

# Machine-learning a virus assembly fitness landscape

Pierre-Philippe Dechant<sup>1</sup> and Yang-Hui He<sup>2</sup>

<sup>1</sup> *Pro Vice Chancellor's Office,  
York St John University, York, YO31 7EX, and  
York Cross-disciplinary Centre for Systems Analysis  
University of York, Heslington YO10 5GE, UK  
ppd22@cantab.net*

<sup>2</sup> *Dept. of Mathematics, City, University of London, EC1V 0HB, UK;  
Merton College, University of Oxford, OX14JD, UK and  
School of Physics, NanKai University, Tianjin, 300071, P.R. China  
hey@maths.ox.ac.uk*

## Abstract

Realistic evolutionary fitness landscapes are notoriously difficult to construct. A recent cutting-edge model of virus assembly consists of a dodecahedral capsid with 12 corresponding packaging signals in three affinity bands. This whole genome/phenotype space consisting of  $3^{12}$  genomes has been explored via computationally expensive stochastic assembly models, giving a fitness landscape in terms of the assembly efficiency. Using latest machine-learning techniques by establishing a neural network, we show that the intensive computation can be short-circuited in a matter of minutes to astounding accuracy.

# 1 Introduction

Two facts about simple viruses have been known for a long time. Firstly, that genetic economy leads to the use of symmetry, such that virus capsids are mostly icosahedral or helical. Secondly, packaging signals, that is secondary structure features in the viral RNA, are often required for encapsidation in viruses with single-stranded genomes. Examples are the origin of assembly sequence in Tobacco Mosaic virus, the psi element in HIV and the TR sequence in MS2. This is an evolutionary advantage, as it ensures vRNA-specific encapsidation and can increase assembly efficiency through a cooperative role of the RNA, which acts as a nucleation site.

More recently, it has been shown that taken together, these two facts suggest that there could be more than one packaging signal, with multiple signals in fact dispersed throughout the genome. This is because the capsid is symmetric, and the packaging signal mechanism functions via interaction between viral RNA and the coat protein (CP). In several cases, this RNA-CP interaction leads to a conformational change in the CP, which only then makes it assembly competent (e.g. TMV and MS2). The picture that emerges is then that there are multiple packaging signals (PS) that recruit CP onto a growing capsid. This reduces the phase space that CP has to search in order to assemble a capsid, resulting in vastly increased assembly efficiency. The details of such a mechanism were found in MS2 and STNV in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Once the details of this mechanism were understood using biochemistry, structural biology, bioinformatics, biophysics and graph theory in these model systems, related mechanisms could be found in clinically relevant viruses [15, 16, 17]. These packaging signals are secondary structure features of the viral genomes where a stemloop in the single-stranded RNA presents a common recognition motif that can bind to CP (see Fig. 1A). The viral genome thus has multiple layers of constraints, by having to code for genes as well as the PS instruction manual. This set of packaging signals can also be repurposed and optimised for the assembly of virus-like particles, which do not share the same genetic constraints as the virus, and could be used e.g. for vaccines, drug delivery or as an anti-viral strategy [18, 19].

An equilibrium model of how to assemble a simplified example of an icosahedral virus, a dodecahedron built from 12 pentagonal faces, was considered in [20] using ODEs. More recently, the multiple dispersed packaging signal paradigm has sparked renewed interest in such a dodecahedral model

[3, 21]. The assembly reaction kinetics was modelled via a set of discrete reactions in a stochastic simulation paradigm based on the Gillespie algorithm [22].

In this model 12 PSs can bind CP, as well as dissociate again, reflecting reversible/equilibrium kinetics. Bound CP can then bind other bound CP, gradually building up a capsid (see Fig. 1B and C). The PSs here have three different bands of binding affinity: weak, medium and strong. These choices correspond to binding energies of 4/8/12 kcal/mol respectively, based on the TR sequence in MS2 which has approximately 12kcal/mol. The binding energy between CPs is much lower, at approximately 2 kcal/mol. This modulation of affinity affects the assembly kinetics, e.g. by providing a nucleation point that starts assembly, or allowing for error-correction via weaker binding elsewhere. The thermodynamics of PS binding and of the number of CP bonds formed then translates into assembly efficiency. This in turn is taken as a proxy for fitness (all other things being equal) – or at least the contribution to the fitness that results from assembly considerations [23].

In [24] the whole space of these  $3^{12}$  genomes (or rather, phenotype profiles) has been explored. The assembly efficiency there is given by the number of capsids that have correctly assembled out of a possible total of 2,000. This efficiency provides a fitness landscape on the 12-dimensional genome space. This is an interesting model that is tractable, in contrast with many other biological systems, as it has a small number of degrees of freedom and is dominated by the symmetry of the capsid. This tractability also allows for the consideration of viral evolution. For instance, mutation of the PS strengths leads to the emergence of a set of related genomes that form a ‘quasispecies’ [25]. One can thus investigate the effect of evolutionary pressures, e.g. those exerted by standard drugs or a novel type of drug that targets packaging signals [26].

This model thus captures many interesting aspects of viral genetics, geometry and assembly. A more realistic model would have more CP building blocks and PSs, e.g. around 60 for MS2 (i.e. one full orbit of the icosahedral group). But the computation time for even these simple genomes and the assembly kinetics that provide the fitness landscape are already considerable. Even other simplified models, e.g. reduced orbits on symmetry axes given by e.g. an icosahedron consisting of 20 triangles with 20 PSs, a rhombic triacontahedron consisting of 30 rhombuses with 30 PSs, or a finer gradation of binding affinity bands are already computationally out of reach.

However, this data set is a perfect example of data that is amenable to a

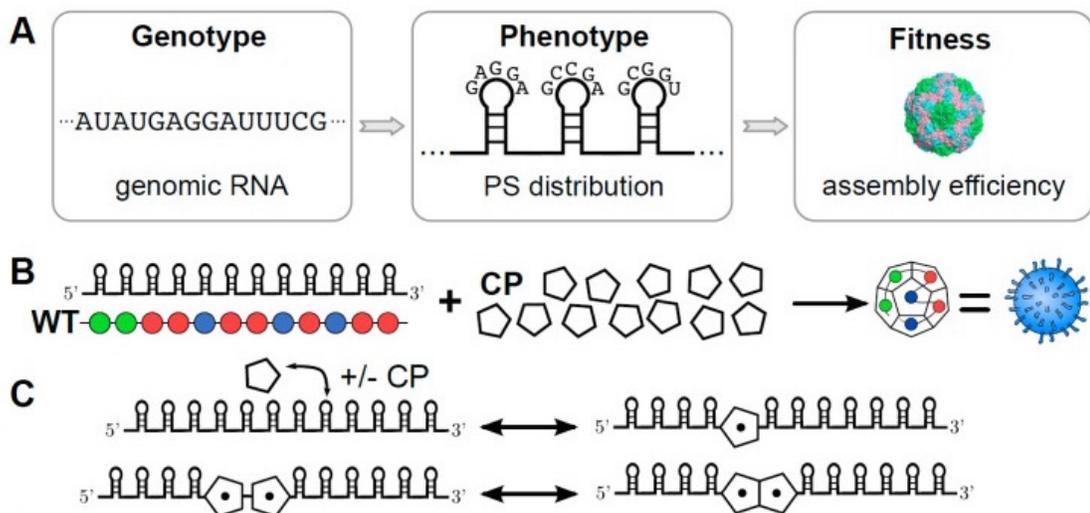


Figure 1: **A** The nucleotide sequence of a virus determines the gene products; however, in addition to this information content the RNA also explores a configuration space of secondary structures. Viruses appear to have evolved to use such motifs to help recruit coat protein with a conserved common recognition motif during assembly. The stability and binding affinity of these packaging signals gives a distinctive profile for viral assembly, which is the phenotype relevant to assembly. Assembly efficiency is the fitness of this phenotype, or at least the contribution to the overall fitness that is determined by aspects of assembly. **B** The genomes in the model consist of twelve packaging signals (PS) that can take weak, medium and strong binding affinities. They successively recruit twelve pentagonal coat proteins, which together form the dodecahedral virion in this model. **C** The stochastic simulation algorithm models several possible reactions. Firstly, packaging signals can bind coat proteins (and fall apart again), and secondly, two coat proteins that have been recruited by packaging signals can bind to each other. The fitness landscape was computed for 2000 virions for each possible genome, making the computations very intensive.

machine-learning approach, since it associates a vector input with a number output. We therefore train a neural network to predict the fitness landscape. The network is trained on a subset of the whole genome space, and validated on the remainder of the data. This proof-of-principle shows that it is very fast for a neural network to learn the inherent patterns within the large degeneracy of the detailed stochastic modelling to predict assembly efficiency fitness for unseen genomes. The danger is that some subtleties of the stochastic modelling are lost, but allowing for computation times many orders of magnitude faster. This approach could thus in future be used to tackle more realistic models such as the ones mentioned above. Stochastic simulations could be used to partially explore these larger genome spaces, calculating assembly fitness in order to provide a training set for a neural network. The rest of the fitness landscape can then be predicted by the artificial intelligence; it is also possible to only compute this fitness if necessary, e.g. when a new genome arises through mutation in a quasispecies model, such that such computation may only be necessary ‘precedurally’.

## 2 Methodology and Results

From a purely mathematical point of view, we have the following problem. Let (weak, medium, strong) be denoted respectively by (1, 2, 3). The input is a vector  $v$  in a 12-dimensional vector space over  $\mathbb{F}_3$ , the field of three elements. The output is an integer (which we treat as a real number) between 0 and 2000, which we can normalise into  $\epsilon \in [0, 1] \subset \mathbb{R}$  by dividing by 2000. The algorithm used by [3, 24] is thus a map

$$f : v \in \mathbb{F}_3^{12} \longrightarrow \epsilon \in [0, 1] . \quad (1)$$

A typical example is

$$\{1, 1, 1, 2, 2, 2, 3, 1, 2, 2, 1, 1\} \longrightarrow \frac{1523}{2000} \simeq 0.7615 \quad (2)$$

The map  $f$  is a computationally intensive one (with individual genome run times between 20 minutes and 12 hours, and cumulative run time of 3-4 weeks on the N8 Polaris high performance computing research cluster, Intel 2.6 GHz Sandy Bridge E5-2670 processors, with a total of 5,312 cores, with a mix of 4 and 16Gb of RAM, <https://n8hpc.org.uk/facilities/>). Nevertheless a brute-force simulation has been performed on the  $3^{12} = 531,441$  possible

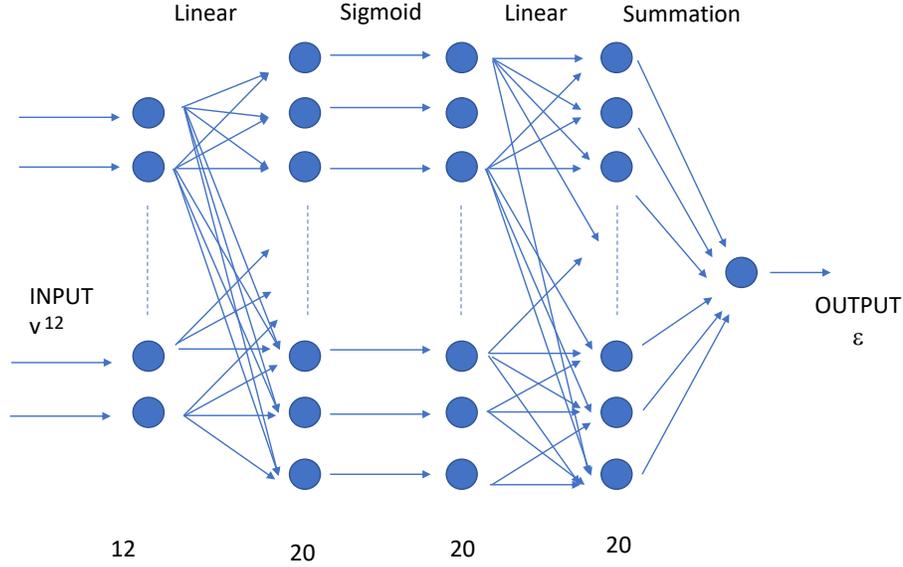


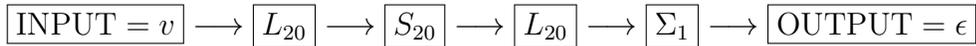
Figure 2: The structure of the neural network used in the calculation.

input values and the efficiency value extracted. This gives us a database of some half a million known cases of the form (2).

Such a problem is perfectly adapted to supervised machine-learning: we know many input values and wish to train some artificial intelligence to associate the input with the known output on some small subset, and use it to predict the output for unseen input. The advantage of this approach is that often approximate results can be attained at reduction in computation time by many orders of magnitude. The paradigm of using machine-learning in algebraic geometry and more general classes of problems in pure mathematics was proposed in [27, 28, 29] to satisfying accuracy, and it is a similar philosophy that we will adopt here.

Let us first try the following specific procedure:

- Take the full data  $D$  of the form (2), of size  $3^{12}$ ;
- Establish the neural network  $\mathcal{N}$ , a 3-layer perceptron



In the above, L means a linear-layer, S, a sigmoid layer and  $\Sigma$ , a summation layer. In particular, the first linear layer  $L_{20}$  is a fully connected layer taking the 12-vector  $v$  to 20 neurons by simply the linear function  $y = wx + b$ . This is then fed into an element-wise sigmoid layer  $\sigma(x) = (1 + e^{-x})^{-1}$  of 20 neurons, followed again by a linear layer, which is then summed to the real number  $\epsilon$  as the output. The schematic of  $\mathcal{N}$  is shown in Figure 2. We have taken this neural network only to illustrate the power of our methodology and have not optimised the hyper-parameters such as 20, nor the network architecture or the choice of the type of neurons.

- Now split  $D$  into a training set  $T$  of 30,000 random samples; the validation set will be the complement  $V = D \setminus T$ . Note that the training data is only about 5.6% of the total data.
- We train  $\mathcal{N}$  with  $T$  and validate on  $V$ .
- As a further check, we create a “fake” validation set  $\tilde{V}$  which has the same inputs as that of  $V$  but with output randomly assigned from the set of correct outputs.

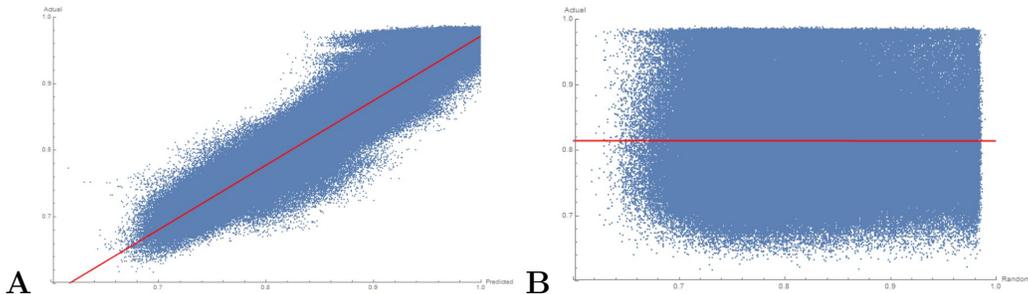


Figure 3: **A** A scatter plot of the predicted and actual value of  $\epsilon$  for the validation data from a 30,000 training sample; **B** scatter plot of a random-prediction versus the actual  $\epsilon$  value.

On an ordinary laptop (Intel Core m5-6Y57 CPU, 1.10GHz, 2 Cores, 4 Logical Processors, with 8Gb of RAM), the training took about 45 seconds,

and the prediction about 10 seconds. The algorithm is implemented on Mathematica [30] and is expected to even faster on the Python package Tensorflow [31]. In other words, the entire computation took under 1 minute with an ordinary notebook as opposed to the many hours it took on a supercomputer.

We present the result in Figure 3. In part **A**, we present a plot of the predicted  $\epsilon$  on the horizontal versus the actual  $\epsilon$  on the vertical. There are 501,441 points. One can see that they cluster near the desired  $y = x$  line, which would mean perfect prediction (note the axis ranges). To give some precise measures, the best fit line is  $y = -0.0262122 + 1.02519x$  with F-statistic  $2.45082 \times 10^6$  and p-value less than  $10^{-10^6}$ . The R-squared value is 0.830151. To double check, we plot the same result for the fake validation set  $\tilde{V}$  in part **B**. It is obvious that the distribution is much less structured and essentially randomly occupies a square. The fit here is  $y = 0.815233 - 0.000947474x$  i.e. practically a constant, with a poor F-statistic of 0.450401 and a poor p-value of 0.502145. The R-squared value is  $8.98217 \times 10^{-7}$ . This is very re-assuring for less than 6% of seen data and total computation time of less than 1 minute on an ordinary laptop.

To give an idea of the prediction, for the  $(1, 1, \dots, 1)$  vector, the net predicts  $\epsilon = 0.87069$ , or 1741. The original value is 200, but that is a singular outlier in the whole data set, which we would not expect the neural network to be able to reproduce. For the  $(2, 2, \dots, 2)$  vector, the net predicts  $\epsilon = 0.834721$ , or 1669; the correct value is 1745. For the  $(3, 3, \dots, 3)$  vector, the net predicts  $\epsilon = 0.673568$ , or 1347; the correct value is 1309.

We will use R-squared, a real number between 0 and 1, as a measure of accuracy of the machine-learning; the closer it is to 1, the better the fit (for a good reference on machine-learning and goodness of fit measure, cf. e.g.[32]). We present the **learning curve** for R-squared to demonstrate the response of the neural network in Figure 4. We split the data into a fraction  $x$  of random samples for training and validate on the complement  $1 - x$ , done for training set from 30,000 to 500,000, in steps of 30,000. The R-squared value is computed for each case as a measure of precision. For each  $x$ , we repeat the random sampling 10 times, for which we get the error bars.

As a comparison, one might imagine that since there is an underlying pattern being machine-learned, a simple regression might suffice. That is, could one fit a hyperplane  $f(x_i) = a_0 + \sum_{i=1}^{12} a_i v_i$  to the data? We perform this over the entire dataset, and find that the best multi-linear regression obtains only  $R^2 = 0.575428$ . Introducing non-linearity and more parameters,

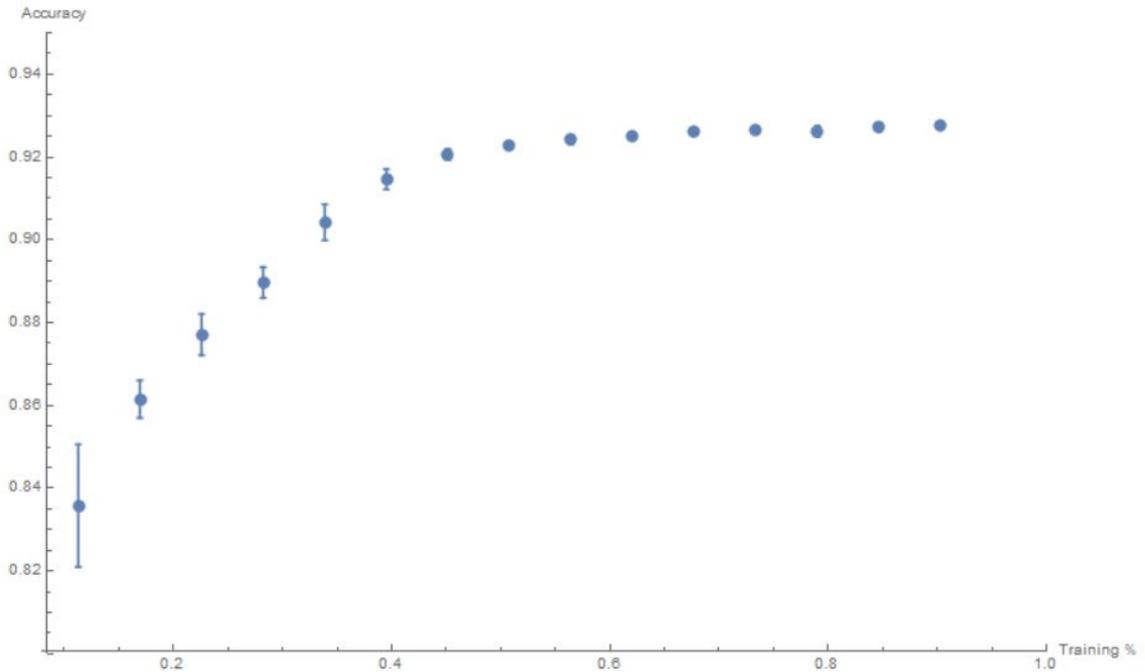


Figure 4: Learning curve for the R-square value versus fraction (seen) training data from 0 to 1.

such as fitting  $f(x_i) = a_0 + \sum_{i=1}^{12} a_i v_i + \sum_{i=1}^{12} b_i v_i^2$  does not do much better, at  $R^2 = 0.665484$ . The inherent complexity (non-linearity) of the problem is therefore best captured by our neural network approach.

## Ethics statement

This work did not use personal data, nor does the research have ethical implications. The research was performed with the highest standards of academic integrity.

## Data accessibility

This work does not have any experimental data.

## Competing interests

The authors have no conflicts of interest to declare.

## Author's contributions

Both authors contributed equally to the paper.

## Acknowledgements

We would like to thank Richard Bingham and Eric Dykeman for making available their code, fitness landscape and figures. We would like to thank Reidun Twarock, Richard Bingham and Eric Dykeman for interesting discussions. YHH is indebted to the STFC for grant ST/J00037X/1.

## Funding statement

This work was not funded by any specific grant.

## References

- [1] Ó. Rolfsson, S. Middleton, I. W. Manfield, S. J. White, B. Fan, R. Vaughan, N. A. Ranson, E. Dykeman, R. Twarock, J. Ford, et al., Direct evidence for packaging signal-mediated assembly of bacteriophage MS2, *Journal of molecular biology* 428 (2) (2016) 431–448.
- [2] O. Rolfsson, K. Toropova, N. Ranson, P. Stockley, Mutually-induced conformational switching of RNA and coat protein underpins efficient assembly of a viral capsid, *J Mol Biol* 401 (2010) 309–322. doi:10.1016/j.jmb.2010.05.058.
- [3] E. C. Dykeman, P. G. Stockley, R. Twarock, Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy, *Proceedings of the National Academy of Sciences* 111 (14) (2014) 5361–5366.
- [4] V. L. Morton, E. C. Dykeman, N. Stonehouse, A. E. Ashcroft, R. Twarock, P. Stockley, The impact of viral RNA on assembly pathway selection, *J Mol Biol* 401 (2010) 298–308. doi:DOI:10.1016/j.jmb.2010.05.059.

- [5] P. G. Stockley, S. J. White, E. Dykeman, I. Manfield, O. Rolfsson, N. Patel, R. Bingham, A. Barker, E. Wroblewski, R. Chandler-Bostock, et al., Bacteriophage MS2 genomic RNA encodes an assembly instruction manual for its capsid, *Bacteriophage* 6 (1) (2016) e1157666.
- [6] P. G. Stockley, R. Twarock, S. E. Bakker, A. M. Barker, A. Borodavka, E. Dykeman, R. J. Ford, A. R. Pearson, S. E. Phillips, N. A. Ranson, et al., Packaging signals in single-stranded RNA viruses: nature’s alternative to a purely electrostatic assembly mechanism, *Journal of biological physics* 39 (2) (2013) 277–287.
- [7] E. C. Dykeman, P. G. Stockley, R. Twarock, Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome, *Journal of molecular biology* 425 (17) (2013) 3235–3249.
- [8] P. G. Stockley, N. A. Ranson, R. Twarock, A new paradigm for the roles of the genome in ssRNA viruses, *Future Virology* 8 (6) (2013) 531–543.
- [9] K. C. Dent, R. Thompson, A. M. Barker, J. A. Hiscox, J. N. Barr, P. G. Stockley, N. A. Ranson, The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release, *Structure* 21 (7) (2013) 1225–1234.
- [10] R. I. Koning, J. Gomez-Blanco, I. Akopjana, J. Vargas, A. Kazaks, K. Tars, J. M. Carazo, A. J. Koster, Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure in situ, *Nature communications* 7.
- [11] J. A. Geraets, E. C. Dykeman, P. G. Stockley, N. A. Ranson, R. Twarock, Asymmetric genome organization in an RNA virus revealed via graph-theoretical analysis of tomographic data, *PLoS Comput Biol* 11 (3) (2015) e1004146.
- [12] N. Patel, E. C. Dykeman, R. H. Coutts, G. P. Lomonosoff, D. J. Rowlands, S. E. Phillips, N. Ranson, R. Twarock, R. Tuma, P. G. Stockley, Revealing the density of encoded functions in a viral RNA, *Proceedings of the National Academy of Sciences* 112 (7) (2015) 2227–2232.

- [13] R. Twarock, P. G. Stockley, RNA-mediated virus assembly: mechanisms and consequences for viral evolution and therapy, *Annual Review of Biophysics*.
- [14] R. Twarock, G. Leonov, P. G. Stockley, Hamiltonian path analysis of viral genomes, *Nature communications* 9 (1) (2018) 2021.
- [15] S. Shakeel, E. C. Dykeman, S. J. White, A. Ora, J. J. Cockburn, S. J. Butcher, P. G. Stockley, R. Twarock, Genomic RNA folding mediates assembly of human parechovirus, *Nature Communications* 8 (1) (2017) 5.
- [16] H. Stewart, R. Bingham, S. White, E. Dykeman, C. Zothner, A. Tuplin, P. G. Stockley, R. Twarock, M. Harris, Identification of novel RNA secondary structures within the hepatitis C virus genome reveals a cooperative involvement in genome packaging, *Scientific reports* 6 (2016) 22952.
- [17] N. Patel, S. J. White, R. F. Thompson, R. Bingham, E. U. Weiß, D. P. Maskell, A. Zlotnick, E. C. Dykeman, R. Tuma, R. Twarock, et al., HBV RNA pre-genome encodes specific motifs that mediate interactions with the viral core protein that promote nucleocapsid assembly, *Nature microbiology* 2 (2017) 17098.
- [18] N. Patel, E. Wroblewski, G. Leonov, S. E. Phillips, R. Tuma, R. Twarock, P. G. Stockley, Rewriting nature’s assembly manual for a ssRNA virus, *Proceedings of the National Academy of Sciences* (2017) 201706951.
- [19] P.-P. Dechant, P. G. Stockley, R. Twarock, A new antiviral approach exploiting insights into packaging signal-mediated assembly.
- [20] A. Zlotnick, To build a virus capsid: an equilibrium model of the self assembly of polyhedral protein complexes, *Journal of molecular biology* 241 (1) (1994) 59–67.
- [21] E. C. Dykeman, P. G. Stockley, R. Twarock, Building a viral capsid in the presence of genomic RNA, *Physical Review E* 87 (2) (2013) 022717.
- [22] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, *The journal of physical chemistry* 81 (25) (1977) 2340–2361.

- [23] E. C. Dykeman, A model for viral assembly around an explicit RNA sequence generates an implicit fitness landscape, *Biophysical Journal* 113 (3) (2017) 506–516.
- [24] R. Twarock, R. J. Bingham, E. C. Dykeman, P. G. Stockley, A modelling paradigm for RNA virus assembly, *Current opinion in virology*.
- [25] M. Eigen, Viruses: evolution, propagation, and defense, *Nutrition reviews* 58 (s1).
- [26] R. J. Bingham, E. C. Dykeman, R. Twarock, RNA virus evolution via a quasispecies-based model reveals a drug target with a high barrier to resistance, *Viruses* 9 (11) (2017) 347.
- [27] Y.-H. He, Deep-Learning the Landscape, [arXiv:1706.02714](https://arxiv.org/abs/1706.02714).
- [28] Y.-H. He, Machine-learning the string landscape, *Phys. Lett. B* 774 (2017) 564–568. doi:10.1016/j.physletb.2017.10.024.
- [29] Y.-H. He, The Calabi-Yau Landscape: from Geometry, to Physics, to Machine-Learning, [arXiv:1812.02893](https://arxiv.org/abs/1812.02893).
- [30] Wolfram Research Inc., *Mathematica*, Version 11.3, champaign, IL, 2018.
- [31] Martín Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, software available from [tensorflow.org](https://tensorflow.org). URL <http://tensorflow.org/>
- [32] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1st Edition, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1994.