Energy Landscapes for a Machine Learning Application to Series Data

Andrew J. Ballard, Jacob D. Stevenson, Ritankar Das, and David J. Wales[a)]
University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Methods developed to explore and characterise potential energy landscapes are applied to the corresponding landscapes obtained from optimisation of a cost function in machine learning. We consider neural network predictions for the outcome of local geometry optimisation in a triatomic cluster, where four distinct local minima exist. The accuracy of the predictions is compared for fits using data from single and multiple points in the series of atomic configurations resulting from local geometry optimisation, and for alternative neural networks. The machine learning solution landscapes are visualised using disconnectivity graphs, and signatures in the effective heat capacity are analysed in terms of distributions of local minima and their properties.

[a)]dw34@cam.ac.uk

I. INTRODUCTION

In this contribution we apply methodology developed for the exploration of potential energy landscapes in molecular science to landscapes defined by an application of machine learning (ML). In this ML framework a cost function is defined by optimisation of the parameters in a fitting procedure, training on selected data. For non-convex formulations the training optimisation will generally have multiple solutions,[1–4] which are analogous to the different local minima or isomers of a molecular system. We have recently found that this analogy can be exploited in an initial study of two distinct ML formulations, namely non-linear regression, and neural network classification.[5] In the present work, we consider neural network fits for series data produced for geometry optimisation of a triatomic cluster. By combining information from different points in the optimisation series into composite data items we introduce correlation into the training data, and examine how this affects the quality of the fits and the predictions for data outside the fit, where the outcomes are also known. Details of the series data used in these computational experiments are given in §II.

Neural networks were investigated in this initial work mainly because of their familiarity in the chemical physics community. Alternative machine learning approaches will be considered in the future, and compared with the present results.

The main objective of the present study is to illustrate the opportunities that exploitation of the general energy landscapes machinery can provide in the context of ML. There is a wealth of experience available from research on potential energy landscapes in chemical physics, especially for atomistic systems,[6–8] which can now be brought to bear on the solution landscapes obtained from ML applications. For example, the emergence of characteristic properties from underlying features of the landscape has revealed common features for systems that reliably self-assemble, or alternatively, exhibit glassy phenomenology.[9] Separate features in the heat capacity as a function of temperature, and distinct relaxation time scales are associated with multifunnel landscapes,[10,11] which have served as benchmarks for global optimisation,[12] enhanced thermodynamic sampling,[13–20] and rare event dynamics.[21–23] In this initial exploration of ML solution landscapes we have immediately identified analogous features in the thermodynamic properties, which are defined from densities of states in the configuration space described by the fitting parameters (§VI). The fundamental connection between ML applications and previous work on potential energy landscapes follows simply by considering the fitting parameters as coordinates that define

a configuration space, and the value of the fitting function that is optimised as an effective energy. Hence we may refer to the objective function value as the 'energy' in the following discussions.

II. SERIES DATA FROM MOLECULAR GEOMETRY OPTIMISATION

The series data used in these investigations were obtained by running energy minimisations from random starting configurations of a triatomic cluster, and saving information from each optimisation step. Here we must carefully distinguish between stationary points for the molecule, and stationary points for the machine learning objective function. We will therefore refer to minima and transition states of the atomic cluster in terms of the *molecular energy landscape*, and for stationary points corresponding to alternative link weights for the neural network we will refer to the *ML landscape*.

For the model interatomic potential there are four distinct molecular local minima, and hence four possible outcomes. The four solutions have different interatomic distances, and these distances converge from the initial random values as each geometry optimisation proceeds. Predicting which molecular minimum (outcome) will be reached from the distances close to convergence should be simple, but predictions are expected to become more difficult if we only consider information from the start of the optimisation. Investigating how much series information is actually needed to predict the outcomes reliably is actually of practical interest, since it would provide criteria for terminating each optimisation and potentially saving significant computer time. This kind of application has previously been discussed by Swersky, Snoek and Adams.[24]

The model employed here was used in earlier studies to visualise the basins of attraction on the molecular energy landscape for different stationary points and comparison of optimisation methods.[25,26] This scheme was recently revisited[27] to compare the performance of a new algorithm[28] with the customised L-BFGS routine implemented in our GMIN[29] and OPTIM[30] programs. L-BFGS corresponds to a limited memory version of the quasi-Newton Broyden,[31] Fletcher,[32] Goldfarb,[33] Shanno,[34] BFGS procedure. All the local minimisations in the present work used the customised L-BFGS algorithm, which proved to be the fastest method.[27] A fixed maximum step size of 0.1 units was employed throughout, which resulted in series of between 11 and 91 steps for a convergence condition of 10^{-6} on the root mean square gradient in reduced units, as defined below.

The interatomic potential was constructed from the pairwise Lennard-Jones form[35] plus a three-

body Axilrod–Teller term:[36]

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

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 parameter Z governs the relative importance of the three-body term, and we used $Z = 2$ throughout. This choice produces a landscape with three distinct linear local minima, and one other minimum for an equilateral triangle geometry. The potential energies are -2.219 and -2.185, respectively, and there are interparticle distances of 1.10876 and 2.21752 in the linear minima, and 1.16875 for the triangle. Here we have adopted a reduced unit system with $\epsilon = \sigma = 1$.

The three linear molecular local minima correspond to structures with each of the three atoms in the central position. These are permutational isomers, and are readily distinguished by the interparticle distances. In contrast, the equilateral triangle minimum with D_{3h} symmetry has only one distinct permutational isomer. The series data collected during local minimisation included all the interatomic distances, the instantaneous potential energy, and the root mean square gradient. Since the result of each minimisation is known at the end, we can also calculate the minimised distance of each instantaneous geometry from the four possible outcomes. However, in this initial survey we have simply used the interparticle distances as input data.

A database of 10,000 local minimisation sequences for the molecular potential energy landscape was collected using random initial distributions of the three atoms in a cube of side length $2\sqrt{3}$ reduced units. Here we simply want to choose an intial volume that is small enough to prevent dissociated states from appearing. The results reported in the following sections used data from 500 of these series for training and 500 for testing; employing larger portions of the database does not lead to any new phenomenology, but increases the run times. The local minimisations involve negligable computational effort, and 10,000 optimisation series were created to provide enough data to ensure that our general conclusions do not depend on the size of the database. The training and testing sets were taken from the first and second halves of the database, respectively, and different sizes were also considered (results omitted for brevity).

The ML landscapes that we characterise correspond to predictions of the four possible outcomes (the four isomers) given data from one or more configurations in each local minimisation sequence. To train the neural networks on single data points we used the three distances r_{12}, r_{13} and r_{23} as inputs associated with the appropriate outcome. We considered single data points taken

from different positions in the series of configurations produced by local minimisation. To provide additional information for fitting we employed the three distances at two or three consecutive configurations from the optimisation series. Of course, it is also possible to choose configurations at different spacings, starting from different points in each series. We simply used the first two and the first three points in this initial study, for specificity.

III. NEURAL NETWORK PREDICTIONS

The ML landscapes we consider here correspond to an artificial neural network (NN), with three, six or nine input nodes corresponding to one, two, or three sets of interparticle distances from each of the local geometry optimisation series of configurations. A single hidden layer with between three and six nodes was employed, together with an output layer of $N_{\rm out} = 4$ nodes (or classes), for the four distinguishable local minima. The outputs were normalised softmax probabilities obtained from the neural network's final layer output $\mathbf{y} = \{y_0, y_1, y_2, y_3\}$ as

$$p_c(\mathbf{X}) = e^{y_c} / \sum_{a=0}^{3} e^{y_a}. \tag{2}$$

For training we considered $N_{\rm data}$ data points, $\mathbf{Z} = \{\mathbf{z}^1, \ldots, \mathbf{z}^{N_{\rm data}}\}$, each with dimension $N_{\rm in}$, e.g. $\mathbf{z}^\alpha = \{z_1^\alpha, \ldots, z_{N_{\rm in}}^\alpha\}$. ML stationary points and landscapes were calculated for a given set of training data points by optimising an objective function formulated as a sum of the cross-entropy defined from the softmax probabilities, and an L2 regularisation term (to prevent overfitting):

$$E(\mathbf{X}; \mathbf{Z}) = -\sum_{d=1}^{N_{\rm data}} \ln p_{c(d)}(\mathbf{X}) + \lambda \mathbf{X}^2, \tag{3}$$

where $c(d)$ is the class label for data point d specified by the training set, and λ is a constant (fixed at 0.0001). The regularization is performed on all parameter degrees of freedom except for the bias nodes of the neural net. When we evaluate $E(\mathbf{X}; \mathbf{Z})$ for data points outside the training set, with $\mathbf{X}$ fixed, the class labels $c(d)$ correspond to the outcomes for the new data, which are also known. Stationary points of the objective function are defined by $N_{\rm NN}$ neural network parameters, which can be collected in a $N_{\rm NN}$ dimensional vector $\mathbf{X}$ for each minimum and transition state.

For correct predictions the probabilities should be unity, so we could write errors as

$$\epsilon^\alpha = 1 - \exp\left(\lambda \mathbf{X}^2 - E(\mathbf{X}; \mathbf{z}^\alpha)\right), \tag{4}$$

where $\mathbf{z}^\alpha$ is a data point containing precise values for the outcome configuration α, with $\alpha = 0$ the equilateral triangle and $\alpha = 1,\ 2,\ 3$ data points for the three linear molecular minima.

IV. EXPLORING THE ML SOLUTIONS LANDSCAPE

The methods employed to survey the ML landscape have been extensively developed and discussed in the context of atomistic systems, including atomic and molecular clusters, biomolecules, and condensed matter. In particular, we characterise local minima and the pathways that connect them via transition states on the ML landscape. These minima and transition states are stationary points of the cost function that is optimised in fitting the neural network parameters to training data, where the gradient vanishes. Here we employ the geometrical definition of a transition state, as a stationary point with a single negative Hessian (second derivative matrix) eigenvalue. Local minima have no negative Hessian eigenvalues. The pathways defined by each transition state were calculated by L-BFGS minimisation following small displacements parallel and antiparallel to the Hessian eigenvector corresponding to the unique negative eigenvalue. The same procedure is used for both the molecular and ML landscapes.

To locate the global minimum on the ML landscape we employed basin-hopping global optimisation.[37–39] Transition state candidates were obtained using the doubly-nudged[40] elastic band[41] method, with accurate refinement based on hybrid eigenvector-following.[42] The resulting ML stationary point databases and connectivity information constitute a kinetic transition network.[7,43,44] Various techniques for growing and refining such networks have been developed in the discrete path sampling framework;[21,22] the present work employed implementations within the open source python-based PELE code.[45]

Useful insight can often be obtained from visualising a transition network using disconnectivity graphs.[10,46] Although quantitative analysis requires additional information, such as densities of states to compute thermodynamic properties and unimolecular rate constants, the overall appearance of the graph may be sufficient to deduce various emergent properties. For example, a single funnel appearance means that that the system belongs to a universality class corresponding to efficient relaxation to the global minimum. In contrast, multiple funnel landscapes, with competing low-lying minima separated by high barriers, can produce distinct relaxation times and features in the heat capacity. In the limit of an exponentially large number of alternative (disordered) solutions glassy phenomenology is expected. Higher order structure defined by cage-breaking rearrangements associated with diffusion has recently been identified for certain structural glass-formers, and was also characterised in the corresponding disconnectivity graphs.[9]

V. RESULTS

Table I summarises the ML stationary point databases calculated in the present work. Neural networks with three, four, five and six nodes in the hidden layer were compared for three different data formats, using the interatomic distances from one, two, and three consecutive points in the configurations obtained from local geometry optimisation on the molecular energy landscape. The first, second, and third points were used, and these training and testing formats are referred to as '1', '$1+2$', and '$1+2+3$'. For comparison, we also compared results obtained for single data points with three hidden nodes at the 10th step, the 5th from last step, and the final step of each series of configurations.

The performance of the minima in the ML training solutions landscape was analysed by evaluating the objective function for a test set of 500 geometry optimisation sequences from outside the training set. The misclassification probabilities were also calculated, revealing that the fraction f of misclassified test set images is generally very low, even for data format '1', where only the initial interatomic distances are used. The misclassification fraction is zero for the test set in each case for both the training set global minimum, and the minimum that gives the lowest residual for the test set. The percentage of minima with non-zero misclassification fractions and the maximum and minimum fractions are summarised in Table III.

We therefore continued the analysis, as in previous work,[5] using the misclassification distance from the global minimum, defined as the fraction of test set images that are misclassified by one minimum but not both. The disconnectivity graphs in Figures 1 and 2 are coloured according to this metric. These figures reveal that the distances are generally small for low-lying solutions, rising to values of between around 0.2 and 0.4 for minima with distinctly higher residual objective functions. The performance of the low-lying ML minima is generally not very different from the global minimum.

The flexibility of models involving more hidden nodes is reflected in the lower values achieved for the fits to training data summarised in Table II. Lower residual errors are also achieved in the single data points corresponding to the 10th, fifth from last ('-5') and final configurations. The results for the latter two data sets are virtually identical, indicating that the geometry optimisations are essentially converged. Hence predicting the outcome becomes straightforward, and this limit is reflected in the simpler appearance of the heat capacity curves, described below. In contrast, the results obtained for configurations at step ten do not correspond to this limit, but give lower

objective functions than for datasets '1', '$1+2$' and '$1+2+3$' when three and four hidden nodes are used to fit these three data formats. However, lower minima are obtained when five and six hidden nodes are used in the latter fits. These lower values are obtained at the expense of much more complicated ML landscapes. In particular, the number of stationary points on the ML landscape increases rapidly with the number of neural network parameters (Table I). Similar growth in complexity with the number of particles is well known in atomistic systems.[47,48] However, the disconnectivity graphs in Figure 1 suggest that most of the solutions still perform similarly to the global minimum in terms of predictive properties.

VI. THERMODYNAMIC PROPERTIES OF THE MACHINE LEARNING LANDSCAPE

All the machinery that has been developed to characterise observable properties of atomistic systems in terms of the underlying potential energy landscape can be carried over to the machine learning environment. In this section we explore properties corresponding to the equilibrium thermodynamics, using the superposition approach,[6,16,47,49–51] where the global canonical partition function, $Z(T)$, is written as a sum over contributions from local minima, $Z_\alpha(T)$. This formulation is exact, but is usually applied using a harmonic approximation for convenience, where the $Z_\alpha(T)$ can be written analytically in terms of eigenvalues of the Hessian (second derivative) matrix:

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

where κ is the number of non-zero eigenvalues, $\mu_{\alpha i}$, for the Hessian matrix of minimum α, $\beta = 1/k_B T$, k_B is the Boltzmann constant, and E_α is the corresponding potential energy. Smaller Hessian eigenvalues, associated with lower curvatures and hence wider basins of attraction in configuration space, correspond to higher configurational entropy. The superposition approach is ergodic by construction, and can also incorporate quantum effects[52] and anharmonicity.[19,49,51–55] Zero Hessian eigenvalues are associated with continuous symmetries of the potential. For the neural networks considered here there is a single zero eigenvalue. Hence the value of κ is one less than the number of neural network parameters, which are given in Table I for all the systems considered in the present work.

The Hessian eigenvalues that appear in equation (5) correspond to the principal curvatures at each of the ML local minima, which define the local density of states and configuration volume for

each of these stationary points. In atomistic systems these properties are related to the vibrational entropy. For the machine learning landscapes considered here this entropic contribution will affect the relative probabilities of locating a particular minimum during the fitting procedure.

Features in the heat capacity have proved to be particularly insightful in previous work. In particular, competition between two alternative relatively low energy morphologies can produce a low temperature peak below the melting transition, which corresponds to a solid-solid transition.[10,11] This structure presents a challenge for global optimisation,[12] global thermodynamic sampling in the presence of broken ergodicity,[13–20] and rare event dynamics.[21–23] The position and magnitude of the corresponding heat capacity peak are governed by the balance between potential energy and entropy in the competing structures, analogous to a first order phase transition. The magnitude of the heat capacity peak increases with the energy and entropy differences of the two phase-like forms.[56]

Remarkably, the heat capacity calculated for these ML landscapes exhibits multiple peaks in most cases. For the simplest example, using just the interparticle distances at the starting configuration and three hidden nodes, there are three peaks (Figure 3). To understand how these peaks emerge from the underlying landscape we have calculated C_V using partial sums in the partition function, including all the minima in order of energy up to a given threshold. The lowest peak at $k_BT \approx 0.2$ is quantitatively reproduced by the two lowest minima. Hence these two solutions are sufficiently different in terms of their entropic characteristics (the associated volumes of configuration space) to produce a small but distinct effective 'latent heat'.[56] The next peak, around $k_BT \approx 9$, requires the lowest 124 minima to appear. There is a large step in function value (of order 10) between minima 124 and 125, so we might interpret this peak in terms of a 'melting' transition between low-lying minima and the more numerous solutions with higher energy and higher entropy up to number 124. The largest peak at $k_BT \approx 20$ is quantitatively reproduced when minimum 153 is included. This particular solution has one unusually small Hessian eigenvalue, and the associated configurational entropy is very large.

In fact, all the heat capacity curves for the different data formats and NN architectures considered in the present work exhibit multiple peaks, aside from the results for the single configuration data formats corresponding to the fifth from last and final points (Figure 4). Here the configurations that constitute the training data are practically converged to the final minima, and predicting the result is straightforward. In practical applications we are more likely to be interested in difficult cases, and the present results suggest that hierarchical structure with groups of solutions

associated with similar energies and individual entropic characteristics are likely to appear. In particular, most panels in Figure 4 reveal a narrow low-temperature peak. This feature probably involves a transition in maximum equilibrium probability from the global minimum to the lowest excited state or states, as for the example considered in detail above.

The height of the largest peak increases systematically as more inputs are included in data formats '1', '$1 + 2$' and '$1 + 2 + 3$'. The temperature range for transitions generally decreases with the number of hidden nodes, and the position of the largest peak moves to lower temperature. This shift is probably a consequence of the larger number of relatively low-lying solutions, which can be seen in the disconnectivity graphs for the datasets '1' based on the initial configurations. If the peak results from a pseudo-first order transition then increasing the entropy difference and decreasing the energy difference between the low energy and high entropy 'phases' will lower the transition temperature.

Although multifunnel potential energy landscapes have been extensively studied for atomistic systems, the analogues that we have characterised here for ML solution landscapes clearly warrant detailed investigation in future work. In particular, the appearance of features in statistical properties such as the heat capacity are determined by the structure of the underlying landscape, and some clearer general principles may emerge. The resulting insight might be helpful in designing and understanding ML frameworks to produce more efficient representations. For example, ensembles of decision trees are exploited in ensemble learning methods such as random forests.[57,58] Here each tree corresponds to a set of parameters for a single model. To obtain better predictive tools it might be helpful to diversify the ensemble of solutions.[59] Since features in the ML heat capacity analogue correspond to equilibria between qualitatively different ML minima, a good strategy might involve combining solutions from each phase-like form. We plan to investigate such possibilities in future work.

VII. CONCLUSIONS

In this contribution we have applied techniques from the potential energy landscapes approach[6] to the landscape defined by a machine learning procedure based on neural networks. This work is a first step in part of a more ambitious plan to develop more fundamental connections between energy landscape theory as developed for chemical physics applications and machine learning. There have recently been many physics-inspired contributions in machine learning, includ-

ing: thermodynamics-based models for rational decision-making;[60] generative models from non-equilibrium simulations,[61] and interpretation of neural network landscapes as spin glasses.[3] Our hope is that such connections can provide new insight into the machine learning systems in question, and the underlying physical theories used to understand them. In paticular, systematic improvements in prediction accuracy, speed, and addressing issues such as overfitting, would be useful. These aims will require interdisciplinary efforts from researchers engaged in the energy landscapes approach and the machine learning community, and we hope that our present results will help to stimulate these interactions. For example, the heat capacity features that we have identified may highlight qualitatively different sorts of locally optimal fits. It may be possible to utilise this diversity in designing weighted combinations of solutions that provide greater accuracy, or improve the prediction of particular outcomes.

It is noteworthy that neural networks of the sort we have considered were employed 30 years ago to fit intermolecular potentials for weakly bound molecular complexes.[62] The prospect of improving such procedures using the energy landscapes approach that has been developed in the intervening period could benefit several areas of contemporary research interest.

VIII. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the EPSRC and the ERC.

REFERENCES

[1]R. Collobert, F. Sinz, J. Weston, and L. Bottou, in “emph“bibinfo■booktitle■Proceedings■of ■■the■23rd■International■Conference■on■MachineŁearning, ICML ’06 (ACM, New York, NY, USA, 2006) pp. 201–208, ISBN 1-59593-383-2, http://doi.acm.org/10.1145/1143844.1143870.

[2]L. Zhao, M. Mammadov, and J. Yearwood, in “emph“bibinfo■booktitle■Data■Mining ■■Workshops■(ICDMW),■2010■IEEE■International■Conference■on (2010) pp. 1281–1288.

[3]A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, CoRR **abs/1412.0233** (2014), http://arxiv.org/abs/1412.0233.

[4]Y. Dauphin, R. Pascanu, Ç. Gülçehre, K. Cho, S. Ganguli, and Y. Bengio, CoRR **abs/1406.2572** (2014), http://arxiv.org/abs/1406.2572.

[5]A. J. Ballard, J. Stevenson, and D. J. Wales(2016).

[6]D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).

[7]D. J. Wales, Curr. Op. Struct. Biol. **20**, 3 (2010).

[8]D. J. Wales, Phil. Trans. Roy. Soc. A **370**, 2877 (2012).

[9]V. K. de Souza and D. J. Wales, J. Chem. Phys. **129**, 164507 (2008).

[10]D. J. Wales, M. A. Miller, and T. R. Walsh, Nature **394**, 758 (1998).

[11]J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **110**, 6896 (1999).

[12]M. T. Oakley, R. L. Johnston, and D. J. Wales, Phys. Chem. Chem. Phys. **15**, 3965 (2013).

[13]J. P. Neirotti, F. Calvo, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **112**, 10340 (2000).

[14]F. Calvo, J. P. Neirotti, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **112**, 10350 (2000).

[15]V. A. Mandelshtam, P. A. Frantsuzov, and F. Calvo, J. Phys. Chem. A **110**, 5326 (2006).

[16]V. A. Sharapov, D. Meluzzi, and V. A. Mandelshtam, Phys. Rev. Lett. **98**, 105701 (2007).

[17]V. A. Sharapov and V. A. Mandelshtam, J. Phys. Chem. A **111**, 10284 (2007).

[18]F. Calvo, Phys. Rev. E **82**, 046703 (2010).

[19]D. J. Wales, Chem. Phys. Lett. **584**, 1 (2013).

[20]R. M. Sehgal, D. Maroudas, and D. M. Ford, J. Chem. Phys. **140**, 104312 (2014).

[21]D. J. Wales, Mol. Phys. **100**, 3285 (2002).

[22]D. J. Wales, Mol. Phys. **102**, 891 (2004).

[23]M. Picciani, M. Athenes, J. Kurchan, and J. Tailleur, J. Chem. Phys. **135**, 034108 (2011).

[24]K. Swersky, J. Snoek, and R. P. Adams, arXiv:1406.3896 [stat.ML](2014).

[25]D. J. Wales, J. Chem. Soc. Faraday Trans. **88**, 653 (1992).

[26]D. J. Wales, J. Chem. Soc. Faraday Trans. **89**, 1305 (1993).

[27]D. Asenjo, J. D. Stevenson, D. J. Wales, and D. Frenkel, J. Phys. Chem. B **117**, 12717 (2013).

[28]E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, Phys. Rev. Lett. **97**, 170201 (Oct 2006), `http://link.aps.org/doi/10.1103/PhysRevLett.97.170201`.

[29]D. J. Wales, "GMIN," (2010), updated February 19, 2010.

[30]D. J. Wales, "Optim: A program for geometry optimisation and pathway calculations," http://www-wales.ch.cam.ac.uk/software.html.

[31]C. G. Broyden, IMA J. Appl. Math. **6**, 76 (1970), http://imamat.oxfordjournals.org/content/6/1/76.full.pdf+html, `http://imamat.oxfordjournals.org/content/6/1/76.abstract`.

[32]R. Fletcher, Comput. J. **13**, 317 (1970), http://comjnl.oxfordjournals.org/content/13/3/317.full.pdf+html, http://comjnl.oxfordjournals.org/content/13/3/317.abstract.

[33]D. Goldfarb, Math. Comp. **24**, 23 (1970), ISSN 00255718, http://www.jstor.org/stable/2004873.

[34]D. F. Shanno, Math. Comp. **24**, 647 (1970), ISSN 00255718, http://www.jstor.org/stable/2004840.

[35]J. E. Jones and A. E. Ingham, Proc. R. Soc. A **107**, 636 (1925).

[36]P. M. Axilrod and E. Teller, J. Chem. Phys. **11**, 299 (1943).

[37]Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **84**, 6611 (1987).

[38]Z. Li and H. A. Scheraga, J. Mol. Struct. **179**, 333 (1988).

[39]D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).

[40]S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **120**, 2082 (2004).

[41]G. Henkelman and H. Jónsson, J. Chem. Phys. **111**, 7010 (1999).

[42]L. J. Munro and D. J. Wales, Phys. Rev. B **59**, 3969 (1999).

[43]F. Noé and S. Fischer, Curr. Op. Struct. Biol. **18**, 154 (2008).

[44]D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique, and F. Fernando, PLoS Comput. Biol. **5**, e1000415 (2009).

[45]"Pele: Python energy landscape explorer, https://github.com/pele-python/pele," https://github.com/pele-python/pele.

[46]O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[47]F. H. Stillinger and T. A. Weber, J. Chem. Phys. **80**, 2742 (1984).

[48]D. J. Wales and J. P. K. Doye, J. Chem. Phys. **119**, 12409 (2003).

[49]D. J. Wales, Mol. Phys. **78**, 151 (1993).

[50]F. H. Stillinger, Science **267**, 1935 (1995).

[51]B. Strodel and D. J. Wales, Chem. Phys. Lett. **466**, 105 (2008).

[52]F. Calvo, J. P. K. Doye, and D. J. Wales, J. Chem. Phys. **115**, 9627 (2001).

[53]J. P. K. Doye and D. J. Wales, J. Chem. Phys. **102**, 9659 (1995).

[54]J. P. K. Doye and D. J. Wales, J. Chem. Phys. **102**, 9673 (1995).

[55]I. Georgescu and V. A. Mandelshtam, J. Chem. Phys. **137**, 144106 (2012).

[56]D. J. Wales and J. P. K. Doye, J. Chem. Phys. **103**, 3061 (1995).

[57]L. Breiman, Machine learning **45**, 5 (2001).

[58]T. K. Ho, Pattern Analysis and Machine Intelligence, IEEE Transactions on **20**, 832 (1998).

	number of hidden nodes			
data	3	4	5	6
1	28, 162, 1504	36, 2559, 10112	44, 4752, 18779	52, 19045, 34052
1+2	37, 310, 2631	48, 1392, 3786	59, 5146, 11816	70, 11997, 20297
1+2+3	46, 336, 987	60, 2565, 4896	74, 10298, 10298	88, 8652, 11192
10	28, 271, 1461			
-5	28, 6, 33			
final	28, 6, 25			

TABLE I. Details of the stationary point databases obtained for the neural network fits of geometry optimisation series data. The number of fitting parameters, local minima, and transition states are tabulated for fits based on selected data points from each local minimisation. '1' refers to the interatomic distances at the starting point, '2' refers to the distances at the second point in the series, etc. '$1+2$' and '$1+2+3$' are fits for the interparticle distances at the first two and the first three configurations, respectively. '10th', '-5' and 'final' refer to the distances at the 10th step, the 5th from last step, and the converged configuration. There are three interparticle distances fitted for all cases except '$1+2$' and '$1+2+3$', which have six and nine distances, respectively.

[59]E. Kleinberg, Annals of Mathematics and Artificial intelligence **1**, 207 (1990).

[60]P. A. Ortega and D. A. Braun, Proc. R. Soc. Lond. A **469** (2013), ISSN 1364-5021, doi: \bibinfo{doi}{10.1098/rspa.2012.0683}.

[61]J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, CoRR **abs/1503.03585** (2015), `http://arxiv.org/abs/1503.03585`.

[62]D. F. R. Brown, M. N. Gibbs, and D. C. Clary, J. Chem. Phys. **105**, 7597 (1996).

	number of hidden nodes			
data	3	4	5	6
1	114.40, 112.98[1], 112.98[1]	81.63, 138.01[1], 107.48[91]	62.43, 135.91, 101.76[5179]	49.40, 136.40[1], 89.86[4733]
1+2	129.96, 167.73[1], 167.73[1]	90.69, 179.39[1], 129.32[256]	67.48, 222.97[1], 119.45[8315]	53.63, 194.35[1], 112.97[13926]
1+2+3	125.17, 224.78[1], 173.57[207]	86.78, 186.68[1], 137.69[1491]	61.23, 218.92[1], 124.13[1588]	46.91, 184.58[1], 119.32[6588]
10	74.16, 153.28[1], 93.97[15]			
-5	1.34, 1.33[1], 1.33[1]			
final	1.34, 1.33[1], 1.33[1]			

TABLE II. Global minimum objective function values for training and testing data. In each case the three numbers correspond to the global minimum for 500 training data points, the objective function for this global minimum evaluated for 500 test data points, and the lowest objective function value for the testing data obtained with any of the local minima for the training data. The indices in square brackets for the testing data refer to the ranking of the local minimum for the fit to the training data, where [1] is the training set global minimum.

	number of hidden nodes			
data	3	4	5	6
1	9.88%, 0.2–0.478	0.93%, 0.2–0.8	0.14%, 0.2–0.478	0.09%, 0.2–0.278
1+2	46.73%, 0.2–0.8	2.81%, 0.2–0.548	3.53%, 0.2–0.276	7.03%, 0.2–0.548
1+2+3	13.08%, 0.2–0.748	1.49%, 0.2–0.548	2.96%, 0.2–0.548	7.29%, 0.2–0.474
10	8.47%, 0.2–0.75			
-5	0%			
final	0%			

TABLE III. Percentage of local minima with non-zero misclassification fractions for the test set data, together with the range of fractional misassignments for the non-zero cases.

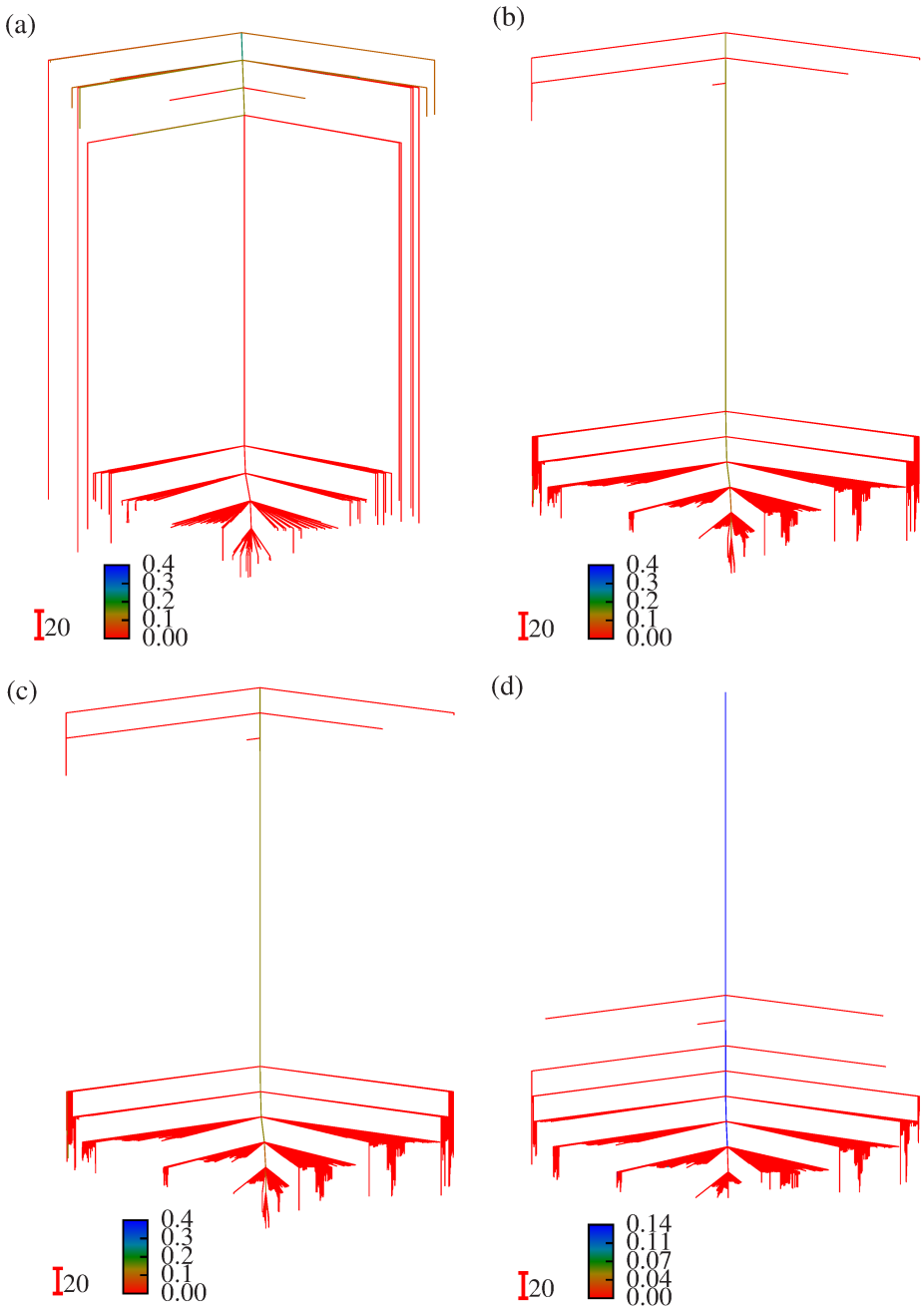

FIG. 1. Disconnectivity graphs for the fitting landscapes of the triatomic cluster geometry optimisation series using data from the initial configuration only with (a) 3, (b) 4, (c) 5, and (d) 6 hidden nodes. The nodes are coloured according to the misclassification distance for the local minima evaluated using training data.

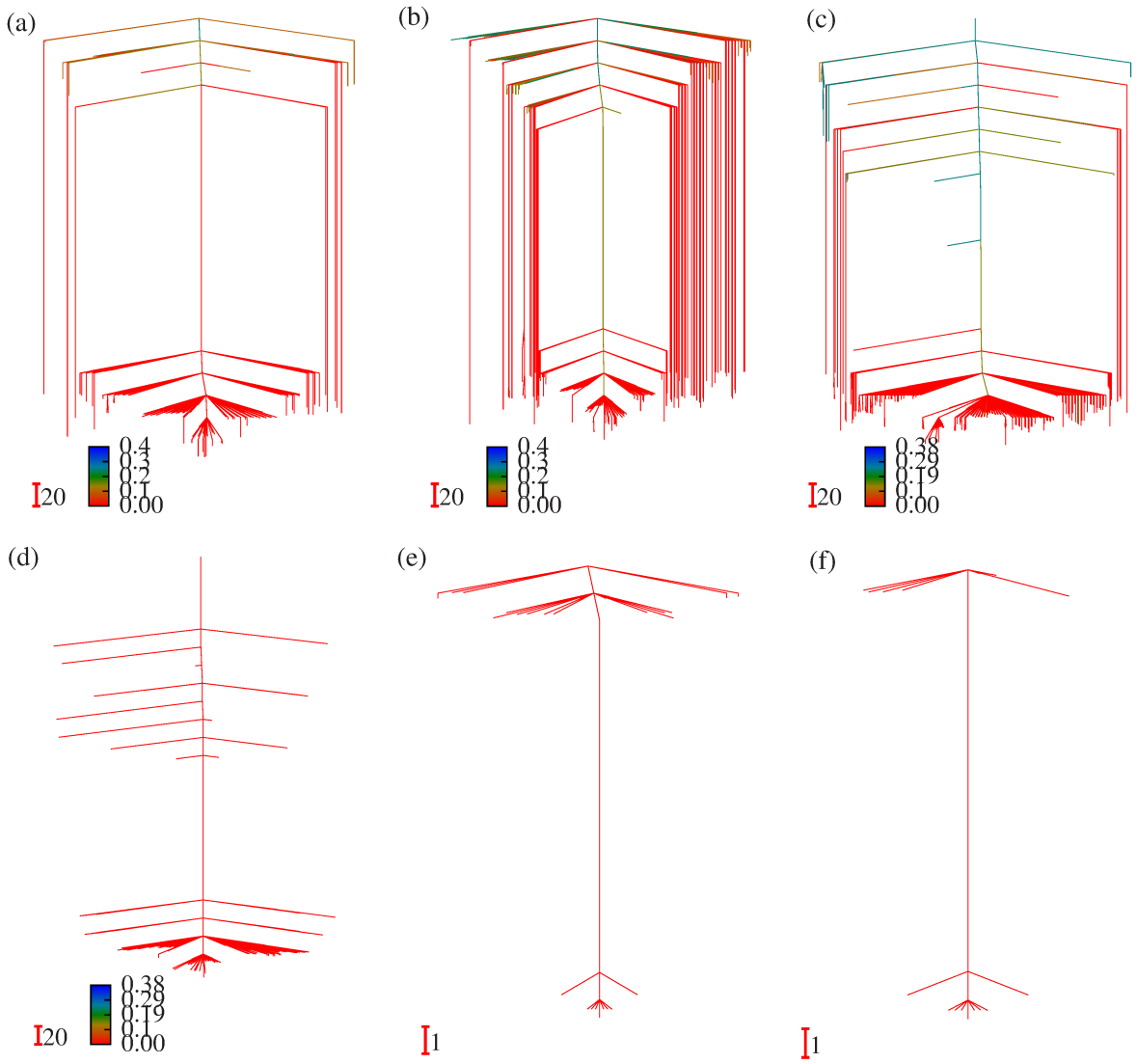

FIG. 2. Disconnectivity graphs for the fitting landscapes of the triatomic cluster geometry optimisation series using 3 hidden nodes and data for (a) the initial configuration (b) the first and second configurations, (c) the first three configurations, (d) the 10th configuration, (e) the 5th configuration from the end, (f) the final configuration. The nodes are coloured according to the misclassification distance for the local minima evaluated using training data.

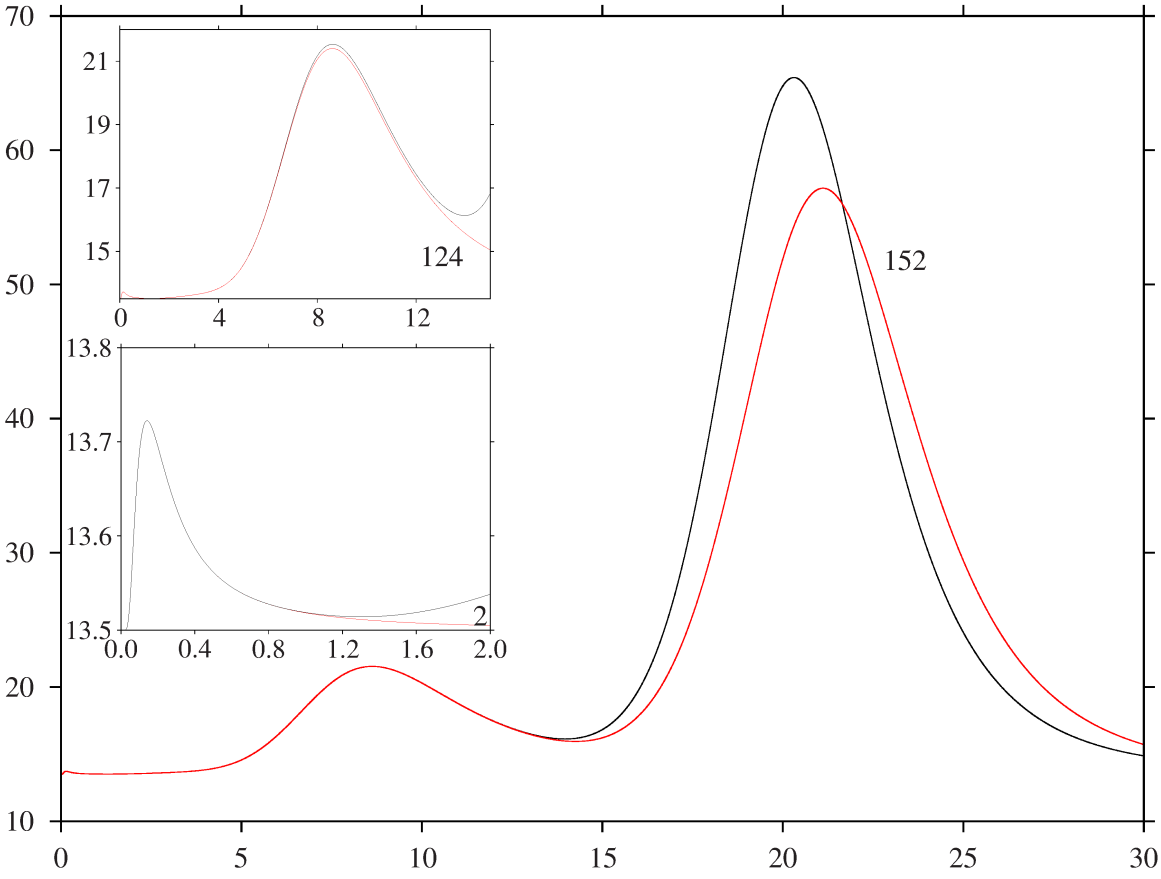

FIG. 3. Heat capacity for the landscape defined for the dataset using only the initial interatomic distances with three hidden nodes. The insets illustrate the convergence of the two low temperature peaks. In each plot the black curve corresponds to C_v calculated from the complete database of minima. 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.

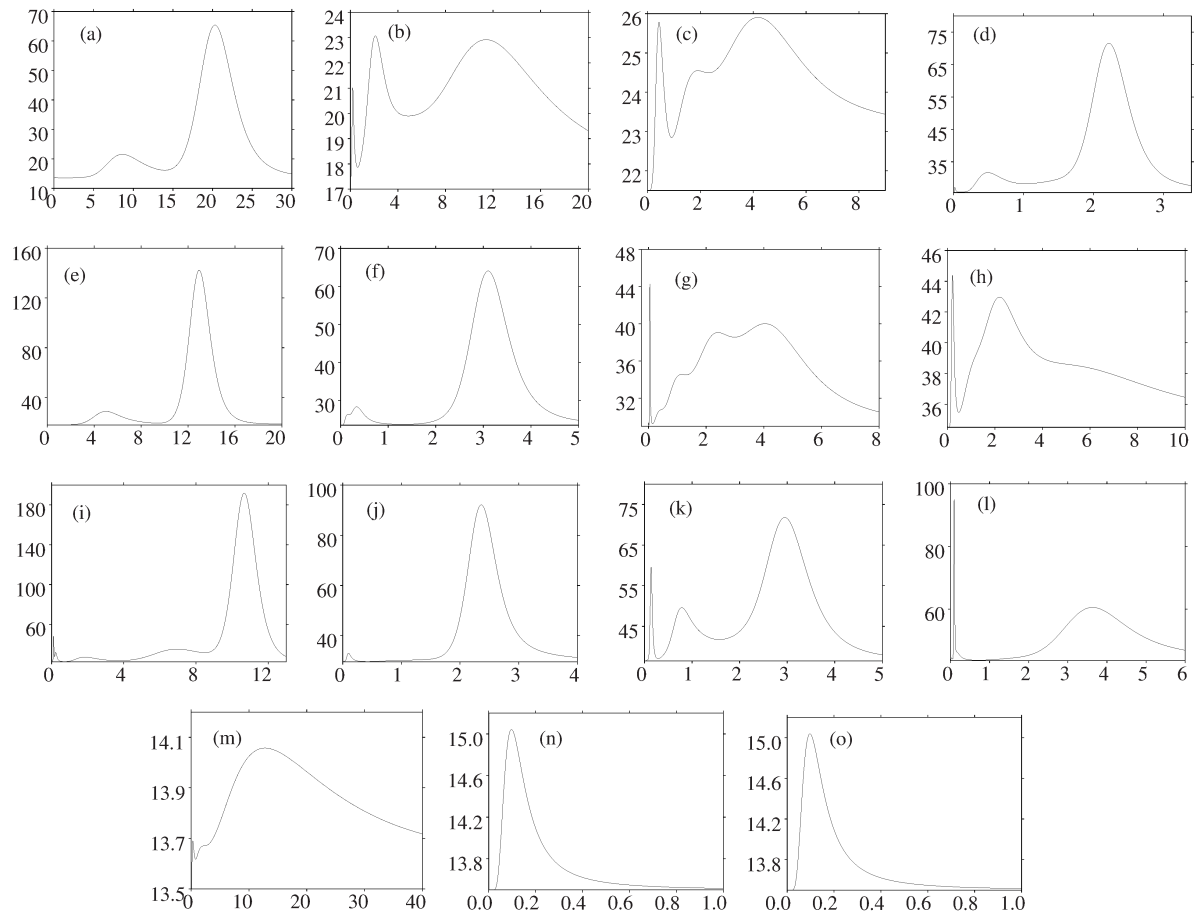

FIG. 4. Heat capacity for all 15 landscapes considered in the present work. (a) to (d) are for data format '1', (e) to (h) for '1 + 2', (i) to (l) for '1 + 2 + 3', in each case for three, four, five and six hidden nodes, respectively. (m), (n) and (o) are for the single point data using the 10th, fifth from last, and final configurations, respectively, each with three hidden nodes. Every panel exhibits multiple heat capacity peaks, aside from the last two, where the data used for training are practically converged, and the predictions become straightforward.