Supporting Information: Topology-based Machine Learning Strategy For Cluster Structure Prediction

Xin Chen,[†] Dong Chen,[†] Mouyi Weng,[†] Yi Jiang,[†] Guo-Wei Wei,[*,‡] and Feng Pan[*,†]

[†]*School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

[‡]*Department of Mathematics, Michigan State University, MI 48824, USA*

E-mail: weig@msu.edu; panfeng@pkusz.edu.cn

CALCULATION DETAILS

Parameters of machine learning models Gradient boosting regression is an iterative ensemble algorithms for approximating a function $F : \mathbb{R}^d \to \mathbb{R}$ from pairs of sample values. The gradient boosting regression in this work is constructed by Sklearn[S1] with learning rate as 0.001, number of estimators as 15000, max depth as 7 and min samples split as 5.

Neural networks (NN) are modeled after the function of neurons in brain. When a neural network has several layers of perceptrons we call it a deep neural network (DNN) and the intermediate layers are known as hidden layers. The DNN here are constructed by Keras[S2] with optimizer as Adam, learning rate as 0.0001, mini-batch size as 128, epoch number as 1000, and architecture as follow. The input is generated following the source code in Github. The topological invariants β_0, β_1 and β_2 in the input are calculated by Ripser,[S3] with max dimension as 2 and thresh as 10.

The Gaussian approximation potential (GAP) model is constructed using a systematic protocol for combining properly scaled many-body descriptors and a radial cutoff of 4 Å, with the software provided by Bartok et al.[S4–S6]

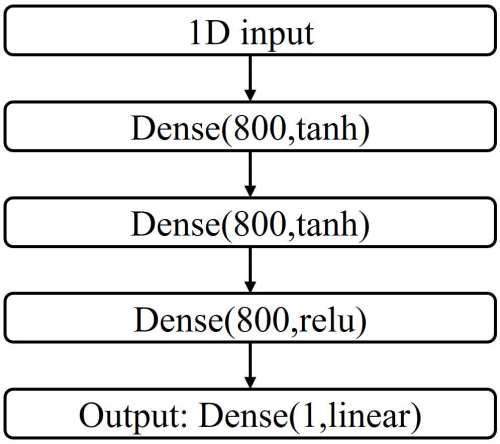

Figure S1: Architecture of the NN.

More specific examples of barcode representation As shown in Fig. S2(a), a flat 'octachedral' Li_6 cluster's barcode does not have β_2, which is different from the barcode of

octachedral Li_6 cluster shown in shown in Fig. 2. Since in the Fig. S2(a), the top and the bottom balls of flat 'octachedral' Li_6 overlap with each other first, which makes it absenting an octachedral structure. Other representations method based on point cloud data may not be able to tell this difference. However, in barcode representation of topological variants, this difference could be captured and expressed clearly by β_2. To further demonstrate the advantage and benefit of barcode representation, Fig. S2(b) illustrates the Li_6 structure of Fig. S2(a) with a small change, the top atom of Li_6 moves 0.1 Åin a certain direction, which is hardly displayed in the picture. Obviously, this change could be captured by bars as shown in the blue circles. Four different Li_20 structures with their barcode representation are shown in Fig. S2(c-f). It is obvious that their main differences are β_1 and β_2, which characterizes more geometric features than β_0 and PPI when the number of atoms is increased.

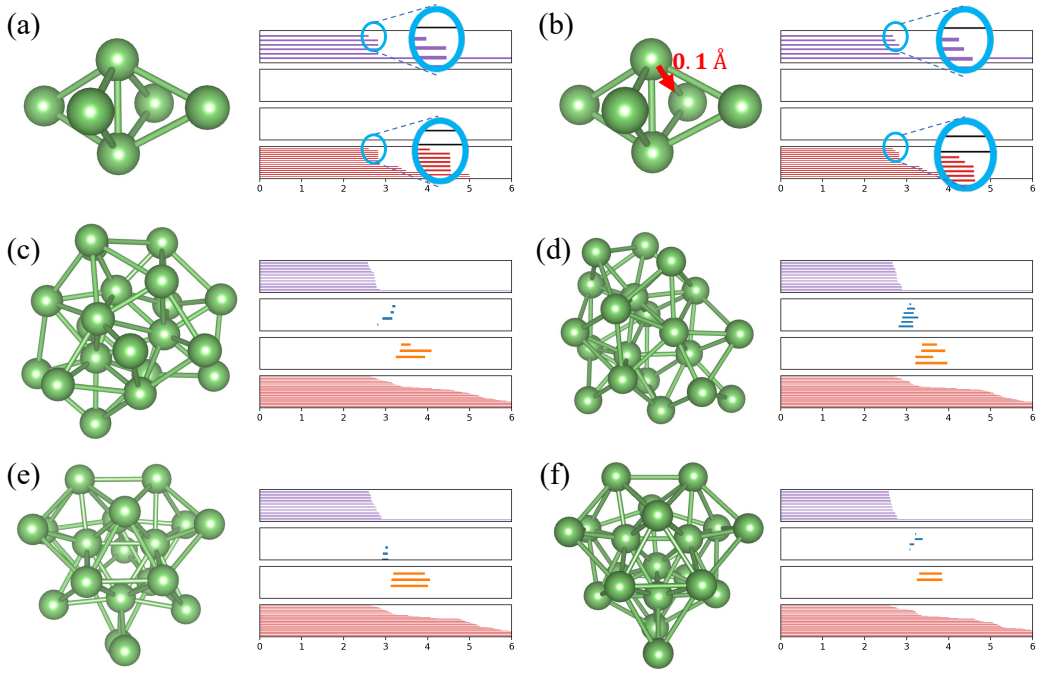

Figure S2: An illustration of barcodes for six clusters. Four panels from top to bottom are β_0, β_1, β_2 and PPI barcodes, respectively. The horizontal axis is the filtration parameter (Å).

Putative globally stable structures of Li_{40} This structure is a perfect 45-atom polyicosahedron with five missing vertex atoms.The polyicosahedral structural packing maintains

a coordination number of 12 for each internal atom. For monoatomic systems, this type of packing can induce bond strain.

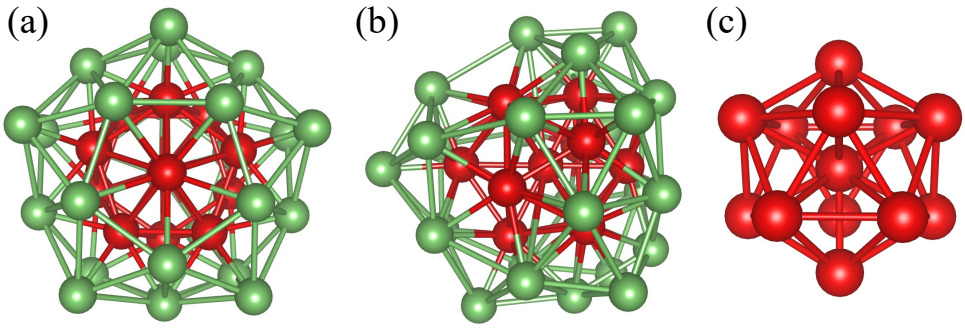

Figure S3: (a) the front view and (b) the side view and (c) internal atoms of the putative global stable structure of Li_{40}.

Electronic supplementary materials All code and data of cluster used in this work were uploaded to Github: https://github.com/ChenXinAtPKU/PHMLBE. The data format are described below. There are 136617 cluster structures used in the work, among them, 130099 structure are $Li_{3\sim10}$ structures. The rest are Li_{20} and Li_{40} structures. In order to make data more feasible and compact, all structure and corresponding binding energy are packed into a file so that one could easily download and extract information from it. Each line of the file corresponds to a cluster. The first number n of the line represents the number of atoms in the cluster, the next $3n$ numbers represent the Cartesian coordinates x, y and z (in Å) of n atoms, respectively. The final number represents the binding energy of the cluster. For example, the cluster data [3.0000 -1.7136 -0.5712 1.7136 0.0000 1.1424 0.0000 1.7136 -0.5712 -1.7136 -0.4024] represents there are three atoms in this cluster, and the first atom locates at (-1.7136, -0.5712, 1.7136), the second atom locates at (0.0000, 1.1424, 0.0000), and the third atom locates at (1.7136, -0.5712, -1.7136). The binding energy of this cluster is -0.4024 eV/atom.

References

(S1) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research* **2011**, *12*, 2825–2830.

(S2) Keras, C. F. GitHub; 2015. 2017.

(S3) Tralie, C.; Saul, N.; Bar-On, R. Ripser. py: A Lean Persistent Homology Library for Python. *J. Open Source Software* **2018**, *3*, 925.

(S4) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

(S5) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(S6) Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120*, 156001.