npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Neural network learns physical rules for copolymer translocation through amphiphilic barriers

Marco Werner[1,2,4]✉, Yachong Guo[2,3,4]✉ and Vladimir A. Baulin[2]✉

Recent developments in computer processing power lead to new paradigms of how problems in many-body physics and especially polymer physics can be addressed. Parallel processors can be exploited to generate millions of molecular configurations in complex environments at a second, and concomitant free-energy landscapes can be estimated. Databases that are complete in terms of polymer sequences and architecture form a powerful training basis for cross-checking and verifying machine learning-based models. We employ an exhaustive enumeration of polymer sequence space to benchmark the prediction made by a neural network. In our example, we consider the translocation time of a copolymer through a lipid membrane as a function of its sequence of hydrophilic and hydrophobic units. First, we demonstrate that massively parallel Rosenbluth sampling for all possible sequences of a polymer allows for meaningful dynamic interpretation in terms of the mean first escape times through the membrane. Second, we train a multi-layer neural network on logarithmic translocation times and show by the reduction of the training set to a narrow window of translocation times that the neural network develops an internal representation of the physical rules for sequence-controlled diffusion barriers. Based on the narrow training set, the network result approximates the order of magnitude of translocation times in a window that is several orders of magnitude wider than the training window. We investigate how prediction accuracy depends on the distance of unexplored sequences from the training window.

npj Computational Materials (2020)6:72 ; https://doi.org/10.1038/s41524-020-0318-5

INTRODUCTION

Polymers are many-body physical objects; in order to describe their equilibrium state and dynamics, it is often required to translate chemical sequence information into free-energy landscapes in three-dimensional space. Rigorous theoretical descriptions can capture only special cases such as homopolymers or multiblock copolymers by following bottom–up approaches that start with interactions on the monomer level or considering the self-similarity of self-avoiding walks on the largest scales. The sequence space available by current polymer chemistry[1–3] or in biopolymers exceeds the limits for closed physical descriptions and is not accessible for complete scans by molecular simulation techniques. A new paradigm of data-driven polymer science is increasingly encouraged by parallel sampling methods[4,5] and the advances in machine learning (ML)[6–14] and has the potential to explore yet undiscovered patterns in sequence–property relationships.

A prominent problem for sequence-controlled polymers is their transport through lipid membranes and biological barriers, which is linked to a wide field of potential biomedical and biotechnological applications. The translocation time of polymer chains through a narrow nano-pore on the scale of one monomer has been described for homopolymers[15,16] by means of scaling relations and, later on, extended the theory to block copolymers[17,18]. As soon as local conformation entropy of the polymer comes into play by widening the pore to a finite diameter and length[19,20], a general expression as a function of the sequence seems challenging in the moment for both charged and uncharged polymers. The absence of a closed analytic theory for sequence-controlled translocation meanwhile does not exclude technical applications of nano-pores for DNA sequencing[21–23].

The picture is similar when considering the translocation of a polymer through a lipid membrane by direct penetration of the membrane's core. Here polymer translocation can be considered as the diffusion of its center of mass along an effective free-energy landscape determined by the self-assembled membrane environment[24–26]. Translocation of homopolymers through bilayer membranes was recently described theoretically by means of propagators as the solution of Edwards equation[27] in good agreement with coarse grained simulations[28]. Simulation results on random copolymers indicate that the main factors for copolymer translocation are their average hydrophobicity as well as their degree of adsorption[29,30] at the membrane–solvent interfaces[31], which shall be reflected in the main modes of their potential of mean force. Experimentally, the passive translocation of synthetic random copolymers into mammalian cells in the absence of cytotoxic effects was discovered[32] and confirmed recently[33,34]. A rigorous theoretical description as a function of sequence, however, is missing to date. The lack of theory does meanwhile not exclude the recent progress in finding artificial cell-penetrating peptides and antimicrobial peptides by high-throughput screening[35–38] that may even outperform evolutionary highly conserved Tat- or Penetratin-based sequences for biomedical application. Wimley et al. found that fine-tuned differences in short-block amphiphilic sequences have a significant effect on peptide translocation rates following rules that seems not obvious at the moment[37]. In turn, sequences leading to optimal points in their biomedical performance can be found in unexpected corners of sequence space that are potentially accessed by sequence–cargo co-evolution[38]. In this work, we make use of a massively parallel sampling of polymer conformations while scanning the full sequence space of a short copolymer and see

[1]Institute Theory of Polymers, Leibniz-Institut für Polymerforschung Dresden e.V., Hohe Straße 6, 01069 Dresden, Germany. [2]Departament d'Enginyeria Química, Universitat Rovira i Virgili, 26 Av. dels Paisos Catalans, 43007 Tarragona, Spain. [3]Kuang Yaming Honors School, Nanjing University, 22 Hankou Road, 210093 Nanjing, China. [4]These authors contributed equally: Marco Werner, Yachong Guo. ✉email: werner-marco@ipfdd.de; yguo@nju.edu.cn; vladimir.baulin@urv.cat

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

that the intricacies of polymer sequence–property relationships can be subtle and unexpected already when considering relatively simple environments.

The laws of physics are, however, normally simple by means of requiring a relatively small number of parameters that can be extracted efficiently from high-dimensional data by ML methods. Artificial neural networks (NNs) have been successfully applied in the dimensionality reduction from chemical monomer composition of polymers to their material properties, such as glass transition temperature[39–43], viscosity[44], solvation free energies[45], and electronic properties[13,14] depending on the repeat units. When addressing long polymer chains, the training and test data are fundamentally limited to a fraction of sequences due to the exponential increase of sequence space. Any restriction or bias in the training data has yet undefined consequences for the NN's projection into unexplored parts of sequence space. Efficient classification and optimization algorithms, such as based on artificial NNs, are in fact "black boxes" and their results therefore need better understanding and explanation. Theoretically, an NN can approximate any continuous mapping given that at least one hidden layer of neurons with sigmoid activation functions is contained[46,47]. Beyond that, the stacking of non-linear filters seems to mark a qualitative difference as compared to shallow ML algorithms such that they may develop internal representations of the input information that correspond to a hierarchy of abstraction levels. The distinguished generalization performance makes so-called deep neural nets particularly efficient when confronted with multiple tasks[48] simultaneously, for instance, in finding quantitative structure–property or structure–activity relationships[49–52]. Recent advances in exploiting NNs for physical problems show that they can help to determine the essential order parameters necessary for predicting a mechanical state in future[10] or classifying a magnetic phases[8].

In this work, we have the luxury to access a complete sequence-to-property map available for training and testing NN algorithms thanks to the graphics processing unit (GPU)-accelerated[4] sampling of random polymer configurations for a given sequence. GPU-accelerated Rosenbluth–Rosenbluth[53] sampling of a copolymer in an external field modeling a lipid membrane allowed us to generate a significant number of configurations for all possible binary sequences for chain length up to $N = 16$. Based on this *unbiased* data, a NN is trained to predict mean first escape times of the polymer through the layer. By systematic selection of a training set, we tested the NN's performance of projection into unseen parts of the complete sequence space.

The rest of the paper is structured as follows: In section "Results," we introduce the sequence-complete sampling data set based on the Rosenbluth–Rosenbluth method for translocation time prediction. By comparison with free-energy estimates of self-avoiding walks near interfaces, we underline the physical meaning and richness of the results. We also analyze the performance of NN based on translocation time prediction for two different training schemes. In section "Discussion," we summarize the results. In section "Methods," we describe the polymer conformation sampling for estimating its translocation time through a membrane as well as the NN model applied.

RESULTS

Rosenbluth–Rosenbluth sampling

Let us consider the inverse mean first escape time $1/\tau$ as a measure for the frequency of translocation of a polymer through the membrane, which is presented in Fig. 1 as a function of the mean hydrophobic fraction along a backbone of $N = 12$ monomers. Results for all sequences are shown and grouped into point clouds centered at the corresponding ratios N_T/N. The point clouds are shaped according to the number, n_b, of blocks of H and

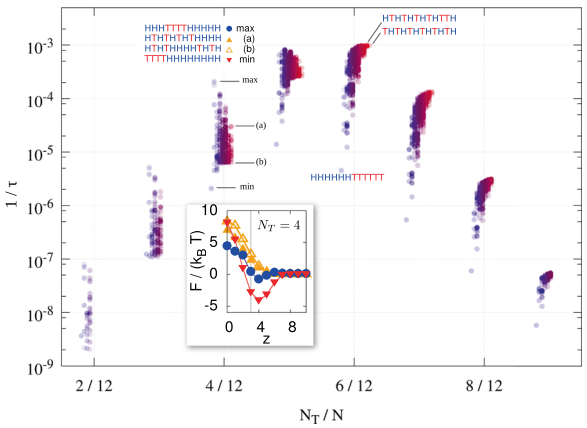

Fig. 1 Translocation frequency vs. fraction of hydrophobic monomers. Inverse mean first escape times (Eq. (7)) as a function of the fraction of hydrophobic monomers of a polymer of length $N = 12$. Results are shown for all sequences containing between two and nine T-type monomers. Results sharing the same number of T-type monomers are spread within windows of width 0.04 along the ordinate according to the number, n_b, of H and T blocks within the sequence. The exact position along the ordinate is calculated as $N_T/N + 0.04 \times [n_b/N - 1/2]$. Results for seven sequences are highlighted by labels.

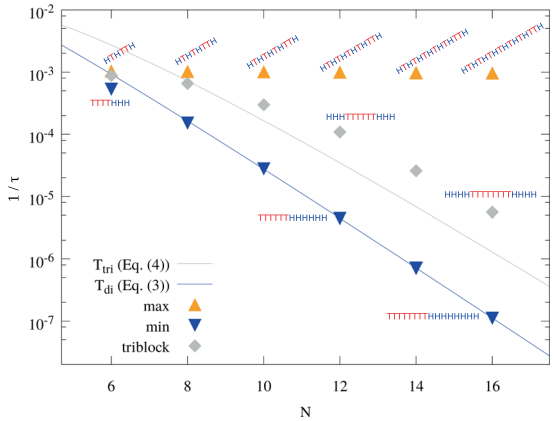

Fig. 2 Translocation frequency vs. chain length. Inverse mean first escape times as a function of the chain length for fractions of hydrophobic monomers of 1/2. We show results for sequences leading to maximal (minimal) inverse escape times.

T species along the sequence in a way that the points on the right hand side of a cloud represent a polymer with a larger number of blocks.

The results in Fig. 1 confirm earlier predictions[28] that a maximum of translocation frequency is found near a point of balanced hydrophobicity of the polymer as given by a balanced fraction of H and T units $N_T/N \sim 1/2$, in case that the typical block size is in the order of the Kuhn segment of the polymer[31].

In Fig. 2, we show the monomer sequences leading to the largest and lowest translocation frequency $1/\tau$ as well as the results for triblock copolymers as a function of chain length for the balanced ratio $N_T/N = 1/2$ of hydrophobic beads. For alternating sequences, the re-scaled translocation frequency (see Eq. (7)) remains in the same order of magnitude showing that the polymers are below the adsorption threshold for the given chain lengths. For polymers that are significantly localized at the membrane–solvent interface, one would expect that the desorption to be the rate-limiting process for translocation. Adsorption effect is clearly visible for diblock copolymers showing a nearly exponential decay of translocation frequency as a function of

chain length. For diblock copolymers, we expect that the desorption of the hydrophobic block from the membrane is the most significant rate-limiting process, and consequently diblock sequences lead to minimal translocation frequencies. It is important to notice that, for hydrophilic blocks larger than the membrane width, the switch of a hydrophilic end from one solvent side to the opposing solvent does only require a limited number of hydrophilic beads to be in contact with the lipid core at the same time, whereas the escape of the hydrophobic block into the solvent requires all monomers of the block to be displaced into solvent environment. Dynamic barriers such as the steric hindrance of the polymer backbone by lipid tails, is, however, not included in the mean-field environment.

In Fig. 2, it becomes visible that the symmetry of the polymer sequence with respect to hydrophilic ends adds an important factor to the desorption probability, in particular when comparing results for triblock copolymers where the longest chains show a more than one decade larger translocation frequency as compared to diblocks. The difference can be understood qualitatively by estimating the adsorption free energy in the strong segregation limit as

$$\Delta F_{\mathrm{ads}}(N) = -c\epsilon N_{\mathrm{T}} + \Delta F_{\mathrm{el}} \tag{1}$$

where c is the average number of favored contacts a T monomer finds in the lipid environment (coordination number) and $\Delta F_{\mathrm{el}} = -k_{\mathrm{B}}T \ln\left[Z_{\mathrm{surf}}/Z_{\mathrm{free}}\right]$ is an elastic contribution due to the reduction of the partition function from Z_{free} to Z_{surf} upon localization at the surface. The partition sum for a self-avoiding walk takes the form[58–60]

$$Z(N) \sim q\mu^N N^{\gamma-1} \tag{2}$$

where q is a non-universal amplitude that may depend on the particular form of short-range interactions and μ is the effective coordination number for the given random walk logic and lattice. The exponent γ depends on the topology of the polymer that is either in free solution or attached to a surface. One applies $\gamma \equiv \gamma_1 \approx 0.678$[61–63] for strands having one end grafted, and $\gamma \equiv \gamma_{11} \approx -0.39$[62,63] for strands having both ends surface attached. The partition sum in free solution scales as $Z_{\mathrm{free}} \sim \mu^N N^{\gamma_0-1}$ with $\gamma_0 \approx 1.1567$[64–66]. Since we further compare only ratios of partition sums for given total chain length, we assume that q- and μ-dependent contributions cancel up to a factor of the order unity.

The probability density to find a symmetric diblock copolymer in bulk solvent as compared to a state adsorbed at an interface as illustrated in Fig. 3 then reads

$$p_{\mathrm{di}}(N) = \exp(\beta\Delta F_{\mathrm{ads}}) = \mathrm{e}^{-\beta c\epsilon N/2}\,\frac{N^{\gamma_0-1}}{(N/2)^{2(\gamma_1-1)}}$$

Now, assuming that the desorption is the rate-limiting process, we write the estimate for the translocation frequency as

$$T_{\mathrm{di}} = T_0 p_{\mathrm{di}} \tag{3}$$

In Fig. 2, we show the results for Eq. (3), where $T_0 = 0.123$ and $c = 19.6$ have been adjusted for obtaining least-squared differences from the diblock Rosenbluth-Rosenbluth sampling (RS) results. The results confirm the dominance of the exponential factor resulting from pair interactions of the hydrophobic block.

The ratio between partition sums for interface-adsorbed diblocks and triblocks allows to project from diblock to triblock predictions for translocation frequencies,

$$T_{\mathrm{tri}} = 2^{2(\gamma_1-1)}\left(\frac{N}{2}\right)^{1-\gamma_{11}} T_{\mathrm{di}} \tag{4}$$

which is plotted in Fig. 2 for comparison. The resulting up-shift catches up to the RS diblock results up to a factor corresponding to a remaining free-energy difference of $1.4k_{\mathrm{B}}T$ that is missing in Eq. (4). Note that we did not consider finite chain length effects in Eq. (2) in scope of this qualitative comparison.

With this discussion in mind, it is interesting to have a look back to Fig. 1 for understanding surprising features observed in the sequence maps of slightly hydrophilic polymers. By the example of a fraction of 4/12 of hydrophobic monomers, we demonstrate that the polymers comprising the shortest amphiphilic blocks (labelled by "(a)" and "(b)" in Fig. 1) are found in a middle range of translocation frequencies, while triblock copolymers similar as those discussed in Figs. 2 and 3 lead to the largest translocation frequencies. A comparison of the free-energy profiles shown as an inset in Fig. 1 underlines the interplay between surface adsorption and hydrophobic/hydrophilic balance that leads to the result. Polymers that contain short blocks ("(a)" and "(b)" in Fig. 1) are mainly subject to an effective free-energy barrier for insertion into the bilayers, which is the rate-limiting factor for translocation. The result reflects the fact that the polymer is effectively hydrophilic and shows negligible surface adsorption effects. Combining T-type monomers into a larger center block, however, allows for anchoring of the polymer at bilayer–solvent interfaces and thereby effectively reduces the rate-limiting repulsion from the membrane environment. On the other hand, for the diblock copolymers with $N_{\mathrm{T}}/N = 4/12$, adsorption at the bilayer–solvent interface turns over to dominate the free-energy profiles and leads to the largest escape times found for the given hydrophilic/hydrophobic ratio.

From the comparison between RS results for τ and previous work[26,28,31] we therefore conclude that the dynamic interpretation of the sampling results is justified.

Machine-learned translocation times

The complete data set generated by GPU-accelerated RS sampling forms a powerful basis for benchmarking the ML-based search for sequences fulfilling given criteria. A network similar to Fig. 4 can be designed in order to predict sequence-determined properties of a polymer[39,44,45]. In this work, we stick to the example of logarithmic translocation times $\log(\tau)$, and refer to a chain length

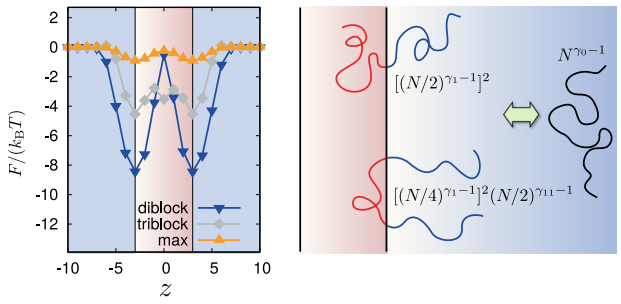

Fig. 3 Free-energy profiles. Free-energy profiles for various polymer architectures ($N = 12$, $N_{\mathrm{T}} = 6$) and corresponding relevant states for estimating desorption probabilities.

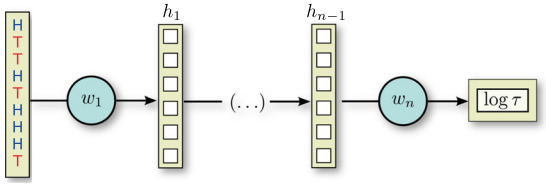

Fig. 4 Architecture of the neural network. Neural network architecture for translocation time prediction of a polymer as a function of H/T sequence.

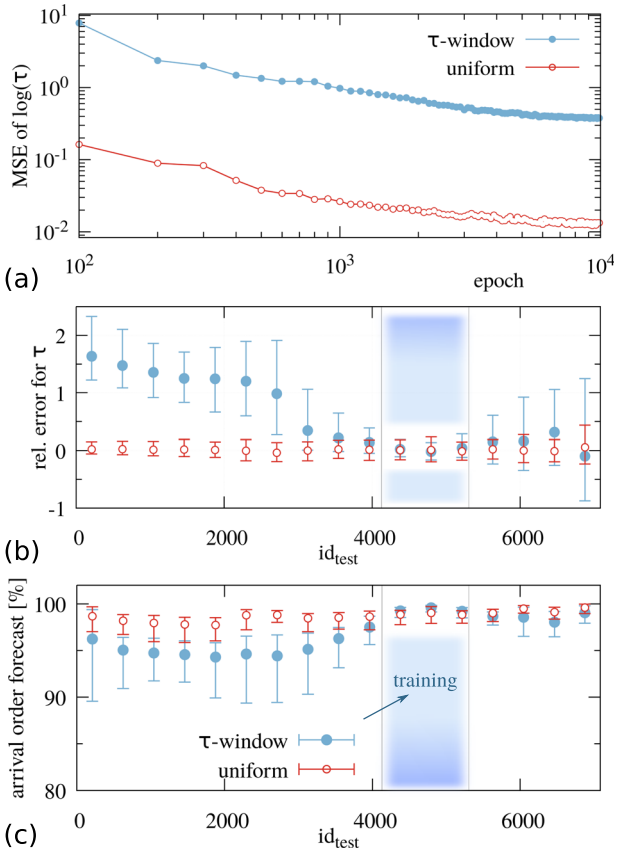

Fig. 5 Performance of the neural network predictions. a Evolution of the mean squared error (MSE) of neural network-based $\log(\tau)$ estimates for the test data set as a function of training epoch. **b** Relative error of the predicted value of τ according to Eq. (5) as a function of sequence index $\mathrm{id}_{\mathrm{test}}$ in the test set. The result is averaged for 17 groups (bins) of sequences along the RS-τ sorted test set ($\mathrm{id}_{\mathrm{test}}$). Error bars denote a confidence interval of 90% by excluding the highest and lowest 5% of the data. The blue box labels the range of sequences of the training set in case of τ-window. **c** Percentage of correctly predicted faster or slower other sequences as a function of sequence index. Here we use the same binning and error bar definition as in **b**.

of $N = 14$ monomers. The total number of sequences excluding the mirror-symmetric ones is $S = 8256$. The fraction of sequences within the training set we fix to $f_{\mathrm{train}} \equiv S_{\mathrm{train}}/S = 1/7$. The ratio between training to the remaining test set results in 1:6. However, we follow two distinct schemes for the distribution of training sequences within the sequence space: In the *uniform* scheme, we define equidistant intervals of size $1/f_{\mathrm{train}}$, along the τ-sorted sequences (id-space), and select the central sequences within each interval as the training set. In contrast, in the τ-window scheme we select every second sequence within a window $S/2 < \mathrm{id} \le S/2 + 2S_{\mathrm{train}}$, where "id" is a τ-sorted unique index in sequence space (section "Methods"). Note that thereby we select sequences within a narrow window in the upper half of translocation times.

In Figs. 5 and 6, we summarize the results of the training, and the performance of the resulting network with respect to the test set. In Fig. 5a, the development of the mean squared error (MSE) between NN- and RS-based $\log(\tau)$ values for all test sequences (unseen) is presented. We note a reliable convergence of MSE values for both uniform and τ-window training sets toward a horizontal line indicating that training was stopped early enough for not running into over-training. In the case of the uniform training set, MSE results typically end up more than one order of magnitude lower as compared to the τ-window training set. The corresponding root mean squared deviation from the expected value typically reduces by a factor of $\sqrt{28} \sim 5$. For the uniform training set, the root mean squared relative deviation from the RS-based $\log(\tau)$ value points to a typical error of 1.0%, whereas for the τ-window training set we observe values of 7.3%.

In Fig. 5b, we show the corresponding mean relative error for the back-converted (not logarithmic) time τ according to

$$\frac{\Delta\tau}{\tau} = \exp[\Delta\log(\tau)] - 1 \tag{5}$$

where $\Delta\log(\tau)$ is the absolute difference between the RS- and NN-based $\log(\tau)$ values. For the uniform training set, the relative error scatters between -23% and $+43\%$ as found for the largest index $\mathrm{id}_{\mathrm{test}}$ (largest τ), whereas for the fastest polymers 90% of sequences stay within an error of -6% to $+15\%$. For this training set, equivalent to a random selection of sequences, such high accuracy of the network prediction is remarkable when seeing that the RS-based values of τ are spread by a maximum factor of $\tau_{\mathrm{max}}/\tau_{\mathrm{min}} \sim 2 \times 10^{11}$. For the τ-window training set, the relative error far away from the training window increases as compared to the uniform set. Nevertheless, as the maximum range of relative errors is found in the interval of $-0.87 \le \Delta\tau/\tau \le 1.25$ for the largest index $\mathrm{id}_{\mathrm{test}}$, we conclude that typically the prediction hits the right order of magnitude for τ despite the fact that we used only the narrow sequence window for training. It is interesting to note that the translocation times of the fastest sequences is typically predicted correctly by a factor of ~3 despite the large distance from the training window. Qualitatively, we expect that, when shifting the τ-window to smaller (larger) values of τ, the prediction accuracy for the smallest τ values will increase (decrease) while accuracy for the largest values of τ will decrease (increase), which is supported by preliminary data (not shown).

Absolute values are not always the main question for the modeled mapping; in some cases it is enough to obtain a decision statement upon the performance of two structures. When comparing two polymer sequences, for instance, we may ask which of those translocates faster. In Fig. 5c, we therefore show the performance of the trained network to give the right answer for this question as a function of sequence $\mathrm{id}_{\mathrm{test}}$. For this purpose, we calculated for each NN result τ_1 for a given test sequence the fraction of all other test sequences leading to an NN output τ_2 that holds the same relation $\tau_1 > \tau_2$ or $\tau_1 < \tau_2$ as the corresponding pair of RS sampling results. In case of uniform training, $98.7^{+1.1}_{-1.6}\%$ of other sequences are correctly attributed as slower or faster (with a confidence of 90%), and for the τ-window training set $96.8^{+2.8}_{-4.5}\%$ of pairs are correctly labeled. For the τ-window training set, the performance far away from the training window is reduced, in particular for sequences with a lower $\mathrm{id}_{\mathrm{test}}$ index. However, the average fraction of correct decisions does not drop below 94.3% for the selected bin size.

By Fig. 5, we therefore demonstrated that a quantitative prediction of translocation times is possible by the applied ML model, and the accuracy depends crucially on the distribution of training sequences.

In Fig. 6, we outline more details of the training result by showing the predicted value of $\log(\tau)$ for the whole test sets of uniform and τ-window in Figs. 6a and 6b, respectively. The monotony of the predicted data points for both training sets follows the base data line despite the scattering of the data as discussed for Fig. 5. In particular, for the τ-window training set, we emphasize that the order of translocation times is predicted correctly for the fastest sequences although the training set covers only a narrow window within the slower half of sequences.

Another interesting observation is the prediction of step-like features in translocation time (arrows in Fig. 6b) as function of $\mathrm{id}_{\mathrm{test}}$ that are reproduced throughout the test set although located outside of the τ-window training range. Thus even the

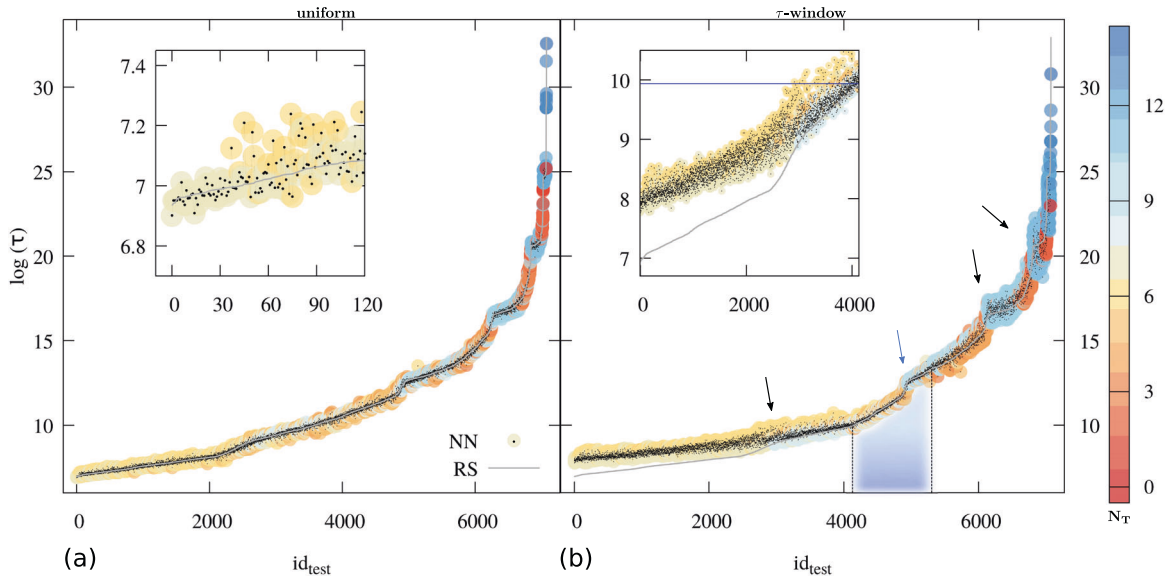

Fig. 6 Unseen data predictions. Neural network prediction (dots) for the mean first escape time τ for unseen data (test set) are compared to RS-based results (gray line). Data are shown as a function of a unique identifier id_{test} for sequences in the test set that is sorted according to the RS-based result for τ. The ratio between training and test set sizes is 1:6. The number of hydrophobic units, N_T, is shown as color-coded halos. In **a**, the uniform training set distributed homogeneously along the full RS τ-sorted sequence list. In **b**, we chose every second sequence within the blue labeled τ-windows range between $id_{test} = 4128$ and $id_{test} = 5305$. Training set sequences are skipped in this plot such that the slope is doubled as compared to the full sequence set within the labeled interval. The insets show details in the fields of lowest τ. The blue horizontal line in **b** inset labels the RS-based τ-value at the lower bound of the τ-window.

relatively simple network seems capable of finding a generic rule that links sequence and translocation time and thereby expresses the rather rich result based on Eqs. (7) and (8) without knowledge of conformation entropy nor the escape times. In view of the generalization performance observed for the τ-window training set, it therefore seems that the network developed an implicit internal representation approximating the mathematical rules linking copolymer sequence and translocation time.

DISCUSSION

We apply a massively parallel sampling of the conformations of amphiphilic copolymers by means of self-avoiding random walks within a given density field representing a model for amphiphilic bilayer membranes. We estimated the free-energy profiles of the polymers composed of hydrophilic (H) and hydrophobic beads (T) with respect to distance from the membrane as a reaction coordinate. We calculated the mean first escape time τ as a measure for polymer translocation time through the model membrane all 2^N binary sequences up to chain length $N \le 16$. Our results confirm that polymer translocation is controlled by a balance of the overall hydrophobicity of the polymer and is inhibited by adsorption at the bilayer–solvent interfaces[26–28,31], which is consistent with the picture for small solutes[67] and larger solid objects such as carbon nanotubes[68].

Amphiphilic polymers at a balanced hydrophobicity show the smallest translocation times when the sequence exposes small repeating amphiphilic features, while longest waiting times are associated with a diblock structure of the whole chain. The different translocation rates between diblock and triblock copolymers as well as their chain-length dependence can be explained qualitatively when comparing adsorption-free energies at the bilayer–solvent interface involving surface-critical exponents. The relatively weak dependence of the translocation time of balanced hydrophobicity small-block alternating copolymers from chain length indicates that local amphiphilic features are only weakly interacting with the bilayer–solvent interfaces and the copolymer effectively resembles a homopolymer chain for which the membrane is energetically transparent. Chain-length dependence in this case is expected to increase when effective monomer association constants are stronger than in the present model. When considering slightly hydrophilic backbones, larger hydrophobic blocks start to become more prominent in sequences leading to smallest translocation times as they promote the association of the net-repulsive backbone with the hydrophobic membrane core.

The extensive database generated by RS sampling has been used to feed a multi-layer artificial NN with four hidden layers in order to explore the capability of so-called deep learning approaches for finding a general rule of how copolymer sequence translates into translocation times through biological barriers. The aim of this work is to test the meaningful interpretation of the "dirty work" of NNs provided by a complete data set of polymer sequences. We demonstrate that, even by using a low fraction 1/7 of uniformly selected training examples as compared to the total number 2^N of binary sequences for $N = 14$, the NN achieves a root mean squared relative deviation in the order of 1% for the logarithmic mean first escape time $\log(\tau)$. In order to test the generalization performance of the network, we implemented a second training scheme, where training examples have been selected from a narrow window of sequences with respect to translocation times τ covering a factor of ≈30 between maximum and minimum translocation times contained in the training set. In this case, the network approximates the order of magnitude of the test data set covering a window being more than nine orders of magnitude wider. We conclude that the NN developed an internal representation of the mathematical rules linking sequence and translocation times, which involve a precise estimate of rate-limiting energy barriers. The network thereby encodes a complex interplay between polymer net hydrophobicity and sequence-dependent adsorption at the bilayer–solvent interfaces that to date can be treated in a theoretically closed form only for special cases as it involves the sequence-dependent polymer conformation entropy and solving the diffusion problem in inhomogeneous

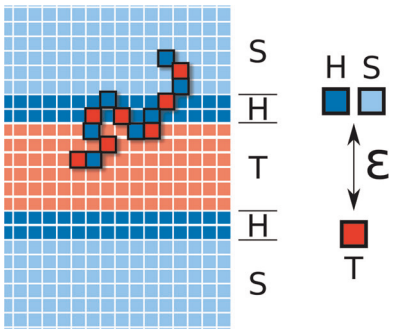

Fig. 7 Simulation model. Illustration of a coarse grained polymer chains as used in our simulations within the grid occupancy of a laterally homogeneous membrane, and repulsive interactions between effectively two components (H,S) and (T).

free-energy landscapes. Our results indicate a systematic decrease of prediction accuracy when moving into unexplored corners of sequence space and challenge future investigation on the relation between training data bias and prediction accuracy.

METHODS

Rosenbluth–Rosenbluth sampling

We consider the diffusive transport of a polymer through a lipid membrane resembling a homogeneous oil slab as shown in Fig. 7. In particular, we are interested in mean first escape time of a polymer through the membrane as a function of length, N, sequence of hydrophilic head (H) and hydrophobic tail monomers (T). Coarse grained polymers are embedded into an external concentration field that represents bilayer membrane on a mean-field level composed of an hydrophilic region (H) and a hydrophobic core (T), as well as solvent (S). The hydrophobic core has a thickness of six lattice units.

Monomers are represented as single-cell occupations on a simple cubic lattice, and bond vectors are taken from a set of 26 vectors with lengths of 1, $\sqrt{2}$, and $\sqrt{3}$ lattice units. Double occupancy of lattice sites is forbidden, and the monomers have excluded volume. This set of static rules corresponds to those of Shaffer's Bond Fluctuation Model[54].

Between hydrophilic sites (H and S), and hydrophobic sites (T), we implement short-range repulsive interactions. We write the internal energies of H and T monomers of the polymer as

$$U_{\mathrm{H}}(\vec{r}) = \epsilon c_{\mathrm{T}}(\vec{r}); \quad U_{\mathrm{T}}(\vec{r}) = \epsilon(c_{\mathrm{S}}(\vec{r}) + c_{\mathrm{H}}(\vec{r})) \tag{6}$$

where $c_x(\vec{r})$ are the number of lattice occupancies by species x on the 26 nearest neighbor sites[55]. In order to keep the model simple, we use only a single interaction parameter defined as $\epsilon = 0.1 k_{\mathrm{B}}T$ with k_{B} being Boltzmann's constant and T the absolute temperature. For the enumeration of c_x, the occupancy of the lattice by a given external concentration field (Fig. 7) is counted, and monomer–monomer contacts are taken into account in a way that contacts with the external field are screened by surrounding monomers. Thereby solvent-induced effects on polymer conformations are represented by the model.

For a given amphiphilic sequence, we aim to calculate the mean first escape time of a polymer between a repulsive boundary at $z = -a$ and an absorbing boundary at $z = +$a, (Fig. 7)[56],

$$\tau = \frac{1}{D}\int_{-a}^{+a} \mathrm{d}z p^{-1}(z) \int_{-a}^{z} \mathrm{d}z' p(z') \tag{7}$$

where D is the diffusion constant of the polymer and $p(z)$ is the probability distribution to find the center of mass of the polymer at a given distance, z, from the bilayer's mid-plane. We define $a = 22$ lattice units, and $D = 1$ (lattice unit)^2, such that the dimensionless number of τ does not include the chain-length dependence of diffusion time.

The probability distribution $p(z)$ is calculated by generating M polymer conformations $\vec{R} = (\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N)$ according to the RS scheme[53]. For each conformation $\vec{R}$, the contact energy $U(\vec{R})$ is calculated according to $U(\vec{R}) = \sum_{i=1}^{N} U_X(\vec{r}_i)$ in units of $k_{\mathrm{B}}T$ according to Eq. (6) depending on the species X of the monomer $X = \mathrm{H}$ or $X = \mathrm{T}$. The center of mass $\bar{z}(\vec{R}) = (1/N)\sum_{i=1}^{N} \vec{r}_i \vec{e}_z$ is evaluated with $\vec{e}_z$ being the lattice unit vector along the membrane's normal direction. The distribution $p(z)$ is then written as

$$p(z) = \tfrac{1}{M}\sum_{i,\bar{z}(\vec{R}_i)\simeq z}^{M} W_i \mathrm{e}^{-\beta U(\vec{R}_i)} \tag{8}$$

where the condition below the sum illustrates that only those conformations contribute whose center of mass is found within a grid distance $(z - 1/2) < \bar{z} \le (z + 1/2)$ from z, and $\beta \equiv 1/(k_{\mathrm{B}}T)$. In Eq. (8), W_i is the Rosenbluth weight of the ith conformation.

For a given sequence of H and T monomers in a polymer backbone, we calculate the mean first escape time according to Eq. (7) based on the generation of $M = 1.5 \times 10^7$ RS-generated chains at uniformly distributed random positions within a periodic lattice of $64 \times 64 \times 64$ lattice sites. The algorithm is implemented for GPUs[4]. In order to analyze how the mean first escape time depends on the amphiphilic sequence of the polymer, we perform the procedure for all 2^N sequences for various degrees of polymerization $N \le 16$.

Artificial NN

We employ a fully connected NN involving tanh-activation as sketched in Fig. 4. The network is composed by 2 hidden layers with 64 nodes each followed by 2 hidden layers with 32 nodes each. The input layer corresponds to a vector of values 0 and 1 representing the considered amphiphilic sequence of hydrophobic (0) and hydrophilic (1) monomers. The output layer consists of one neuron whose output is compared to the RS-based τ value for this sequence. The total network depth is $n = 5$, where only the output activation, $\tanh[\sum_{i=1}^{32}(w_{n,i}h_{n-1,i} + b)]$, includes a bias, b. Since absolute values of τ spread over several orders of magnitude, we perform the training with respect to its logarithm. The RS-based values of $\log(\tau)$ are further linearly normalized and centralized into an interval $[-0.9, 0.9]$ by defining $I(\log(\tau)) = 1.8 \times [(\log(\tau) - \log(\tau_{\min}))/(\log(\tau_{\max}) - \log(\tau_{\min})) - \frac{1}{2}]$ in order to be conveniently expressible by the tanh-activation output. NN-based estimates for $\log(\tau)$ are obtained by the back projection, I^{-1}, of output neuron activations.

All weights are initialized with uniform random numbers in an interval $[-0.3, 0.3]$. The feed-forward (ff) back-propagation (bp)[57] algorithm is employed for training. Error bp is performed after each ff cycle for a randomly selected sequence taken from the training set (stochastic gradient descent). The squared difference between the resulting activation of the output neuron and the RS-based $I(\log(\tau))$ value is used as the cost function for weight and bias adjustment. We set the initial training rate to $\eta = 0.02$, which gets reduced by a factor of (1/1.3) every 10^3 epochs in order to avoid frustration or early over-training effects. One epoch is defined as the average number of ff-bp cycles per sequence–τ pair. We set the total number of epochs to 10^4.

For each sequence, we define an unique integer identifier, $1 \le \mathrm{id} \le S$, that is sorted according to the RS-based τ value. A lower id means a lower τ. The whole of S sequences is divided into a training set of size S_{train} and a test set of the size $S_{\mathrm{test}} = S - S_{\mathrm{train}}$. For the test set, we define a unique identifier $\mathrm{id}_{\mathrm{test}}$ for each sequence that is the analog to id for the total sequence space. The index $\mathrm{id}_{\mathrm{test}}$ labels sequences that are unseen by the network during training.

DATA AVAILABILITY

The data used in the paper are available from the authors upon request.

CODE AVAILABILITY

The source code of the programs used in this paper is available from the authors upon request.

Received: 21 May 2019; Accepted: 18 March 2020;
Published online: 05 June 2020

REFERENCES

1. Lutz, J.-F., Ouchi, M., Liu, D. R. & Sawamoto, M. Sequence-controlled polymers. *Science* **341**, 1238149 (2013).
2. Lutz, J.-F. Defining the field of sequence-controlled polymers. *Macromol. Rapid Commun.* **38**, 1700582 (2017).
3. Rahman, M. A. et al. Macromolecular-clustered facial amphiphilic antimicrobials. *Nat. Commun.* **9**, 5231 (2018).

4. Guo, Y. & Baulin, V. A. GPU implementation of the Rosenbluth generation method for static Monte Carlo simulations. *Comput. Phys. Commun.* **216**, 95–101 (2017).
5. Ren, Y. & Müller, M. Kinetics of pattern formation in symmetric diblock copolymer melts. *J. Chem. Phys.* **148**, 204908 (2018).
6. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
7. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
8. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431 (2017).
9. Wei, Q., Melko, R. G. & Chen, J. Z. Y. Identifying polymer states by machine learning. *Phys. Rev. E* **95**, 032504 (2017).
10. Iten, R., Metger, T., Wilming, H., delRio, L. & Renner, R. Discovering physical concepts with neural networks. *Phys. Rev. Lett.* **124**, 010508 (2020).
11. AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Syst.* **8**, 292–301 (2019).
12. Hoffmann, C., Menichetti, R., Kanekal, K. H. & Bereau, T. Controlled exploration of chemical space by machine learning of coarse-grained representations. *Phys. Rev. E* **100**, 033302 (2019).
13. Wilbraham, L., Sprick, R. S., Jelfs, K. E. & Zwijnenburg, M. A. Mapping binary copolymer property space with neural networks. *Chem. Sci.* **10**, 4973–4984 (2019).
14. St.John, P. C. et al. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **150**, 234111 (2019).
15. Muthukumar, M. Polymer translocation through a hole. *J. Chem. Phys.* **111**, 10371 (1999).
16. Muthukumar, M. Translocation of a confined polymer through a hole. *Phys. Rev. Lett.* **86**, 3188–3191 (2001).
17. Muthukumar, M. Theory of sequence effects on DNA translocation through proteins and nanopores. *Electrophoresis* **23**, 1417–1420 (2002).
18. Mirigian, S., Wang, Y. & Muthukumar, M. Translocation of a heterogeneous polymer. *J. Chem. Phys.* **137**, 064904 (2012).
19. Wong, C. T. A. & Muthukumar, M. Polymer translocation through a cylindrical channel. *J. Chem. Phys.* **128**, 154903 (2008).
20. Sun, L.-Z., Wang, C.-H., Luo, M.-B. & Li, H. Trapped and non-trapped polymer translocations through a spherical pore. *J. Chem. Phys.* **150**, 024904 (2019).
21. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl Acad. Sci. USA* **93**, 13770–13773 (1996).
22. Li, J., Gershow, M., Stein, D., Brandin, E. & Golovchenko, J. A. DNA molecules and configurations in a solid-state nanopore microscope. *Nat. Mater.* **2**, 611–615 (2003).
23. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
24. Katz, Y. & Diamond, J. M. A method for measuring nonelectrolyte partition coefficients between liposomes and water. *J. Membr. Biol.* **17**, 69–86 (1974).
25. Diamond, J. M. & Katz, Y. Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water. *J. Membr. Biol.* **17**, 121–154 (1974).
26. Sommer, J.-U., Werner, M. & Baulin, V. A. Critical adsorption controls translocation of polymer chains through lipid bilayers and permeation of solvent. *Europhys. Lett.* **98**, 18003 (2012).
27. Werner, M., Bathmann, J., Baulin, V. A. & Sommer, J.-U. Thermal tunneling of homopolymers through amphiphilic membranes. *ACS Macro Lett.* **6**, 247–251 (2017).
28. Werner, M., Sommer, J.-U. & Baulin, V. A. Homo-polymers with balanced hydrophobicity translocate through lipid bilayers and enhance local solvent permeability. *Soft Matter* **8**, 11714–11722 (2012).
29. Stepanow, S., Bauerschafer, U. & Sommer, J. U. Adsorption of polymers at interfaces and extended defects. *Phys. Rev. E* **54**, 3899–3905 (1996).
30. Soteros, C. E. & Whittington, S. G. The statistical mechanics of random copolymers. *J. Phys. A Math. Gen.* **37**, R279 (2004).
31. Werner, M. & Sommer, J.-U. Translocation and induced permeability of random amphiphilic copolymers interacting with lipid bilayer membranes. *Biomacromolecules* **16**, 125–135 (2015).
32. Goda, T., Goto, Y. & Ishihara, K. Cell-penetrating macromolecules: direct penetration of amphipathic phospholipid polymers across plasma membrane of living cells. *Biomaterials* **31**, 2380–2387 (2010).
33. Goda, T., Ishihara, K. & Miyahara, Y. Critical update on 2-methacryloyloxyethyl phosphorylcholine (MPC) polymer science. *J. Appl. Polym. Sci.* **132**, 41766 (2015).
34. Goda, T. et al. Translocation mechanisms of cell-penetrating polymers identified by induced proton dynamics. *Langmuir* **35**, 8167–8173 (2019).
35. Marks, J. R., Placone, J., Hristova, K. & Wimley, W. C. Spontaneous membrane-translocating peptides by orthogonal high-throughput screening. *J. Am. Chem. Soc.* **133**, 8995–9004 (2011).
36. Kauffman, W. B., Fuselier, T., He, J. & Wimley, W. C. Mechanism matters: a taxonomy of cell penetrating peptides. *Trends Biochem. Sci.* **40**, 749–764 (2015).
37. Fuselier, T. & Wimley, W. C. Spontaneous membrane translocating peptides: the role of leucine-arginine consensus motifs. *Biophys. J.* **113**, 835–846 (2017).
38. Kauffman, W. B., Guha, S. & Wimley, W. C. Synthetic molecular evolution of hybrid cell penetrating peptides. *Nat. Commun.* **9**, 2568 (2018).
39. Joyce, S. J., Osguthorpe, D. J., Padgett, J. A. & Price, G. J. Neural network prediction of glass-transition temperatures from monomer structure. *J. Chem. Soc. Faraday Trans.* **91**, 2491–2496 (1995).
40. Ulmer II, C. W., Smith, D. A., Sumpter, B. G. & Noid, D. I. Computational neural networks and the rational design of polymeric materials: the next generation polycarbonates. *Comput. Theor. Polym. Sci.* **8**, 311–321 (1998).
41. Mattioni, B. E. & Jurs, P. C. Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks. *J. Chem. Inf. Comput. Sci.* **42**, 232–240 (2002).
42. Duce, C., Micheli, A., Starita, A., Tiné, M. R. & Solaro, R. Prediction of polymer properties from their structure by recursive neural networks. *Macromol. Rapid Commun.* **27**, 711–715 (2006).
43. Duce, C., Micheli, A., Solaro, R., Starita, A. & Tiné, M. R. Recursive neural networks prediction of glass transition temperature from monomer structure: an application to acrylic and methacrylic polymers. *J. Math. Chem.* **46**, 729–755 (2009).
44. Molina, J., Laroche, A., Richard, J.-V., Schuller, A.-S. & Rolando, C. Neural networks are promising tools for the prediction of the viscosity of unsaturated polyester resins. *Front. Chem.* **7**, 375 (2019).
45. Bernazzani, L., Duce, C., Micheli, A., Mollica, V. & Tiné, M. R. Quantitative structure-property relationship (QSPR) prediction of solvation gibbs energy of bifunctional compounds by recursive neural networks. *J. Chem. Eng. Data* **55**, 5425–5428 (2010).
46. Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Netw.* **2**, 183–192 (1989).
47. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Syst.* **2**, 303–314 (1989).
48. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
49. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Multi-task neural networks for QSAR predictions. Preprint at https://arxiv.org/abs/1406.1231 (2014).
50. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55**, 263–274 (2015).
51. Ramsundar, B. et al. Massively multitask networks for drug discovery. Preprint at https://arxiv.org/abs/1502.02072 (2015).
52. Hughes, T. B., Dang, N. L., Miller, G. P. & Swamidass, S. J. Modeling reactivity to biological macromolecules with a deep multitask network. *ACS Cent. Sci.* **2**, 529–537 (2016).
53. Rosenbluth, M. N. & Rosenbluth, A. W. Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.* **23**, 356–359 (1955).
54. Shaffer, J. S. Effects of chain topology on polymer dynamics: bulk melts. *J. Chem. Phys.* **101**, 4205 (1994).
55. Dotera, T. & Hatano, A. The diagonal bond method: a new lattice polymer model for simulation study of block copolymers. *J. Chem. Phys.* **105**, 8413–8427 (1996).
56. Pontryagin, L., Andronov, A. & Vitt, A. On the statistical investigation of dynamic systems. *Zh. Eksper. Teor. Fiz.* **3**, 165–180 (1933).
57. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533 (1986).
58. Grassberger, P. Monte Carlo simulation of 3D self-avoiding walks. *J. Phys. A Math. Gen.* **26**, 2769 (1993).
59. Duplantier, B. Polymer network of fixed topology: renormalization, exact critical exponent γ in two dimensions, and $d = 4 - \epsilon$. *Phys. Rev. Lett.* **57**, 941–944 (1986).
60. De Gennes, P.-G. *Scaling Concepts in Polymer Physics*, 1st edn (Cornell University Press, Ithaca, NY, 1979).
61. Hegger, R. & Grassberger, P. Chain polymers near an adsorbing surface. *J. Phys. A Math. Gen.* **27**, 4069 (1994).
62. Grassberger, P. Simulations of grafted polymers in a good solvent. *J. Phys. A Math. Gen.* **38**, 323 (2005).
63. Clisby, N., Conway, A. R. & Guttmann, A. J. Three-dimensional terminally attached self-avoiding walks bridges. *J. Phys. A Math. Theor.* **49**, 015004 (2016).
64. Hsu, H.-P., Nadler, W. & Grassberger, P. Scaling of star polymers with 1-80 arms. *Macromolecules* **37**, 4658–4663 (2004).
65. Schram, R. D., Barkema, G. T. & Bisseling, R. H. Exact enumeration of self-avoiding walks. *Theory Exp.* **2011**, P06019 (2011).
66. Clisby, N., Liang, R. & Slade, G. Self-avoiding walk enumeration via the lace expansion. *J. Phys. A Math. Theor.* **40**, 10973 (2007).
67. Marrink, S.-J. & Berendsen, H. J. C. Permeation process of small molecules across lipid membranes studied by molecular dynamics simulations. *J. Phys. Chem.* **100**, 16729–16738 (1996).

68. Pogodin, S. & Baulin, V. A. Can a carbon nanotube pierce through a phospholipid bilayer? *ACS Nano* **4**, 5293–5300 (2010).

ACKNOWLEDGEMENTS

M.W. and V.A.B. gratefully thank the EU's Marie Curie Actions under European Union 7th Framework Programme (FP7), Initial Training Network SNAL Grant No. 608184. Y.G. acknowledges funding from the NSFC grant number 11804151 and FRFCU 14380017. V.A.B. acknowledges financial assistance from the Ministerio de Ciencia, Innovación y Universidades of the Spanish Government (CTQ2017-84998-P).

AUTHOR CONTRIBUTIONS

M.W. and V.A.B. conceived the idea, Y.G. designed and performed Rosenbluth Sampling method calculations, M.W. designed artificial neural network and performed neural network and Rosenbluth calculations. M.W. wrote the manuscript with contribution of all authors. All the authors participated in discussions and contributed materially in finalizing the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to M.W., Y.G. or V.A.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.