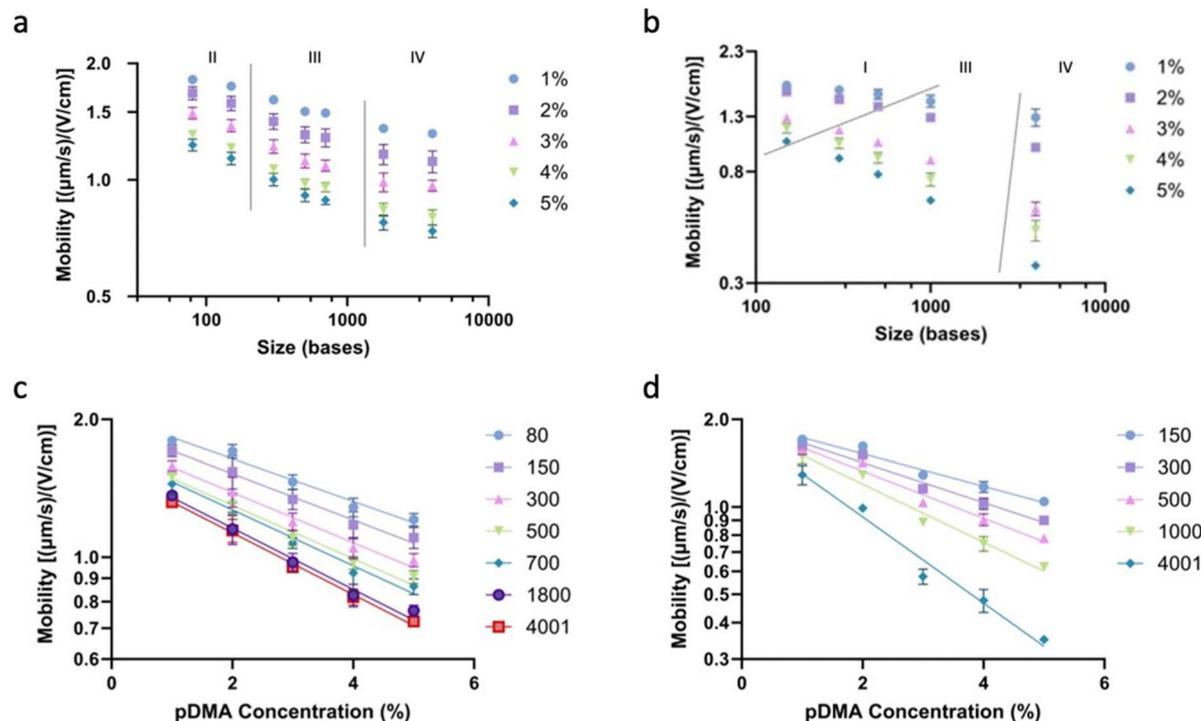

Fig. 2 Double-logarithmic mobility *vs.* size plots for (a) dsRNA and (b) ssRNA at polymer concentrations of 1%, 2%, 3%, 4%, and 5%, at an electric field strength of 417.4 V cm^{-1} where the lines are used to semi-qualitative define the different regimes. The same data is visualized as Ferguson plots for (c) dsRNA and (d) ssRNA and are exponentially fitted to demonstrate potential Ogston motion.

this model limitation may also apply to dsRNA of short lengths in tested condition. This may be due to the many conformations RNA can take on, which may not necessarily be spherical and instead consist of loops, overhangs, and/or other secondary/tertiary structure that are further complicated by the effect of the electric field on stretching and mobility.

Three distinct regions can be seen for dsRNA in Fig. 2a, and with our pore size analysis and the poor fitting of the Ogston model, the first region of the data points may be better described as what Heller concludes as a transition region where is greater than pore size, but the molecules are too still to reptate.[34] As gel concentration increases, this distinction is less prominent, and there seems to be a linear decrease in mobility in Fig. 2a until a plateau in region IV is reached.

Unlike Ogston-like motion, the reptation model was developed for larger DNA due to the assumption that the spherical coil would be too large to fit through the pores of the matrix undeformed and would instead migrate head-first in a "snake-like" motion through "tubes" formed by the polymeric pore networks.[45] Later improved by researchers such as Slater and Viovy, the model, still challenging to utilize mathematically, was modified to account for larger nucleic acid sizes, high electric fields, and the dynamic nature of uncrosslinked polymer networks (eqn (4)).[46]

Regime III can be correlated with regions of reptation without orientation, as seen by the linear decrease on the double logarithm of mobility as a function of fragment length (Fig. 2a and b). Another representation of this delineated region can be found in ESI (Fig. S3†). As eqn (4) demonstrates, the first term dominates for molecules below a critical size (N^* = N_k), and mobility is inversely proportional to fragment size. However, as the N_k becomes larger than N^*, a plateau mobility is reached. The "reptation with orientation" regime that describes separation failure is thought to be partly due to the electric forces leading long fragments to choose a tube of consecutive pores that do not follow a random walk-in space, and therefore resulting mobility is independent of size.[37] The reptation regime has also been explained by other authors as countering the effects of increased charged residues against increased solvent friction due to large nucleic acid size, the latter of which can also be attributed to collisions and subsequent transient dragging of polymer chains.[15,16] This plateau can be seen around 1000 bp for dsRNA (Fig. 2a) and seems to be reached a little later for ssRNA/mRNA in the same condition as shown in the region demarked regime IV in Fig. 2b. The higher critical sizes of single-stranded nucleic acids agree with previous findings.[41] Similar to what was previously noted for ssDNA and dsDNA, this critical size seems to increase with decreasing polymer concentration for ssRNA but remained constant for dsRNA,[34,42] but this is hard to confirm due to gaps in data from sample limitations. However, other studies have found the transition between regimes to be dependent on solution concentration only for RNA and not for ssDNA, which is a potential explanation for the short-lived secondary structures that ssRNA can make, resulting in increased stiffness.[41] As expected, the resolving power for larger frag-

ment sizes is poor in very low gel concentrations, but separation also fails earlier as gel concentration is increased, shown mainly by ssRNA. However, there was a consistent increase in peak width as gel concentration increased for both dsRNA and ssRNA.

Comparison and differentiation of dsRNA and ssRNA

Given that new developments in RNA therapeutics involve a combination of dsRNA and ssRNA/mRNA, it is essential to analyze potential differences in mobility patterns for identification. Given that the unique mix results in a preference for not denaturing the ssRNA due to the possible unraveling of the double-stranded helix of dsRNA, our findings will provide insights into mobility that may be complicated by secondary structures that are typically linearized and subsequently reported in the literature.

Fig. 2a and b demonstrates that the difference in mobility for a single fragment between different gel % is relatively constant for dsRNA but not for single-stranded RNA, which is consistent with the finding of Heller in terms of dsDNA and ssDNA.[34] With extraction and transformation of some data points, a semilogarithmic plot (Fig. 3) was graphed to directly assess the dependence of mobility on pore size for fragment sizes 500 and 4001, which should respectively fall under reptation without orientation and reptation with orientation. The dependence of dsRNA mobility on pore size was similar for both fragment sizes, with slopes of approximately 0.47. However, a considerable difference was seen for the RNA counterpart, suggesting that future developed models or analytical methods may be able to utilize a similar mobility dependence for dsRNA fragments of this size range but not for ssRNA to predict elution time or design experiment parameters.

To analyze the potential difference in mobility between dsRNA and ssRNA in terms of gel concentration, the mobility of both RNA types was plotted against fragment size for each gel condition (Fig. 4). Fig. 4a demonstrates that at low gel concentrations (1% and 2%), the mobility of ssRNA and dsRNA are very similar across all fragment sizes (Fig. 4a), but as gel
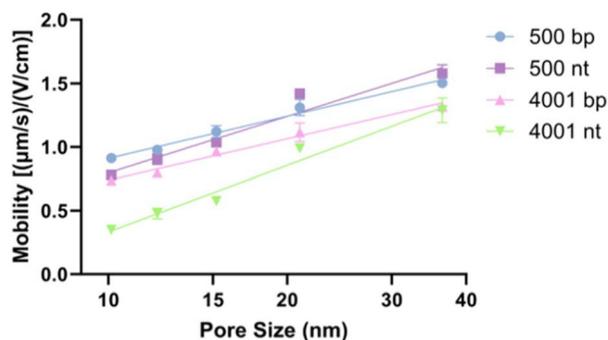

Fig. 3 Dependence of RNA electrophoretic mobility on pore size. dsRNA of 500 bp and 4001 bp, in blue and pink, respectively. ssRNA of 500 nt and 4001 nt, in purple and green, respectively.
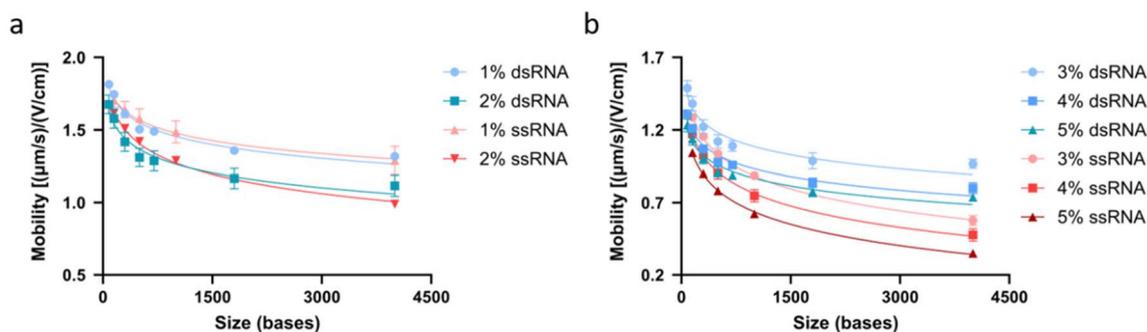
Fig. 4 Mobility *vs.* fragment size of dsRNA and ssRNA in (a) 1–2% gel and (b) 3–5% gel.

percentages increase, there is a clear increase in dsRNA mobility compared to that of ssRNA (Fig. 4b). This difference becomes more pronounced as fragment size increases. The mobility of dsRNA and ssRNA in 1% gel is approximately equal (<5% difference). For gel concentrations greater than 2%, the difference in mobility increases between dsRNA and ssRNA as fragment size increases above ~500 bases for each gel percentage (Fig. 4b), before which the mobility difference is less than 10%. Additionally, as gel percentages increase from 1% to 5%, the difference in mobility for the largest recorded fragment also increases from 2.2% to 60.5%. This trend was also seen when analyzing raw migration time, demonstrating the potential manipulation of higher gel percentages to increase the sep-

aration between dsRNA and ssRNA. This manipulation of gel concentration, and therefore pore size, may be a simple target for high throughput separation of RNA mixture products that then allow for identification, such as in the context of quality control operations.

Impact of chemical modifications on RNA mobility

Using pseudouridine (Ψ) instead of uridine in the IVT mRNA has been critical to therapeutic mRNA efficacy and stability.[22,23] With the reliance on modified mRNA for future therapeutics, it will be essential to characterize if these modifications create a significant impact on the mobility and subsequent characterization of RNA as structurally, Ψ can alter
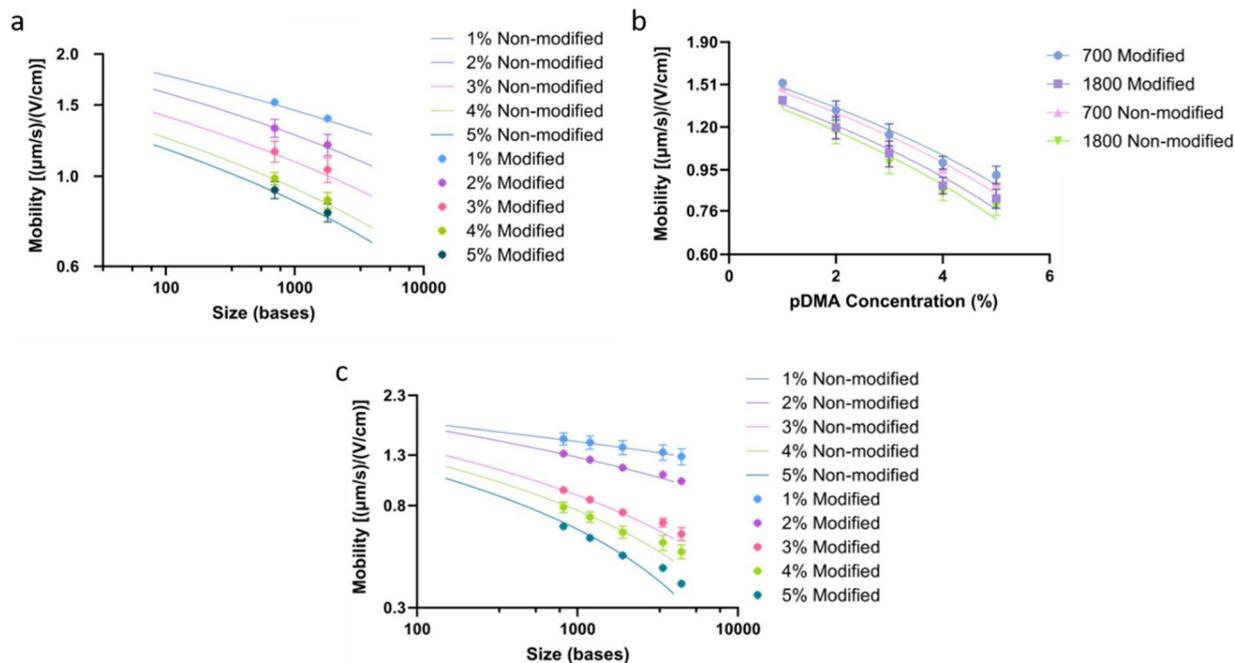
Fig. 5 Mobility comparison between the non-chemically modified and chemically modified fragments at 1, 2, 3, 4, and 5% pDMA concentrations. (a) Double logarithmic plot of mobility *vs.* size with the modified dsRNA fragments represented by markers overlayed onto the line of best fit formed from the mobility of its non-modified counterpart. (b) Comparison of modified and non-modified dsRNA mobility of equal sizes in varying PDMA concentrations. (c) Double logarithmic plot of mobility *vs.* size with the modified ssRNA fragments represented by markers overlayed onto the line of best fit formed from the mobility of its non-modified counterpart.

RNA structure by improving base-pairing, base stacking, and contributing to making the backbone more rigid (through a network of hydrogen bonding interactions).[23]

The migration (Fig. S1a and S2a†) and effective mobility (Fig. 5a and b) of modified dsRNA are very similar to those of its non-modified counterpart, and the difference between mobility is not statistically significant (Fig. 5a and b). Although a similar direct comparison cannot be made for ssRNA due to sample constraints, the migration profiles (Fig. S1b and S2b†) are similar, and the effective mobility of modified RNA fits nicely into the trends seen by its non-modified counterpart (Fig. 5c). The type and extent of modification RNA undergoes will vary depending on the desired application, but these findings demonstrate that resulting electrophoretic analysis may not be significantly impacted.

Machine learning modeling for electrophoretic mobility prediction

Following a thorough experimental study of RNA mobility in microfluidic electrophoresis, there lacks a comprehensive model that can provide clear prediction of the observed phenomena. While applications of the pre-existing models allowed for some insights and connections into biophysical mechanisms of RNA traveling in semi-dilute polymer networks, they are insufficient to provide accurate predictions of the characteristics of the nucleic acids such as size, type, and potential detection time. These are parameters often essential in creating analytical methods to provide adequate separation, resolution, and identification of nucleic acid mixtures, causing there to be significant experimental trial and error to find optimal separation parameters.

While the BRF model allows for some descriptive characterization of mobilities, the actual mathematical equation is very difficult to utilize in practice for the prediction of nucleic acid migratory behavior. Therefore, the use of ANNs were employed, including a PINNs,[33] where the networks are trained using a modified loss function, which includes the governing equation and the training data. Table 1 demonstrates the results of the developed PINN, where the predicted values of n_b obtained for ssRNA and dsRNA test samples are compared against the ground truth, represented by the n_b on the axes.

Similarly, in Table 2, we highlight the predicted migration time of ssRNA and dsRNA test samples and contrast them to the ground truth.

Results from Tables 1 and 2 demonstrate the feasibility of utilizing artificial neural networks, specifically PINNs, to predict nucleic acid characteristics such as migration time and size based on other known parameters with decent accuracy. These values could aid in method development in ways such as guiding decisions on assay parameters to prevent signal peaks overlap, differentiating nucleic acids of different types, characterizing nucleic acid size with limited ladder sample, automating devices, *etc*.

Table 3 summarizes the relative $\mathcal{L}_2$ error (%) computed for five test samples (details provided in Table 4) for both the frameworks of ANNs and both tasks. Test samples refer to the cases that the network was not provided with during the training. The test samples were chosen randomly from the available dataset. For both tasks, the PINNs model performs better than the data-driven model, and both frameworks perform better at task 1 than task 2.

As expected, given the relatively limited data set, the PINNs model significantly outperforms the data-driven model, as it can compensate for the limited data through the physical equations that govern the experiments. This application of PINNs highlights its potential for electrophoretic analysis, which could far exceed the analysis of single- and double-stranded RNA molecules. Finally, we computed the value of from the PINNs framework, and report $\alpha = 192.77$. This is essentially an inverse problem, where the network can predict the value given the minimization of the residual of the govern-

Table 1 Comparative results for the predicted lengths of test samples of ssRNA and dsRNA for the PINNs framework for task 1

	1	2	3	4	5
	ssRNA				
500	486.9	486.9	486.9	486.9	486.9
1000	988.7	988.7	988.7	988.7	988.7
2000	1992.3	1992.3	1992.3	1992.3	1992.3
3000	2995.8	2995.8	2995.8	2995.8	2995.8
4000	4002.1	4002.1	4002.1	4002.1	4002.1
	dsRNA				
500	521.4	521.4	521.4	521.4	521.4
1000	1008.6	1008.6	1008.6	1008.6	1008.6
2000	1984.3	1984.3	1984.3	1984.3	1984.3
3000	2996.8	2996.8	2996.8	2996.8	2996.8
4000	3996.8	3996.8	3996.8	3996.8	3996.8

The vertical axes labeled n_b represents the ground truth.

Table 2 Comparative results for the migration time of test samples of ssRNA and dsRNA for the PINNs framework for task 2

	ssRNA					dsRNA				
	1	2	3	4	5	1	2	3	4	5
500	24.7 (24.3)	28.6 (27.0)	38.2 (37.0)	45.9 (42.5)	53.6 (49.1)	23.2 (25.2)	25.8 (28.6)	32.7 (33.2)	38.3 (38.5)	43.8 (41.3)
1000	27.1 (25.8)	32.6 (29.7)	46.3 (43.3)	57.3 (51.5)	68.2 (61.7)	23.8 (26.5)	27.1 (30.8)	35.3 (35.9)	41.8 (42.0)	48.3 (45.2)
2000	29.8 (27.9)	36.2 (34.3)	55.2 (55.4)	69.8 (67.6)	84.3 (86.5)	24.6 (28.1)	28.3 (32.9)	37.6 (38.5)	45.1 (45.5)	52.5 (49.1)
3000	31.2 (28.8)	39.4 (36.2)	60.0 (60.0)	76.5 (73.9)	93.0 (95.6)	24.9 (28.9)	28.9 (34.2)	38.9 (40.1)	46.8 (47.5)	54.8 (51.4)
4000	32.3 (29.4)	41.3 (37.5)	63.8 (63.2)	81.8 (78.3)	99.8 (102.1)	25.2 (29.5)	29.3 (35.1)	39.7 (41.2)	47.9 (49.0)	56.2 (53.0)

The value in brackets is the ground truth.

Table 3 Relative $\mathcal{L}_2$ error (%) computed for five test samples for both frameworks of ANNs for both tasks

	Task 1	Task 2
Data-driven	9.56	15.12
PINNs	0.44	1.1

Table 4 Details of the test samples used to evaluate the accuracy of the ML models

Samples	n_b	g_c	M_w	t_p	l_p	m_t
1	1000	0.01	320 659	1	2	25.84166667
2	500	0.04	160 409	1	2	42.46833333
3	1800	0.03	1 154 118	2	64	38.9425
4	80	0.02	51 598	2	64	22.89
5	700	0.02	449 018	1	2	29.855

ing equation. We have employed the data generated using the data augmentation approach in the PINNs model. In engineering and biomedical problems, data collection using numerical or physical experiments is often expensive and time-consuming. This study demonstrates that ANNs can be a useful tool for determining complex relationships where governing solutions are not clearly known or when there is insufficient information regarding the relationship between input and outputs, as seen through our discussion on nucleic acid electrophoretic mobility.

Conclusion

An in-depth study of the electrokinetic behavior and resulting mobility patterns of dsRNA and ssRNA was conducted, with the desired application of informing future assay protocols and diagnostics. Characterization of RNA will be crucial given the recent advancements in RNA therapeutic development and the subsequent need for quality assurance, primarily due to the potential for RNA degradation during the formulation and storage processes. While current microfluidic chip-based devices for capillary gel electrophoresis face the limitations of providing high-resolution for large RNA, our study was able to characterize RNA mobility for up to around 4000 bp, which includes the average mRNA length of the human genome as well as the majority of current non-self-amplifying constructs.[47,48] The Ogston-like sieving motion did not describe our data well, indicating limited applicability to RNA of both single-stranded and double-stranded nature above 80 bases. Our resulting mobility *vs.* size graphs may provide guidance for future RNA mobility studies in semi-dilute polymer solution as well as references for mobility of long RNA, dsRNA, and modified RNA, which have not been well studied previously. Points of differentiation between dsRNA and ssRNA mobility in entangled polymer solution were identified and may be coupled with other identification methods, as described by our previous

work, for complete characterization. The potential to exploit increased gel concentrations to exacerbate mobility differences between dsRNA and ssRNA was presented. Chemically modified pseudouridine constructs demonstrated no statistical mobility difference from the non-modified strands, indicating that the developed mobility models could potentially be applied to modified and non-modified RNA molecules alike.

Following a thorough experimental study of RNA mobility in microfluidic electrophoresis, it was clear that there lacks a comprehensive model that can be utilized to provide clear prediction and/or observation of RNA mobility under varying conditions. Real experimental conditions often result in additional complications that static equations are not able to adapt or account for. Due to the limitations of current models and the importance of the prediction of RNA size and migration time for assay development, ANNs were developed to with the intent of guiding future therapeutic develop and analysis with decreased experimental trial and error for method development. It was believed that with adaption of the model through modulation of the parameter and hidden layers of the neural network, adequate prediction may be made. This was proven by the low margin of error of final predications of the PINN.

Limitations of the developed model include generalizability only to the ranges and conditions used to train the model. While the tested RNA lengths can account for the sizes of most current RNA therapies, the rapidly developing field of RNA therapeutics could soon include far greater sizes. The validation of the model was done with limited samples due to sample constraints, but we believe that the developed model proves such methodology is feasible for the prediction of complex relationships seen in microfluidic electrophoresis and with biopharmaceutical development. As with our discussion on chemically modified RNA, we hope that future studies with greater variety of RNA samples can better validate and strengthen our discussion of predictive models and their applicability.

The findings presented in this study serve as an example of how ANNs, particularly PINNs, can be used to complement limited data sets for assessing the electrophoretic behavior of single- and double-stranded RNA molecules. In this study, using PINNs improved the relative error from 9.56% using the data-driven model in task 1 to 0.44% and from 15.12% to 1.1% in task 2. The developed PINN has applications in streamlining development of RNA therapeutics analytical methods through its ability to determine optimal run conditions for given mRNA constructs lengths, preventing potential overlapping of RNA target strand with potential impurities, determining contaminant (*i.e.*, dsRNA or truncated ssRNA) length given the run conditions and migration time, *etc*. This added capability can have significant implications for the biopharmaceutical industry, where machine learning can help streamline the development process, which can involve countless permutations within a given product.

Author contributions

These authors contributed equally: N. S. L., A. C. D. P. Conceptualization, N. S. L., A. C. D. P., and S. G.; data curation, A. C. D. P. and N. S. L.; formal analysis, N. S. L., A. C. D. P., and S. G.; funding acquisition, A. T., A. C. D. P., and S. G.; investigation, A. C. D. P., N. S. L., and M. V.; methodology, N. S. L., A. C. D. P. and S. G.; project administration, A. C. D. P. and A. T.; resources, A. T.; software, S. G.; supervision, A. C. D. P. and A. T.; validation, A. C. D. P., N. S. L., and S. G.; visualization, A. C. D. P., N. S. L., and S. G.; writing – original draft, N. S. L., A. C. D. P., and S. G.; writing – review & editing, N. S. L., A. C. D. P., S. G., and A. T.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

The online version of this article contains ESI† available to authorized users. The ESI† include the following: (1) ESI PDF file including "Additional background/methods: artificial neural networks", "Fig. S1: Non-chemically modified RNA migration *vs.* size", "Fig. S2: Chemically modified RNA migration *vs.* size", and "Fig. S3: Mobility regimes" and (2) an additional file including the raw experimental data and associated calculated mobility for all experiments.

Acknowledgements

References

1 X. Ou, P. Chen, X. Huang, S. Li and B.-F. Liu, Microfluidic chip electrophoresis for biochemical analysis, *J. Sep. Sci.*, 2020, **43**, 258–270.

2 M. A. A. Ragab and E. I. El-Kimary, Recent Advances and Applications of Microfluidic Capillary Electrophoresis: A Comprehensive Review (2017–Mid 2019), *Crit. Rev. Anal. Chem.*, 2021, **51**, 709–741.

3 J. Raffaele, J. W. Loughney and R. R. Rustandi, Development of a microchip capillary electrophoresis method for determination of the purity and integrity of mRNA in lipid nanoparticle vaccines, *Electrophoresis*, 2022, **43**, 1101–1106.

4 A. C. De Peña, N. Li, M. Vaduva, L. Bwanali and A. Tripathi, A microfluidic electrophoretic dual dynamic staining method for the identification and relative quantitation of dsRNA contaminants in mRNA vaccines, *Analyst*, 2023, **148**, 3758–3767.

5 G. G. Diaz-Armas, A. P. Cervantes-Gonzalez, R. Martinez-Duarte and V. H. Perez-Gonzalez, Electrically driven microfluidic platforms for exosome manipulation and characterization, *Electrophoresis*, 2022, **43**, 327–339.

6 S. Marczak, K. Richards, Z. Ramshani, E. Smith, S. Senapati, R. Hill, D. B. Go and H. C. Chang, Simultaneous isolation and preconcentration of exosomes by ion concentration polarization, *Electrophoresis*, 2018, DOI: **10.1002/elps.201700491**.

7 J. Ghanam, *et al.*, DNA in extracellular vesicles: from evolution to its current application in health and disease, *Cell Biosci.*, 2022, **12**, 1–13.

8 S. Ayala-Mar, J. Donoso-Quezada, R. C. Gallo-Villanueva, V. H. Perez-Gonzalez and J. González-Valdez, Recent advances and challenges in the recovery and purification of cellular exosomes, *Electrophoresis*, 2019, **40**, 3036–3049.

9 T. Kirby, mRNA vaccine technology for a multivalent flu vaccine, *Lancet Infect. Dis.*, 2023, **23**, 157.

10 G. Sanyal, Development of functionally relevant potency assays for monovalent and multivalent vaccines delivered by evolving technologies, *npj Vaccines*, 2022, **7**, 1–10.

11 M. N. Gerold, *et al.*, Analytical Performance of a Multiplexed Microarray Assay for Rapid Identification and Quantification of a Multivalent mRNA Vaccine, *Vaccines*, 2024, **12**, 1144.

12 J.-L. Viovy and T. Duke, DNA electrophoresis in polymer solutions: Ogston sieving, reptation and constraint release, *Electrophoresis*, 1993, **14**, 322–329.

13 A. G. Ogston, The spaces in a uniform random suspension of fibres, *Trans. Faraday Soc.*, 1958, **54**, 1754–1757.

14 G. W. Slater and J. Noolandi, New Biased-Reptation Model For Charged Polymers, *Phys. Rev. Lett.*, 1985, **55**, 1579–1582.

15 A. E. Barron, W. M. Sunada and H. W. Blanch, Capillary electrophoresis of DNA in uncrosslinked polymer solutions: Evidence for a new mechanism of DNA separation, *Biotechnol. Bioeng.*, 2000, **52**, 259–270.

16 C. Liu, X. Xu, Q. Wang and J. Chen, Mathematical model for DNA separation by capillary electrophoresis in entangled polymer solutions, *J. Chromatogr. A*, 2007, **1142**, 222–230.

17 J. A. Luckey and L. M. Smith, A model for the mobility of single-stranded DNA in capillary gel electrophoresis, *Electrophoresis*, 1993, **14**, 492–501.

18 E. Stellwagen, Y. Lu and N. C. Stellwagen, Unified description of electrophoresis and diffusion for DNA and other polyions, *Biochemistry*, 2003, **42**, 11745–11750.

19 C. Liu, *et al.*, Analysis of small interfering RNA by capillary electrophoresis in hydroxyethylcellulose solutions, *Electrophoresis*, 2015, **36**, 1651–1657.

20 H. Na, B.-H. Kang, J. Ku, Y. Kim and K.-H. Jeong, On-chip Paper Electrophoresis for Ultrafast Screening of Infectious Diseases, *BioChip J.*, 2021, **15**, 305–311.

21 I.-C. Yeh and G. Hummer, Diffusion and Electrophoretic Mobility of Single-Stranded RNA from Molecular Dynamics Simulations, *Biophys. J.*, 2004, **86**, 681–689.

22 E. Dolgin, The tangled history of mRNA vaccines, *Nature*, 2021, **597**, 318–324.

23 P. Morais, H. Adachi and Y.-T. Yu, The Critical Contribution of Pseudouridine to mRNA COVID-19 Vaccines, *Front. Cell Dev. Biol.*, 2021, **9**, 789427.

24 N. Pardi, M. J. Hogan and D. Weissman, Recent advances in mRNA vaccine technology, *Curr. Opin. Immunol.*, 2020, **65**, 14–20.

25 K. Karikó, *et al.*, Incorporation of Pseudouridine Into mRNA Yields Superior Nonimmunogenic Vector With Increased Translational Capacity and Biological Stability, *Mol. Ther.*, 2008, **16**, 1833–1840.

26 J. Zou, Y. Han and S.-S. So, Overview of Artificial Neural Networks, in *Artificial Neural Networks: Methods and Applications*, ed. D. J. Livingstone, Humana Press, Totowa, NJ, 2009, pp. 14–22.

27 J. G. Greener, S. M. Kandathil, L. Moffat and D. T. Jones, A guide to machine learning for biologists, *Nat. Rev. Mol. Cell Biol.*, 2022, **23**, 40–55.

28 Y. Lecun, Y. Bengio and G. Hinton, Deep learning, *Nature*, 2015, **521**, 436–444.

29 W. S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.*, 1943, **5**, 115–133.

30 M. Raissi, P. Perdikaris and G. E. Karniadakis, Machine learning of linear differential equations using Gaussian processes, *J. Comput. Phys.*, 2017, **348**, 683–693.

31 S. H. Rudy, S. L. Brunton, J. L. Proctor and J. N. Kutz, Data-driven discovery of partial differential equations, *Sci. Adv.*, 2017, **3**, DOI: 10.1126/sciadv.1602614.

32 G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang and L. Yang, *Nat. Rev. Phys.*, 2021, **3**, 422–440.

33 M. Raissi, P. Perdikaris and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving non-linear partial differential equations, *J. Comput. Phys.*, 2019, **378**, 686–707.

34 C. Heller, Separation of double-stranded and single-stranded DNA in polymer solutions: I. Mobility and separation mechanism, *Electrophoresis*, 1999, **20**, 1962–1977.

35 G. Oliver, C. Simpson, M. B. Kerby, A. Tripathi and A. Chauhan, Electrophoretic migration of proteins in semi-dilute polymer solutions, *Electrophoresis*, 2008, **29**, 1152–1163.

36 G. W. Slater, J. Noolandi and A. Eisenberg, Radius of gyration of charged reptating chains in electric fields, *Macromolecules*, 1991, **24**, 6715–6720.

37 G. W. Slater, J. Rousseau, J. Noolandi, C. Turmel and M. Lalande, Quantitative analysis of the three regimes of DNA electrophoresis in agarose gels, *Biopolymers*, 1988, **27**, 509–524.

38 J. A. Abels, F. Moreno-Herrero, T. van der Heijden, C. Dekker and N. H. Dekker, Single-molecule measurements of the persistence length of double-stranded RNA, *Biophys. J.*, 2005, **88**, 2737–2744.

39 H. Chen, *et al.*, Ionic strength-dependent persistence lengths of single-stranded RNA and DNA, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 799–804.

40 E. Herrero-Galán, *et al.* Mechanical Identities of RNA and DNA Double Helices Unveiled at the Single-Molecule Level, *J. Am. Chem. Soc.*, 2013, **135**, 122–131.

41 T. I. Todorov and M. D. Morris, Comparison of RNA, single-stranded DNA and double-stranded DNA behavior during capillary electrophoresis in semi-dilute polymer solutions, *Electrophoresis*, 2002, **23**, 1033–1044.

42 T. Lu, L. J. Klein, S. Ha and R. R. Rustandi, High-Resolution capillary electrophoresis separation of large RNA under non-aqueous conditions, *J. Chromatogr. A*, 2020, **1618**, 460875.

43 M. Chung, D. Kim and A. E. Herr, Polymer sieving matrices in microanalytical electrophoresis, *Analyst*, 2014, **139**, 5635–5654.

44 H. N. Clos and H. Engelhardt, Separations of anionic and cationic synthetic polyelectrolytes by capillary gel electrophoresis, *J. Chromatogr. A*, 1998, **802**, 149–157.

45 O. J. Lumpkin, P. Déjardin and B. H. Zimm, Theory of gel electrophoresis of DNA, *Biopolymers*, 1985, **24**, 1573–1593.

46 G. W. Slater, DNA gel electrophoresis: The reptation model (s), *Electrophoresis*, 2009, **30**, S181–S187.

47 M. Bhattacharya, *et al.*, Bioengineering of Novel Non-Replicating mRNA (NRM) and Self-Amplifying mRNA (SAM) Vaccine Candidates Against SARS-CoV-2 Using Immunoinformatics Approach, *Mol. Biotechnol.*, 2022, **64**, 510–525.

48 NCBI News: Spring 2003 | Human Reference Sequence. https://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring03/human.html.