Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Supporting information: Neural Network Embeddings based Similarity Search Method for Atomistic Systems

Yilin Yang,[a‡] Mingjie Liu,[a‡] and John R. Kitchin[*a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Contents

1 Hyperparamters

1.1 GemNet

Table 1 Hyperparameters for the GemNet model used in our work. More details about the hyperparameters can be found in the OC20 GitHub repository.

Hyperparameter	Value	Hyperparameter	Value
num_spherical	7	num_radial	128
num_blocks	3	emb_size_atom	128
emb_size_edge	128	emb_size_trip	64
emb_size_rbf	16	emb_size_cbf	16
emb_size_bil_trip	64	num_before_skip	1
num_after_skip	2	num_concat	1
num_atom	3	cutoff	6.0
max_neighbors	50	rbf	gaussian
envelope	polynomial	cbf	spherical_harmonics
extensive	true	oft_graph	false
output_init	heOrthogonal	activation	silu
regress_forces	true	direct_forces	true

[a] *Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213; E-mail: jkitchin@andrew.cmu.edu*
‡ These authors contributed equally to this work.

1.2 FAISS Index

Table 2 Hyperparameters for the Faiss IndexIVFPQ method. More details about the hyperparameters can be found in Faiss wiki.

Hyperparameter	Value
Coarse Quantizer	IndexFlatL2
d (dimension)	128
nlists (number of centroids in coarse quantizer)	7000
m (number of subvectors for division)	128
nbits (encoding size for a subvector)	8

1.3 Flare GPR Model

Table 3 Hyperparameters for the Flare GPR model.

Hyperparameter	Value
cutoff radius	3.7
descriptors	two body and three body
kernel	square kernel
length scale	0.5
energy noise	0.005
force noise	0.005
stress noise	0.1

2 Supplementary Examples for ANN Search

2.1 ANN Search Examples in QM9

Here we attach additional examples for searches done in QM9.

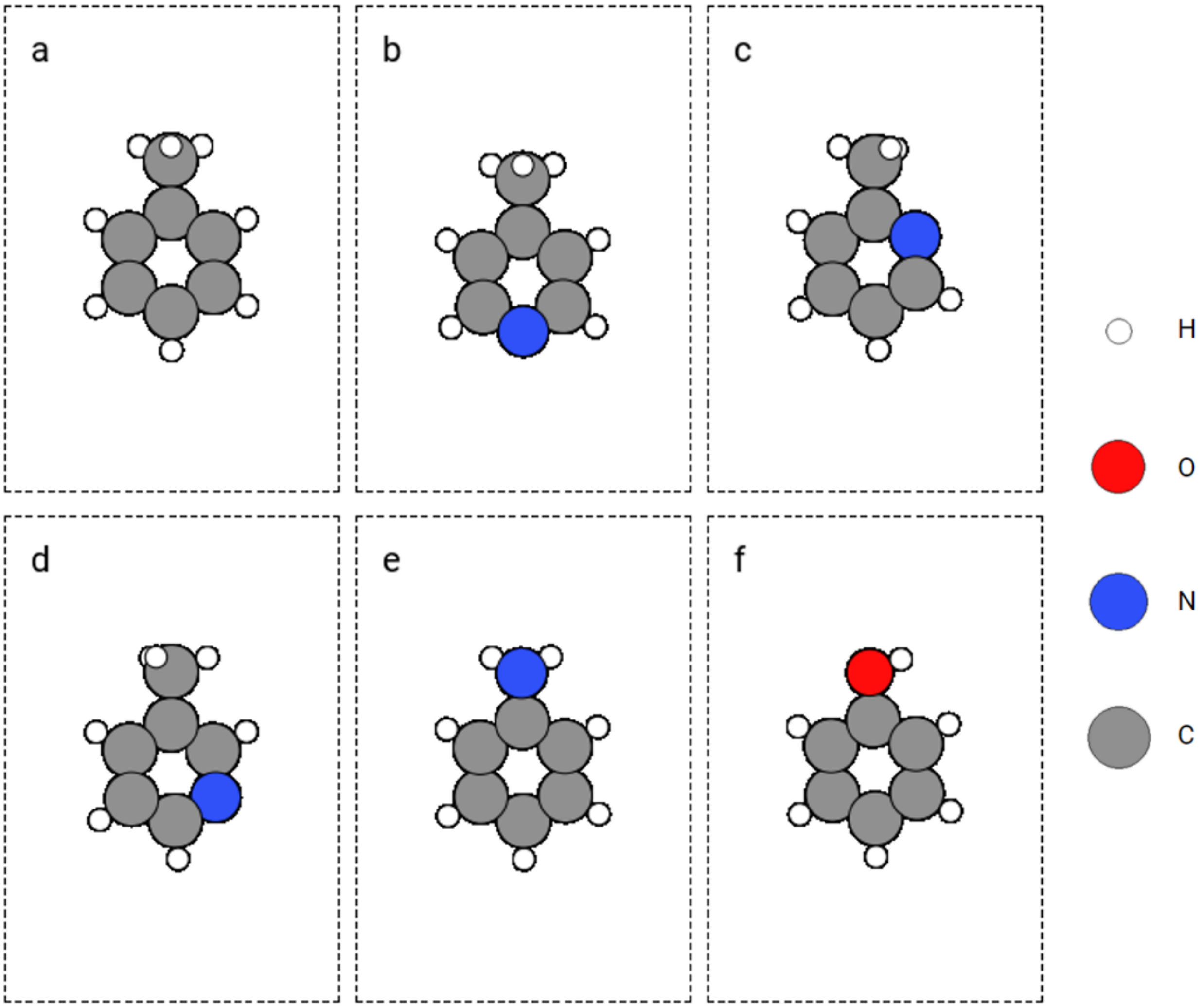

Fig. 1 Similar molecules (b to f) retrieved from querying toluene molecule (a).

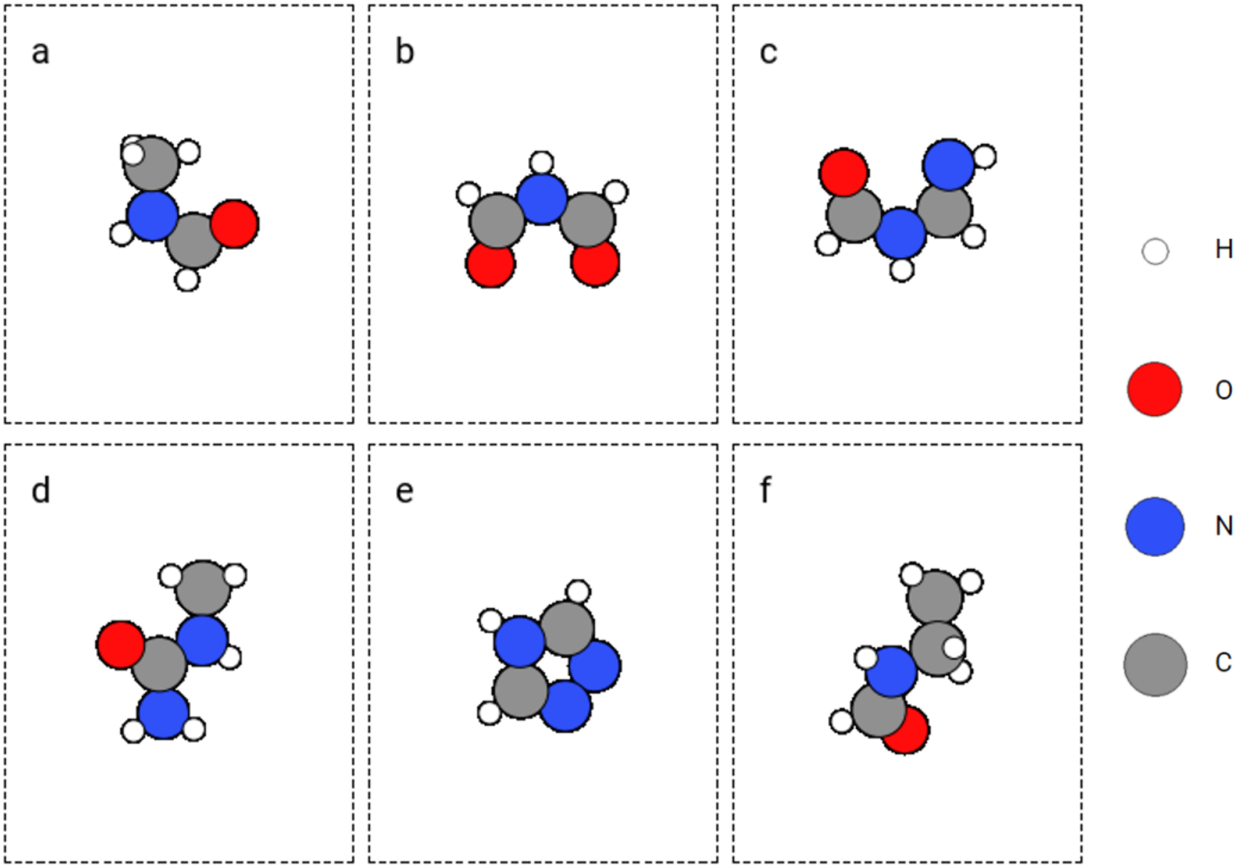

Fig. 2 Similar -N(H)- substructure (b to f) retrieved from querying -N(H)- substructure (a).

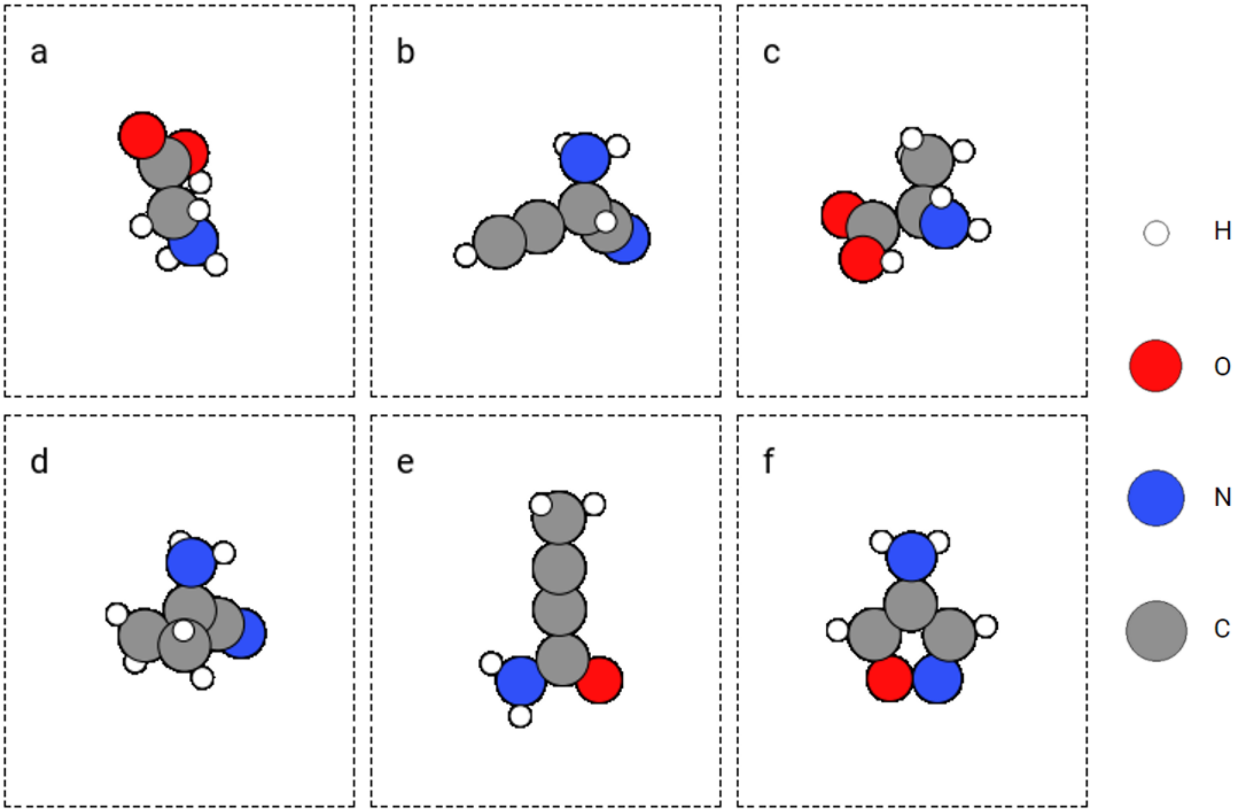

Fig. 3 Similar -NH_2 substructure (b to f) retrieved from querying -NH_2 substructure (a).

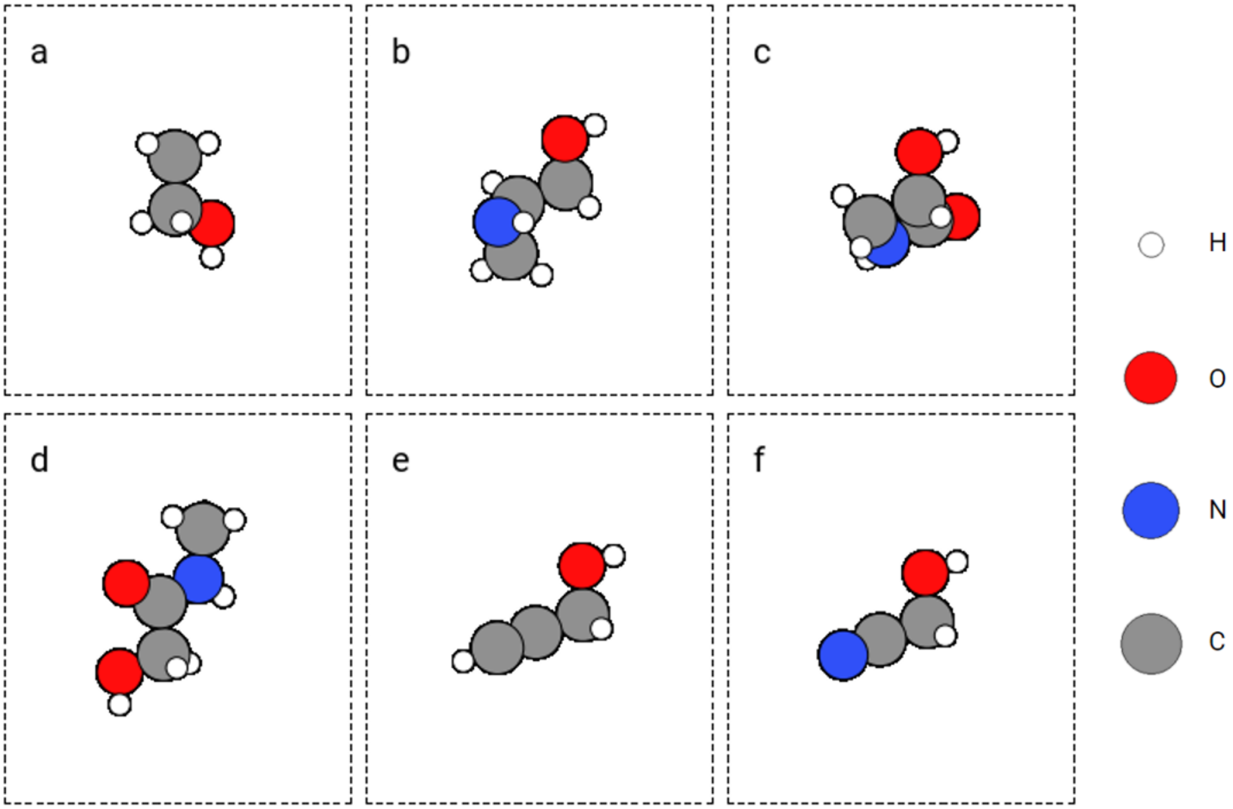

Fig. 4 Similar -OH substructure (b to f) retrieved from querying -OH substructure (a).

2.2 ANN Search Example in Materials Project

Here, we present another example using GemNet embedding and ANN to search for similar bulk environment in the Materials Project database. The query and searched atoms are shown in Figure 5. The query palladium atom (atom 2 in Figure 5 a) and two other palladium atoms (atom 3 and 4) form a hollow site and there is a zinc atom (atom 0) on this hollow site. The searched palladium atoms are all similar to the query palladium atom in terms of the atom arrangement and element type.

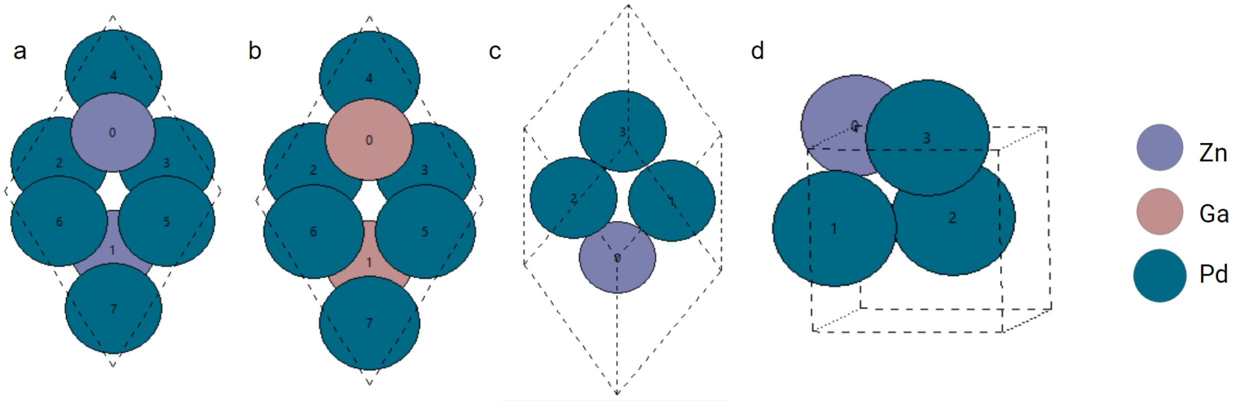

Fig. 5 Top 10 nearest atoms to a palladium query atom in the Materials Project dataset. Atom 2 in figure a is the query atom. Atom 2 to 7 in figure b, atom 1, 2 in figure c, and atom 1, 3 in figure d are the searched atoms.

2.3 Supporting Configurations for the OC20 Case

Here, we attached supporting configurations for the examples shown in the paper including the zoomed-in local configurations for the search results and the randomly selected configurations.

2.3.1 Search for O

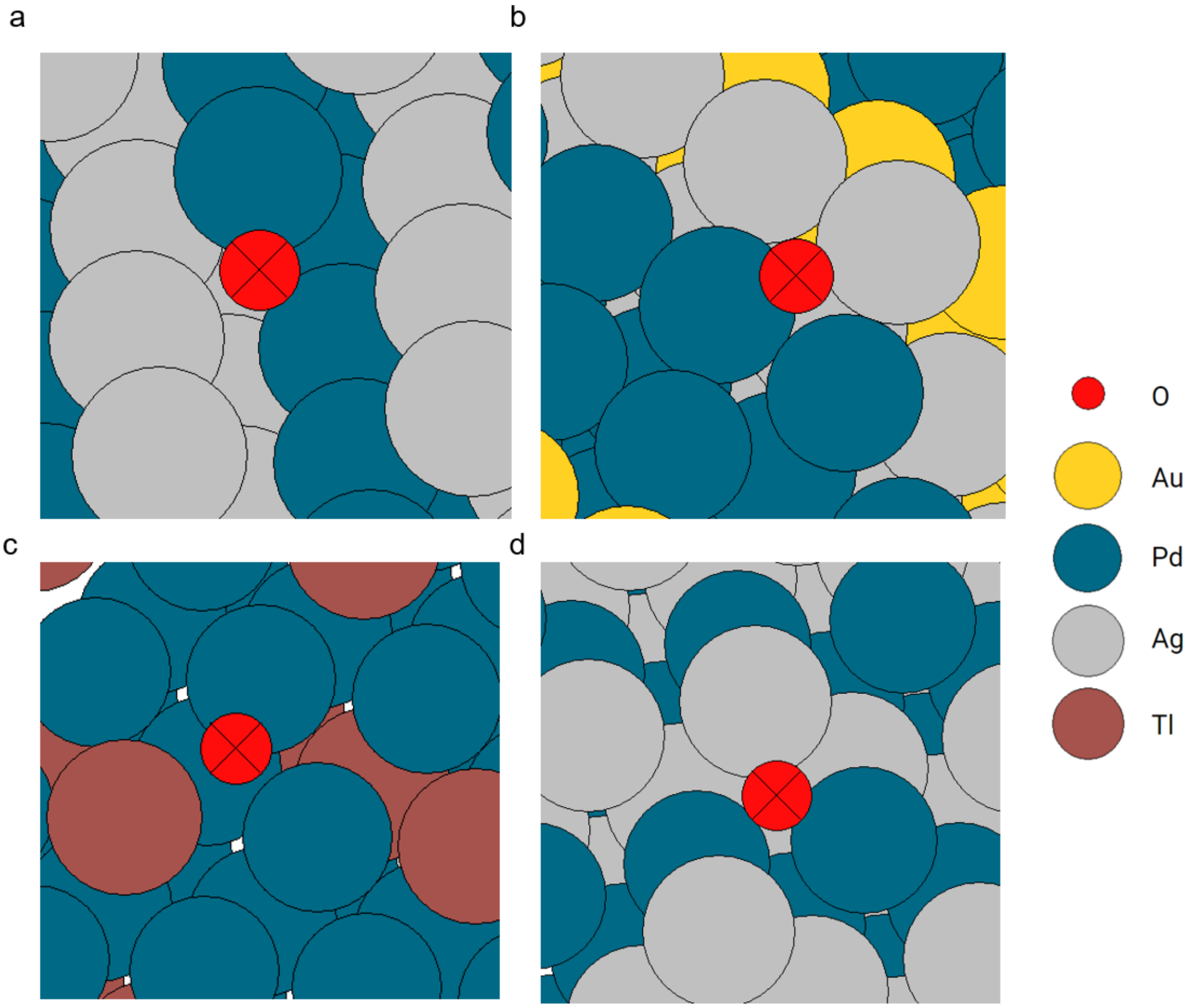

Fig. 6 Zoomed-in local configurations for the oxygen search example.

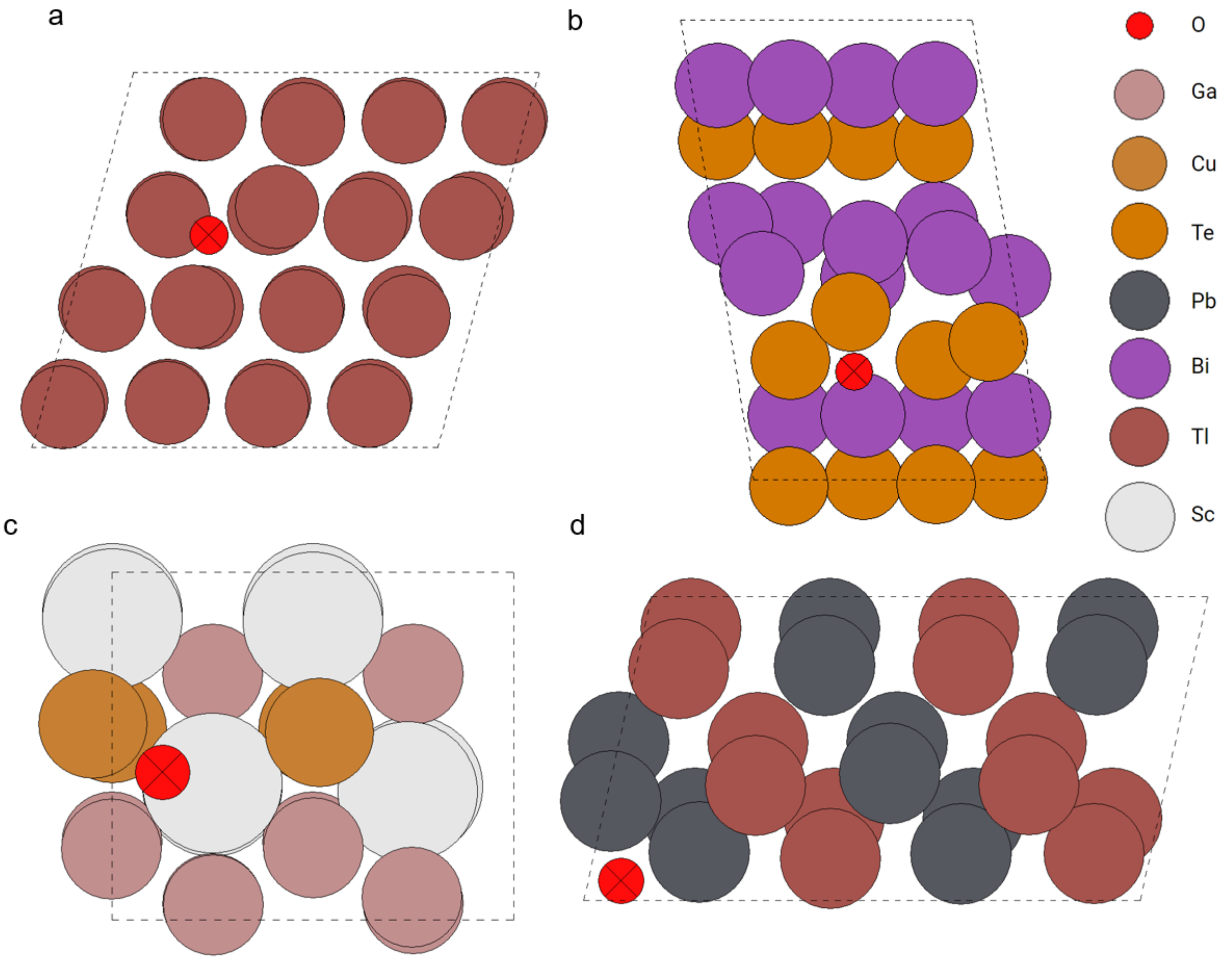

Fig. 7 Four randomly selected oxygen atoms in the OC20 dataset.

2.3.2 Search for C_2H_2

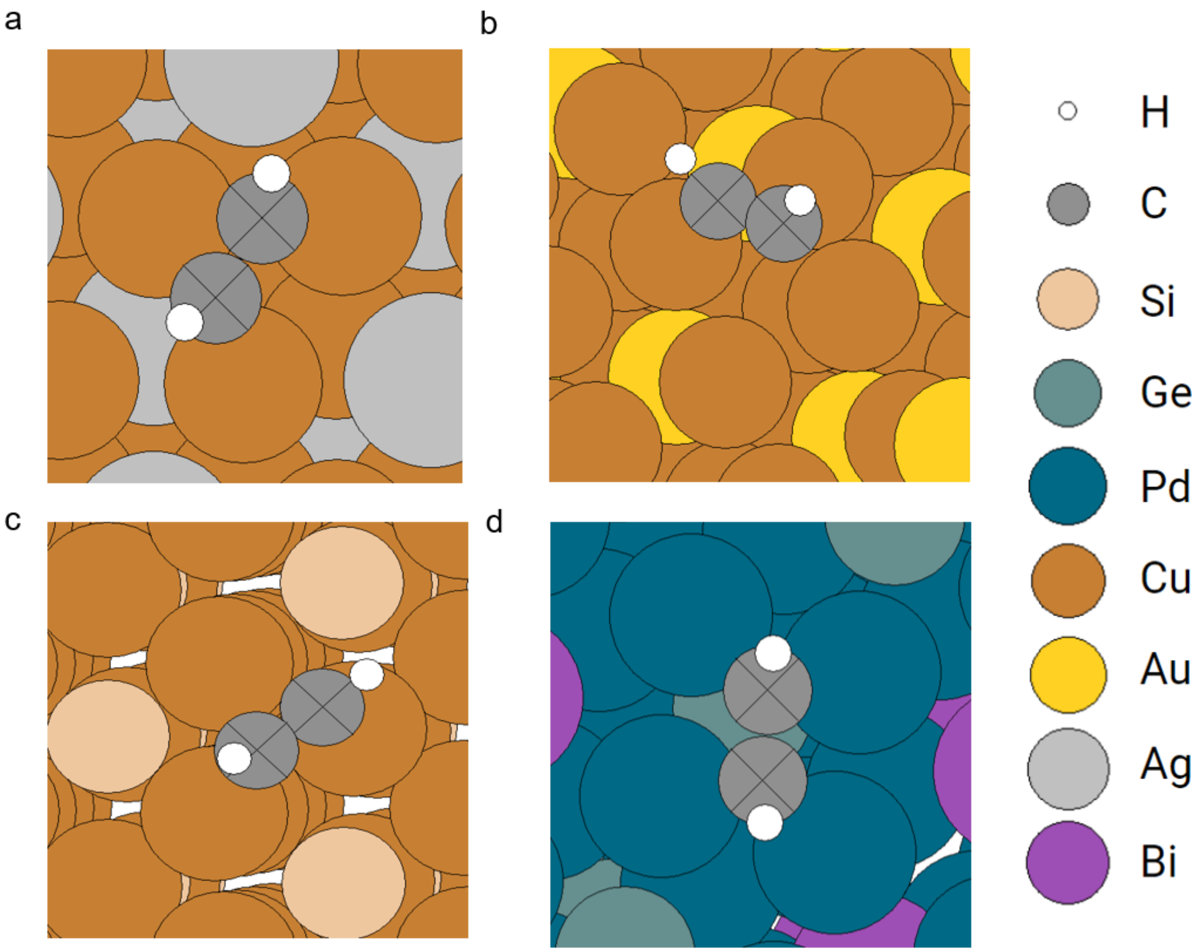

Fig. 8 Zoomed-in local configurations for the acetylene search example.

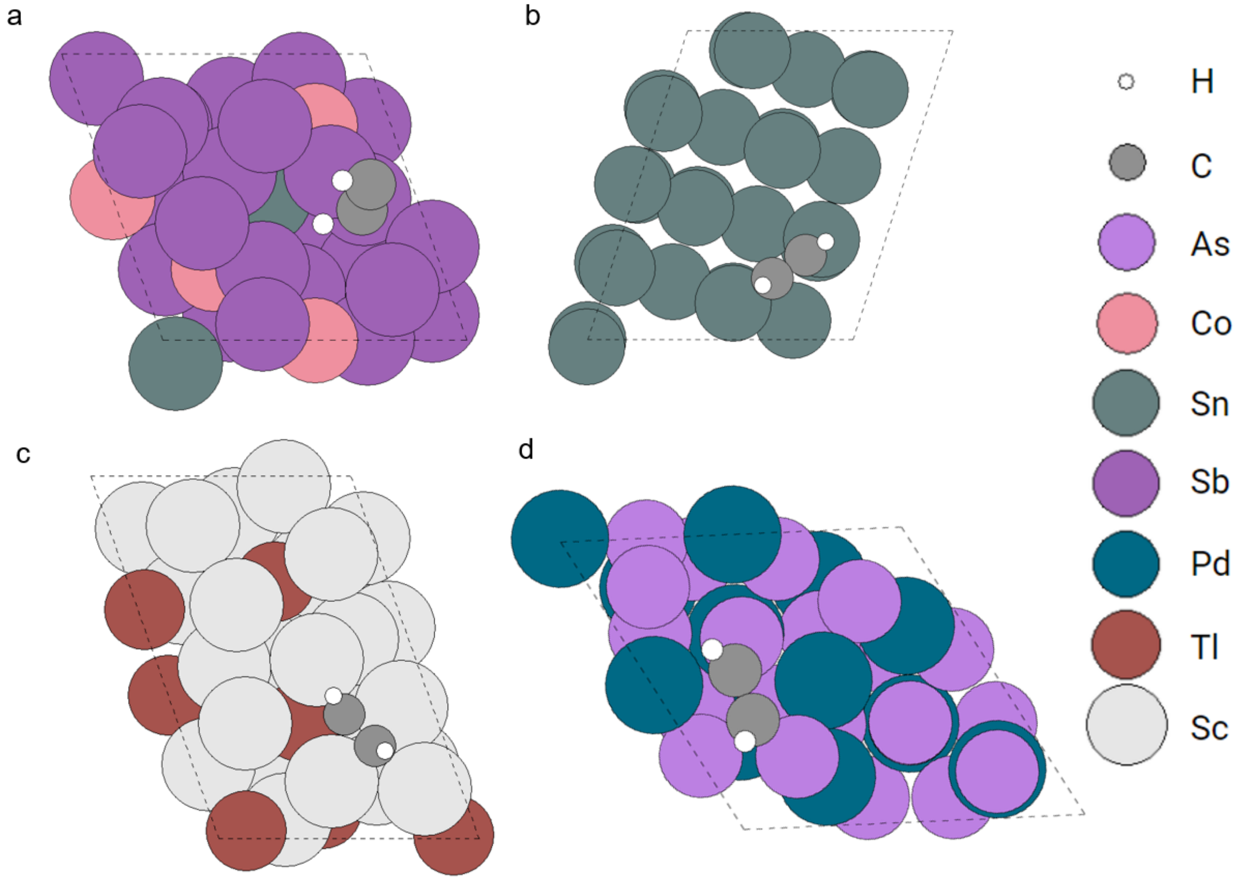

Fig. 9 Four randomly selected acetylene adsorption systems in the OC20 dataset.

3 Data availability

The data generated in this work is too large to completely share directly (it exceeds 80GB in total). Instead, we have created a data archive at `https://doi.org/10.1184/R1/19968323`. This archive contains about 20GB of data that were used in the QM9 and Materials Project examples. The OC20 dataset resulted in over 60 GB of index and other data. This data can be generated following the examples contained in the data archive.