Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting information: Site-Net: Using global self-attention and real-space supercells to capture long-range interactions in crystal structures

Michael Moran, Michael W. Gaultois,* Vladimir V. Gusev, Matthew J. Rosseinsky

June 30, 2023

Construction of roughly cubic supercells with a limited number of atoms

Site-net operates by recursively aggregating sites with their local environment, so it is critical that the local environment of an atomic site is well-behaved at radial distances examined by the attention heads. Accordingly, there is a soft requirement to provide each atomic site in the crystal the largest local environment possible, so a supercell is created to explicitly include longer range interactions with images of the minimal $P1$ unit cell.

Ideally, there should be no edge effects or finite size effects owing to the construction of the supercell; the local environment of any atomic site within the minimal $P1$ unit cell should be equivalent to looking out into the infinite crystal structure. In practice, the size of the model is limited by computational resources, and the distance at which these edge effects begin to contribute is defined here as the self-intersection limit, which is half the shortest distance from any site to its own image in a neighbouring unit cell (or supercell). As an example, an attention head operating on an atomic site at the centre of an orthorhombic unit cell would examine all interactions out to the edge of the unit cell, after which there are no interactions to consider for attention along that direction, resulting in a self-intersection limit equal to half the shortest unit cell parameter. It is possible that overtraining could result from the model learning the edge effects if making use of interactions beyond this range.

To maximize the self-intersection limit, all crystal structures are transformed to the largest possible supercell that is approximately cubic and contains an appropriate number of atoms – fewer than a specified limit. We present below a simple algorithm for this task. To explore how the supercells behave at different size limits, we build supercells of 100, 300, 500, 1000, and 2000 atoms for crystal structures within the Matbench band gap prediction dataset, and examine key features relevant to Site-Net (figs. S1 to S4). For every crystal structure, we change the maximum size limit of the supercell and then examine (a) the resulting number of atoms (fig. S2), (b) the self-intersection limit at which edge-effects will begin to contribute (fig. S3), and (c) the deviation of the supercell from an ideal cube (fig. S4). Importantly, all features become more well-behaved with increasing supercell size, and a greater portion of the dataset becomes available for training owing to the ability to describe crystal structures with more atoms (fig. S1).

We now formally describe the transformation creating an approximately cubic supercell. To begin, let V^P be the matrix where each row is a basis vector of the minimal $P1$ unit cell. We perform Gram–Schmidt orthogonalization procedure on V^P to obtain the decomposition $V^P = RQ$, where R is the upper triangular matrix and Q is an orthogonal matrix, *i.e.*, $QQ^T = I$. Note that normalisation is not performed, thus R has values of 1 on the diagonal entries. Given that $R^{-1}V^P = Q$, the transformation R^{-1} creates an orthogonal unit cell. This is the orthogonalization component of our transformation.

Next, Q is used to compute the scaling component of the supercell transformation, denoted as S. The shortest unit cell basis vector from Q is iteratively incremented until any further incrementation would bring the number of atoms in the unit cell above the specified limit (500 in this work); S is the diagonal matrix that performs the scaling of Q, and the i-th diagonal entry $s_{i,i}$ encodes the number of times the i-th lattice vector should be repeated to form the supercell.

Finally, by combining R^{-1} and S into SR^{-1}, we compute a transformation matrix that converts the minimal $P1$ unit cell V^P into an approximately cubic supercell denoted V^S. Note that R is invertible, as it has only values of 1 on the diagonal, so its determinant is 1, and its inverse is also an upper triangular matrix. The combined transformation SR^{-1} will have non-zero values on the diagonal; however, the matrix entries are likely to be non-integer, which is problematic because only non-singular integer matrices are guaranteed to create a valid supercell. Accordingly, the entries of SR^{-1} are rounded to the nearest non-zero integer to obtain the cubic supercell lattice parameters C^S, where $V^S = SR^{-1}V^P$. This whole procedure ensures the supercell V^S is valid, approximately cubic, and contains less than a specified number of atoms.

Importantly, performing the orthogonalization and scaling together is much more versatile than performing these transformations serially, as rounding matrix elements does not need to be performed until the end of the process. Effectively, performing both operations together means the minimal $P1$ unit cell V^P can be tiled along directions not parallel to the unit cell basis vectors. Accordingly, the final supercell is expected to be a better approximation of a cube, which will have the largest possible self-intersection limit.

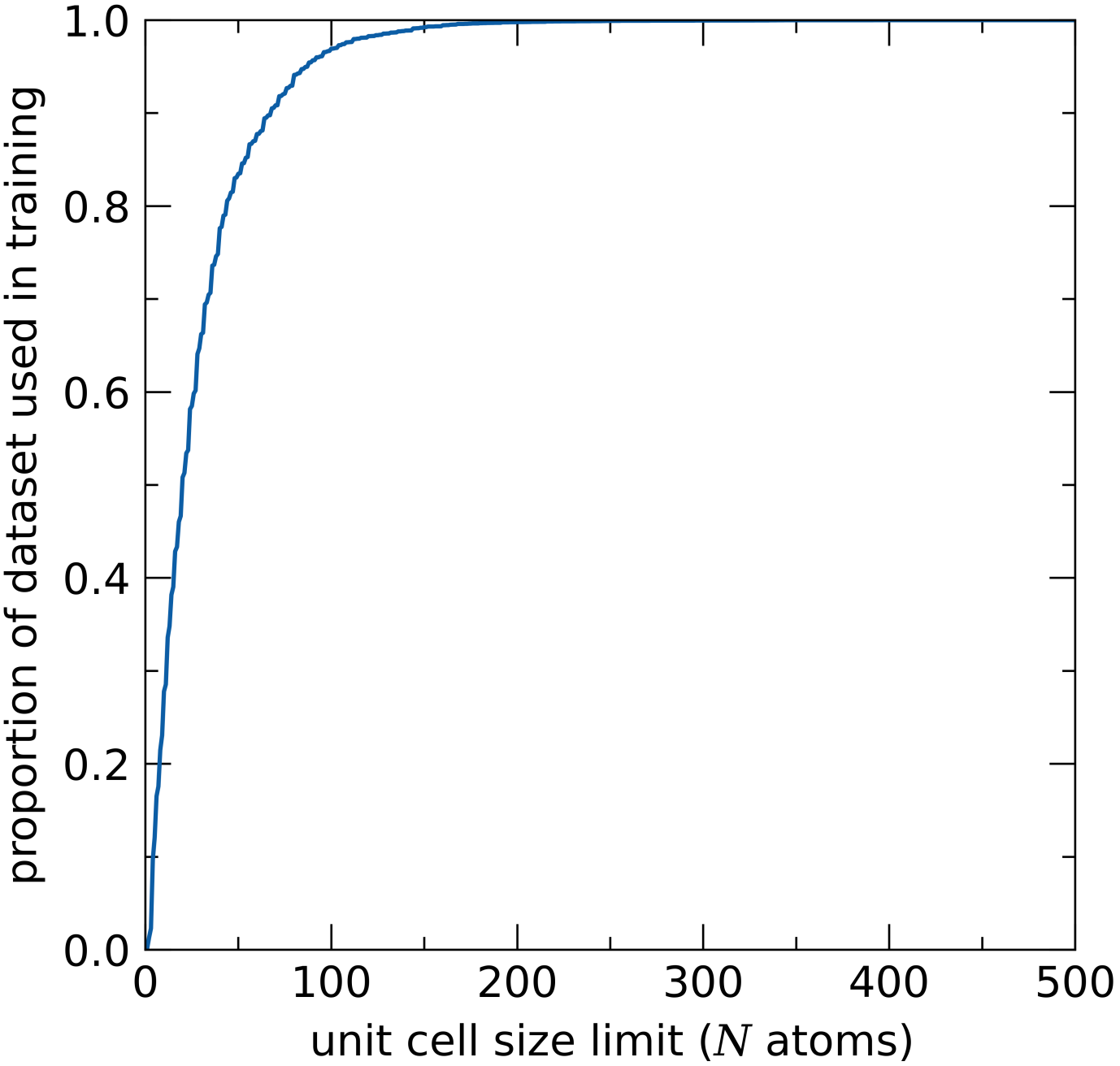

Figure S1 The proportion of crystal structures preserved in the training data of the Matbench band gap prediction task as a function of unit cell size limit (*i.e.*, the maximum number of atoms in the unit cell). Some crystal structures are lost from the training data as a function of the maximum number of atoms that are considered; the cutoff of 500 atoms used in this work allows using all the crystal structures for training, and a cutoff of 100 atoms allows using 97% of the crystal structures for training. For models with a lower cutoff limit, no crystals are excluded from the test dataset when Site-Net is run in inference mode.

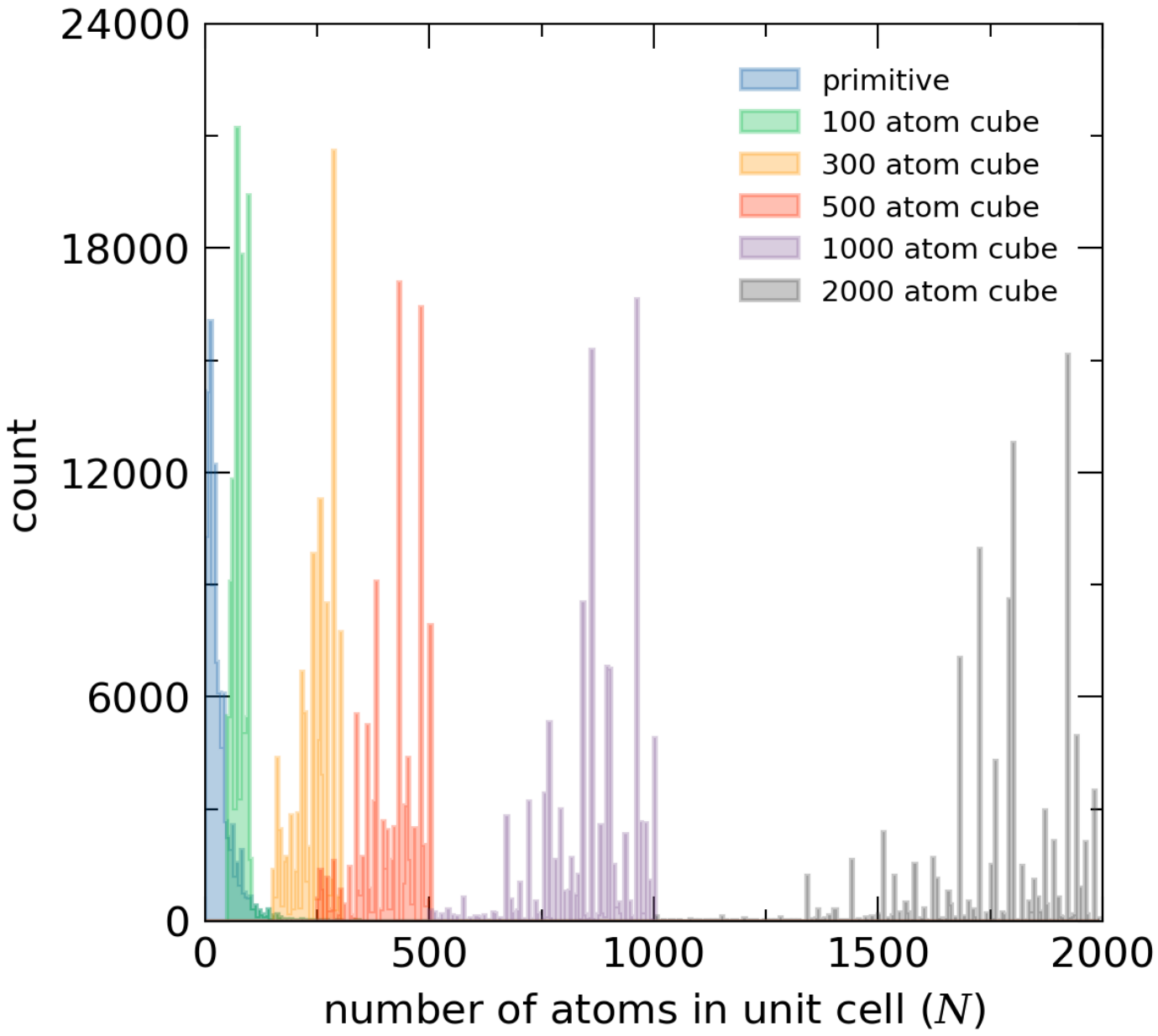

Figure S2 The number of atoms in the unit cells or supercells of crystal structures in the Matbench band gap prediction task for different unit cell size limits (*i.e.*, the maximum number of atoms in the unit cell). For a given supercell size limit, the minimum number of atoms in the unit cell will be $N/2$, and the maximum will be N.

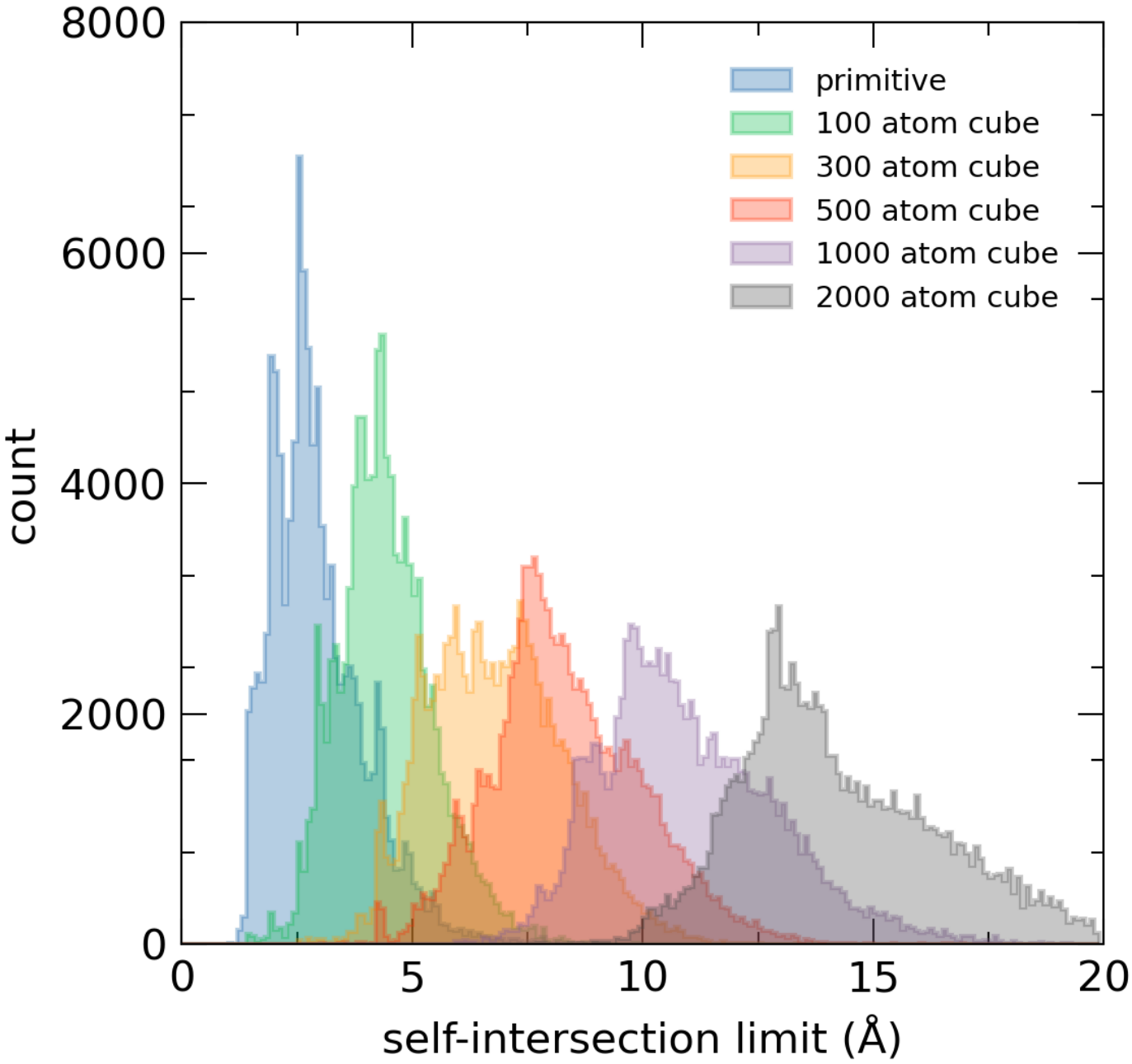

Figure S3 The self-intersection limit for unit cells or supercells of crystal structures in the Matbench band gap prediction task at different unit cell size limits. The self-intersection limit is defined here as half the minimum distance from any atomic site in the unit cell to its mirror image in a neighbouring unit cell. For orthorhombic unit cells, the self-intersection limit is half the shortest unit cell parameter. Radial distances longer than the self-intersection limit will lead to finite size effects, which will become more significant with distance. Larger supercells increase the self-intersection limit, and enable the examination of longer range interactions. Ultimately, the local geometry must be somewhat distorted and this is a trade off that comes the complete attention, geometric correctness is sacrificed to capture the periodicity.

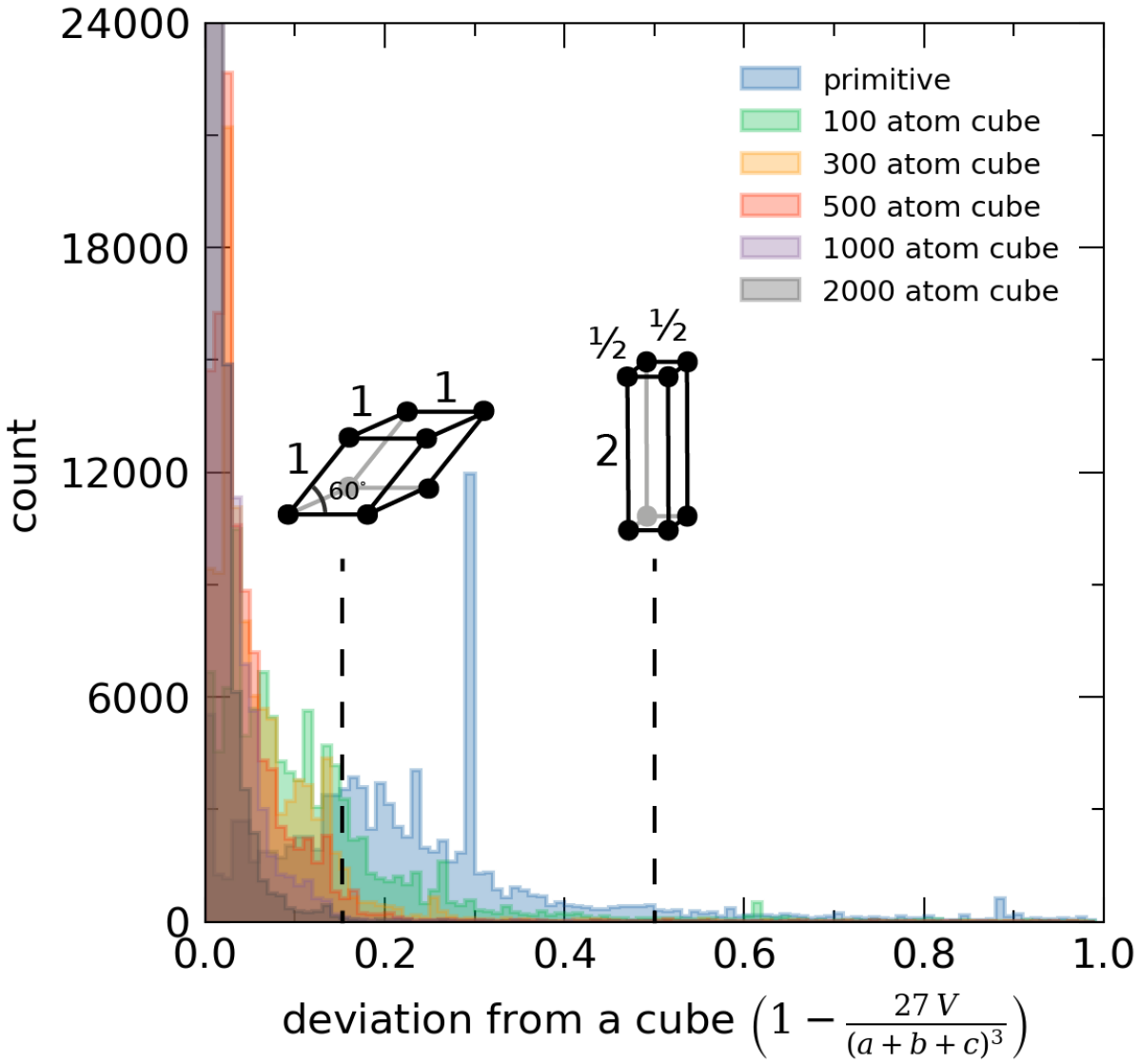

Figure S4 The deviation from a cube for unit cells or supercells of crystal structures in the Matbench band gap prediction task at different unit cell size limits. The deviation is defined here as one minus the ratio between the volume of a unit cell and the analogous cubic unit cell with a side length equal to the average of the unit cell parameters. A more isotropic cell ensures the local environment of an atomic site is well-behaved at longer radial distances, and should minimize finite size effects. With increasing number of atoms allowed in the supercell, the distribution of cell shapes approaches a cube.

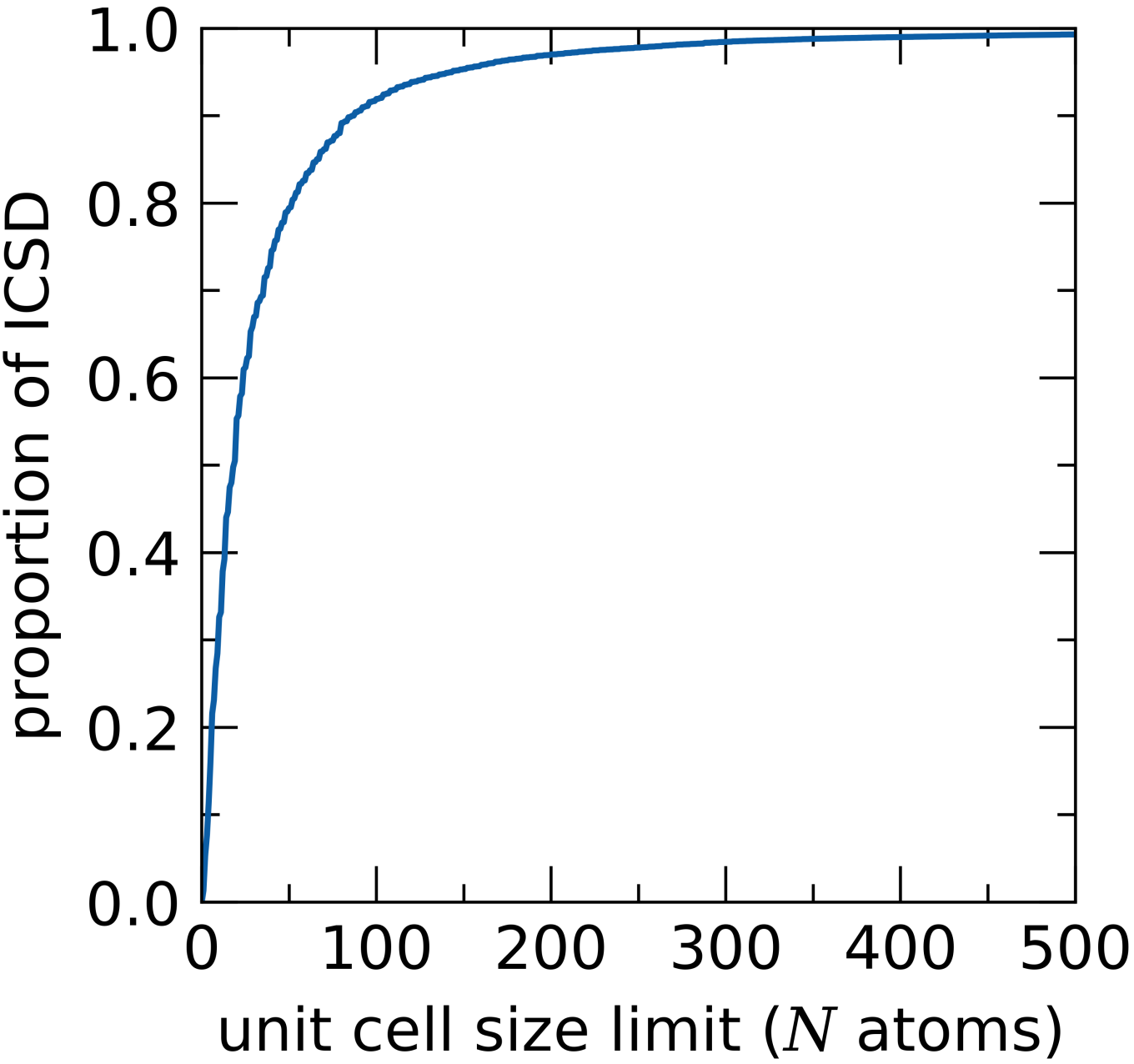

Figure S5 The proportion of all crystal structures in the ICSD as a function of unit cell size limit (*i.e.*, the maximum number of atoms in the unit cell). The cutoff limit of 500 atoms used in this work would allow training on 99% of the crystal structures in the ICSD.

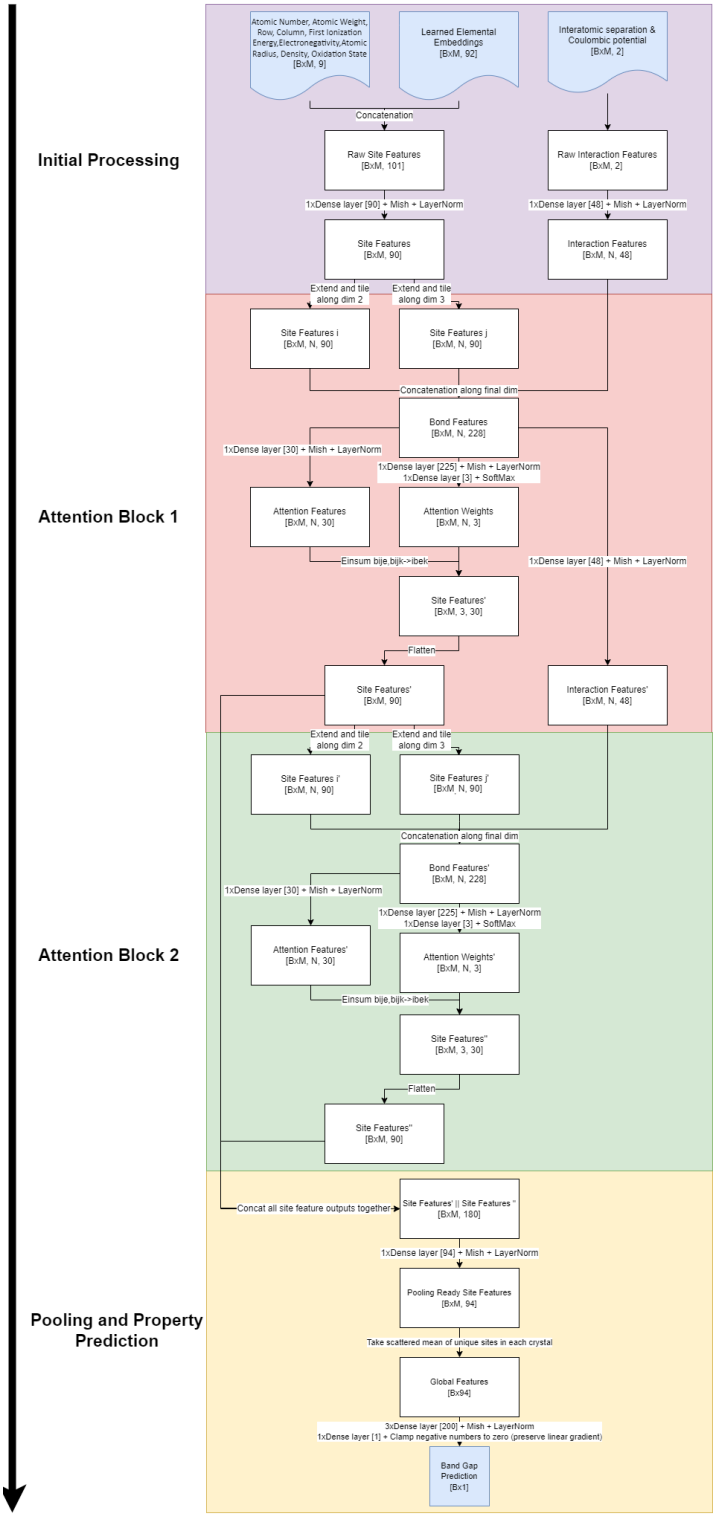

Figure S6 The complete Site-Net model graph for the benchmarking model trained on band gap, with all hyperparameters shown numerically. B, N, and M represent the batch size, the number of atomic sites in the unit cell, and the number of sites in the primitive unit cell respectively. These are dictated by the data structure during training, and can be arbitrarily set during inference. For simplicity, only the output dimension size for neural net layers is shown. Dense layers are always applied to the rightmost rank of the tensor.