PNAS
www.pnas.org

Supplementary Information for

Archetypal Landscapes for Deep Neural Networks

P.C. Verpoort, A.A. Lee, D.J. Wales

Philipp C. Verpoort.
E-mail: pcv22@cam.ac.uk

This PDF file includes:

Supplementary text
Figs. S1 to S5
Tables S1 to S2
Caption for Movie S1
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Movie S1

Supporting Information Text

In this supporting information, we explain the data set used for training and testing in the main contribution. Sec. S1 outlines the $LJAT_3$ classification problem that was employed and the data sets that were generated. Sec. S2 describes the visualisation of solutions that is used throughout our main report. We also tabulate mean values for uphill and downhill barriers for individual transition states and for the pathway with the lowest maximum transition state energy connecting other minima to the global minimum.

S1. Predicting the Outcome of Geometry Optimisation for an Atomic Cluster

This benchmarking problem has been used in several previous contributions that employed neural network fits with single hidden layers (1–3). This work investigated how the corresponding machine learning landscapes and predictions varied with the number of nodes and the number of training data, including the effect of memory in sequences of molecular configurations. The system is a triatomic cluster bound by pairwise Lennard-Jones (4) and three-body Axilrod–Teller (5) terms, parameterised so that there are three permutational isomers of a linear minimum, distinguished by the central atom, and one additional minimum for an equilateral triangle with D_{3h} symmetry. The total potential energy for this $LJAT_3$ cluster is

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

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, and Z is a parameter that weights the contribution of the three-body term. For $Z = 2$ the linear minima have potential energy $V = -2.219\,\varepsilon$, and the triangle lies slightly higher with $V = -2.185\,\varepsilon$. For the triangle $r_{12} = r_{13} = r_{23} = 1.16875\,\sigma$, and in the linear minima the nearest-neighbour distances are both 1.10876σ.

The aim of this multinomial logistic regression problem is to predict which of the four local minima a geometry optimisation will find, given some information about initial or intermediate configurations in terms of the interparticle distances. To generate data we consider starting geometries constructed from randomly distributing the three atoms in a cube of side length L. The initial values of r_{12} and r_{13} were employed as the input data for all the tests conducted in the present work, and we considered two datasets, the first (D1) for 10,000 minimisations for a cube with $L = 2\sqrt{3}\,\sigma$, and the second (D2) for 200,000 minimisations with $L = 1.385\,\sigma$. We used half of each dataset for training and half for testing, where appropriate. Dataset D1 was employed in previous work, while D2 was generated for the present investigation. Each local minimisation employed the same LBFGS minimisation routine described in **Methods**, and the convergence condition was taken as 10^{-6} for the root mean square gradient in reduced units of ε/σ.

The molecular configuration is completely characterised by three interparticle distances, r_{12}, r_{13}, and r_{23}. If we supply sufficient training data with these three inputs, predicting the outcome of minimisation can be essentially perfect. The problem is then equivalent to learning the basin of attraction for each local minimum, which is a well-defined volume of configuration space for steepest-descent minimisation (6, 7).

By restricting the input data to r_{12} and r_{13}, and omitting r_{23}, we make the prediction problem harder. For the linear minima with atom 2 or atom 3 in the middle, r_{13} and r_{12}, respectively, are much larger than for the triangle. However, these distances are only about 5% different in the triangle and the linear minimum with atom 1 in the central position. The basins of attraction of the triangle and this third linear minimum therefore overlap significantly in the space defined by r_{12} and r_{13}. The best predictions we can achieve will therefore occur when we have converged the relative probabilities of finding these two structures as a function of r_{12} and r_{13}.

The same considerations will apply for larger molecules: if we sample the whole configuration space sufficiently, we should be able to predict which basin of attraction any starting structure corresponds to,

and the corresponding local minimum. Otherwise, the best we can do is to learn the relative probabilities, averaged over the missing degrees of freedom. Hence we find the current benchmark appealing because of the ability to generate arbitrary amounts of training and testing data, because of the clear physical interpretations, and because of the practical importance of the configuration volumes themselves. For example, the volume of basins of attraction provides measures of configurational entropy, which has been applied to analyse granular packings (8). Reliable prediction for the outcome of minimisation would enable us to reduce the time required for such calculations by stopping earlier, with a weaker convergence threshold on the magnitude of the gradient (9).

S2. Visualisation of Solutions

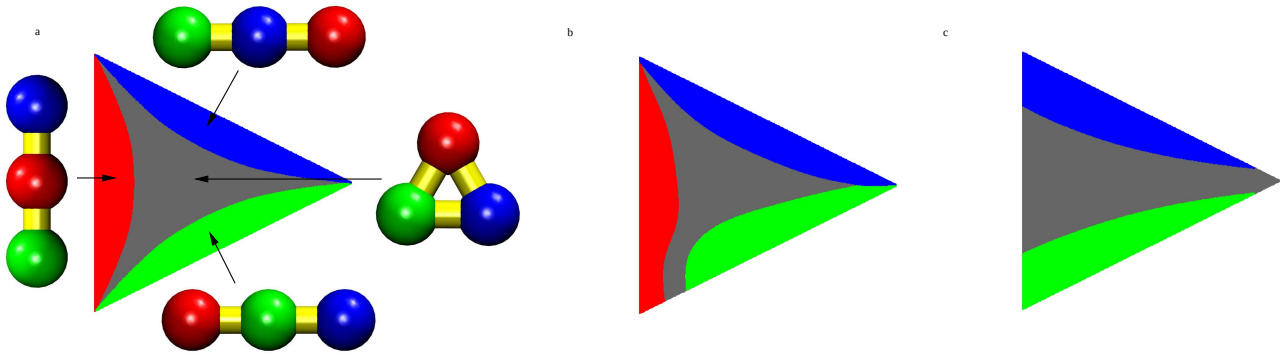

Fig. S1. Graphical representation of the $LJAT_3$ classification problem. **(a)** Colored according to the true outcome determined by geometry optimization for the $LJAT_3$ cluster. The four optimal atomic configurations are associated with their corresponding basins of attraction. **(b)** Colored according to the predictions for the global minimum of a single hidden layer neural network with 3 hidden nodes and 100 training data confined in the plane $\mathcal{R}'$ (AUC 0.98 from corresponding test set). **(c)** Colored according to the predictions for the global minimum of a single hidden layer neural network with 10 hidden nodes trained on 100,000 training data in $\mathcal{R}$ (AUC 0.79 from corresponding test set).

Some insight into the different local minima in the cost function for a given neural network and training data can be obtained graphically for the $LJAT_3$ prediction problem (3, 10). We construct a two-dimensional projection of coordinates in the plane $r_{12}+r_{13}+r_{23}=3r_e$ from the three-dimensional space $\{r_{12},r_{13},r_{23}\}$, where $r_e=2^{1/6}$ is the equilibrium bond length in a dimer and in the equilateral triangle minimum. The orthogonal unit vectors $\hat{\mathbf{v}}_1=(1,1,-2)/\sqrt{6}$ and $\hat{\mathbf{v}}_2=(1,-1,0)/\sqrt{2}$ lie in this plane and are perpendicular to the $\{1,1,1\}$ direction. We define projected coordinates $x=(r_{12}+r_{13}-2r_{23})/\sqrt{6}$ and $y=(r_{12}-r_{13})/\sqrt{2}$. For a regular 700×700 grid with $-\sqrt{3}r_e<x,y<\sqrt{3}r_e$, we solve for r_{12}, r_{13}, and r_{23} with $r_{12}+r_{13}+r_{23}=3r_e$, which gives 79524 geometrically feasible (x,y) points, and the associated values of $\{r_{12},r_{13},r_{23}\}$, and Cartesian coordinates. The 79524 values of r_{12} and r_{13} constitute a third dataset D3. The feasible geometries are distributed over a triangle in (x,y) space, where the centre of each edge corresponds to a linear geometry with two distances of $3r_e/4$ and one of $3r_e/2$, and each vertex corresponds to two atoms coincident and the third at a distance of $3r_e/2$. The equilateral triangular minimum maps to $(x,y)=(0,0)$.

The pixels on the (x,y) grid are coloured according to the minimum with highest predicted probability when the associated configuration is used as input data for any given neural network. If the equilateral triangle has the highest probability the pixel is gray, while the three linear minima with atoms 1, 2, and 3 in the centre are coloured red, green and blue, respectively. Since we are omitting r_{23} from all the training data, we anticipate that predictions will be significantly perturbed from previous calculations that included all three distances as inputs (3). The target result is given by the known outcomes obtained by energy minimisation with the same colour scheme (Fig. S1a). The basins of attraction for the three linear minima are symmetrically disposed along the three edges of the triangle in the (x,y) projection, while the remaining basin for the D_{3h} minimum has three-fold symmetry in this space.

Although this graphical representation only includes a subset of configurations in one plane defined in the three-dimensional space $\{r_{12},r_{13},r_{23}\}$, comparison with the target reference pattern provides a very useful indication of how well any particular neural network performs. It can be used for any set of weights

in any of the neural networks we consider, including transition states and all the configurations along the pathways that connect the local minima. An illustrated profile for pathways leading to the global minimum is shown in Fig. S3, and a movie with frames constructed from all the pathway configurations is available as Supplementary Information. The capability to visualise cuts through a testing data set in terms of the evolution in the predictive capabilities might prove useful in understanding how to construct better fits in future work.

S3. Area Under Curve

To quantify the prediction capabilities of any given local minimum we calculated the area under curve (AUC) for receiver operating characteristic (ROC) plots of the true positive rate, $T_{\rm pr}$, against the false positive rate, $F_{\rm pr}$, as a function of the threshold probability, P, for predicting convergence to the equilateral triangle. These rates are defined as

$$\begin{aligned} T_{\rm pr}(\mathbf{W};P) &= \sum_{d=1}^{N_{\rm data}} \delta_{c(d),0}\,\Theta\Big(p_0(\mathbf{W})-P\Big) \Big/ \sum_{d=1}^{N_{\rm data}} \delta_{c(d),0}\,, \\ F_{\rm pr}(\mathbf{W};P) &= \sum_{d=1}^{N_{\rm data}} (1-\delta_{c(d),0})\,\Theta\Big(p_0(\mathbf{W})-P\Big) \Big/ \sum_{d=1}^{N_{\rm data}} (1-\delta_{c(d),0})\,, \end{aligned} \qquad [2]$$

where Θ is the Heaviside step function and δ is the Kronecker delta. The AUC value is then

$$\mathrm{AUC}(\mathbf{W}) = \int_0^1 T_{\rm pr}(\mathbf{W};P)\ \mathrm{d}F_{\rm pr}(\mathbf{W};P)\,, \qquad [3]$$

and was obtained numerically.

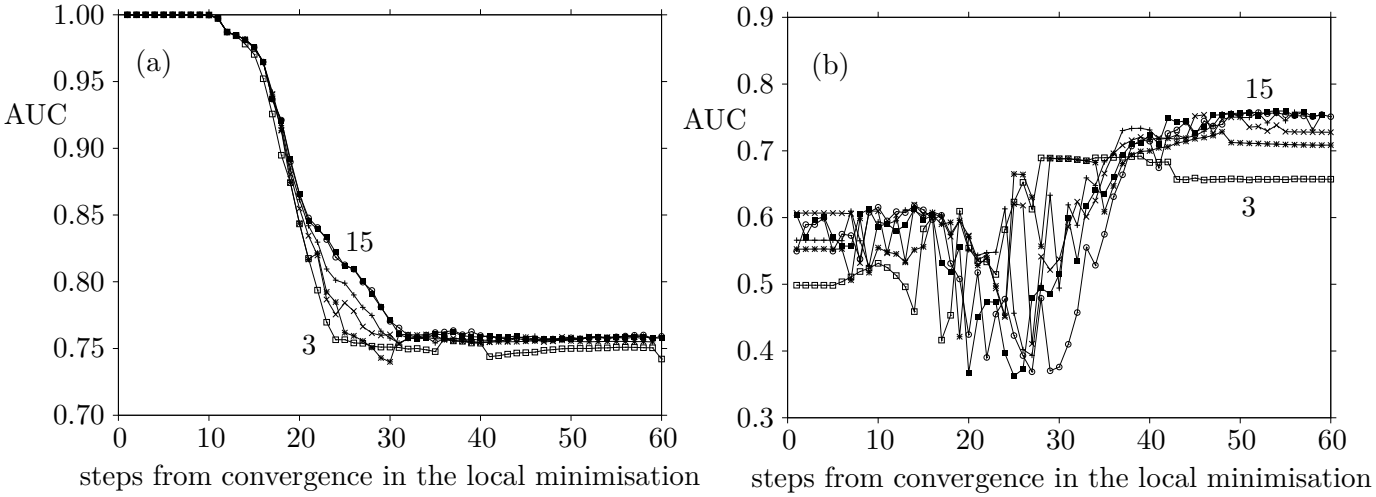

Fig. S2. AUC values obtained for the global minimum fit to 5000 training data in database D1 with networks containing a single hidden layer and 3, 4, 5, 6, 10 and 15 nodes. Global optimisation was performed for 5000 (r_{12}, r_{13}) pairs as a function of the number of steps to convergence in the geometry optimisation, as in previous work (2). (a) Results for the 5000 testing data in database D1 for (r_{12}, r_{13}) pairs at the same number of steps to convergence as in each fit. (b) Results for the 79524 test data in database D3. The plots for 3 and 15 hidden nodes are indicated in each case, and the AUC values generally increase with the number of nodes as the training configurations approach the random initial configurations at the maximum number of steps from convergence.

We performed additional global optimisation runs for single hidden layers with 3, 4, 5, 6, 10 and 15 nodes using configurations corresponding to saved (r_{12}, r_{13}) data along the 5000 training minimisation sequences in database D1. As in previous work (2), we find that the AUC values for the minima obtained in training, and for configurations in the 5000 testing sequences in D1 at the same position in the minimisation, improve systematically as the geometry optimisations approach convergence (Fig. S2a). However, if we apply the solutions to the testing data in database D3 the best AUC values of around 0.75 correspond to fits using

the starting configurations, i.e. the random (r_{12}, r_{13}) training data (Fig. S2b). Not surprisingly, training only on configurations close to the four equilibrium geometries of the cluster produces best fits that do not generalise as well to different configurations.

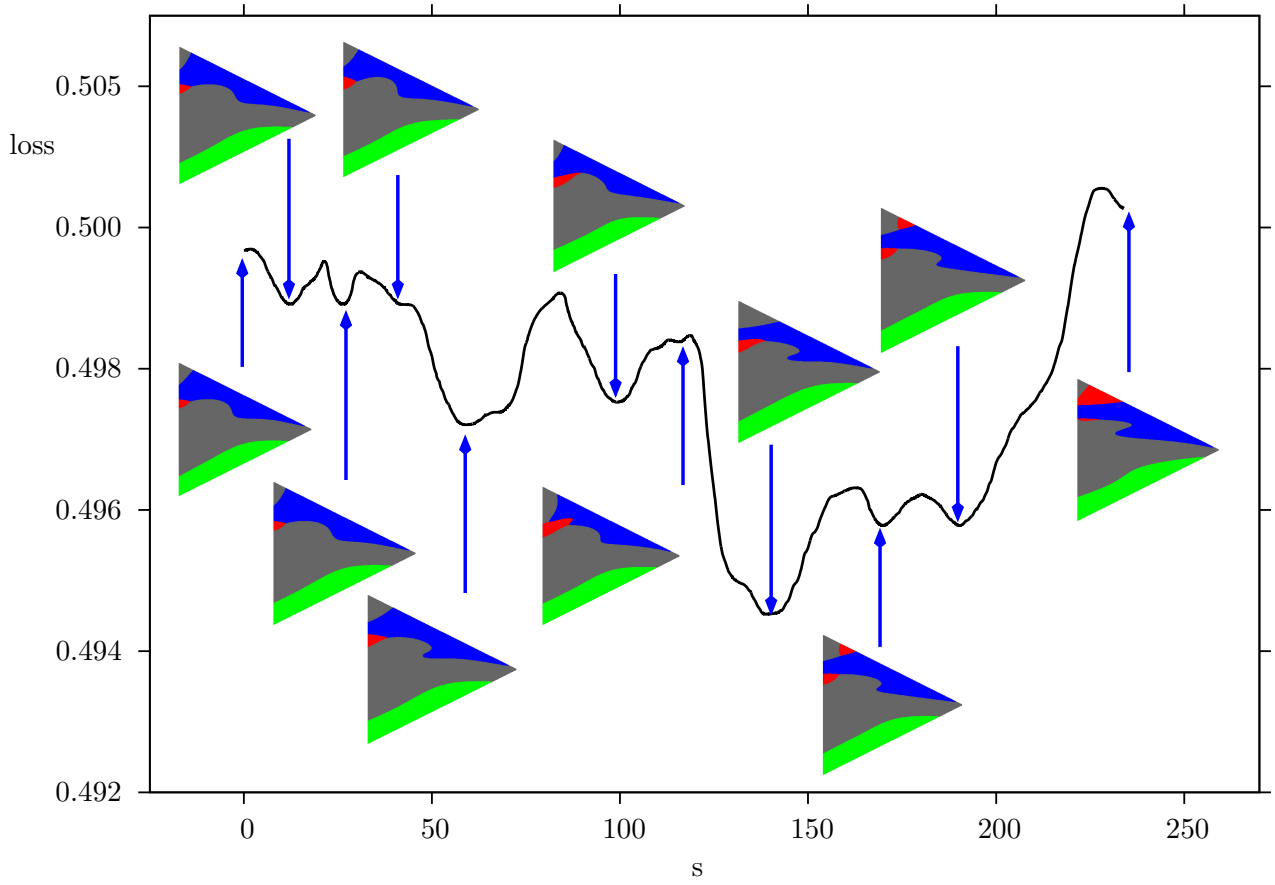

Fig. S3. Loss profile for a pathway involving eleven minima and ten transition states for the 1HL network trained on 250 data for $LJAT_3$ geometry optimisations. A graphical representation of the predictions for the D3 test set is indicated for each minimum. The global minimum is the eighth in the sequence. The horizontal axis corresponds to the integrated path length, s, which is calculated by treating the variable weights in the neural network according to a Euclidean distance metric.

Table S1. Average uphill and downhill barriers for all the transition states and directly connected minima located in training, excluding degenerate rearrangements (7, 11). The second table is the mean of the barrier divided by the loss difference between the two minima, yielding a dimensionless parameter.

	mean barriers					
	$H = 1$		$H = 2$		$H = 3$	
N_{data}	uphill	downhill	uphill	downhill	uphill	downhill
100	0.114×10^{-1}	0.531×10^{-2}	0.155×10^{-1}	0.468×10^{-2}	0.158×10^{-1}	0.489×10^{-2}
1000	0.985×10^{-4}	0.196×10^{-4}	0.239×10^{-2}	0.853×10^{-3}	0.314×10^{-2}	0.978×10^{-3}
2000	0.704×10^{-4}	0.415×10^{-4}	0.126×10^{-2}	0.534×10^{-3}	0.144×10^{-2}	0.489×10^{-3}
10000	0.109×10^{-3}	0.455×10^{-4}	0.466×10^{-3}	0.186×10^{-3}	0.748×10^{-3}	0.274×10^{-3}
100000	0.603×10^{-4}	0.432×10^{-4}	0.185×10^{-3}	0.973×10^{-4}	0.775×10^{-3}	0.198×10^{-3}

	mean barriers divided by loss difference of minima					
	$H = 1$		$H = 2$		$H = 3$	
N_{data}	uphill	downhill	uphill	downhill	uphill	downhill
100	10.70	9.70	6.683	5.683	7.612	6.612
1000	2.131	1.131	5.767	4.767	4.915	3.915
2000	2.743	1.743	6.570	5.570	4.466	3.466
10000	3.167	2.167	7.952	6.952	5.672	4.672
100000	13.359	12.359	6.588	5.588	2.096	1.096

Table S2. Average downhill barrier to the global minimum for all the other minima located in training, and average of the downhill barrier divided by the loss difference between the two minima (scaled column).

	$H = 1$		$H = 2$		$H = 3$	
N_{data}		scaled		scaled		scaled
100	0.105×10^{-3}	0.0178	0.276×10^{-2}	0.0783	0.324×10^{-2}	0.0781
1000	0.298×10^{-5}	0.0905	0.357×10^{-3}	0.0562	0.540×10^{-3}	0.0462
2000	0.589×10^{-4}	0.9777	0.101×10^{-3}	0.0676	0.566×10^{-4}	0.0294
10000	0.415×10^{-5}	0.0597	0.316×10^{-4}	0.4342	0.324×10^{-4}	0.0477
100000	0.663×10^{-5}	0.3301	0.332×10^{-4}	0.3019	0.216×10^{-4}	0.0286

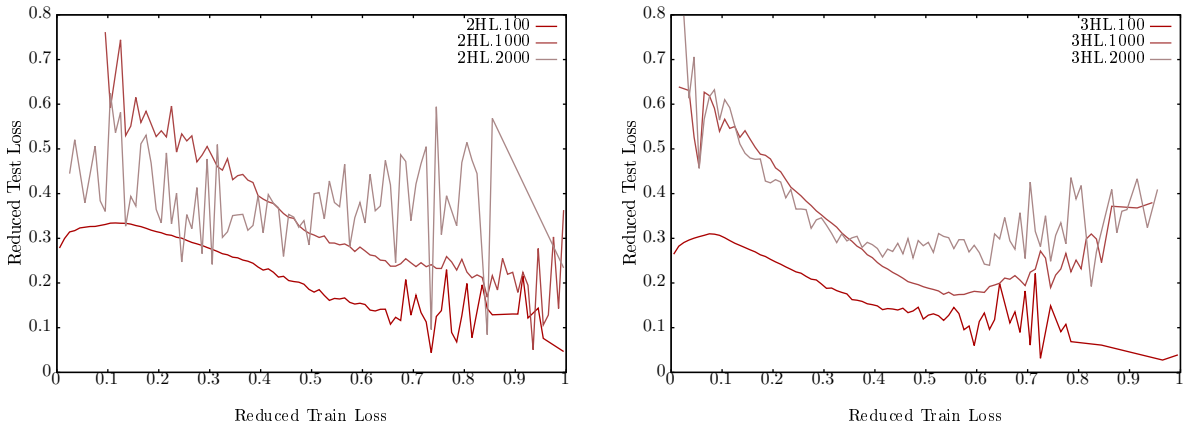

Fig. S4. Reduced test loss plotted against reduced train loss of minima of the the LJAT loss function landscapes for $H = 2, 3$ and $N_{\mathrm{data}} = 100, 1000, 2000$. The train loss is divided into 100 intervals and the test loss is averaged over all minima found with train loss in the interval. The reduced train (test) loss is defined as $L_{\mathrm{red}}(L) = \frac{L - L_{\mathrm{min}}}{L_{\mathrm{max}} - L_{\mathrm{min}}}$, where L_{max} is the maximal and L_{min} is the minimal train (test) loss value in the corresponding database of minima. The graph shows how for the average test loss increases towards the bottom of the train loss landscape for $N_{\mathrm{data}} = 1000, 2000$, as one would normally expect. For $N_{\mathrm{data}} = 100$ however, the average test loss seems to be decreasing again at the bottom of the train loss landscape.

S4. Example Disconnectivity Graphs for a Structural Glass-Former

Two examples of disconnectivity graphs obtained for model structural glass-formers are shown in Figure S5 for comparison with the loss function landscapes illustrated for neural networks. The system in question is a binary mixture of particles interacting via a Lennard-Jones potential (4) modelled with periodic boundary conditions and 60 or 256 particles in the supercell, described as BLJ_{60} and BLJ_{256}, respectively. Here, BLJ_{60} contains 48 type A and 12 type B particles, while BLJ_{256} contains 204 A and 52 B particles, and the results correspond to a number density of σ_{AA}^{-3}, where $2^{1/6}\sigma_{AA}$ is the pair equilibrium separation for two A particles. The corresponding pair well depth is ϵ_{AA}. Choosing $\sigma_{AA} = 1$ and $\epsilon_{AA} = 1$ defines a system of reduced units, and the additional parameters are $\sigma_{AB} = 0.8$, $\sigma_{BB} = 0.88$, $\epsilon_{AB} = 1.5$, and $\epsilon_{BB} = 0.5$. (15) The pairwise interactions were shifted and truncated according to the Stoddard-Ford scheme to assure continuous energy and first derivatives. The database for BLJ_{60} contains over 11000 minima, and the database for BLJ_{256} has 2500.

The potential energy landscapes in Figure S5 exhibit hierarchical structure, which appears to be common to other structural glasses (16–18), with numerous low-lying amorphous configurations separated by high barriers or order $30k_BT_g$ for glass transition temperature T_g. Full details of the database construction and analysis can be found in the original reports (12–14).

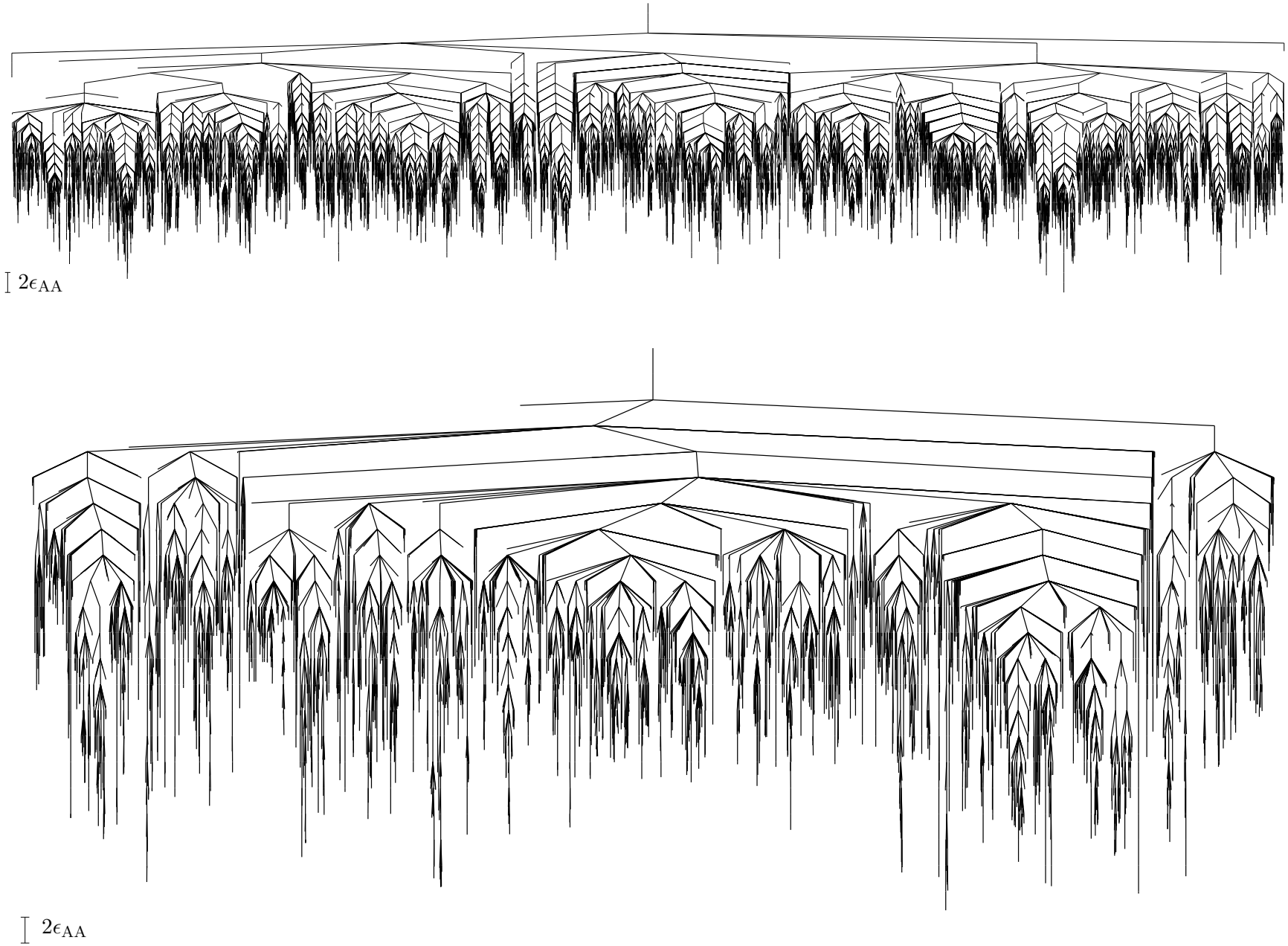

Fig. S5. Example disconnectivity graphs for binary Lennard-Jones systems containing 60 atoms, BLJ_{60} (top), and 256 atoms, BLJ_{256} (bottom), in periodically repeated supercells.(12–14)

Movie S1. Graphical representation of the predictions for the D3 test set along the pathway shown in Fig. S3.

References

1. A. J. Ballard, J. D. Stevenson, R. Das, and D. J. Wales. Energy landscapes for a machine learning application to series data. *J. Chem. Phys.*, 144(12):124119, 2016.
2. R. Das and D. J. Wales. Machine learning prediction for classification of outcomes in local minimisation. *Chem. Phys. Lett.*, 667:158 – 164, 2017. .
3. D. J. Wales. Exploring energy landscapes. *Ann. Rev. Phys. Chem.*, 69:401–425, 2018. .
4. J. E. Jones and A. E. Ingham. On the calculation of certain crystal potential constants, and on the cubic crystal of least potential energy. *Proc. R. Soc. A*, 107:636–653, 1925.
5. B. M. Axilrod and E. Teller. Interaction of the van der waals type between three atoms. *J. Chem. Phys.*, 11:299, 1943.
6. P. G. Mezey. *Potential Energy Hypersurfaces*. Elsevier, Amsterdam, 1987.
7. D. J. Wales. *Energy Landscapes*. Cambridge University Press, Cambridge, 2003.
8. Stefano Martiniani, K. Julian Schrenk, Jacob D. Stevenson, David J. Wales, and Daan Frenkel. Turning intractable counting into sampling: Computing the configurational entropy of three-dimensional jammed packings. *Phys. Rev. E*, 93:012906, 2016.
9. K. Swersky, J. Snoek, and R. P. Adams. Freeze-thaw bayesian optimization. *arXiv:1406.3896 [stat.ML]*, 2014.
10. A. J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J. D. Stevenson, and D. J. Wales. Energy landscapes for machine learning. *Phys. Chem. Chem. Phys.*, 19:12585–12603, 2017. .
11. R. E. Leone and P. v. R. Schleyer. *Angew. Chem. Int. Ed. Engl.*, 9:860, 1970.
12. V. K. de Souza and D. J. Wales. Energy landscapes for diffusion: Analysis of cage-breaking processes (13 pages). *J. Chem. Phys.*, 129:164507, 2008.
13. V. K. de Souza and D. J. Wales. Connectivity in the potential energy landscape for binary lennard-jones systems (12 pages). *J. Chem. Phys.*, 130:194508, 2009.
14. V. K. de Souza, J. D. Stevenson, S. P. Niblett, J. D. Farrell, and D. J. Wales. Defining and quantifying frustration in the energy landscape: Applications to atomic and molecular clusters, biomolecules, jammed and glassy systems. *J. Chem. Phys.*, 146(12):124103, 2017. .
15. W. Kob and H. C. Andersen. Scaling behavior in the beta-relaxation regime of a supercooled lennard-jones mixture. *Phys. Rev. Lett.*, 73:1376–1379, 1994.
16. S. P. Niblett, V. K. de Souza, J. D. Stevenson, and D. J. Wales. Dynamics of a molecular glass former: Energy landscapes for diffusion in ortho-terphenyl. *J. Chem. Phys.*, 145(2):024505, 2016. .
17. S. P. Niblett, M. Biedermann, D. J. Wales, and V. K. de Souza. Pathways for diffusion in the potential energy landscape of the network glass former sio2. *The Journal of Chemical Physics*, 147(15):152726, 2017. .
18. S. P. Niblett, V. K. de Souza, R. L. Jack, and D. J. Wales. Effects of random pinning on the potential energy landscape of a supercooled liquid. *The Journal of Chemical Physics*, 149(11):114503, 2018. .