Supplementary Information

Atomic energies from a convolutional neural network

Xin Chen[†§], *Mathias S. Jørgensen*[§], *Jun Li*[†], *Bjørk Hammer*[§]

[†]Department of Chemistry and Laboratory of Organic Optoelectronics & Molecular Engineering of the Ministry of Education, Tsinghua University, Beijing 100084, China

[§]Interdisciplinary Nanoscience Center (iNANO) and Department of Physics and Astronomy, Aarhus University, Aarhus DK-8000, Denmark

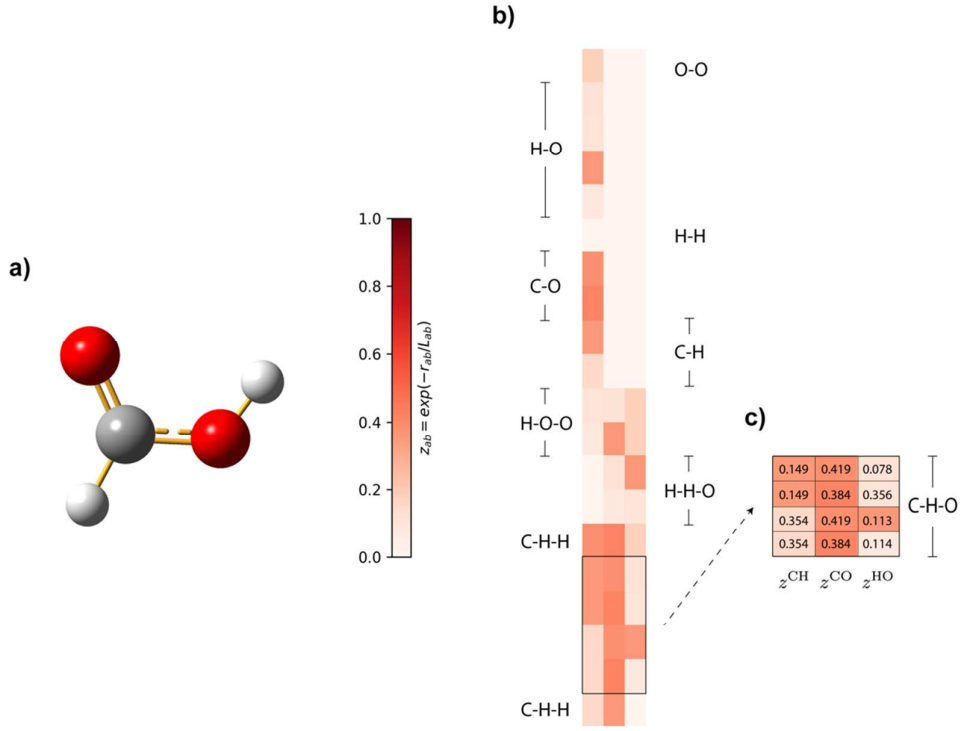

Figure S1. a) The HCOOH molecule and **b)** is the input feature matrix of the 2-body and 3-body terms. Each element of this matrix represents a pairwise interaction. This matrix has 20 ($C_3^5 + C_2^5$) rows and 3 (C_2^3) columns. The corresponding k-body term for the rows are listed. **c)** is the feature matrix of the Carbon-Hydrogen-Oxygen (CNO) terms.

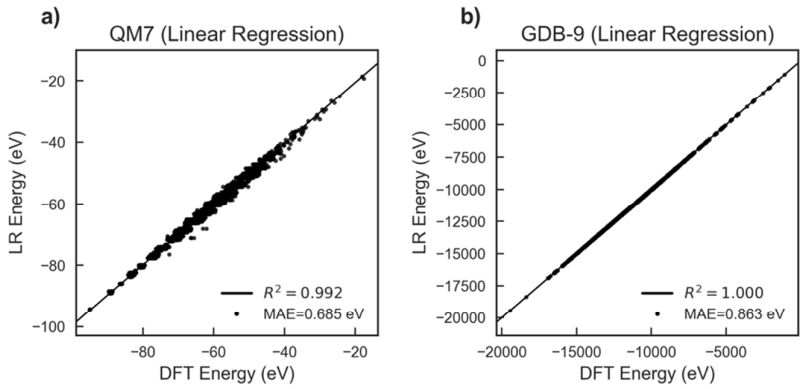

Figure S2. Linear regression (LR) results on QM7 (a) and GDB-9 (b). The LR energy of each molecule is determined by the following linear equation: $E^{LR} = \sum n^A E^A$, where E^{LR} is the predicted energy, n^A represents the number of atom A (C, H, O, N, S for QM7 and C, H, O, N, F for GDB-9) in a molecule and E^A is the learnable parameter for atom type A.

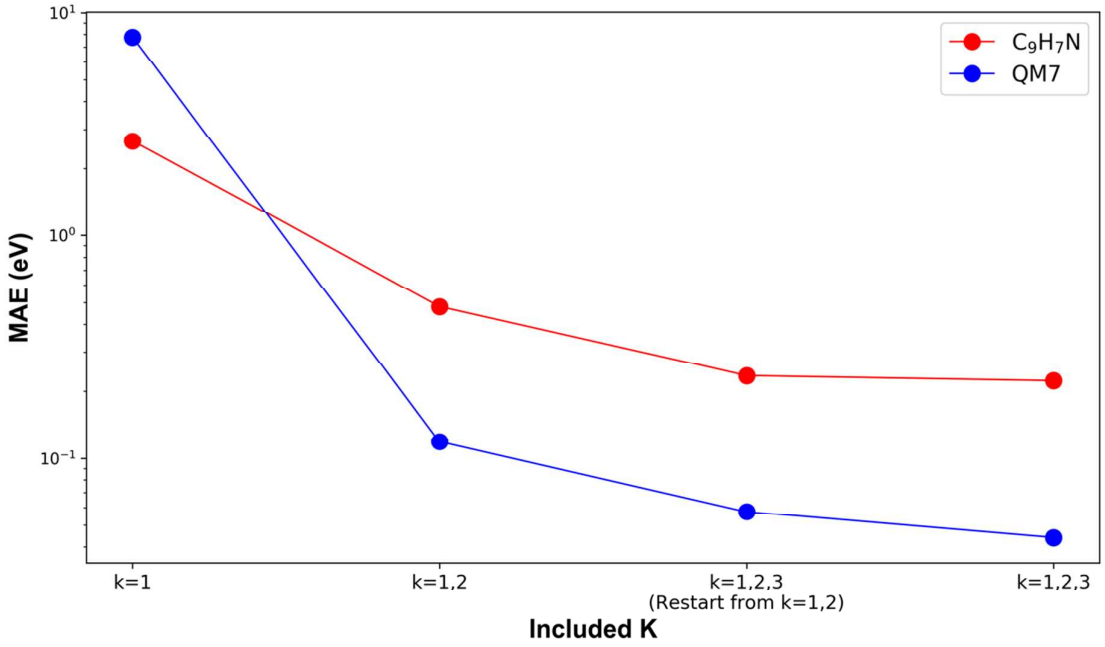

Figure S3. Comparison of the testing mean absolute errors (MAE) of models with different ranked k-body (k=1,2,3) terms on the C_9H_7N (PBE) dataset and QM7 dataset. Including the 3-body terms can significantly reduce the MAE. The training settings are described in Table S2.

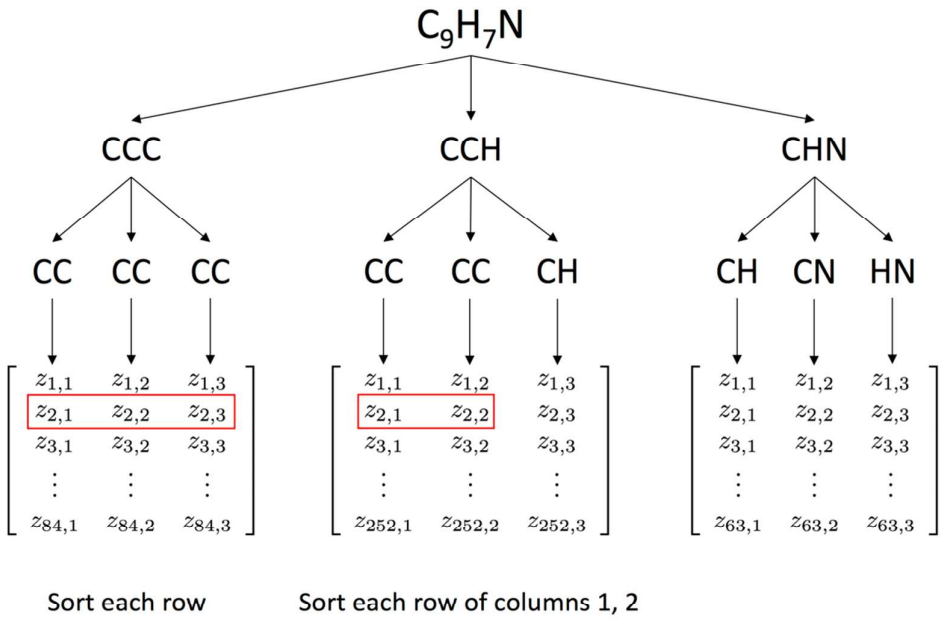

Figure S4. The conditional sorting scheme. For CHN, since each column represents a unique bond (C-C, C-H, C-N), sorting is not needed. For CCH, every row of the first two columns (both represent C-C bonds) shall be sorted and for the CCC every row needs to be sorted.

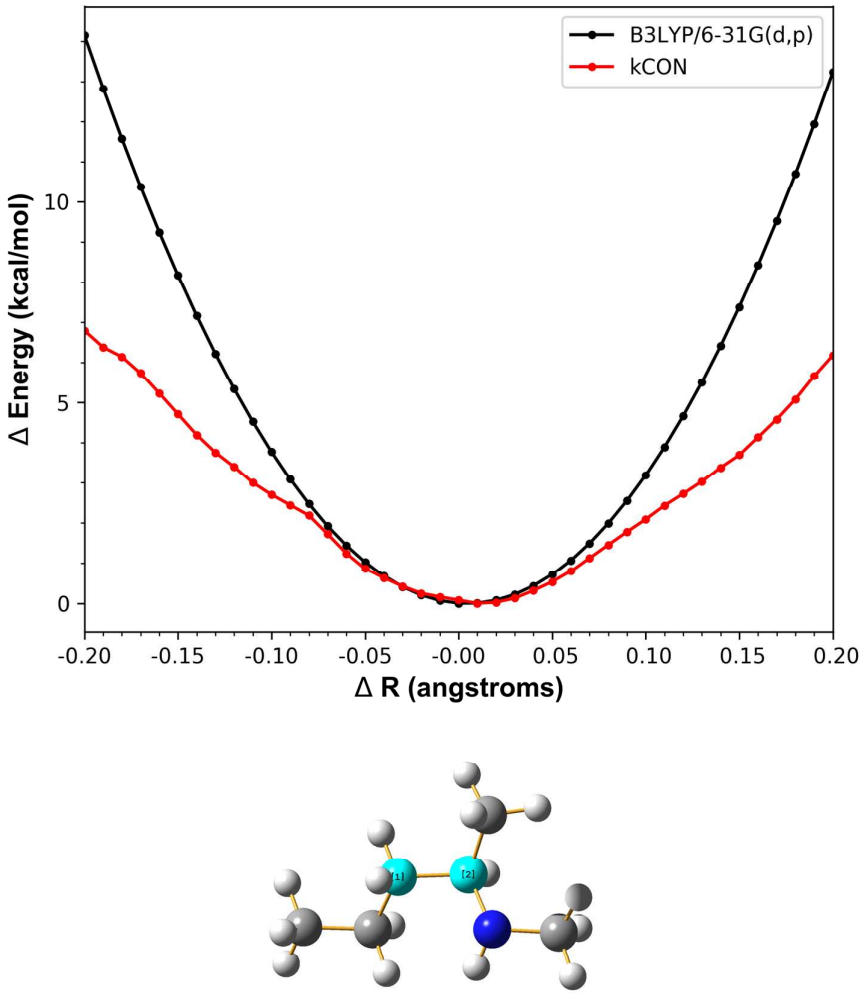

Figure S5. Potential energy surface scan of the C-C bond stretching. The kCON energies are calculated with the QM7 model. The DFT energies are calculated with Gaussian 09, Revision D.01 at the B3LYP/6-31G(d,p) level[1-2].

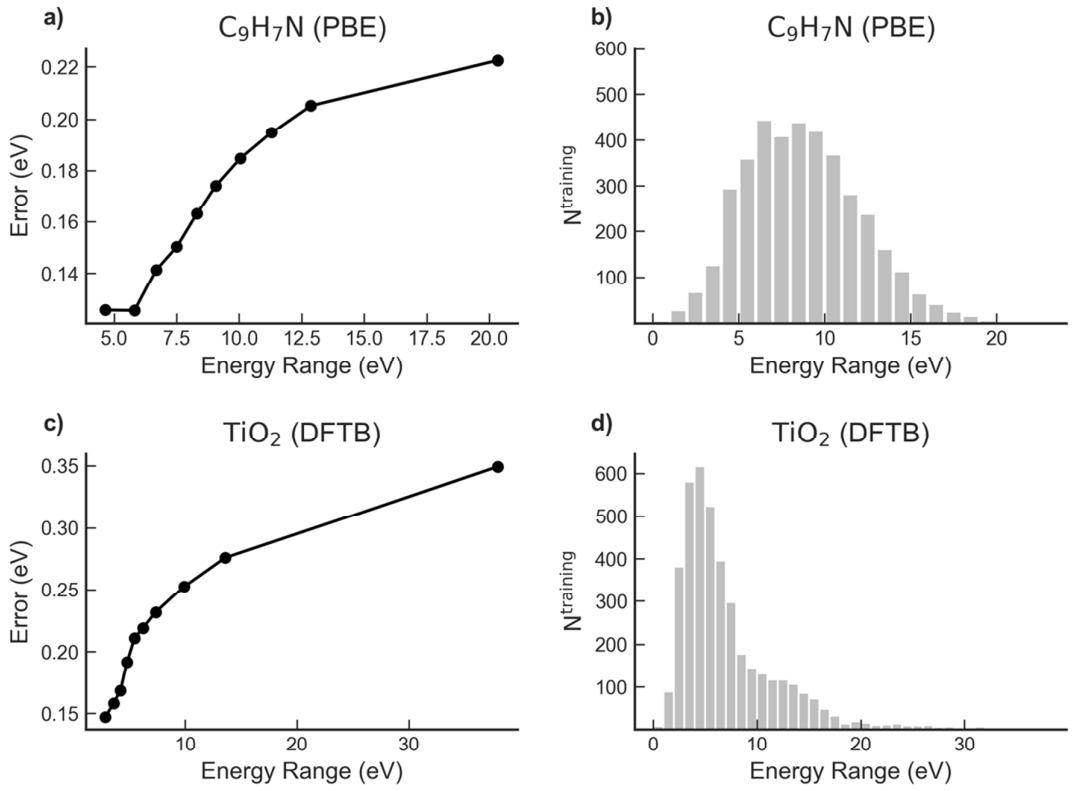

Figure S6. a) and **b)** are the accumulative MAEs on the datasets of C_9H_7N and TiO_2. **c)** and **d)** demonstrate the energy distribution of the C_9H_7N and TiO_2 isomers used for training

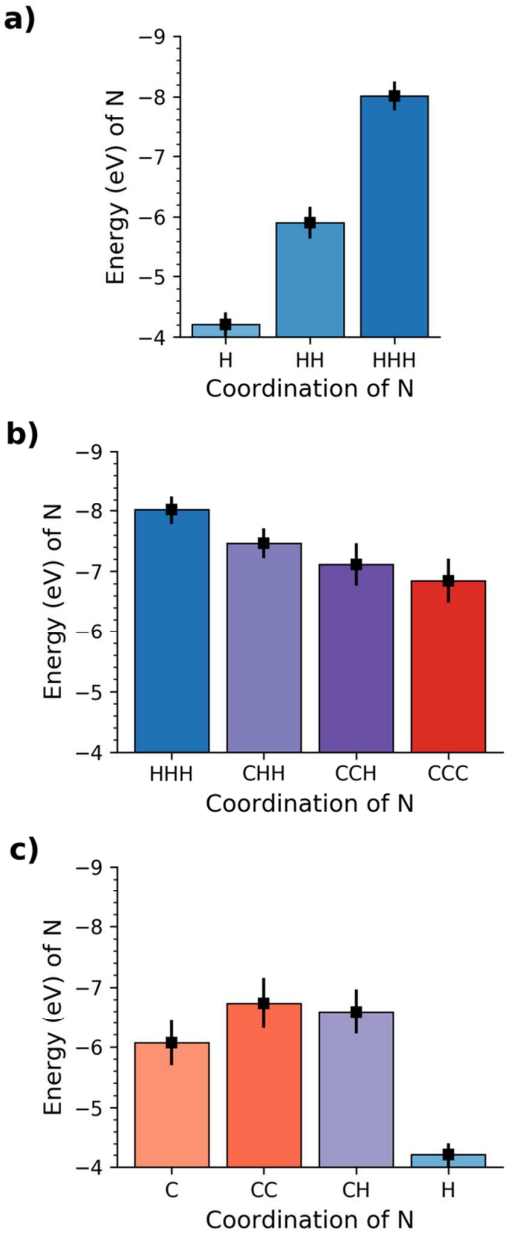

Figure S7. The distribution of the atomic energy of the nitrogen as a function of the coordination in all structures of the C_9H_7N dataset. **a)** showing the nitrogen energy in radical hydrogen species (H, HH) and ammonia (HHH). **b)** comparing the nitrogen energy in primary amine groups (CHH) with other high-coordination species (CCC and CCH). **c)** demonstrating the remaining groups and nitrogen with CC coordination tends to be very stable because CC often involves aromatic compounds similar to the global minimum.

Name	Shape	Number of weights
CCC/Input	[50, 1, 84, 3]	
CCC/Hidden1	[50, 1, 84, 128]	
CCC/Hidden1/weights	[1, 1, 3, 128]	384
CCC/Hidden1/bias	[128]	128
CCC/Hidden2	[50, 1, 84, 128]	
CCC/Hidden2/weights	[1, 1, 128, 64]	8192
CCC/Hidden2/bias	[64]	64
CCC/Hidden3	[50, 1, 84, 64]	
CCC/Hidden3/weights	[1, 1, 64, 32]	2048
CCC/Hidden3/bias	[32]	32
CCC/Hidden4	[50, 1, 84, 32]	
CCC/Output/weights	[1, 1, 32, 1]	32
CCC/Output/bias	[0]	0
CCC/Total		10880

Table S1. The shapes of the layers of the 3-body CNN for interactions of CCC. The batch size is 50. The number of CCC interactions for organic compounds of composition C_9H_7N is $\mathbf{C}_3^9 = 84$ so that the third dimensions, from CCC/Input to CCC/Hidden5 are all 84. The output layer does not have any bias unit.

Dataset		QM7 (DFT)	GDB-9 (DFT)	C_9H_7N (DFT)	TiO_2 (DFTB)	C_9H_7N (DFTB)
Layer Sizes		128,64,32	32,64,64,32	128,64,32	128,64,32	60,80,90,60
Batch Size		50	100	50	50	60
Learning Rate	Initial	0.001	0.0004	0.0008	0.0005	0.0002
	Decay Function	Exponential		Exponential	Exponential	
	Decay Step	5000		25000	25000	
	Decay Rate	0.96		0.90	0.90	

Table S2. The kCON training settings.

The Datasets

The Evolutionary Algorithm (EA)[3-4] implemented in the Atomic Simulation Environment (ASE)[5] is used to generate the datasets of C_9H_7N (DFTB)[6] and anatase TiO_2(001) (DFTB)[6]. The population size N^{pop} is chosen to be 20. The structures, including the initial populations, are optimized with the density functional tight binding (DFTB)[7] theory using DFTB+[8]. For C_9H_7N, the bond parameters are from Gaus et al.[9] and for TiO_2 we use Dolgonos's[10]. For C_9H_7N the structures are optimized in a $25 \times 25 \times 10$ Å unit cell including only forces in the x- and y- directions to produce two-dimensional structures, and for TiO_2 the lattice parameters are $a = 3.74$ Å, $b = 9.47$ Å and only x- and y- directions are periodic. The force threshold for all local optimizations is set to be 0.05 eV/Å/atom. The similarity of any two structures is calculated with the following equation[6]:

$$s_{ab} = \sum_{A-B} \left(\frac{N_A + N_B}{2N} \cdot \frac{\sum_n \left| f^a_{A-B}(n) - f^b_{A-B}(n) \right|}{\frac{1}{2} \sum_n \left| f^a_{A-B}(n) + f^b_{A-B}(n) \right|} \right)$$

where A, B represent types of atoms, N is the total number of atoms and f_{A-B} is the sorted list of interatomic distances of atom pairs of type A-B. Two structures are considered similar if $s_{ab} < 0.02$. Only unique structures are present in the datasets. Both the C_9H_7N (DFTB) and anatase TiO_2(001) (DFTB) datasets have 5000 structures. The C_9H_7N (DFTB) dataset is re-optimized with GPAW[11-12] using PBE[13] in LCAO[14] mode and DZP basis set to obtain a C_9H_7N DFT dataset. The C_9H_7N DFT dataset has 4845 unique structures. The QM7[15-16] and GDB-9[17] datasets were published by Rupp et al. QM7 is a subset of the huge database GDB-13[15], and has 7165 different stable organic molecules of up to 23 atoms (C, H, N, O, S) and the GDB-9 dataset contains 133,885 stable organic molecules of up to 29 atoms (C, H, N, O, F). For all datasets, we randomly select 80% of the total samples for training and 20% for validation.

1. Gaussian 09, Revision D.01, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell; Montgomery, J. K.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian, Inc., Wallingford CT, 2013.
2. Kim, K.; Jordan, K. D. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.* **1994,** *98*, 10089-10094.
3. Vilhelmsen, L. B.; Hammer, B. A genetic algorithm for first principles global structure optimization of supported nano structures. *J. Chem. Phys.* **2014,** *141*, 044711.
4. Vilhelmsen, L. B.; Hammer, B. Identification of the Catalytic Site at the Interface Perimeter of Au Clusters on Rutile TiO2(110). *Acs Catal* **2014,** *4*, 1626-1631.
5. Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dulak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Jensen, P. B.; Kermode, J.; Kitchin, J. R.; Kolsbjerg, E. L.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Maronsson, J. B.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiotz, J.; Schutt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z. H.; Jacobsen, K. W. The atomic simulation environment-a Python library for working with atoms. *J. Phys. Condens. Matter.* **2017,** *29*, 273002.
6. Jorgensen, M. S.; Groves, M. N.; Hammer, B. Combining evolutionary algorithms with clustering toward rational global structure optimization at the atomic scale. *J. Chem. Theroy Comput.* **2017,** *13*, 1486–1493.
7. Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B.* **1998,** *58*, 7260-7268.
8. Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J Phys Chem A* **2007,** *111*, 5678-5684.
9. Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theroy Comput.* **2013,** *9*, 338-354.
10. Dolgonos, G.; Aradi, B.; Moreira, N. H.; Frauenheim, T. An Improved Self-Consistent-Charge Density-Functional Tight-Binding (SCC-DFTB) Set of Parameters for Simulation of Bulk and Molecular Systems Involving Titanium. *J. Chem. Theroy Comput.* **2010,** *6*, 266-278.
11. Mortensen, J. J.; Hansen, L. B.; Jacobsen, K. W. Real-space grid implementation of the projector augmented wave method. *Phys. Rev. B.* **2005,** *71*.
12. Enkovaara, J.; Rostgaard, C.; Mortensen, J. J.; Chen, J.; Dułak, M.; Ferrighi, L.; Gavnholt, J.; Glinsvad, C.; Haikola, V.; Hansen, H. A.; Kristoffersen, H. H.; Kuisma, M.; Larsen, A. H.; Lehtovaara, L.; Ljungberg, M.; Lopez-Acevedo, O.; Moses, P. G.; Ojanen, J.; Olsen, T.; Petzold, V.; Romero, N. A.; Stausholm-Møller, J.; Strange, M.; Tritsaris, G. A.; Vanin, M.; Walter, M.; Hammer, B.; Häkkinen, H.; Madsen, G. K. H.; Nieminen, R. M.; Nørskov, J. K.; Puska, M.; Rantala, T. T.; Schiøtz, J.; Thygesen, K. S.; Jacobsen, K. W. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys. Condens. Matter.* **2010,** *22*, 253202.

13. Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996,** *77*, 3865-3868.
14. Larsen, A. H.; Vanin, M.; Mortensen, J. J.; Thygesen, K. S.; Jacobsen, K. W. Localized atomic basis set in the projector augmented wave method. *Phys. Rev. B.* **2009,** *80*, 195112.
15. Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009,** *131*, 8732–8733.
16. Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012,** *108*, 058301.
17. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **2014,** *1*, 140022.