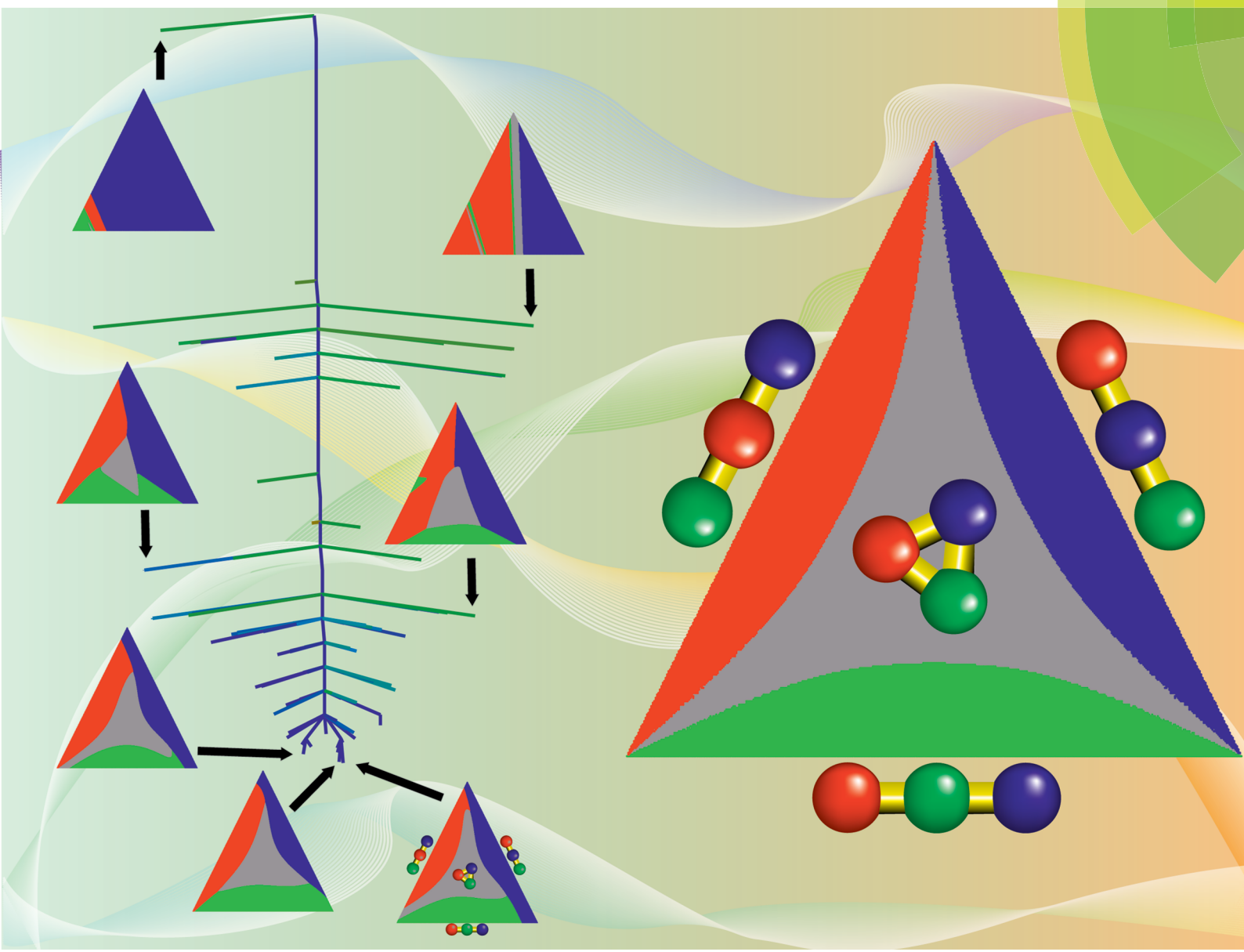
Volume 19 | Number 20 | 28 May 2017 | Pages 12561–13388

PCCP

Physical Chemistry Chemical Physics

rsc.li/pccp



ISSN 1463-9076

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

David J. Wales *et al.*

Energy landscapes for machine learning

PCCP

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2017, **19**, 12585

Energy landscapes for machine learning

Andrew J. Ballard,[a] Ritankar Das,[a] Stefano Martiniani,[a] Dhagash Mehta,[b] Levent Sagun,[c] Jacob D. Stevenson[d] and David J. Wales*[a]

Received 19th February 2017,
Accepted 20th March 2017

DOI: 10.1039/c7cp01108c

rsc.li/pccp

Machine learning techniques are being increasingly used as flexible non-linear fitting and prediction tools in the physical sciences. Fitting functions that exhibit multiple solutions as local minima can be analysed in terms of the corresponding machine learning landscape. Methods to explore and visualise molecular potential energy landscapes can be applied to these machine learning landscapes to gain new insight into the solution space involved in training and the nature of the corresponding predictions. In particular, we can define quantities analogous to molecular structure, thermodynamics, and kinetics, and relate these emergent properties to the structure of the underlying landscape. This Perspective aims to describe these analogies with examples from recent applications, and suggest avenues for new interdisciplinary research.

I. Introduction

Optimisation problems abound in computational science and technology. From force field development to thermodynamic sampling, bioinformatics and computational biology, optimisation methods are a crucial ingredient of most scientific disciplines.[1] Geometry optimisation is a key component of the potential energy landscapes approach in molecular science, where emergent properties are predicted from local minima and the transition states that connect them.[2,3] This formalism has been applied to a wide variety of physical systems, including atomic and molecular clusters, biomolecules, mesoscopic models, and glasses, to understand their structural properties, thermodynamics and kinetics. The methods involved in the computational potential energy landscapes approach amount to optimisation of a non-linear function (energy) in a high-dimensional space (configuration space). Machine learning problems employ similar concepts: the training of a model is performed by optimising a cost function with respect to a set of adjustable parameters.

Understanding how emergent observable properties of molecules and condensed matter are encoded in the underlying potential energy surface is a key motivation in developing the theoretical and computational aspects of energy landscape research. The fundamental insight that results has helped to unify our understanding of how structure-seeking systems, such as 'magic number' clusters, functional biomolecules, crystals and self-assembling structures, differ from amorphous materials and landscapes that exhibit broken ergodicity.[3–6] Rational design of new molecules and materials can exploit this insight, for example associating landscapes for self-assembly with a well defined free energy minimum that is kinetically accessible over a wide range of temperature. This structure, where there are no competing morphologies separated from the target by high barriers, corresponds to an 'unfrustrated' landscape.[2,5,7,8] In contrast, designs for multifunctional systems, including molecules with the capacity to act as switches, correspond to multifunnel landscapes.[9,10]

In this Perspective we illustrate how the principles and tools of the potential energy landscape approach can be applied to machine learning (ML) landscapes. Some initial results are presented, which indicate how this view may yield new insight into ML training and prediction in the future. We hope our results will be relevant for both the energy landscapes and machine learning communities. In particular, it is of fundamental interest to compare these ML landscapes to those that arise for molecules and condensed matter. The ML landscape provides both a means to visualise and interpret the cost function solution space and a computational framework for quantitative comparison of solutions.

II. The energy landscape perspective

The potential energy function in molecular science is a surface defined in a (possibly very high-dimensional) configuration space, which represents all possible atomic configurations.[2,11] In the potential energy landscape approach, this surface is divided into basins of attraction, each defined by the steepest-descent pathways that lead to a particular local minimum.[2,11] The mapping from a continuous multidimensional surface to local minima

[a] *University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, UK. E-mail: dw34@cam.ac.uk*
[b] *Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, IN, USA*
[c] *Mathematics Department, Courant Institute, New York University, NY, USA*
[d] *Microsoft Research Ltd, 21 Station Road, Cambridge, CB1 2FB, UK*

can be very useful. In particular, it provides a route to prediction of structure and thermodynamics.[2] Similarly, transitions between basins can be characterised by geometric transition states (saddle points of index one), which lie on the boundary between one basin and another.[2] Including these transition states in our description of the landscape produces a kinetic transition network,[3,12,13] and access to dynamical properties and 'rare' events.[14–17] The pathways mediated by these transition states correspond to processes such as molecular rearrangements, or atomic migration. For an ML landscape we can define the connectivity between minima that represent different locally optimal fits to training data in an analogous fashion. To the best of our knowledge, interpreting the analogue of a molecular rearrangement mechanism for the ML landscape has yet to be explored.

Construction of a kinetic transition network[3,12,13] also provides a convenient means to visualise a high-dimensional surface. Disconnectivity graphs[4,18] represent the landscape in terms of local minima and connected transition states, reflecting the barriers and topology through basin connectivity. The overall structure of the disconnectivity graph can provide immediate insight into observable properties:[4] a single-funnelled landscape typically represents a structure-seeking system that equilibrates rapidly, whereas multiple funnels indicate competing structures or morphologies, which may be manifested as phase transitions and even glassy phenomenology. Locating the global minimum is typically much easier for single funnel landscapes.[19]

The decomposition of a surface into minima and transition states is quite general and can naturally be applied to systems that do not correspond to an underlying molecular model. In particular, we can use this strategy for machine learning applications, where training a model amounts to minimisation of a cost function with respect to a set of parameters. In the language of energy landscapes, the machine learning cost function plays the role of energy, and the model parameters are the 'coordinates' of the landscape. The minimised structures represent the optimised model parameters for different training iterations. The transition states are the index one saddle points of the landscape.[20]

Energy landscape methods[2] could be particularly beneficial to the ML community, where non-convex optimisation has sometimes been viewed as less appealing, despite supporting richer models with superior scalability.[21] The techniques described below could provide a useful computational framework for exploring and visualising ML landscapes, and at the very least, an alternative view to non-convex optimisation. The first steps in this direction have recently been reported.[22–24] The results may prove to be useful for various applications of machine learning in the physical sciences. Examples include fitting potential energy surfaces, where neural networks have been used extensively[25–32] for at least 20 years.[33–35] Recent work includes prediction of binding affinities in protein–ligand complexes,[36] applications to the design of novel materials,[37,38] and refinement of transition states[39] using support vector machines.[40]

In the present contribution we illustrate the use of techniques from the energy landscapes field to several ML examples, including non-linear regression, and neural network classification. When surveying the cost function landscapes, we employed the same techniques and algorithms as for the molecular and condensed matter systems of interest in the physical sciences: specifically, local and global minima were obtained with the basin-hopping method[41–43] using a customised LBFGS minimisation algorithm.[44] Transition state searches employed the doubly-nudged[45,46] elastic band[47,48] approach and hybrid eigenvector-following.[49,50] These methods are all well established, and will not be described in detail here. We used the python-based energy landscape explorer pele,[51] with a customised interface for ML systems, along with the GMIN,[52] OPTIM,[53] and PATHSAMPLE[54] programs, available for download under the Gnu General Public Licence.

III. Prediction for classification of outcomes in local minimisation

Neural networks (NN) have been employed in two previous classification problems that analyse the underlying ML landscape, namely predicting which isomer results from a molecular geometry optimisation[23] and for patient outcomes in a medical diagnostic context.[24] Some of the results from the former study will be illustrated here, and we must carefully distinguish isomers corresponding to minima of a molecular potential energy landscape from the ML landscape of solutions involved in predicting which of the isomers will result from geometry optimisation starting from a given molecular configuration. We must also distinguish this classification problem from *ab initio* structure prediction: the possible outcomes of the geometry optimisation must be known in advance, either in terms of distinct isomers, or the range that they span in terms of potential energy or appropriate structural order parameters for larger systems. The ability to make predictions that are sufficiently reliable could produce significant savings in computational effort for applications that require repeated local minimisation. Examples include basin-sampling for calculating global thermodynamic properties in systems subject to broken ergodicity,[55] construction of kinetic transition networks,[56] and methods to estimate the volume of basins of attraction for jammed packings, which provide measures of configurational entropy in granular packings.[57] Here the objective would be to terminate the local minimisation as soon as the outcome could be predicted with sufficient confidence to converge the properties of interest.[23,58]

The test system considered here is a simple triatomic cluster with four distinguishable local minima that represent the possible outcomes for local minimisation. This system has previously served as a benchmark for visualising the performance of different geometry optimisation approaches.[59–61] The potential energy, V, is a sum of pairwise terms, corresponding to the Lennard-Jones form,[62] and the three-body Axilrod–Teller function,[63] which represents an instantaneous induced dipole–induced dipole interaction:

$$V = 4\varepsilon \sum_{i<j} \left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} \right] + \gamma \sum_{i<j<k} \left[\frac{1 + 3\cos\theta_1 \cos\theta_2 \cos\theta_3}{\left(r_{ij} r_{ik} r_{jk}\right)^3} \right], \quad (1)$$

where θ_1, θ_2 and θ_3 are the internal angles of the triangle formed by atoms i, j, k. r_{ij} is the distance between atoms i and j. The influence of the three-body term is determined by the magnitude of the parameter γ, and we use $\gamma = 2$ in reduced units, where the equilateral triangle ($V = -2.185\varepsilon$) competes with three permutational isomers of a linear minimum ($V = -2.219\varepsilon$). In the triangle the bond length is 1.16875σ, and in the linear minima the distance from the centre atom to its neighbours is 1.10876σ.

The objective of our machine learning calculations for this system is a classification, to predict which of the four minima a local minimisation would converge to, given data for one or more configurations. The data in question could be combinations of the energy, gradient, and geometrical parameters for the structures in the optimisation sequence. Our initial tests, which are mostly concerned with the structure of the ML solution landscape, employed the three interparticle separations r_{12}, r_{13} and r_{23} as data.[23] Inputs were considered for separations corresponding to the initial geometry, and for one or more configurations in the minimisation sequence.

A database of 10 000 energy minimisations, each initiated from different atomic coordinates distributed in a cube of side length $2\sqrt{3}\sigma$, was created using the customised LBFGS routine in our OPTIM program[53] (LBFGS is a limited memory quasi-Newton Broyden,[64] Fletcher,[65] Goldfarb,[66] Shanno,[67] scheme). Each minimisation was converged until the root mean square gradient fell below 10^{-6} reduced units, and the outcome (one of the four minima) was recorded.

The neural networks used in the applications discussed below all have three layers, corresponding to input, output, and hidden nodes.[68] A single hidden layer has been used successfully in a variety of previous applications.[69] A bias was added to the sum of weights used in the activation function for each hidden node, w_j^{bh}, and each output node, w_i^{bo}, as illustrated in Fig. 1. For inputs we consider $\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^{N_{\mathrm{data}}}\}$, where N_{data} is the number of minimisation sequences in the training or test set, each of which has dimension N_{in}, so that $\mathbf{x}^\alpha = \left\{x_1^\alpha, \ldots, x_{N_{\mathrm{in}}}^\alpha\right\}$. For this example there are four outputs corresponding to the four possible local minima (as in Fig. 1), denoted by y_i^{NN}, with

$$y_i^{\mathrm{NN}} = w_i^{\mathrm{bo}} + \sum_{j=1}^{N_{\mathrm{hidden}}} w_{ij}^{(1)} \tanh\left[w_j^{\mathrm{bh}} + \sum_{k=1}^{N_{\mathrm{in}}} w_{jk}^{(2)} x_k\right], \qquad (2)$$

for a given input $\mathbf{x}$ and link weights $w_{ij}^{(1)}$ between hidden node j and output i, and $w_{jk}^{(2)}$ between input k and hidden node j.

The four outputs were converted into softmax probabilities as

$$p_c^{\mathrm{NN}}(\mathbf{W};\mathbf{X}) = \mathrm{e}^{y_c^{\mathrm{NN}}} \Big/ \sum_{a=0}^{3} \mathrm{e}^{y_a^{\mathrm{NN}}}. \qquad (3)$$

This formulation is designed to reduce the effect of outliers.

In each training phase we minimise the cost (objective) function, $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X})$, with respect to the variables $w_{ij}^{(1)}$, $w_{jk}^{(2)}$, w_j^{bh} and w_i^{bo}, which are collected into the vector of weights $\mathbf{W}$. An L^2 regularisation term is included in $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X})$, corresponding to a sum of squares of the independent variables, multiplied by a constant coefficient λ, which is chosen in advance and fixed.

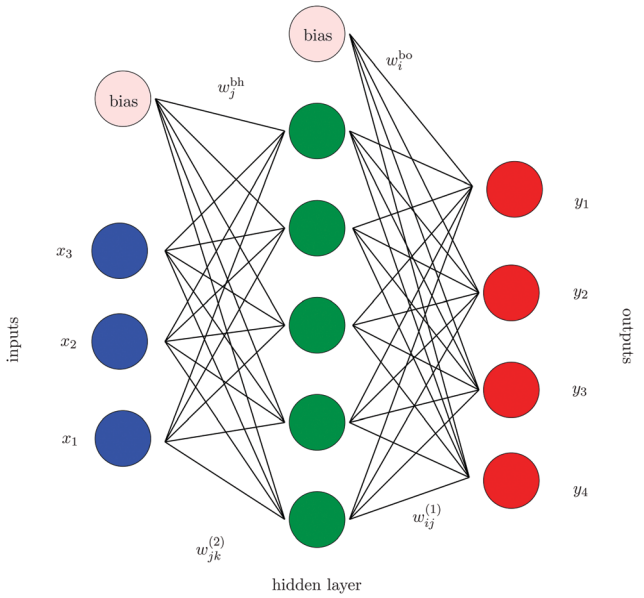


Fig. 1 Organisation of a three-layer neural network with three inputs, five hidden nodes, and four outputs. The training variables are the link weights, $w_{jk}^{(2)}$, $w_{ij}^{(1)}$, and the bias weights, w_j^{bh} and w_i^{bo}.

This term is intended to prevent overfitting; it disfavours large values for individual variables. We have considered applications where the regularisation is applied to all the variables, and compared the results with a sum that excludes the bias weights. Regularising over all the variables has the advantage of eliminating the zero Hessian eigenvalue that otherwise results from an additive shift in all the w_i^{bo}. Such zero eigenvalues are a consequence of continuous symmetries in the cost function (Noether's theorem). For molecular systems such modes commonly arise from overall translation and rotation, and are routinely dealt with by eigenvalue shifting or projection using the known analytical forms for the eigenvectors.[2,70] Larger values of the parameter λ simplify the landscape, reducing the number of minima. This result corresponds directly with the effect of compression for a molecular system,[71] which has been exploited to accelerate global optimisation. A related hypersurface deformation approach has been used to treat graph partitioning problems.[72]

For each LBFGS geometry optimisation sequence, d, with N_{in} inputs collected into data item $\mathbf{x}^d$, we know the actual outcome or class label, $c(d) = 0$, 1, 2 or 3, corresponding to the local minimum at convergence. The networks were trained using either 500 or 5000 of the LBFGS sequences, chosen at random with no overlap, by minimising

$$E^{\mathrm{NN}}(\mathbf{W};\mathbf{X}) = -\frac{1}{N_{\mathrm{data}}} \sum_{d=1}^{N_{\mathrm{data}}} \ln p_{c(d)}^{\mathrm{NN}}(\mathbf{W};\mathbf{X}) + \lambda \mathbf{W}^2. \qquad (4)$$

Results were compared for different values of λ and in some cases for regularisation excluding the bias weights. These formulations, including analytic first and second derivatives with respect to $\mathbf{W}$, have been programmed in our GMIN global optimisation program[52] and in our OPTIM code for analysing stationary points and pathways.[53] $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X})$ was minimised

using the same customised LBFGS routine that was employed to create the database of minimisation sequences for the triatomic cluster.

In the testing phase the variables $\mathbf{W}$ are fixed for a particular local minimum of $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X}_{\mathrm{train}})$ obtained with the training data, and we evaluate $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X}_{\mathrm{test}})$ for 500 or 5000 of the minimisation sequences outside the training set. The results did not change significantly between the larger and smaller training and testing sets.

We first summarise some results for ML landscapes corresponding to input data for the three interparticle distances at each initial random starting point. The number of local minima obtained[23] was 162, 2559, 4752 and 19 045 for three, four, five and six hidden nodes, respectively, with 1504, 10 112, 18 779 and 34 052 transition states. The four disconnectivity graphs are shown in Fig. 2. In each case the vertical axis corresponds to $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X}_{\mathrm{train}})$, and branches terminate at the values for particular local minima. At a regular series of thresholds for E^{NN} we perform a superbasin analysis,[18] which segregates the ML solutions into disjoint sets. Local minima that can interconvert *via* a pathway where the highest transition state lies below the threshold are in the same superbasin. The branches corresponding to different sets or individual minima merge at the threshold energy where they first belong to a common superbasin. In this case we have coloured the branches according to the misclassification index, discussed further in Section V, which is defined as the fraction of test set images that are misclassified by the minimum in question or the global minimum, but not both. All the low-lying minima exhibit small values, meaning that they perform much like the global minimum. The index rises to between 0.2 and 0.4 for local minima with higher values of $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X}_{\mathrm{train}})$. These calculations were performed using the pele[51] ML interface for the formulation in eqn (2), where regularisation excluded the weights for the bias nodes.[23]

When more hidden nodes are included the dimensionality of the ML landscape increases, along with the number of local minima and transition states. This observation is in line with well known results for molecular systems: as the number of atoms and configurational degrees of freedom increases the number of minima and transition states increases exponentially.[73,74] The residual error reported by $E^{\mathrm{NN}}(\mathbf{W};\mathbf{X}_{\mathrm{train}})$ decreases as more parameters are included, and so there is a trade-off between the complexity of the landscape and the quality of the fit.[23]

The opportunities for exploiting tools from the potential energy landscape framework have yet to be investigated systematically. As an indication of the properties that might be of interest we now illustrate a thermodynamic analogue corresponding to the heat capacity, C_{V}. Peaks in C_{V} are particularly insightful, and in molecular systems they correspond to phase-like transitions. Around the transition temperature the occupation probability shifts between local minima with qualitatively different properties in terms of energy and entropy: larger peaks correspond to greater differences.[75] For the ML landscape we would interpret such features in terms of fitting solutions with different properties. Understanding how and why the fits differ could be useful in combining solutions to produce better predictions. Here we simply illustrate some representative results, which suggest that ML landscapes may support an even wider range of behaviour than molecular systems.

To calculate the C_{V} analogue we use the superposition approach[2,73,76–79] where the total partition function, $Z(T)$, is written as a sum over the partition functions $Z_\alpha(T)$, for all the local minima, α. This formulation can be formally exact, but is usually applied in the harmonic approximation using normal mode analysis to represent the vibrational density of states. The normal mode frequencies are then related to the eigenvalues of the Hessian (second derivative) matrix, $\mu_\alpha(i)$, for local minimum α:

$$Z(T) = \sum_\alpha Z_\alpha(T) \approx \sum_\alpha \frac{\mathrm{e}^{-\beta E_\alpha}}{(\beta/2\pi)^{\kappa/2} \prod_{i=1}^{\kappa} \mu_\alpha(i)^{1/2}}, \qquad (5)$$

here κ is the number of non-zero eigenvalues, $\beta = 1/k_{\mathrm{B}}T$, k_{B} is the Boltzmann constant, and $E^{\mathrm{NN}}(\mathbf{W}_\alpha;\mathbf{X}_{\mathrm{train}})$ is the objective (loss) function for minimum α. T plays the role of temperature in this picture, with $C_{\mathrm{V}}(T) = (1/k_{\mathrm{B}}T^2)\partial^2 \ln Z(T)/\partial\beta^2$.

Many molecular and condensed matter systems exhibit a C_{V} peak corresponding to a first order-like melting transition, where the occupation probability shifts from a relatively small number of low energy minima, to a much larger, entropically favoured, set of higher energy structures, which are often more disordered. However, low temperature peaks below the melting temperature, corresponding to solid–solid transitions, can also occur. These features are particularly interesting, because they suggest the presence of competing low energy morphologies, which may represent a challenge for structure prediction,[80] and lead to broken ergodicity,[55,79,81–86] and slow interconversion rates that constitute 'rare events'.[14,87,88] Initial surveys of the C_{V} analogue for ML landscapes, calculated using analytical second derivatives of $E^{\mathrm{NN}}(\mathbf{W}_\alpha;\mathbf{X})$, produced plots with multiple peaks, suggesting richer behaviour than for molecular systems.

One example is shown in Fig. 3, which is based on the ML solution landscape for a neural network with three hidden nodes and inputs corresponding to the three interparticle distances at the initial geometries of all the training optimisation sequences. These results were obtained using the pele[51] ML interface for the neural network formulation in eqn (2), where regularisation did not include the weights for the bias nodes.[23] The superposition approach provides a clear interpretation for the peaks, which we achieve by calculating C_{V} from partial sums over the ML training minima, in order of increasing $E^{\mathrm{NN}}(\mathbf{W}_\alpha;\mathbf{X}_{\mathrm{train}})$. The first peak around $k_{\mathrm{B}}T \approx 0.2$ arises from competition between the lowest two minima. The second peak around $k_{\mathrm{B}}T \approx 9$ is reproduced when the lowest 124 minima are included, and the largest peak around $k_{\mathrm{B}}T \approx 20$ appears when we sum up to minimum 153. The latter solution exhibits one particularly small Hessian eigenvalue, producing a relatively large configurational entropy contribution, which increases with T. The harmonic approximation will break down here, but nevertheless serves to highlight the qualitative difference in character of minima with exceptionally small curvatures.

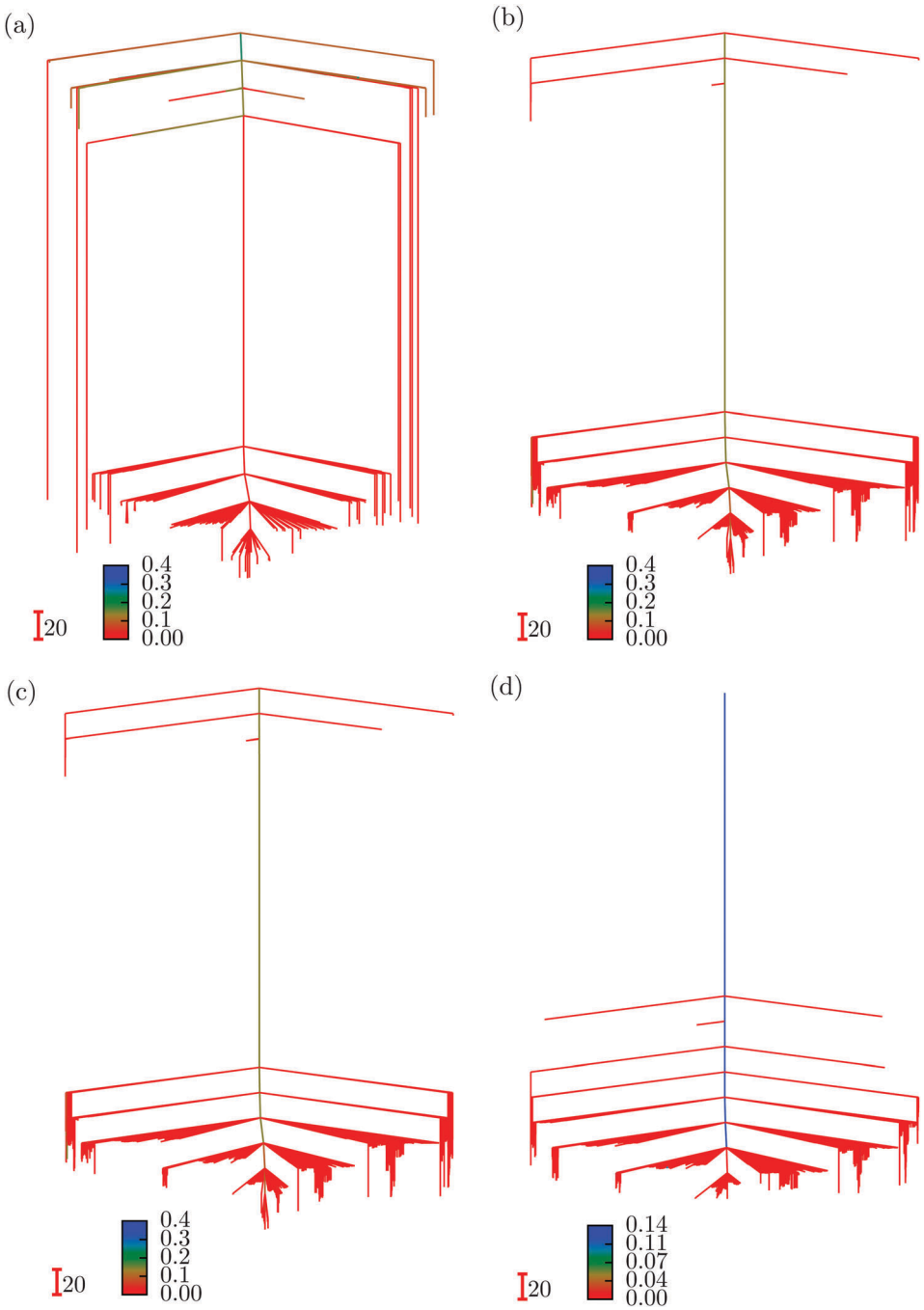


Fig. 2 Disconnectivity graphs for the fitting landscapes of a triatomic cluster. Three inputs were used for each minimisation sequence, corresponding to the three interatomic distances in the initial configuration. These graphs are for neural networks with (a) 3, (b) 4, (c) 5, and (d) 6 hidden nodes. The branches are coloured according to the misclassification distance for the local minima evaluated using training data, as described in Section V.

In molecular systems competition between alternative low energy structures often accounts for C_V peaks corresponding to solid–solid transitions, and analogues of this situation may well exist in the ML scenario. Some systematic shifts in the C_V analogue could result from the density of local minima on the ML landscape (the landscape entropy[89–92]). Understanding such effects might help to guide the construction of improved predictive tools from combinations of fitting solutions. Interpreting the ML analogues of molecular structure and interconversion rates between minima might also prove to be insightful in future work.

A subsequent study investigated the quality of the predictions using two of the three interatomic distances, r_{12} and r_{13}, and the effects of memory, in terms of input data from successive configurations chosen systematically from each geometry optimisation sequence.[93] The same database of LBFGS minimisation sequences for the triatomic cluster was used here, divided randomly into training and testing sets of equal size (5000 sequences each).

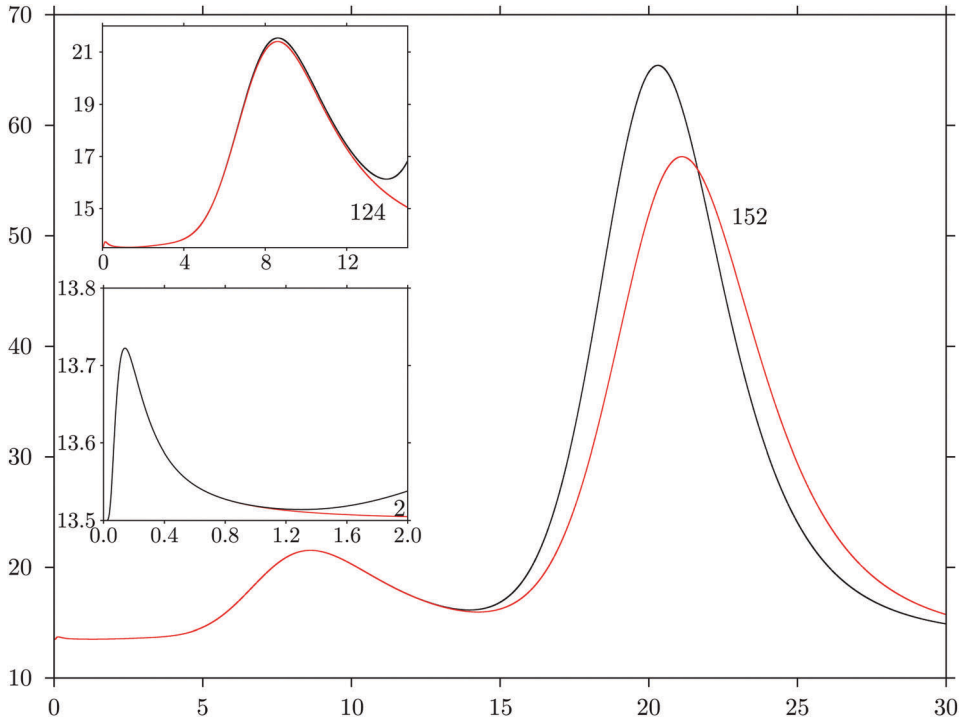


Fig. 3 Heat capacity analogue for the ML landscape defined for the dataset using only the three initial interatomic distances with three hidden nodes. The insets illustrate the convergence of the two low temperature peaks. In each plot the black curve corresponds to C_V calculated from the complete database of minima. The red curves labelled '2', '124' and '152' correspond to C_V calculated from truncated sums including only the lowest 2, 124, and 152 minima, respectively.

The quality of the classification prediction into the four possible outcomes can be quantified using the area under curve (AUC) for receiver operating characteristic (ROC) plots.[69] ROC analysis began when radar receiver operators needed to distinguish signals corresponding to aircraft from false readings, including flocks of birds. The curves are plots of the true positive rate, T_{pr}, against the false positive rate, F_{pr}, as a function of the threshold probability, P, for making a certain classification. Here, P is the threshold at which the output probability $p_0^{\mathrm{NN}}(\mathbf{W};\mathbf{X})$ is judged sufficient to predict that a minimisation would converge to the equilateral triangle, so that

$$T_{\mathrm{pr}}(\mathbf{W};\mathbf{X};P)=\sum_{d=1}^{N_{\mathrm{data}}}\delta_{c(d),0}\Theta\left(p_0^{\mathrm{NN}}(\mathbf{W};\mathbf{X})-P\right)\Big/\sum_{d=1}^{N_{\mathrm{data}}}\delta_{c(d),0},$$

$$F_{\mathrm{pr}}(\mathbf{W};\mathbf{X};P)=\sum_{d=1}^{N_{\mathrm{data}}}\left(1-\delta_{c(d),0}\right)\Theta\left(p_0^{\mathrm{NN}}(\mathbf{W};\mathbf{X})-P\right)\Big/\sum_{d=1}^{N_{\mathrm{data}}}\left(1-\delta_{c(d),0}\right), \quad (6)$$

where Θ is the Heaviside step function and δ is the Kronecker delta. The area under the curve can then be obtained by numerical integration of

$$\mathrm{AUC}(\mathbf{W};\mathbf{X})=\int_0^1 T_{\mathrm{pr}}(\mathbf{W};\mathbf{X};P)\mathrm{d}F_{\mathrm{pr}}(\mathbf{W};\mathbf{X};P). \quad (7)$$

AUC(**W**;**X**) can be interpreted as the probability that for two randomly chosen data inputs, our predictions will discriminate between them correctly. AUC values between 0.7 and 0.8 are usually considered 'fair', 0.8 to 0.9, 'good', and 0.9 to 1 'excellent'.

Fig. 4 shows the AUC values obtained with the LBFGS database when the input data consists of r_{12} and r_{13} values at different points in the minimisation sequence. Here the horizontal axis corresponds to s, the number of steps from convergence, and the AUC values therefore tend to unity when s is small, where the configurations are close to the final minimum. Each panel in the figure includes two plots, which generally coincide quite closely. The plot with the lower AUC value is the one obtained with the global minimum located for $E^{\mathrm{NN}}(\mathbf{W}_s;\mathbf{X}_{\mathrm{train}})$ for configurations s steps for convergence. The AUC value in the other plot is always greater than or equal to the value for the global minimum, since it is the maximum AUC calculated for all the local minima characterised with the same neural net architecture and any s value. Minimisation sequences that converge in fewer than s steps are padded with the initial configuration for larger s values, which is intended to present a worst case scenario.

Fig. 4 shows that the prediction quality decays in a characteristic fashion as configurations move further from the convergence limit. It also illustrates an important result, namely that the performance of the global minimum obtained with the training data is never surpassed significantly by any of the other local minima, when all these fits are applied to the test data. Hence the global minimum is clearly a good place to start if we wish to make predictions, or perhaps construct classification schemes based on more than one local minimum of the ML

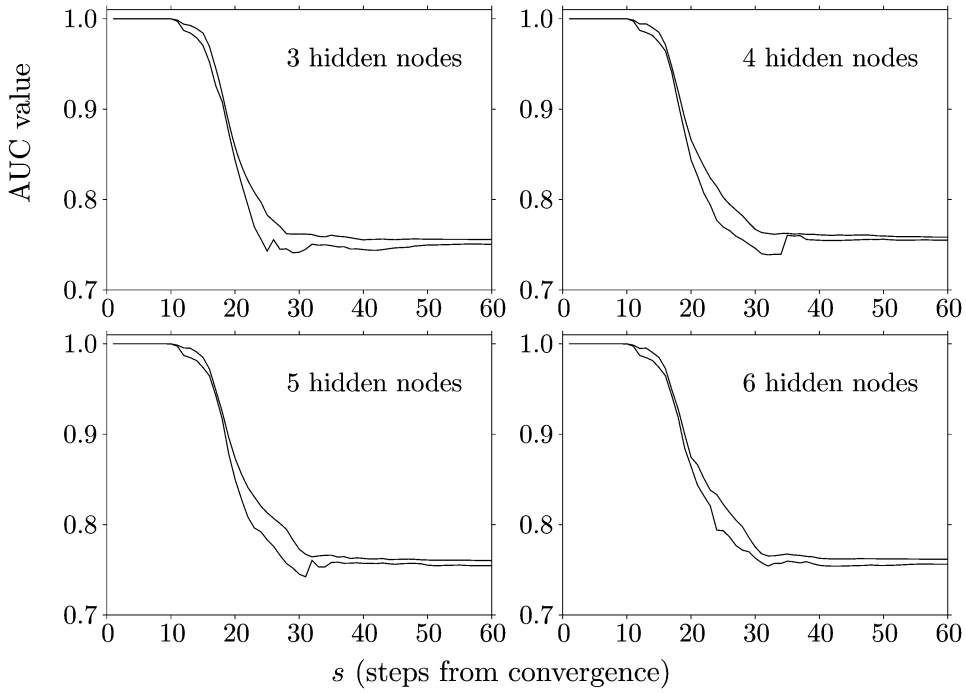


Fig. 4 AUC values for 5000 minimisation sequences in the LBFGS testing set, evaluated using the parameters obtained for the global minimum neural network fit with 5000 training sequences and $\lambda = 10^{-4}$. The four panels correspond to 3, 4, 5 or 6 hidden nodes, as marked, and the lower curve corresponds to the global minimum of $E^{\mathrm{NN}}(\mathbf{W}_\alpha;\mathbf{X}_{\mathrm{train}})$ with $\mathbf{X}_{\mathrm{train}}$ containing r_{12} and r_{13} at a single configuration in each minimisation sequence, located s steps from convergence. Each panel has a second plot of the highest AUC value for the test data attained with any local minimum obtained in training having the same number of inputs and hidden nodes, including results for all the λ values considered and for all values of s, from 1 to 80. The AUC value for the global minimum with $\lambda = 10^{-4}$ and the configuration in question is included in this set, but can be exceeded by one of the many local minima obtained over the full range of λ and s. Beyond s around 60 the plots are essentially flat.

landscape obtained in training. The corresponding fits and AUC calculations were all rerun to produce Fig. 4 using the fitting function defined in eqn (2) and regularisation over all variables, for comparison with the simplified bias weighting employed in ref. 93. There is no significant difference between our results for the two schemes.

IV. Non-linear regression

Regression is perhaps the most well-known task in machine learning, referring to any process for estimating the relationships between dependent and independent variables. As we show in this section, even a relatively simple non-nonlinear regression problem leads to a rich ML landscape. As in the standard regression scenario, we consider a set of N_{data} data points $D = ((x_1,t_1),\ldots,(x_{N_{\mathrm{data}}},t_{N_{\mathrm{data}}}))$, and a model $y(x;\boldsymbol{q})$ that we wish to fit to D by adjusting M parameters $\boldsymbol{q} = (q_1,q_2,\ldots,q_M)$. In this example, we investigate the following non-linear model:

$$y(x;\boldsymbol{q}) = \mathrm{e}^{-q_1 x}\sin(q_2 x + q_3)\sin(q_4 x + q_5). \tag{8}$$

Our regression problem is one-dimensional (x is a scalar), with a five-dimensional vector $\boldsymbol{q}$ that parameterises the model.

We performed regression on the above problem, with a dataset D consisting of $N_{\mathrm{data}} = 100$ data points with x_i values sampled uniformly in $[0,3\pi]$, and corresponding t_i values given by our model with added Gaussian white noise (mean zero, $\sigma = 0.02$):

$$t_i = y(x_i;\boldsymbol{q}^\star) + \text{noise}, \tag{9}$$

with a particular *ad hoc* parameter choice $\boldsymbol{q}$ = (0.1, 2.13, 0.0, 1.34, 0.0). The cost function we minimise is a standard sum of least squares:

$$E(\boldsymbol{q}) = \sum_{i=1}^{N_{\mathrm{data}}} [t_i - y(x_i;\boldsymbol{q})]^2. \tag{10}$$

The objective of this regression problem is to find a best-fit mapping from input (x) to target variables (t). Intuitively, we expect minimisation of eqn (10) with respect to the parameters $\boldsymbol{q}$ to yield an optimal value $\boldsymbol{q} \approx \boldsymbol{q}^\star$. However, since E is a non-convex function of $\boldsymbol{q}$, there are multiple solutions to the equation $\nabla E(\boldsymbol{q}) = 0$ and hence the outcome depends on the minimisation procedure and the initial conditions.

We explored the landscape described by $E(\boldsymbol{q})$, and display the various solutions in Fig. 5 alongside our data and $y(x;\boldsymbol{q} = \boldsymbol{q}^\star)$. In this case, the global minimum is a fairly accurate representation of the solution used to generate D. However, 88 other solutions were found which do not accurately represent the data, despite being valid local minima. In Fig. 6 we show the disconnectivity graph for $E(\boldsymbol{q})$. Here the vertical axis corresponds to $E(\boldsymbol{q})$, and branches terminate at the values for corresponding local minima, as in Section III. The graph shows that the energy of

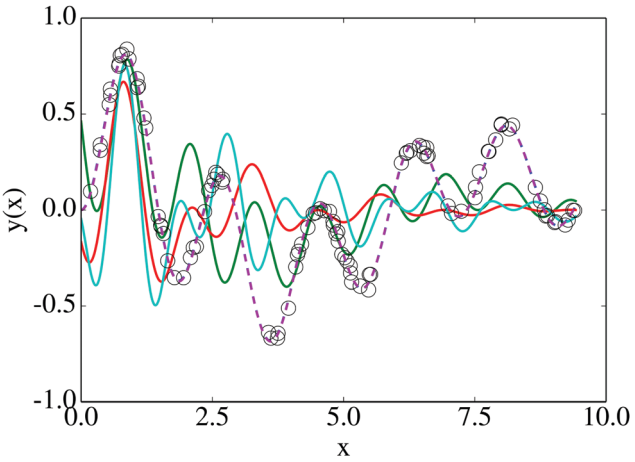


Fig. 5 Results from nonlinear regression for the model given by eqn (8): the global minimum of the cost function (dashed line) is plotted with various local minima solutions (solid lines) and the data used for fitting (black circles). The model used to generate the data is indistinguishable from the curve corresponding to the global minimum.

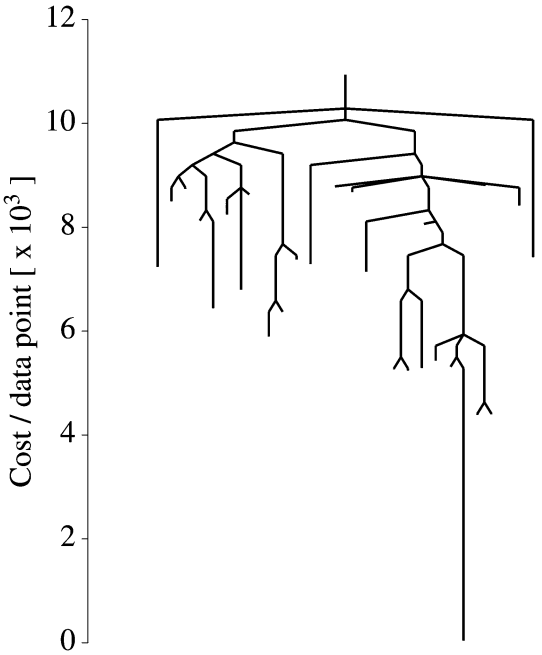


Fig. 6 Disconnectivity graph for a nonlinear regression cost function [eqn (8)]. Each branch terminates at a local minimum at the value of $E(\mathbf{q})$ for that minimum. By following the lines from one minimum to another, one can read off the energy barrier on the minimum energy path connecting them.

the global minimum solution is separated from the others by an order of magnitude, and is clearly the best fit to our data (dotted curve in Fig. 5). The barriers in this representation can be interpreted in terms of transformations between different training solutions in parameter space, and could be indicative of the distinctiveness of the minima they connect. The minima of this landscape were found by the basin-hopping method,[41–43] as described in Section II.

V. Digit recognition using a neural network

The next machine learning landscape we explore here is another artificial neural network, this time trained for digit recognition on the MNIST dataset.[94] Our network architecture consists of 28×28 input nodes corresponding to input image pixels, a single hidden layer with 10 nodes, and a softmax output layer of 10 nodes, which represent the 10 digit classes. This model contains roughly 8000 adjustable parameters, which quantify the weight of given nodes in activating one of their successors. The cost function we optimise for this classification example is the same multinomial logistic regression function that is described above in Section III, which is standard for classification problems. Here 'logistic' means that the dependent variable (outcome) is a category, in this case the assignment of the image for a digit, and 'multinomial' means that there are more than two possible outcomes. An L^2 regularisation term was again added to the cost function, as described in Section III. Unless otherwise mentioned, all the results described below are for a regularisation coefficient of $\lambda = 0.1$.

The neural network defined above is quite small, and is not intended to compete with well-established models trained on MNIST.[95] Rather, our goal in this Perspective is to gain insight into the landscape. This aim is greatly assisted by using a model that is smaller, yet still behaves similarly to more sophisticated implementations. To converge the disconnectivity graph, Fig. 7, in particular the transition states, the model was trained on $N_{\text{data}} = 1000$ data points. The results assessing performance, Fig. 8 and 9, were tested on $N_{\text{data}} = 10\,000$ images.

We explored the landscape of this network for several different values of the L^2 regularisation parameter λ. The graph shown

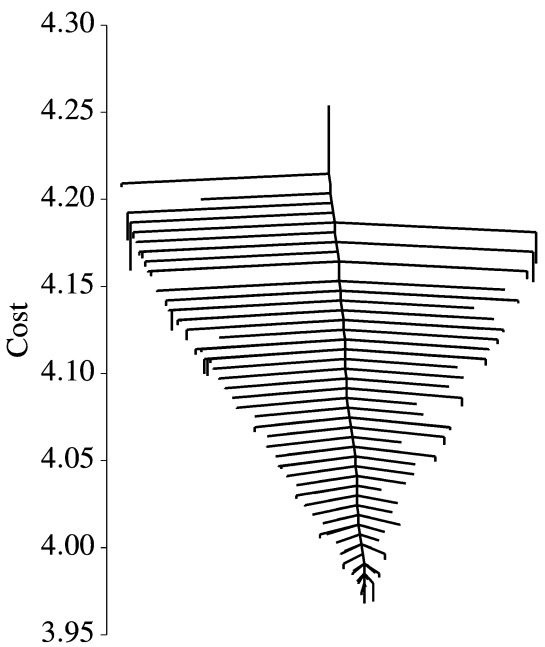


Fig. 7 Disconnectivity graph of neural network ML solutions for digit recognition.

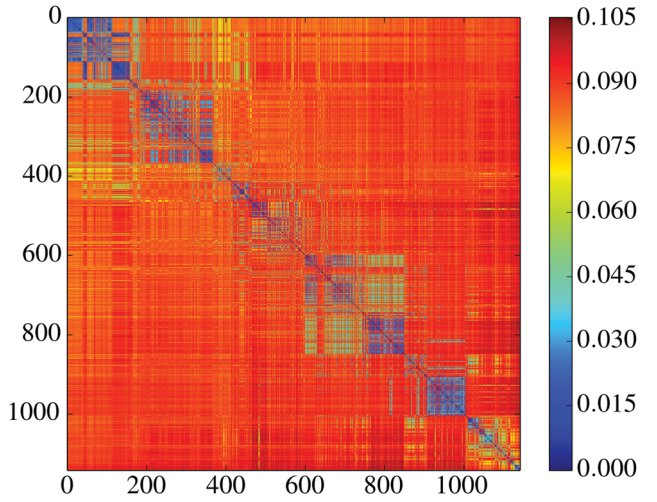


Fig. 8 Misclassification heat map for various solutions to the digit recognition problem on the ML landscape. This map displays the degree of similarity of any pair of minima based upon correct test set classification. See text for details.

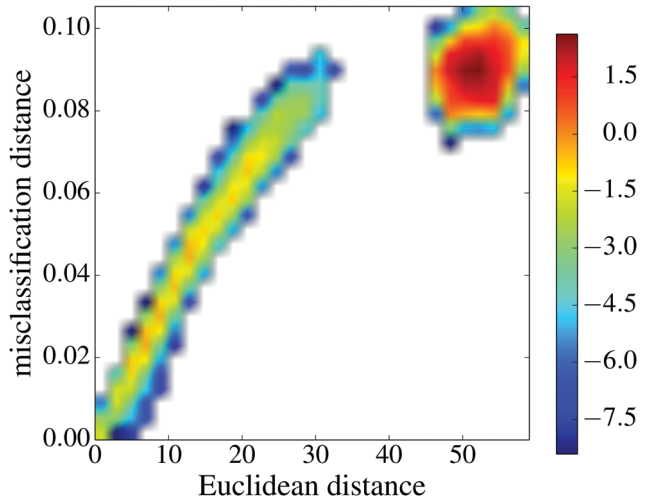


Fig. 9 Joint probability density plot of minimum–minimum Hamming distance and geometric distance in parameter space. The distance metric in parameter space is correlated with the misclassification distance between minima. The probability density is coloured on a log scale.

in Fig. 7, with $\lambda = 0.01$, is representative of all the others: we observe a single funnel landscape, where, in contrast to the nonlinear regression example, all of the minima occur in a narrow range of energy (cost) values. This observation is consistent with recent work suggesting that the energies of local minima for neural networks are spaced exponentially close to the global minimum[96,97] with the number of variables (number of optimised parameters or dimensionality) of the system.

We next assess the performance profile of the NN minima by calculating the misclassification rate on an independent test set. Judging by average statistics the minima seem to perform very similarly: the fraction f of misclassified test set images is comparable for most of them, with $0.133 \leq f \leq 0.162$ (mean $\bar{f} = 0.148$, standard deviation $\sigma_f = 0.0043$). This observation is also consistent with previous results where local minima were found to perform quite similarly in terms of test set error.[98,99] When looking beyond average statistics, however, we uncover more interesting behaviour. To this end we introduce a misclassification distance ℓ between pairs of minima, which we define as the fraction of test set images that are misclassified by one minimum but not both.[100] A value $\ell_{ij} = 0$ implies that all images are classified in the same way by the two minima; a value $\ell_{ij} = \ell_{ij}^{\max} = f_i + f_j$ implies that every misclassified image by i was correctly classified by j, and every misclassified image by j was correctly classified by i. In Fig. 8 we display the matrix ℓ_{ij}, which shows that the minima cluster into groups that are self-similar, and distinct from other groups. So, although all minima perform almost identically when considering the misclassification rate alone, their performance looks quite distinct when considering the actual sets of misclassified images. We hypothesise that this behaviour is due to a saturated information capacity for our model. This small neural network can only encode a certain amount of information during the training process. Since there are many training images, there is much more information to encode than it is possible to store. The clustering of minima in Fig. 8 then probably reflects the differing information content that each solution retains. Here it is important to remember that each of the minima were trained on the same set of images; the distinct minima arise solely from the different starting configurations prior to optimisation.

The misclassification similarity can be understood from the underlying ML landscape. We investigated correlations between misclassification distance and distance in parameter space. In Fig. 9 we display the joint distribution of the misclassification distance and Euclidean distance (L^2 norm) between the parameter values for each pair of minima. We see that for a range of values these two measures are highly correlated, indicating that the misclassification distance between minima is determined by their proximity on the underlying landscape. Interestingly, for very large values of geometric distance there is a large (yet seemingly random) misclassification distance.

The seemingly random behaviour could possibly be the result of symmetry with respect to permutation of neural network parameters. There exist a large number of symmetry operations for the parameter space that leave the NN prediction unchanged, yet would certainly change the L^2 distance with respect to another reference point. A more rigorous definition of distance (currently unexplored), would take such symmetries into account. There are at least two such symmetry operations.[68] The first of these results from the antisymmetry of the tanh activation function: inverting the sign of all weights and bias leading into a node will lead to an inverted output from that node. The second symmetry is due to the arbitrary labelling of nodes: swapping the labels of nodes i and j within a given hidden layer will leave the output of a NN unchanged.

VI. Network analysis for machine learning landscapes

The landscape expressed in terms of a connected database of local minima and transition states can be analysed in terms of

network properties.[101–103] The starting point of such a description is the observation that for a potential energy function with continuous degrees of freedom, each minimum is connected to other minima through steepest-descent paths mediated by transition states. Hence, one can construct a network[104] in a natural way, where each minimum corresponds to a node. If two minima are connected *via* a transition state then there is an edge between them. In this preliminary analysis we consider an unweighted and undirected network; this approach will be extended in the future to edge weights and directions that are relevant to kinetics. In this initial analysis we are only interested in whether or not two minima are connected, and multiple connections make no difference. The objective of working with unweighted and undirected networks is to first focus on the global structure of the landscape, providing the foundations for analysis of how emergent thermodynamic and kinetic properties are encoded in future work.

After constructing the network we can analyse properties such as average shortest path length, diameter, clustering coefficients, node degrees and their distribution. For an unweighted and undirected network, the shortest path between a pair of nodes is the path that passes through the fewest edges. The number of edges on the shortest path is then the shortest path length between the pair of nodes, and the average shortest path length is the average of the shortest path lengths over all the pairs of nodes. The diameter of a network is the path length of the longest shortest path. For the network of minima, the diameter of the network is the distance, in terms of edges, between the pair of minima that are farthest apart. The node degree is the number of directly connected neighbours.

For the non-linear regression model, we found 89 minima and 121 transition states. The resulting network consists of 89 nodes and 113 edges (Fig. 10). Although this is a small network we use it to introduce the analysis. The average node degree is 2.54 and the average shortest path length is 6.0641. The network diameter, *i.e.* the longest shortest path, is 15. Hence, on average a minimum is around 6 steps away from any other minimum, and the pair farthest apart are separated by 15 steps. Thus, a minimum found by a naive numerical minimisation procedure may be on average 6 steps, and in the worst case 15 steps, from the global minimum. Both the average shortest path and network diameter of this network are significantly larger than for a random network of an equivalent size.[105]

The network of minima for the neural network model in Section V has 142 nodes and 643 edges (Fig. 11). The nodes have on average 9.06 nearest neighbours, the average shortest path is 3.18, and the network diameter is 8. Hence, a randomly selected minimum is on an average only 3 steps away from any other minimum in terms of minimum-transition state-minimum triples. Therefore, on average, the global minimum is only 3 steps away from any other local minimum. The networks of minima defined by molecules such as Lennard-Jones clusters, Morse clusters, and the Thomson problem have also been shown to have small (of order $O(\log[\text{number of minima}])$) average shortest path lengths, meaning that any randomly picked local minimum is only a few steps away from the global minimum.[101–103,106] Moreover, these networks exhibit small-world behaviour,[107] *i.e.* the average shortest path lengths of these networks are small and similar to equivalent size random networks, whereas their clustering coefficients are significantly larger than those of equivalent size random networks. We have conjectured that the small-world properties of networks of minima are closely related to the single-funnel nature of the corresponding landscapes.[108–110] Some networks also exhibit scale-free properties,[111] where the node-degrees follow a power-law distribution. In such networks only a few nodes termed hubs have most of the connections, while most other nodes have only a small number.

The benefits of analysing network properties of machine learning landscapes may be two-fold. One can attempt to construct more efficient and tailor-made algorithms to find the global minimum of machine learning problems by exploiting these network properties, for example, by navigating the 'shortest path' to the global minimum. We also obtain a quantitative description of the distance between a typical minimum from the best fit solution.

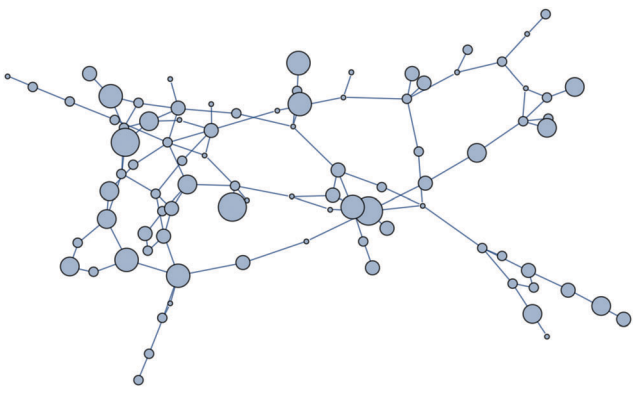


Fig. 10 Network of minima for the nonlinear regression cost function [eqn (8)]. Each node corresponds to a minimum. There is an edge between two minima if they are connected by at least one transition state. The size of the nodes is proportional to the number of neighbours.

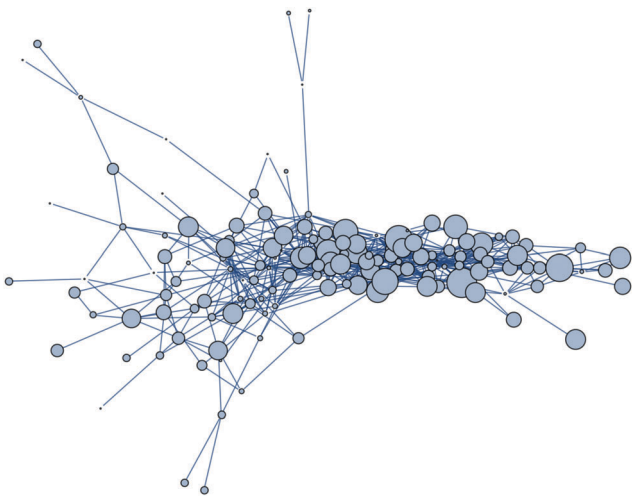


Fig. 11 Network of minima for the NN model cost function [eqn (8)] applied to digit recognition. The size of the nodes is proportional to the number of neighbours.

In the future, we plan to study further properties of the networks of minima for a variety of artificial neural networks and test the small-world and scale-free properties. We hope that these results may be useful in constructing algorithms to find the global minimum of non-convex machine learning cost functions.

VII. The *p*-spin model and machine learning

Many machine learning problems are solved *via* some kind of optimisation that makes use of gradient based methods, such as the stochastic gradient descent. Such algorithms utilise the local geometry of the landscape, and eventually iterations stop progressing when the norm of the (stochastic) gradients approach zero. This leads to the following general question: for a real valued function, what values do the critical points have? The answer certainly depends on the structure of the function we have at hand. In this section, we will examine two classes of functions: the Hamiltonian of the *p*-spin spherical glass, and the loss function that arises in the optimisation of a simple deep learning problem. In this section, the deep learning problem is the same as the one introduced in Section V with a larger hidden layer and zero regularisation. The functions have very different structures, and at first sight they do not appear to resemble one another in any meaningful way. However, we find that the two systems may exhibit similar characteristics in terms of the values of their critical points, in spite of the apparent differences.

A. Concentration of critical points of the *p*-spin model

The Hamiltonian of a mean-field spin glass is a homogeneous polynomial of a given degree, where the coefficients of the polynomial describe the nature and strength of the interaction between the spins. Since the polynomial is homogeneous, a common choice is to restrict the variables (spins) to the unit sphere. We investigate what can be said about the minimisation problem if we choose the coefficients of this polynomial at random and independently from the standard normal distribution.

First, we define the notation for the rest of the section. We consider real valued polynomials, $H(\mathbf{w})$, where $\mathbf{w}$ is the vector of variables of H; the degree of the polynomial H is p. We define the dimension of the polynomial by the length of the vector $\mathbf{w}$, so if $\mathbf{w} = (w_1,\ldots,w_N)$ then H is an N-dimensional polynomial. A degree p polynomial is homogeneous polynomial if it satisfies the following condition for any real t:

$$H(t\mathbf{w}) = H((tw_1,\ldots,tw_N)) = t^p H(\mathbf{w}) \tag{11}$$

Finally, a degree p polynomial of N variables will have N^p coefficients (some of which may be zero). The coefficients will be denoted $x_{i_1},\ldots,i_p$, where each index runs from 1 to N.

Having defined the notation, we now clarify the connection between polynomials and spin glasses. Suppose the vector, $\mathbf{w} = (w_1,\ldots,w_N)$, describes the states of N Ising spins that are +1 or −1. Then $\sum w_i^2 = N$, so that the distance to the origin is $\sqrt{N}$. The continuous analogue of this model is a hypercube embedded in a sphere of radius $\sqrt{N}$. Therefore, we can interpret the Hamiltonian of a spherical, p-body, spin glass by a homogeneous polynomial of degree p. This formulation for spin systems has been studied extensively in ref. 112–114. From here on, we will explicitly denote the dimension and the degree of the polynomial in a subscript.

The simplest case is when the degree is $p = 1$ and the polynomial (Hamiltonian) becomes

$$H_{N,1}(\mathbf{w}) = \sum_{i=1}^{N} w_i x_i, \tag{12}$$

where the spins $(w_1,\ldots,w_n) \equiv \mathbf{w} \in \mathbb{R}^N$ are constrained to the sphere *i.e.* $\sum_{i=1}^{N} w_i^2 = N$, where N is the number of spins. The coefficients $x_i \sim \mathcal{N}(0,1)$ are independent and identically distributed standard normal random variables. For $p = 1$ there exist only two stationary points, one minimum and one maximum.

When $p = 2$ the polynomial becomes

$$H_{N,2}(\mathbf{w}) = \sum_{i,j=1}^{N} w_i w_j x_{ij}. \tag{13}$$

This is a simple quadratic form with $2N$ stationary points located at the eigenvectors of the matrix $\mathbf{X}$ with elements $\mathbf{X}_{ij} \equiv x_{ij}$, with values (energies) equal to the corresponding eigenvalues.[115,116]

The picture is rather different when we look at polynomials with degree $p > 2$. When $p = 3$ the polynomial becomes

$$H_{N,3}(\mathbf{w}) = \frac{1}{N}\sum_{i,j,k=1}^{N} w_i w_j w_k x_{ijk}. \tag{14}$$

The normalisation factor $1/N$ for coupling coefficients $x_{ijk} \sim \mathcal{N}(0, 1)$, is chosen to make the extensive variables scale with N. In other words, when $\sum_{i=1}^{N} w_i^2 = N$, the variance of the Hamiltonian is proportional to N. With this convenient choice of normalization, the results of Auffinger *et al.*[112] show that $H_{N,3}(\mathbf{w})$ has exponentially many stationary points, including exponentially many local minima (see ref. 117 for a complementary numerical study). In Fig. 12 we show the distribution of minima for the normalised $H_{N,3}(\mathbf{w})$ obtained by gradient descent for various system sizes, N, and for a single realisation of the coefficients. Since the variance of the Hamiltonian scales with N, dividing by N enables us to compare energies for systems with different dimensions. The initial point is chosen uniformly at random from the sphere. The step size is constant throughout the descent, until the norm of the gradient becomes smaller than 10^{-6}. For small N the energies of the minima are broadly scattered. However as the number of spins increases, the distribution concentrates around a threshold. Further details of the calculations can be found in Sagun *et al.*[99] (and ref. 118 for a complementary study).

To obtain a more precise picture for the energy levels of critical points, we will focus on the level sets of the polynomial

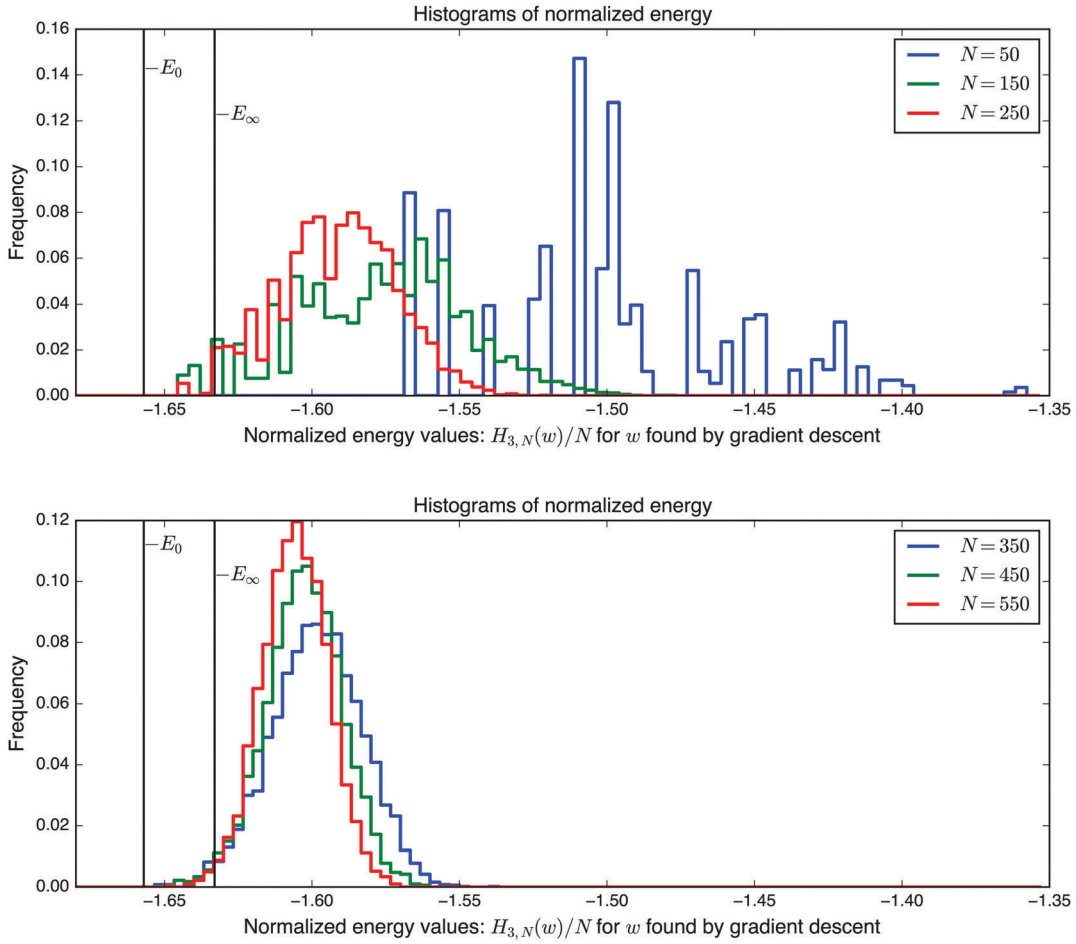


Fig. 12 Histogram for the energies of points found by gradient descent with the Hamiltonian defined in eqn (14), comparing low-dimensional and high-dimensional systems. $-E_0$ denotes the ground state, and $-E_\infty$ denotes the large N limit of the level for the bulk of the local minima.

and count the number of critical points in the interior of a given level set. Here, the polynomial is assumed to be non-degenerate (its Hessian has nonzero eigenvalues). Then, a critical point is defined by a point whose gradient is the zero vector, and a critical point of index k has exactly k negative Hessian eigenvalues. Finally, our description of the number of critical points will be in the exponential form and in the asymptotic, large N, limit.

Following Auffinger *et al.*, let $A_u = \{\mathbf{w} \in \mathbb{R}^N : H_{N,3}(\mathbf{w}) < u$ and $\sum_{i=1}^{N} w_i^2 = N\}$ be the set of points in the domain of the function whose values lie below u. Also, let $C_k(A_u)$ denote the number of stationary (critical) points with index k in the set A_u. Hence $C_k(A_u)$ counts the number of critical points with values below u. Then, the main theorem of ref. 112 produces the asymptotic expectation value $\mathbb{E}$ for the number of stationary points below energy u:

$$\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{E}(C_k(A_u)) = \Theta_k(u) \qquad (15)$$

where Θ is the complexity function, explicitly given in ref. 112. Note that the complexity function Θ is non-decreasing and it is flat above some level. This result indicates that there are no more finite index critical points at high levels, or to be more precise, it is far less probable to find them. We denote this level as $-E_\infty$. The second crucial quantity is when the complexity becomes negative. This level has the property that there are no more critical points of a specified index below it. We denote this level by $-E_k$, where k is the given index. For example, there are no more local minima below level $-E_0$, which in turn means that the ground state is bounded from below by $-E_0$. In particular, Θ approaches a constant for $u > -E_\infty = 2\sqrt{2/3} \approx -1.633$ and is bounded from below by $-E_0 \approx -1.657$. We therefore have a lower bound for the value of the ground state, and all stationary points exist in the energy band $-E_0 \le u \le -E_\infty$.

An inspection of the complexity function provides insight about the geometry at the bottom of the energy landscape (Fig. 13). The ground state is roughly at $u = -1.657$. For $u \ge -E_\infty$ we do not see any local minima, because they all have values that are within the band $-E_0 \le u \le -E_\infty$. Moreover, in the same band the stationary points of index $k > 0$ are outnumbered by local minima. In fact, this result holds hierarchically for all critical points of finite index (recall that the result is asymptotic so that by finite we mean fixing the index first and then taking the limit $N \to \infty$). If we denote the

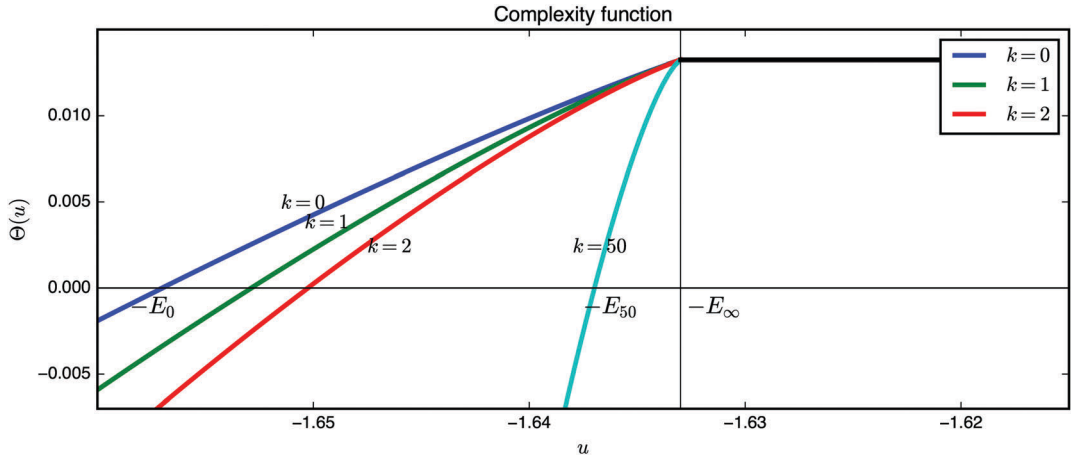


Fig. 13 Plots of the complexity function Θ_k defined in eqn (15) for local minima, and saddles of index 1, 2 and 50. In the band $(-E_0, -E_{50})$ there are only critical points with indices {1, 2,..., 49}.

x-axis intercept of the corresponding complexity function Θ_k as $-E_k$, with

$$\Theta_k(-E_k) = 0 \quad \text{for} \quad k = 1, 2, \ldots \tag{16}$$

Below the level $-E_k$ the function only has critical points of index strictly less than k. This is consistent with the 'glassiness' or 'frustration' that one would expect for such a system: a random quench is most likely to locate a minimum around the $-E_\infty$ threshold and to find a lower energy minimum numerous saddle points need to be overcome. This result suggests the following scenario for finding local minima below the threshold. First identify an initial local minimum through some minimisation algorithm. Since these points are dominant at the $-E_\infty$ threshold, probabilistically speaking, the algorithm will locate one around this value. Now we wish to jump over saddle points to reach local minima with lower energies. Since the number of saddles is much less than the number of local minima below the threshold, it may take a lot longer to find them. This feature of the landscape could make finding the global minimum a relatively difficult task. However, since basin-hopping[41–43] removes downhill barriers, this approach might still be effective, depending on the organisation of the landscape. Testing the performance of basin-hopping for such landscapes is an interesting future research direction. On the other hand, if the band $(-E_0, -E_\infty)$ is narrow, which is the case for the spherical 3-spin glass Hamiltonian described above, then it may be sufficient to locate the first local minimum and stop there, since further progress is unlikely.

This scenario holds for the p-spin Hamiltonian with any $p \geq 3$ where the threshold for the number of critical points is obtained asymptotically in the limit $N \to \infty$. To demonstrate that it holds for reasonably small N Fig. 14 shows the results for the $p = 3$ case with increasing dimensions. The concentration of local minima near the threshold increases rather quickly.

In Fig. 15 we show disconnectivity graphs for $p = 3$ spin spherical glass models with sizes $N \in [50, 100]$ and fixed coefficients $x_{ijk} \sim \mathcal{N}(0,1)$. The landscape appears to become more frustrated already for small N, and the range over which local minima are found narrows with increasing system size, in agreement with the concentration phenomenon described in ref. 112.

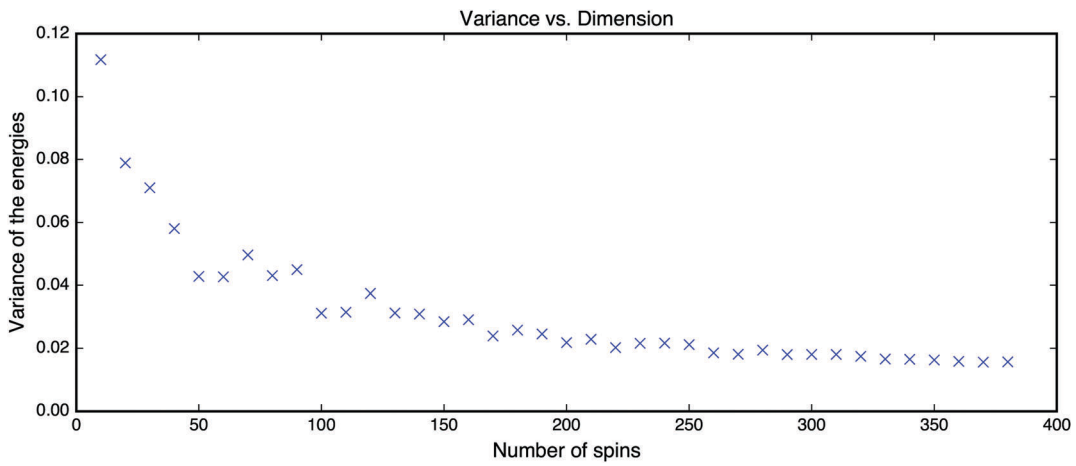


Fig. 14 Empirical variance of energies found by the gradient descent *vs.* the number of spins of the Hamiltonian defined in eqn (14).

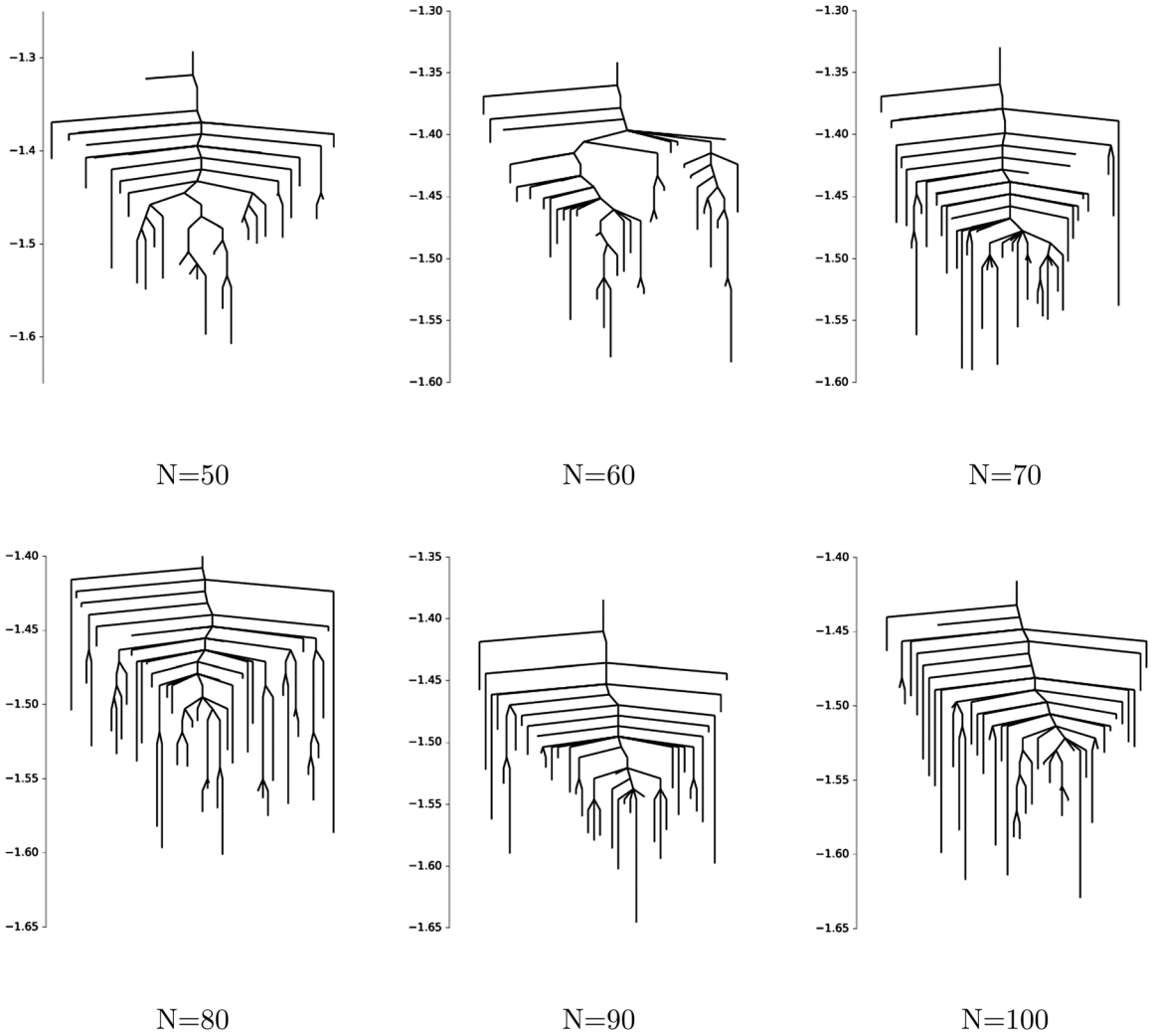


Fig. 15 Disconnectivity graphs for $p = 3$ spin spherical glass models of size $N \in [50, 100]$. Each disconnectivity graph refers to a single realisation of the coefficients $x_{ijk} \sim \mathcal{N}(0,1)$. Frustration in the landscape is already visible for small system sizes, and the minima appear to concentrate over a narrowing band of energy values for as N increases.

Interestingly, the concentration of stationary points phenomenon does not seem to be limited to the p-spin system. Related numerical results[99] show that the same effect is observed for a slightly modified version of the random polynomial, $\hat{H}_{N,3}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \mathbf{w}^{(3)}) = \sum_{i,j,k=1}^{N} w_i^{(1)} w_j^{(2)} w_k^{(3)} x_{ijk}$ defined on the three-fold product of unit spheres. We do not yet have an analogue of the above theorem for $\hat{H}$, so guidance from numerical results is particularly useful.

B. Machine learning landscapes and concentration of stationary points

The concept of complexity for a given function, as defined by the number and the nature of critical points on a given subset of the domain, gives rise to a description of the landscape as outlined above. If the energy landscape of machine learning problems is complex in this specific sense, we expect to see similar concentration phenomena in the optimisation of the corresponding loss functions. In fact, it is not straightforward to construct an analogue of the θ function as in eqn (15). However, we can empirically check whether optimisation stalls at a level above the ground state, as for the homogeneous polynomials with random coefficients described above.

Let $D := (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be n data points with input $\boldsymbol{x}_i \in \mathbb{R}^N$, and label $y \in \mathbb{R}$; and let $G(\boldsymbol{w}, \boldsymbol{x}) = \hat{y}$ describe the output that approximates the label y parametrized by $\boldsymbol{w} \in \mathbb{R}^M$. Further, let $\ell(G(\boldsymbol{w}, \boldsymbol{x}), y) = \ell(\hat{y}, y)$ be a non-negative loss function. Once we fix the dataset, we can focus on the averaged loss described by

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \ell(G(\boldsymbol{w}, \boldsymbol{x}_i), y_i) \tag{17}$$

The function in eqn (17) is non-negative, but it is not obvious where the ground state is located, and an empirical study could be inconclusive. The following procedure fixes this problem. (1) Create two identical models and split the training data in half. (2) Using the first half of the data, train the first network, thereby obtaining a point $w^\star$ with a small value of $L(w^\star)$.

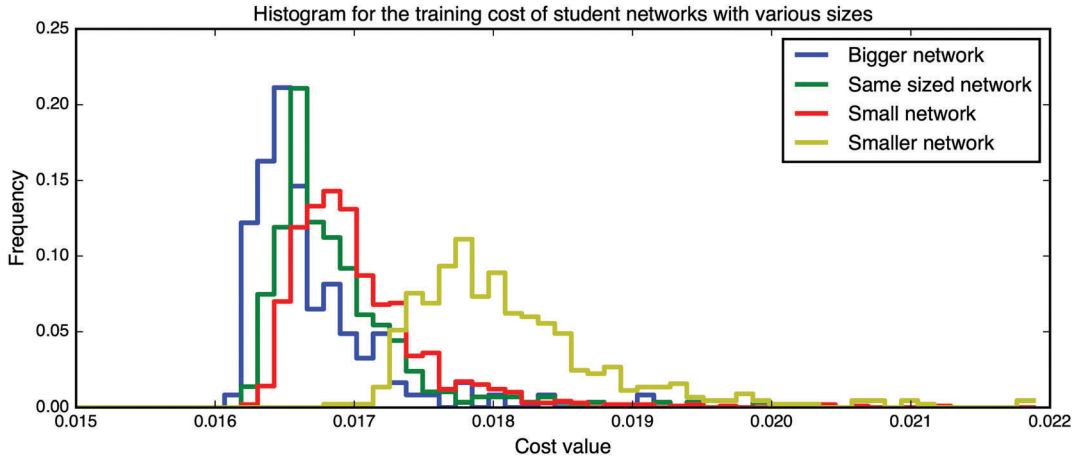


Fig. 16 Training on linear, fully-connected MNIST networks (student networks) of various sizes. The labels for the network sizes are relative to the first network that is used to create new labels. For the larger and equally sized networks the ground state is known to lie at zero, yet the training stalls at a value around 0.016.

(3) Using w^*, create new labels for the second half of the data, replacing the true labels with the output of the first model $G(\boldsymbol{w}^*)$. (4) Using these new data pairs, $(\boldsymbol{x},G(\boldsymbol{w}^*))$ train the second network. This procedure ensures that the loss function for the second network over the new dataset has configurations that have exactly value zero. Simply by finding a copy of the first network, $\boldsymbol{w}^*$, the loss for the second network will be $L(\boldsymbol{w}^*) = 0$. In fact, due to the permutation symmetry in the parameters (see Fig. 1) the loss value remains zero for all the points in the correct permutations of $\boldsymbol{w}^*$. Now the optimisation on the second loss function has a known ground state at zero, and we can check empirically whether optimisation stalls above that level (Fig. 16).[99]

We emphasise that the similarity between the p-spin Hamiltonian and the machine learning loss function lies only in the concentration phenomena, perhaps because the two functions share some underlying structure. It is likely that this observation of concentration in the two systems, spin glasses and deep learning, are due to different reasons that are peculiar to their organisation. It is also possible that such concentration behaviour is universal, in the sense that it can be observed in various non-convex and high dimensional systems for which the two systems above are just examples. However, if we wish to describe the ML landscape in terms of glassy behaviour, we might seek justification through the non-linearities in neural networks (the hidden layer in Fig. 1). In some sense, the non-linearity of a neural network is the key that introduces competing forces in which neurons can be inhibitory or exhibitory. This behaviour may introduce a similar structure to the signs of interaction coefficients in spins, introducing frustration. We note that these interpretations are rather speculative at present. Another problem that requires further research is identification of all critical points, not only the ones with low index. A more systematic way to identify the notion of complexity is through finding all of the critical points of a given function. A further challenge lies in the degeneracy of the systems in deep learning. The random polynomials that we considered have non-degenerate Hessian eigenvalue spectra for stationary points at the bottom of the landscape. However, a typical neural network has most Hessian eigenvalues near zero.[119]

Recently, a complementary study of the minima and saddle points of the p-spin model has been initiated using the numerical polynomial homotopy continuation method,[120,121] which guarantees to find all the minima, transition states and higher index saddles for moderate sized systems.[116,117] An advantage of this approach is that one can also analyse the variances of the number of minima, transition states, *etc.*, providing further insight into the landscape of the p-spin model. A study for larger values of p, analysing the interpretation in terms of a deep learning network, is in progress.[118]

VIII. Basin volume calculations: quantifying the energy landscape

The enumeration of stationary points in the energy landscape provides a direct measure of complexity. This quantity is directly related to the landscape entropy[89–92] (to be distinguished from the well entropy associated with the vibrational modes of each local minimum[122]) and is crucial for understanding the emergent dynamics and thermodynamics. In this context other important questions include determining the level of the stationary points (as we discussed in Section VII) and the volume of their basins of attraction. These volumes are of great practical interest because they provide an *a priori* measure of the probability of finding a particular minimum following energy minimisation. This probability is particularly important within the context of non-convex optimisation problems and machine learning, where an established protocol to quantify the landscape and the *a priori* outcome of learning is lacking.

The development of general methods for enumerating the number of accessible microstates in a system, and ideally their individual probabilities, is therefore of great general interest. As discussed in Section VII, for a few specific cases there exist methods – either analytical or numerical – capable of producing

exact estimates of these numbers. However, these techniques are either not sufficiently general or simply not practical for problems of interest. To date, at least two general and practical computational approaches have been developed. 'Basin-sampling'[55] employs a model anharmonic density of states for local minima organised in bins of potential energy, and has been applied to atomic clusters, including benchmark systems with double funnel landscapes that pose challenging issues for sampling. The mean basin volume (MBV) method developed by Frenkel and co-workers[57,123–125] is similar in spirit to basin-sampling, but is based on thermodynamic integration and, being completely general, requires no assumptions on the shape of the basins (although thus far all examples are limited to the enumeration of the minima). MBV has been applied in the context of soft sphere packings and has facilitated the direct computation of the entropy in two[123,124] and three[57] dimensions. Furthermore, the technique has allowed for a direct test of the Edwards conjecture in two dimensions,[126] suggesting that only at unjamming – when the system changes from a fluid to a solid, which is the density of practical significance for many granular systems – the entropy is maximal and all packings are equally probable.

Despite the high computational cost, the MBV underlying principle is straightforward. Assuming that the total configuration volume $\mathcal{V}$ of the energy landscape is known (simply $\mathcal{V} = V^N$ for an ideal gas of non-interacting atoms), if we can estimate the mean basin volume of all states, the number of minima is simply

$$\Omega = \frac{\mathcal{V}}{\langle v_{\mathrm{basin}} \rangle}, \tag{18}$$

where $\langle v_{\mathrm{basin}} \rangle$ is the unbiased average volume of the basins of attraction. We distinguish the biased from the unbiased distribution of basin volumes because, when generating the minima following minimisation from uniformly sampled points in $\mathcal{V}$, they will be sampled in proportion to the volume of the basin of attraction, and therefore the observed distribution of v_{basin} is biased. A detailed discussion of the unbiasing procedure for jammed soft-sphere packings is given in ref. 57. The Boltzmann-like entropy of the system is then simply $S_{\mathrm{B}} = \ln \Omega - \ln N!$. Similarly, from knowledge of the biased (observed) distribution of basin volumes v_i alone, one can compute the Gibbs-like (or Shannon) entropy $S_{\mathrm{G}} = -\sum_{i=1}^{\Omega} p_i \ln p_i - \ln N!$, where $p_i = v_i/\mathcal{V}$ is the relative probability for minimum i.

The computation of the basin volume is performed by thermodynamic integration. In essence, we perform a series of Markov chain Monte Carlo random walks within the basin applying different biases to the walkers and, from the distributions of displacements from a reference point (usually the minimum), compute the dimensionless free energy difference between a region of known volume and that of an unbiased walker. In other words

$$f_{\mathrm{basin}} = f_{\mathrm{ref}} + (\hat{f}_{\mathrm{basin}} - \hat{f}_{\mathrm{ref}}) \tag{19}$$

where the dimensionless free energy is $f = -\ln v$ and the hat refers to quantities estimated up to an additive constant by the free energy estimator of choice, either Frenkel–Ladd[57,127] or the multi-state Bennet acceptance ratio method (MBAR).[125,128] The high computational cost of these calculations is due to the fact that, in order to perform a random walk in the body of the basin, a full energy minimisation is required to check whether the walker has overstepped the basin boundary.

Recently the approach has been validated when the dynamics determining the basin of attraction are stochastic in nature,[129] which is precisely the situation encountered in the training by stochastic optimisation of neural networks and other non-convex machine learning problems. The extension of these techniques to machine learning is another exciting prospect, as it would provide a general protocol for quantifying the machine learning landscape and establishing, for instance, whether the different solutions to learning occur with different probabilities and, if so, what their distribution is. This characterisation of the learning problem could help to develop better models, as well as better training algorithms.

IX. Conclusions

In this Perspective we have applied theory and computational techniques from the potential energy landscapes field[2] to analyse problems in machine learning. The multiple solutions that can result from optimising fitting functions to training data define a machine learning landscape,[23] where the cost function that is minimised in training takes the place of the molecular potential energy function. This machine learning landscape can be explored and visualised using methodology transferred directly from the potential energy landscape framework. We have illustrated how this approach can be used through examples taken from recent work, including analogies with thermodynamic properties, such as the heat capacity, which reports on the structure of the equilibrium properties of the solution space as a function of a fictitious temperature parameter. The interpretation of ML landscapes in terms of analogues of molecular structure and transition rates is an intriguing target for future work.

Energy landscape methods may provide a novel way of addressing one of the most intriguing questions in the machine learning research, namely why does machine learning work so well? One way to ask this question more quantitatively is to investigate why we can usually find a good candidate for the global minimum of a machine learning cost function relatively quickly, even in the presence of so many local minima. The present results suggest that the landscape for a range of models are single-funnel-like, *i.e.* the largest basin of attraction is that of the global minimum, and the downhill barriers that separate it from local minima are relatively small. This organisation facilitates rapid relaxation to the global minimum for global optimisation techniques, such as basin-hopping. Another possible explanation is that many local minima provide fits that are competitive with the global minimum.[96,97,99] In fact, these two scenarios are compatible, so that global optimisation leads us downhill on the landscape, where we encounter local minima that provide predictions or classifications of useful accuracy.

The ambition to develop more fundamental connections between machine learning disciplines and computational chemical physics could be very productive. For example, there have recently been many physics-inspired contributions in machine learning, including thermodynamics-based models for rational decision-making,[130] generative models from non-equilibrium simulations.[131] The hope is that such connections can provide better intuition about the machine learning problems in question, and perhaps also the underlying physical theories used to understand them.

Acknowledgements

It is a pleasure to acknowledge discussions with Prof. Daan Frenkel, Dr Victor Ruehle, Dr Peter Wirnsberger, Prof. Gérard Ben Arous, and Prof. Yann Lecun. This research was funded by EPSRC grant EP/I001352/1, the Gates Cambridge Trust, and the ERC. DM was in the Department of Applied and Computational Mathematics and Statistics when this work was performed, and his current affiliation is Department of Systems, United Technologies Research Center, East Hartford, CT, USA.

References

1 S. T. Chill, J. Stevenson, V. Ruehle, C. Shang, P. Xiao, J. D. Farrell, D. J. Wales and G. Henkelman, *J. Chem. Theory Comput.*, 2014, **10**, 5476.
2 D. J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, 2003.
3 D. J. Wales, *Curr. Opin. Struct. Biol.*, 2010, **20**, 3.
4 D. J. Wales, M. A. Miller and T. R. Walsh, *Nature*, 1998, **394**, 758.
5 D. J. Wales, *Philos. Trans. R. Soc., A*, 2005, **363**, 357.
6 V. K. de Souza and D. J. Wales, *J. Chem. Phys.*, 2008, **129**, 164507.
7 J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Proteins*, 1995, **21**, 167.
8 J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes, *Annu. Rev. Phys. Chem.*, 1997, **48**, 545.
9 D. Chakrabarti and D. J. Wales, *Soft Matter*, 2011, **7**, 2325.
10 Y. Chebaro, A. J. Ballard, D. Chakraborty and D. J. Wales, *Sci. Rep.*, 2015, **5**, 10386.
11 P. G. Mezey, *Potential Energy Hypersurfaces*, Elsevier, Amsterdam, 1987.
12 F. Noé and S. Fischer, *Curr. Opin. Struct. Biol.*, 2008, **18**, 154.
13 D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique and F. Fernando, *PLoS Comput. Biol.*, 2009, **5**, e1000415.
14 D. J. Wales, *Mol. Phys.*, 2002, **100**, 3285.
15 C. Dellago, P. G. Bolhuis and D. Chandler, *J. Chem. Phys.*, 1998, **108**, 9236.
16 D. Passerone and M. Parrinello, *Phys. Rev. Lett.*, 2001, **87**, 108302.
17 W. E, W. Ren and E. Vanden-Eijnden, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2002, **66**, 052301.
18 O. M. Becker and M. Karplus, *J. Chem. Phys.*, 1997, **106**, 1495.
19 J. P. K. Doye, M. A. Miller and D. J. Wales, *J. Chem. Phys.*, 1999, **110**, 6896.
20 J. N. Murrell and K. J. Laidler, *Trans. Faraday Soc.*, 1968, **64**, 371.
21 R. Collobert, F. Sinz, J. Weston and L. Bottou, *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, ACM, New York, NY, USA, 2006, pp. 201–208.
22 M. Pavlovskaia, K. Tu and S.-C. Zhu, Mapping the Energy Landscape of Non-convex Optimization Problems, in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, ed. X.-C. Tai, E. Bae, T. F. Chan and M. Lysaker, 10th International Conference, EMMCVPR 2015, Springer International Publishing, Hong Kong, China, 2015, pp. 421–435.
23 A. J. Ballard, J. D. Stevenson, R. Das and D. J. Wales, *J. Chem. Phys.*, 2016, **144**, 124119.
24 R. Das and D. J. Wales, *Phys. Rev. E*, 2016, **93**, 063310.
25 S. Manzhos and T. Carrington, *J. Chem. Phys.*, 2006, **125**, 084109.
26 S. Houlding, S. Y. Liem and P. L. A. Popelier, *Int. J. Quantum Chem.*, 2007, **107**, 2817.
27 J. Behler, R. Martoňák, D. Donadio and M. Parrinello, *Phys. Rev. Lett.*, 2008, **100**, 185501.
28 C. M. Handley, G. I. Hawe, D. B. Kell and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2009, **11**, 6365.
29 J. Behler, S. Lorenz and K. Reuter, *J. Chem. Phys.*, 2007, **127**, 014705.
30 J. Li and H. Guo, *J. Chem. Phys.*, 2015, **143**, 214304.
31 K. Shao, J. Chen, Z. Zhao and D. H. Zhang, *J. Chem. Phys.*, 2016, **145**, 071101.
32 C. M. Handley and P. L. A. Popelier, *J. Phys. Chem. A*, 2010, **114**, 3371.
33 T. B. Blank, S. D. Brown, A. W. Calhoun and D. J. Doren, *J. Chem. Phys.*, 1995, **103**, 4129.
34 D. F. R. Brown, M. N. Gibbs and D. C. Clary, *J. Chem. Phys.*, 1996, **105**, 7597.
35 H. Gassner, M. Probst, A. Lauenstein and K. Hermansson, *J. Phys. Chem. A*, 1998, **102**, 4596.
36 P. J. Ballester and J. B. O. Mitchell, *Bioinformatics*, 2010, **26**, 1169.
37 A. W. Long and A. L. Ferguson, *J. Phys. Chem. B*, 2014, **118**, 4228.
38 B. A. Lindquist, R. B. Jadrich and T. M. Truskett, *J. Chem. Phys.*, 2016, **145**, 111101.
39 Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K. R. Muller and G. Henkelman, *J. Chem. Phys.*, 2012, **136**, 174101.
40 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273.
41 Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 6611.
42 Z. Li and H. A. Scheraga, *J. Mol. Struct.*, 1988, **179**, 333.
43 D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111.
44 J. Nocedal, *Math. Comput.*, 1980, **35**, 773.
45 S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, 2004, **120**, 2082.

46 S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, 2004, **121**, 6689.
47 G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901.
48 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978.
49 L. J. Munro and D. J. Wales, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 3969.
50 Y. Zeng, P. Xiao and G. Henkelman, *J. Chem. Phys.*, 2014, **140**, 044115.
51 Pele: Python energy landscape explorer, https://github.com/pele-python/pele.
52 D. J. Wales, *GMIN: a program for basin-hopping global optimisation, basin-sampling, and parallel tempering*, http://www-wales.ch.cam.ac.uk/software.html.
53 D. J. Wales, *OPTIM: a program for geometry optimisation and pathway calculations*, http://www-wales.ch.cam.ac.uk/software.html.
54 D. J. Wales, *PATHSAMPLE: a program for generating connected stationary point databases and extracting global kinetics*, http://www-wales.ch.cam.ac.uk/software.html.
55 D. J. Wales, *Chem. Phys. Lett.*, 2013, **584**, 1.
56 F. Noé, D. Krachtus, J. C. Smith and S. Fischer, *J. Chem. Theory Comput.*, 2006, **2**, 840.
57 S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales and D. Frenkel, *Phys. Rev. E*, 2016, **93**, 012906.
58 K. Swersky, J. Snoek and R. P. Adams, 2014, arXiv:1406.3896 [stat.ML].
59 D. J. Wales, *J. Chem. Soc., Faraday Trans.*, 1992, **88**, 653.
60 D. J. Wales, *J. Chem. Soc., Faraday Trans.*, 1993, **89**, 1305.
61 D. Asenjo, J. D. Stevenson, D. J. Wales and D. Frenkel, *J. Phys. Chem. B*, 2013, **117**, 12717.
62 J. E. Jones and A. E. Ingham, *Proc. R. Soc. A*, 1925, **107**, 636.
63 P. M. Axilrod and E. Teller, *J. Chem. Phys.*, 1943, **11**, 299.
64 C. G. Broyden, *J. Inst. Math. Its Appl.*, 1970, **6**, 76.
65 R. Fletcher, *Comput. J.*, 1970, **13**, 317.
66 D. Goldfarb, *Math. Comput.*, 1970, **24**, 23.
67 D. F. Shanno, *Math. Comput.*, 1970, **24**, 647.
68 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
69 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2009.
70 M. Page and J. W. McIver, *J. Chem. Phys.*, 1988, **88**, 922.
71 J. P. K. Doye, *Phys. Rev. E*, 2000, **62**, 8753.
72 F. H. Stillinger and T. A. Weber, *J. Stat. Phys.*, 1988, **52**, 1429.
73 F. H. Stillinger and T. A. Weber, *Science*, 1984, **225**, 983.
74 D. J. Wales and J. P. K. Doye, *J. Chem. Phys.*, 2003, **119**, 12409.
75 D. J. Wales and J. P. K. Doye, *J. Chem. Phys.*, 1995, **103**, 3061.
76 D. J. Wales, *Mol. Phys.*, 1993, **78**, 151.
77 F. H. Stillinger, *Science*, 1995, **267**, 1935.
78 B. Strodel and D. J. Wales, *Chem. Phys. Lett.*, 2008, **466**, 105.
79 V. A. Sharapov, D. Meluzzi and V. A. Mandelshtam, *Phys. Rev. Lett.*, 2007, **98**, 105701.
80 M. T. Oakley, R. L. Johnston and D. J. Wales, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3965.
81 J. P. Neirotti, F. Calvo, D. L. Freeman and J. D. Doll, *J. Chem. Phys.*, 2000, **112**, 10340.
82 F. Calvo, J. P. Neirotti, D. L. Freeman and J. D. Doll, *J. Chem. Phys.*, 2000, **112**, 10350.
83 V. A. Mandelshtam, P. A. Frantsuzov and F. Calvo, *J. Phys. Chem. A*, 2006, **110**, 5326.
84 V. A. Sharapov and V. A. Mandelshtam, *J. Phys. Chem. A*, 2007, **111**, 10284.
85 F. Calvo, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2010, **82**, 046703.
86 R. M. Sehgal, D. Maroudas and D. M. Ford, *J. Chem. Phys.*, 2014, **140**, 104312.
87 D. J. Wales, *Mol. Phys.*, 2004, **102**, 891.
88 M. Picciani, M. Athenes, J. Kurchan and J. Tailleur, *J. Chem. Phys.*, 2011, **135**, 034108.
89 F. Sciortino, W. Kob and P. Tartaglia, *J. Phys.: Condens. Matter*, 2000, **12**, 6525.
90 T. V. Bogdan, D. J. Wales and F. Calvo, *J. Chem. Phys.*, 2006, **124**, 044102.
91 G. Meng, N. Arkus, M. P. Brenner and V. N. Manoharan, *Science*, 2010, **327**, 560.
92 D. J. Wales, *ChemPhysChem*, 2010, **11**, 2491.
93 R. Das and D. J. Wales, *Chem. Phys. Lett.*, 2017, **667**, 158.
94 Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, *Proc. IEEE*, 1998, **86**, 2278.
95 See the MNIST database: http://yann.lecun.com/exdb/mnist.
96 Y. Dauphin, R. Pascanu, Ç. Gülçehre, K. Cho, S. Ganguli and Y. Bengio, CoRR abs/1406.2572, 2014.
97 A. J. Bray and D. S. Dean, *Phys. Rev. Lett.*, 2007, **98**, 150201.
98 A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous and Y. LeCun, CoRR abs/1412.0233, 2014.
99 L. Sagun, V. U. Güney, G. Ben Arous and Y. LeCun, ICLR2015 Workshop Contribution, 2014, arXiv:1412.6615.
100 The misclassification distance can also be viewed as the Hamming distance between misclassification vectors of the two minima in question.
101 J. P. K. Doye, *Phys. Rev. Lett.*, 2002, **88**, 238701.
102 J. P. K. Doye and C. P. Massen, *J. Chem. Phys.*, 2005, **122**, 084105.
103 D. Mehta, J. Chen, D. Z. Chen, H. Kusumaatmaja and D. J. Wales, *Phys. Rev. Lett.*, 2016, **117**, 028301.
104 M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, 2010.
105 S. H. Strogatz, *Nature*, 2001, **410**, 268.
106 J. W. R. Morgan, D. Mehta and D. J. Wales, to appear.
107 D. J. Watts and S. H. Strogatz, *Nature*, 1998, **393**, 440.
108 D. J. Wales, J. P. K. Doye, M. A. Miller, P. N. Mortenson and T. R. Walsh, *Adv. Chem. Phys.*, 2000, **115**, 1.
109 J. P. K. Doye, *Phys. Rev. Lett.*, 2002, **88**, 238701.
110 J. M. Carr and D. J. Wales, *J. Phys. Chem. B*, 2008, **112**, 8760.
111 A.-L. Barabási and R. Albert, *Science*, 1999, **286**, 509.
112 A. Auffinger, G. Ben Arous and J. Černý, *Commun. Pure Appl. Math.*, 2013, **66**, 165.
113 A. Auffinger and G. B. Arous, *et al.*, *Ann. Probab.*, 2013, **41**, 4214.

114 A. Auffinger and W.-K. Chen, 2017, arXiv:1702.08906, arXiv preprint.
115 Y. V. Fyodorov and P. Le Doussal, *J. Stat. Phys.*, 2014, **154**, 466.
116 D. Mehta, J. D. Hauenstein, M. Niemerg, N. J. Simm and D. A. Stariolo, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2015, **91**, 022133.
117 D. Mehta, D. A. Stariolo and M. Kastner, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2013, **87**, 052143.
118 D. Mehta, H. Sidky, Y. Dauphin and J. W. Whitmer, to appear.
119 L. Sagun, L. Bottou and Y. LeCun, 2016, arXiv:1611.07476, arXiv preprint.
120 A. J. Sommese and C. W. Wampler, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific, 2005, vol. 99.
121 D. Mehta, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2011, **84**, 025702.
122 L. Berthier and D. Coslovich, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 11668.
123 N. Xu, D. Frenkel and A. J. Liu, *Phys. Rev. Lett.*, 2011, **106**, 245502.
124 D. Asenjo, F. Paillusson and D. Frenkel, *Phys. Rev. Lett.*, 2014, **112**, 098002.
125 S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales and D. Frenkel, *Phys. Rev. E*, 2016, **94**, 031301.
126 S. Martiniani, K. J. Schrenk, K. Ramola, B. Chakraborty and D. Frenkel, 2016, arXiv:1610.06328, arXiv preprint.
127 D. Frenkel and A. J. C. Ladd, *J. Chem. Phys.*, 1984, **81**, 3188.
128 M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.
129 D. Frenkel, K. J. Schrenk and S. Martiniani, 2016, arXiv: 1612.06131, arXiv preprint.
130 P. A. Ortega and D. A. Braun, *Proc. R. Soc. London, Ser. A*, 2013, **469**, DOI: 10.1098/rspa.2012.0683.
131 J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, CoRR abs/1503.03585, 2015.