# Supplementary Information for "Minimizing the training dataset for efficient crystal structure prediction by graph neural network"

Chuan-Nan Li,[1,2,3,a)] Hanpu Liang,[3,a)] Si-Yuan Xu[3],[3] Xie Zhang,[4,b)] Su-Huai Wei,[3,c)] and Zijing Lin[1,2,d)]

[1)] *Department of Physics, University of Science and Technology of China, Hefei 230026, China*

[2)] *Hefei National Laboratory, University of Science and Technology of China, Hefei 230088, China*

[3)] *Beijing Computational Science Research Center, Beijing, 100193, China*

[4)] *School of Materials Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China*

[a)] These authors contributed equally to this work.

[b)] Electronic mail: xie.zhang@nwpu.edu.cn

[c)] Electronic mail: suhuaiwei@csrc.ac.cn

[d)] Electronic mail: zjlin@ustc.edu.cn

Here we estimate the maximum efficiency improvement of graph neural network (GNN)[1] when applied to the direct-space asymmetric unit (DAU) instead of unit cell. We consider the situation in which $N$ atoms are uniformly dispersed in unit cell. We can determine the value of $N$ by summing the product of the number of atoms in DAU and its corresponding multiplicity, i.e., $N = \sum_i^K \mathrm{num}_i * \mathrm{mul}_i$, where the multiplicity is the number of equivalent positions in the unit cell. Since the inner area of the DAU generally contains most of the atoms ($\mathrm{num}_K$) and has the highest multiplicity ($\mathrm{mul}_K$), we can approximate $N$ as $\mathrm{num}_K * \mathrm{mul}_K$. Meanwhile, the computational cost of GNN is nearly linear depend on the number of input atoms, the maximum efficiency improvement thus can be estimated by the highest multiplicity of each space group. The results are listed in Tables S1–S5.

TABLE S1. Maximum efficiency improvement of GNN applied in 2D materials.

| Space Group | Maximum Improvement | Space Group | Maximum Improvement |
|---|---|---|---|
| 1 (*P1*) | 1 | 10 (*P4*) | 4 |
| 2 (*P211*) | 2 | 11 (*P4mm*) | 8 |
| 3 (*P1m1*) | 2 | 12 (*P4gm*) | 8 |
| 4 (*P1g1*) | 2 | 13 (*P3*) | 3 |
| 5 (*C1m1*) | 4 | 14 (*P3m1*) | 6 |
| 6 (*P2mm*) | 4 | 15 (*P31m*) | 6 |
| 7 (*P2mg*) | 4 | 16 (*P6*) | 6 |
| 8 (*P2gg*) | 4 | 17 (*P6mm*) | 12 |
| 9 (*C2mm*) | 8 | | |

TABLE S2. Maximum efficiency improvement of GNN applied in 3D materials.

| Space Group | Maximum Improvement | Space Group | Maximum Improvement |
|---|---|---|---|
| 1 (*P1*) | 1 | 12 (*C2/m*) | 8 |
| 2 (*P-1*) | 2 | 13 (*P2/c*) | 4 |
| 3 (*P2*) | 2 | 14 (*P21/c*) | 4 |
| 4 (*P21*) | 2 | 15 (*C2/c*) | 8 |
| 5 (*C2*) | 4 | 16 (*P222*) | 4 |
| 6 (*Pm*) | 2 | 17 (*P2221*) | 4 |
| 7 (*Pc*) | 2 | 18 (*P21212*) | 4 |
| 8 (*Cm*) | 4 | 19 (*P212121*) | 4 |
| 9 (*Cc*) | 4 | 20 (*C2221*) | 8 |
| 10 (*P2/m*) | 4 | 21 (*C222*) | 8 |
| 11 (*P21/m*) | 4 | 22 (*F222*) | 16 |

TABLE S3. Maximum efficiency improvement of GNN applied in 3D materials.

| Space Group | Maximum Improvement | Space Group | Maximum Improvement |
|---|---|---|---|
| 23 (*I222*) | 8 | 57 (*Pbcm*) | 8 |
| 24 (*I212121*) | 8 | 58 (*Pnnm*) | 8 |
| 25 (*Pmm2*) | 4 | 59 (*Pmmn*) | 8 |
| 26 (*Pmc21*) | 4 | 60 (*Pbcn*) | 8 |
| 27 (*Pcc2*) | 4 | 61 (*Pbca*) | 8 |
| 28 (*Pma2*) | 4 | 62 (*Pnma*) | 8 |
| 29 (*Pca21*) | 4 | 63 (*Cmcm*) | 16 |
| 30 (*Pnc2*) | 4 | 64 (*Cmca*) | 16 |
| 31 (*Pmn21*) | 4 | 65 (*Cmmm*) | 16 |
| 32 (*Pba2*) | 4 | 66 (*Cccm*) | 16 |
| 33 (*Pna21*) | 4 | 67 (*Cmma*) | 16 |
| 34 (*Pnn2*) | 4 | 68 (*Ccca*) | 16 |
| 35 (*Cmm2*) | 8 | 69 (*Fmmm*) | 32 |
| 36 (*Cmc21*) | 8 | 70 (*Fddd*) | 32 |
| 37 (*Ccc2*) | 8 | 71 (*Immm*) | 16 |
| 38 (*Amm2*) | 8 | 72 (*Ibam*) | 16 |
| 39 (*Abm2*) | 8 | 73 (*Ibca*) | 16 |
| 40 (*Ama2*) | 8 | 74 (*Imma*) | 16 |
| 41 (*Aba2*) | 8 | 75 (*P4*) | 4 |
| 42 (*Fmm2*) | 16 | 76 (*P41*) | 4 |
| 43 (*Fdd2*) | 16 | 77 (*P42*) | 4 |
| 44 (*Imm2*) | 8 | 78 (*P43*) | 4 |
| 45 (*Iba2*) | 8 | 79 (*I4*) | 8 |
| 46 (*Ima2*) | 8 | 80 (*I41*) | 8 |
| 47 (*Pmmm*) | 8 | 81 (*P-4*) | 4 |
| 48 (*Pnnn*) | 8 | 82 (*I-4*) | 8 |
| 49 (*Pccm*) | 8 | 83 (*P4/m*) | 8 |
| 50 (*Pban*) | 8 | 84 (*P42/m*) | 8 |
| 51 (*Pmma*) | 8 | 85 (*P4/n*) | 8 |
| 52 (*Pnna*) | 8 | 86 (*P42/n*) | 8 |
| 53 (*Pmna*) | 8 | 87 (*I4/m*) | 16 |
| 54 (*Pcca*) | 8 | 88 (*I41/a*) | 16 |
| 55 (*Pbam*) | 8 | 89 (*P422*) | 8 |
| 56 (*Pccn*) | 8 | 90 (*P4212*) | 8 |

TABLE S4. Maximum efficiency improvement of GNN applied in 3D materials.

| Space Group | Maximum Improvement | Space Group | Maximum Improvement |
|---|---|---|---|
| 91 (*P4122*) | 8 | 125 (*P4/nbm*) | 16 |
| 92 (*P41212*) | 8 | 126 (*P4/nnc*) | 16 |
| 93 (*P4222*) | 8 | 127 (*P4/mbm*) | 16 |
| 94 (*P42212*) | 8 | 128 (*P4/mnc*) | 16 |
| 95 (*P4322*) | 8 | 129 (*P4/nmm*) | 16 |
| 96 (*P43212*) | 8 | 130 (*P4/ncc*) | 16 |
| 97 (*I422*) | 16 | 131 (*P42/mmc*) | 16 |
| 98 (*I4122*) | 16 | 132 (*P42/mcm*) | 16 |
| 99 (*P4mm*) | 8 | 133 (*P42/nbc*) | 16 |
| 100 (*P4bm*) | 8 | 134 (*P42/nnm*) | 16 |
| 101 (*P42cm*) | 8 | 135 (*P42/mbc*) | 16 |
| 102 (*P42nm*) | 8 | 136 (*P42/mnm*) | 16 |
| 103 (*P4cc*) | 8 | 137 (*P42/nmc*) | 16 |
| 104 (*P4nc*) | 8 | 138 (*P42/ncm*) | 16 |
| 105 (*P42mc*) | 8 | 139 (*I4/mmm*) | 32 |
| 106 (*P42bc*) | 8 | 140 (*I4/mcm*) | 32 |
| 107 (*I4mm*) | 16 | 141 (*I41/amd*) | 32 |
| 108 (*I4cm*) | 16 | 142 (*I41/acd*) | 32 |
| 109 (*I41md*) | 16 | 143 (*P3*) | 3 |
| 110 (*I41cd*) | 16 | 144 (*P31*) | 3 |
| 111 (*P-42m*) | 8 | 145 (*P32*) | 3 |
| 112 (*P-42c*) | 8 | 146 (*R3*) | 9 |
| 113 (*P-421m*) | 8 | 147 (*P-3*) | 6 |
| 114 (*P-421c*) | 8 | 148 (*R-3*) | 18 |
| 115 (*P-4m2*) | 8 | 149 (*P312*) | 6 |
| 116 (*P-4c2*) | 8 | 150 (*P321*) | 6 |
| 117 (*P-4b2*) | 8 | 151 (*P3112*) | 6 |
| 118 (*P-4n2*) | 8 | 152 (*P3121*) | 6 |
| 119 (*I-4m2*) | 16 | 153 (*P3212*) | 6 |
| 120 (*I-4c2*) | 16 | 154 (*P3221*) | 6 |
| 121 (*I-42m*) | 16 | 155 (*R32*) | 18 |
| 122 (*I-42d*) | 16 | 156 (*P3m1*) | 6 |
| 123 (*P4/mmm*) | 16 | 157 (*P31m*) | 6 |
| 124 (*P4/mcc*) | 16 | 158 (*P3c1*) | 6 |

TABLE S5. Maximum efficiency improvement of GNN applied in 3D materials.

| Space Group | Maximum Improvement | Space Group | Maximum Improvement |
|---|---|---|---|
| 159 (*P31c*) | 6 | 195 (*P23*) | 12 |
| 160 (*R3m*) | 18 | 196 (*F23*) | 48 |
| 161 (*R3c*) | 18 | 197 (*I23*) | 24 |
| 162 (*P-31m*) | 12 | 198 (*P213*) | 12 |
| 163 (*P-31c*) | 12 | 199 (*I213*) | 24 |
| 164 (*P-3m1*) | 12 | 200 (*Pm-3*) | 24 |
| 165 (*P-3c1*) | 12 | 201 (*Pn-3*) | 24 |
| 166 (*R-3m*) | 36 | 202 (*Fm-3*) | 96 |
| 167 (*R-3c*) | 36 | 203 (*Fd-3*) | 96 |
| 168 (*P6*) | 6 | 204 (*Im-3*) | 48 |
| 169 (*P61*) | 6 | 205 (*Pa-3*) | 24 |
| 170 (*P65*) | 6 | 206 (*Ia-3*) | 48 |
| 171 (*P62*) | 6 | 207 (*P432*) | 24 |
| 172 (*P64*) | 6 | 208 (*P4232*) | 24 |
| 173 (*P63*) | 6 | 209 (*F432*) | 96 |
| 174 (*P-6*) | 6 | 210 (*F4132*) | 96 |
| 175 (*P6/m*) | 12 | 211 (*I432*) | 48 |
| 176 (*P63/m*) | 12 | 212 (*P4332*) | 24 |
| 177 (*P622*) | 12 | 213 (*P4132*) | 24 |
| 178 (*P6122*) | 12 | 214 (*I4132*) | 48 |
| 179 (*P6522*) | 12 | 215 (*P-43m*) | 24 |
| 180 (*P6222*) | 12 | 216 (*F-43m*) | 96 |
| 181 (*P6422*) | 12 | 217 (*I-43m*) | 48 |
| 182 (*P6322*) | 12 | 218 (*P-43n*) | 24 |
| 183 (*P6mm*) | 12 | 219 (*F-43c*) | 96 |
| 184 (*P6cc*) | 12 | 220 (*I-43d*) | 48 |
| 185 (*P63cm*) | 12 | 221 (*Pm-3m*) | 48 |
| 186 (*P63mc*) | 12 | 222 (*Pn-3n*) | 48 |
| 187 (*P-6m2*) | 12 | 223 (*Pm-3n*) | 48 |
| 188 (*P-6c2*) | 12 | 224 (*Pn-3m*) | 48 |
| 189 (*P-62m*) | 12 | 225 (*Fm-3m*) | 192 |
| 190 (*P-62c*) | 12 | 226 (*Fm-3c*) | 192 |
| 191 (*P6/mmm*) | 24 | 227 (*Fd-3m*) | 192 |
| 192 (*P6/mcc*) | 24 | 228 (*Fd-3c*) | 192 |
| 193 (*P63/mcm*) | 24 | 229 (*Im-3m*) | 96 |
| 194 (*P63/mmc*) | 24 | 230 (*Ia-3d*) | 96 |

TABLE S6. The structural validity and average energy of initial structures of four different methods (ML-PSO, ML-DFT, DFT-GA, and SCCOP) on 27 typical compounds. Each method has repeated 5 times to generate initial samples. The solid line means that the average energy is greater than zero. The DAUGNN of SCCOP is randomly initialized.

| | | Structural Validity (%) | | | Average Energy (eV/atom) | | |
|---|---|---|---|---|---|---|---|
| Compounds | Dimension | ML-PSO(DFT) | DFT-GA | SCCOP | ML-PSO(DFT) | DFT-GA | SCCOP |
| BN | 2 | 89.5 | 100.0 | 100.0 | -5.0038 | -0.3862 | -4.5068 |
| C | 2 | 67.2 | 100.0 | 100.0 | -4.2315 | -0.6811 | -5.0535 |
| FeSe | 2 | 93.9 | 100.0 | 100.0 | -0.2252 | -0.0009 | -2.5340 |
| $MoTe_2$ | 2 | 85.3 | 100.0 | 100.0 | -0.0057 | 0.0000 | -2.6888 |
| $Mo_2C$ | 2 | 100.0 | 100.0 | 100.0 | 0.0000 | -0.0360 | -3.7200 |
| SiC | 2 | 65.4 | 100.0 | 100.0 | -1.8279 | -0.0580 | -4.1524 |
| $Ti_2C$ | 2 | 100.0 | 100.0 | 100.0 | 0.0000 | -0.1611 | -4.0651 |
| BaO | 3 | 100.0 | 100.0 | 100.0 | -2.5802 | -2.0223 | -3.5338 |
| BeS | 3 | 100.0 | 100.0 | 100.0 | -2.7016 | -1.3387 | -3.1908 |
| C | 3 | 100.0 | 100.0 | 100.0 | -5.0743 | -4.2250 | -5.2224 |
| CaO | 3 | 100.0 | 100.0 | 100.0 | -2.7499 | -2.3678 | -3.6545 |
| $CaTiO_3$ | 3 | 100.0 | 100.0 | 100.0 | -4.7407 | -3.4935 | -4.5581 |
| CdO | 3 | 99.6 | 100.0 | 100.0 | -1.7696 | -0.8799 | -2.3478 |
| CdS | 3 | 100.0 | 100.0 | 100.0 | -1.3625 | -0.6330 | -1.9324 |
| CsCl | 3 | 100.0 | 100.0 | 100.0 | -1.4466 | -1.1234 | -2.4130 |
| GaAs | 3 | 99.7 | 100.0 | 100.0 | -2.3411 | -1.2963 | -3.2288 |
| KF | 3 | 100.0 | 100.0 | 100.0 | -2.0759 | -1.8304 | -2.7469 |
| LiF | 3 | 100.0 | 100.0 | 100.0 | -3.0137 | -2.5299 | -3.0642 |
| MgO | 3 | 100.0 | 100.0 | 100.0 | -2.7630 | -2.1304 | -3.1916 |
| MgS | 3 | 99.6 | 100.0 | 100.0 | -1.7695 | -1.1438 | -2.6839 |
| NaCl | 3 | 100.0 | 100.0 | 100.0 | -2.0633 | -1.1435 | -2.3143 |
| $PbMoO_3$ | 3 | 100.0 | 100.0 | 100.0 | -2.4754 | -2.2387 | -4.3484 |
| $PbZrO_3$ | 3 | 100.0 | 100.0 | 100.0 | -1.8184 | -2.4638 | -4.6649 |
| RbCl | 3 | 100.0 | 100.0 | 100.0 | -2.5289 | -1.0804 | -2.3926 |
| SrS | 3 | 100.0 | 100.0 | 100.0 | -3.1306 | -1.2857 | -3.1246 |
| $TiB_2$ | 3 | 100.0 | 100.0 | 100.0 | -2.7240 | -2.9971 | -4.9640 |
| ZnO | 3 | 87.2 | 100.0 | 100.0 | -1.2925 | -1.1868 | -2.6299 |
| Rate | — | 18/27 | **27/27** | **27/27** | 4/27 | 0/27 | **23/27** |

TABLE S7. The structural validity and average energy of initial structures of four different methods (ML-PSO, ML-DFT, DFT-GA, and SCCOP) on 27 typical compounds. Each method has repeated 5 times to generate initial samples. The solid line means that the average energy is greater than zero. The DAUGNN of SCCOP is pretrained by databse.

| | | Structural Validity (%) | | | Average Energy (eV/atom) | | |
|---|---|---|---|---|---|---|---|
| Compounds | Dimension | ML-PSO(DFT) | DFT-GA | SCCOP | ML-PSO(DFT) | DFT-GA | SCCOP |
| BN | 2 | 89.5 | 100.0 | 100.0 | -5.0038 | -0.3862 | -6.1222 |
| C | 2 | 67.2 | 100.0 | 100.0 | -4.2315 | -0.6811 | -6.3948 |
| FeSe | 2 | 93.9 | 100.0 | 100.0 | -0.2252 | -0.0009 | -2.7905 |
| $MoTe_2$ | 2 | 85.3 | 100.0 | 100.0 | -0.0057 | 0.0000 | -2.1225 |
| $Mo_2C$ | 2 | 100.0 | 100.0 | 100.0 | 0.0000 | -0.0360 | -5.2784 |
| SiC | 2 | 65.4 | 100.0 | 100.0 | -1.8279 | -0.0580 | -5.4964 |
| $Ti_2C$ | 2 | 100.0 | 100.0 | 100.0 | 0.0000 | -0.1611 | -5.0393 |
| BaO | 3 | 100.0 | 100.0 | 100.0 | -2.5802 | -2.0223 | -4.1975 |
| BeS | 3 | 100.0 | 100.0 | 100.0 | -2.7016 | -1.3387 | -3.7173 |
| C | 3 | 100.0 | 100.0 | 100.0 | -5.0743 | -4.2250 | -6.1803 |
| CaO | 3 | 100.0 | 100.0 | 100.0 | -2.7499 | -2.3678 | -4.5083 |
| $CaTiO_3$ | 3 | 100.0 | 100.0 | 100.0 | -4.7407 | -3.4935 | -5.3116 |
| CdO | 3 | 99.6 | 100.0 | 100.0 | -1.7696 | -0.8799 | -2.9449 |
| CdS | 3 | 100.0 | 100.0 | 100.0 | -1.3625 | -0.6330 | -2.2391 |
| CsCl | 3 | 100.0 | 100.0 | 100.0 | -1.4466 | -1.1234 | -2.5226 |
| GaAs | 3 | 99.7 | 100.0 | 100.0 | -2.3411 | -1.2963 | -3.3658 |
| KF | 3 | 100.0 | 100.0 | 100.0 | -2.0759 | -1.8304 | -3.1067 |
| LiF | 3 | 100.0 | 100.0 | 100.0 | -3.0137 | -2.5299 | -3.8424 |
| MgO | 3 | 100.0 | 100.0 | 100.0 | -2.7630 | -2.1304 | -4.1469 |
| MgS | 3 | 99.6 | 100.0 | 100.0 | -1.7695 | -1.1438 | -3.2129 |
| NaCl | 3 | 100.0 | 100.0 | 100.0 | -2.0633 | -1.1435 | -2.7214 |
| $PbMoO_3$ | 3 | 100.0 | 100.0 | 100.0 | -2.4754 | -2.2387 | -5.0351 |
| $PbZrO_3$ | 3 | 100.0 | 100.0 | 100.0 | -1.8184 | -2.4638 | -5.3016 |
| RbCl | 3 | 100.0 | 100.0 | 100.0 | -2.5289 | -1.0804 | -2.5808 |
| SrS | 3 | 100.0 | 100.0 | 100.0 | -3.1306 | -1.2857 | -3.6581 |
| $TiB_2$ | 3 | 100.0 | 100.0 | 100.0 | -2.7240 | -2.9971 | -5.6665 |
| ZnO | 3 | 87.2 | 100.0 | 100.0 | -1.2925 | -1.1868 | -3.3529 |
| Rate | — | 18/27 | **27/27** | **27/27** | 0/27 | 0/27 | **27/27** |

TABLE S8. Comparison of the time cost and search accuracy of four CSP methods for 27 typical compounds.

| Methods | Dimension | Accuracy | Time (min) |
|---|---|---|---|
| ML-PSO | 2 | 0/7 | 8.09 |
| ML-DFT | 2 | 7/7 | 12.47 |
| DFT-GA | 2 | 7/7 | 95.19 |
| SCCOP | 2 | **7/7** | **7.06** |
| ML-PSO | 3 | 6/20 | 2.04 |
| ML-DFT | 3 | 16/20 | 12.43 |
| DFT-GA | 3 | 20/20 | 115.69 |
| SCCOP | 3 | **20/20** | **4.14** |

Figs. S1 and S2 illustrate the search process for 27 compositions. For each composition, the left figure shows the energy descent process, and the right figure displays its density of states (DOS).

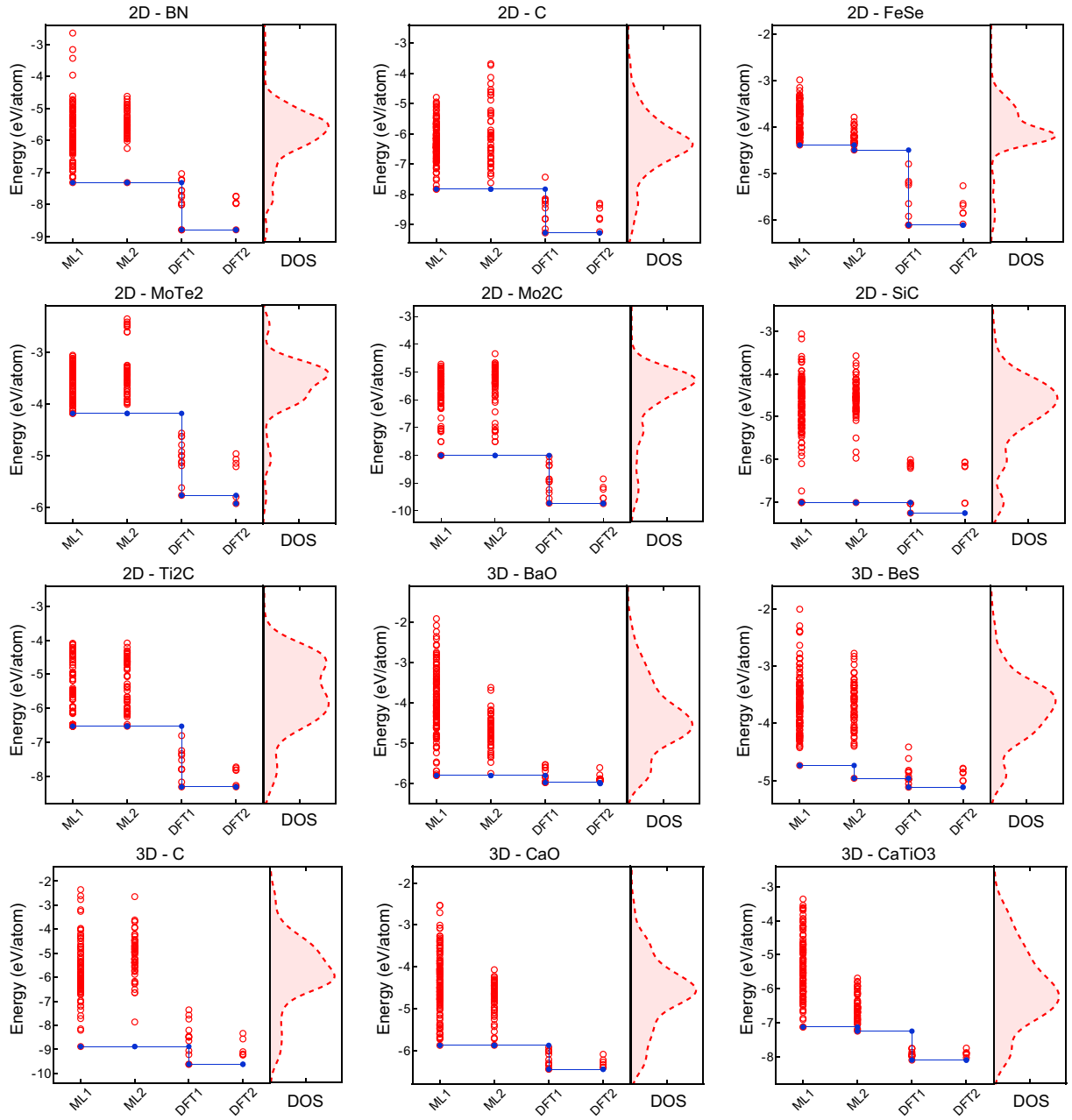


Fig. S1. Structure search process of SCCOP on 2D and 3D materials. ML1 and ML2 are the machine learning (ML) structure search and optimization, respectively, DFT1 and DFT2 are the structure optimization by VASP.

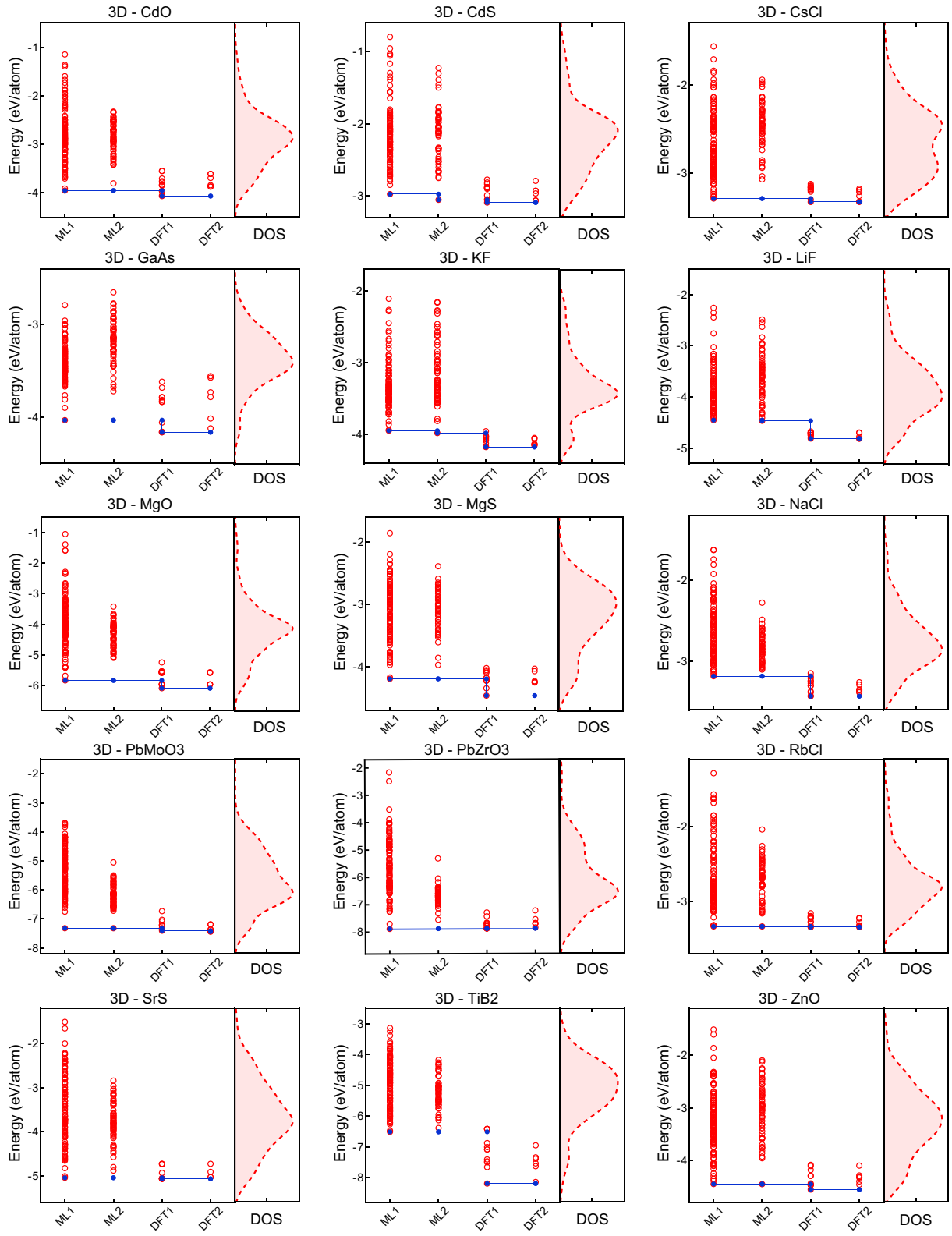


Fig. S2. Structure search process of SCCOP on 2D and 3D materials. ML1 and ML2 are the machine learning (ML) structure search and optimization, respectively, DFT1 and DFT2 are the structure optimization by VASP.

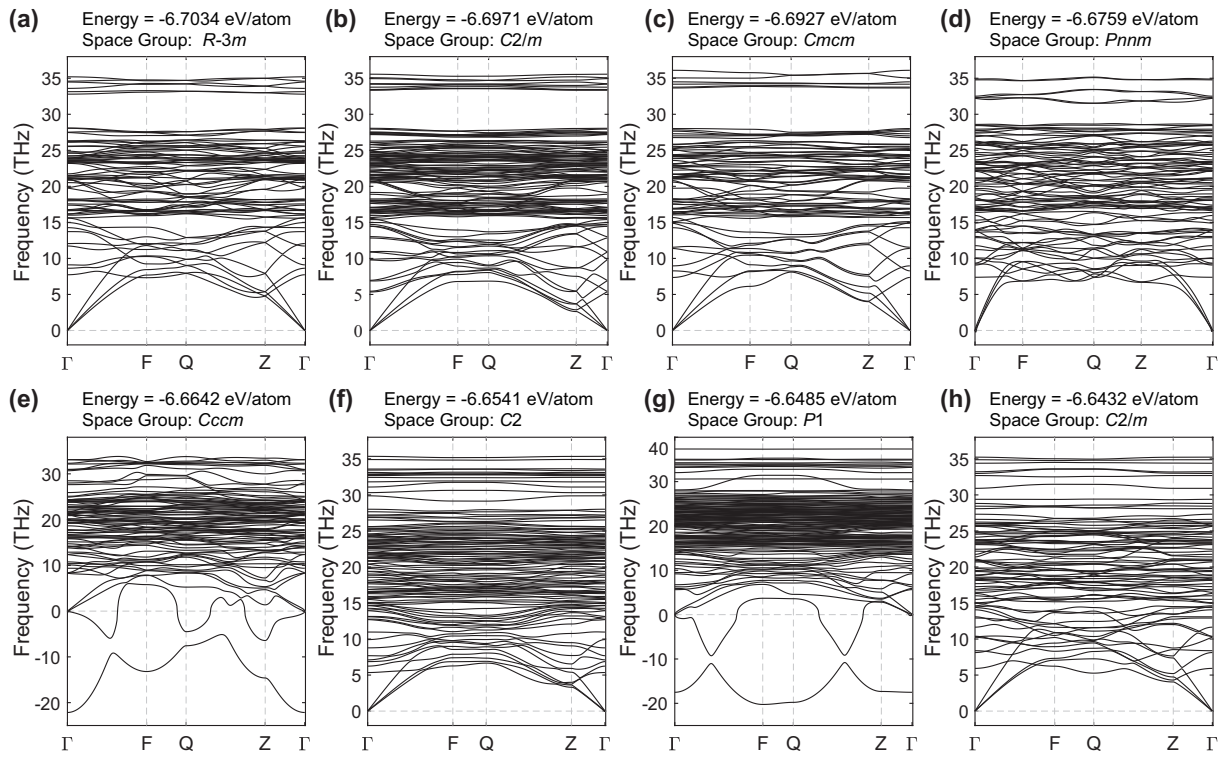


Fig. S3. Phonon spectrum of boron allotropes discovered by SCCOP.

To generate the lattice parameters $\boldsymbol{L}$, which include the lengths of three lattice vectors and the angles between them, we fit the values $(a, b, c, \alpha, \beta, \gamma)$ using statistical distributions obtained from databases. In most cases, we use the Gamma distribution $\Gamma(\alpha, \beta)$, where $\alpha$ is the shape parameter and $\beta$ is the scale parameter. This distribution is suitable for generating lattice parameters due to its non-negative nature. However, in some cases, fitting with a Gamma distribution is challenging, and we use a normal distribution $\mathcal{N}(\mu, \sigma^2)$ instead, where $\mu$ and $\sigma$ are the mean and standard deviation of the normal distribution, respectively. Additionally, we also employ a mixture of two distributions, such as $p_1 \cdot \mathcal{N}(\mu_1, \sigma_1^2)$, $p_2 \cdot \mathcal{N}(\mu_2, \sigma_2^2)$, or $p_1 \cdot \Gamma(\alpha_1, \beta_1)$, $p_2 \cdot \Gamma(\alpha_2, \beta_2)$. This means that each distribution is selected with a probability of $p_1$ and $p_2$, respectively, and the total probability density is the sum of the probabilities of each individual distribution. For 2D materials, we utilize three databases: JARVIS-DFT[2], C2DB[3], and 2DMATPedia[4]. The distributions of lattice parameters for 2D materials are presented in Table S9, taking into consideration that the value of $c$ depends on the thickness of the vacuum layer. For 3D materials, we utilize the OQMD[5] database, and the distributions of lattice parameters for 3D materials are shown in Table S10.

TABLE S9. Distribution of $(a, b, c, \alpha, \beta, \gamma)$ for different space groups in 2D materials.

| Space Group | $a$ | $b$ | $c$ | $\alpha$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|---|---|
| 1–2 | $\Gamma(9.21, 0.53)$ | $\Gamma(7.04, 1.06)$ | — | 90° | 90° | $\begin{cases} 0.58 \cdot \mathcal{N}(107.12, 8.59) \\ 0.42 \cdot \mathcal{N}(73.34, 6.21) \end{cases}$ |
| 3–9 | $\Gamma(10.23, 0.43)$ | $\Gamma(8.51, 0.88)$ | — | 90° | 90° | 90° |
| 10–12 | $\Gamma(12.72, 0.35)$ | $b = a$ | — | 90° | 90° | 90° |
| 13–17 | $\begin{cases} 0.4 \cdot \Gamma(24.82, 0.26) \\ 0.6 \cdot \Gamma(57.08, 0.06) \end{cases}$ | $b = a$ | — | 90° | 90° | 120° |

TABLE S10. Distribution of $(a, b, c, \alpha, \beta, \gamma)$ for different space groups in 3D materials.

| Space Group | $a$ | $b$ | $c$ | $\alpha$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|---|---|
| 1–2 | $\Gamma(11.14, 0.54)$ | $\Gamma(14.91, 0.42)$ | $\Gamma(9.31, 0.75)$ | $\Gamma(19.67, 5.48)$ | $\Gamma(21.63, 4.76)$ | $\Gamma(6.80, 12.82)$ |
| 3–15 | $\Gamma(8.93, 0.73)$ | $\Gamma(7.51, 0.90)$ | $\Gamma(9.99, 0.83)$ | 90° | $\begin{cases} 0.85 \cdot \mathcal{N}(108.24, 8.77) \\ 0.15 \cdot \mathcal{N}(74.80, 4.97) \end{cases}$ | 90° |
| 16–74 | $\Gamma(5.84, 1.05)$ | $\Gamma(6.95, 0.93)$ | $\Gamma(6.40, 1.09)$ | 90° | 90° | 90° |
| 75–142 | $\Gamma(17.21, 0.27)$ | $b = a$ | $\Gamma(11.64, 0.69)$ | 90° | 90° | 90° |
| 143–194 | $\Gamma(35.58, 0.13)$ | $b = a$ | $\Gamma(20.78, 0.24)$ | 90° | 90° | 120° |
| 195–230 | $\Gamma(17.24, 0.27)$ | $b = a$ | $c = a$ | 90° | 90° | 90° |

For the prediction of area calculated by lattice parameters $(a, b, \gamma)$, the formula[6] $\hat{S}_{\text{opt}} = A\sum_i r_i + B\sum_i X_i + C\sum_i Ea_i + D$ is used, where $A = 3.19$, $B = -0.01$, $C = 0.25$ and $D = 1.18$. The relative error is fitted by a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the $\mu = -0.10$ and $\sigma = 0.28$, $\hat{S}_{\text{min}} \leq S_{\text{opt}} \leq \hat{S}_{\text{max}}$ holds in 95% of cases, where $\hat{S}_{\text{min}} = \frac{1}{1+\mu+2\sigma} \cdot \hat{S}_{\text{opt}}$ and $\hat{S}_{\text{max}} = \frac{1}{1+\mu-2\sigma} \cdot \hat{S}_{\text{opt}}$.

For the space utilization rate, which is defined as $\sum_i S_i / S_{\text{opt}} \cdot 100\%$. We collect structures from databases: JARVIS-DFT, C2DB, and 2DMATPedia, and only keep the structures with thickness less than 4Å. The distribution of space utilization rate of 2D materials is shown in Fig. S4, and the $\mu = 0.56$ and $\sigma = 0.18$, This means that the volume inequality $\widetilde{S}_{\text{min}} \leq S_{\text{opt}} \leq \widetilde{S}_{\text{max}}$ holds with 95% probability, where $\widetilde{S}_{\text{min}} = \frac{1}{\mu+2\sigma} \cdot \sum_i S_i$ and $\widetilde{S}_{\text{max}} = \frac{1}{\mu-2\sigma} \cdot \sum_i S_i$.

Therefore, by combining the two bounds from distribution of relative error and space utilization rate, we obtain the tighter bounds for constraining area of 2D materials,

$$\max(\hat{S}_{\text{min}}, \widetilde{S}_{\text{min}}) \leq S_{\text{opt}} \leq \min(\hat{S}_{\text{max}}, \widetilde{S}_{\text{max}})$$

Finally, the lattice parameters $(a, b, \gamma)$ are generated repeatedly by the statistical distribution until the area constraint is satisfied.

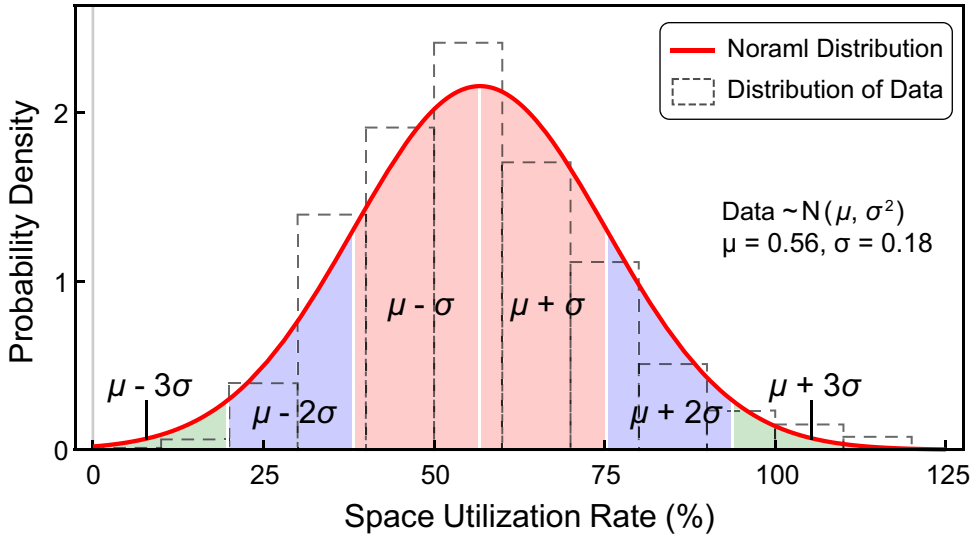


Fig. S4. Distribution of space utilization for 2D materials.

For 2D materials, we collected structures from three databases: JARVIS-DFT, C2DB, and 2DMATPedia. After removing ill-converged crystals, the full database contains 10,751 materials. For 3D materials, we collected structures from the OQMD database. After removing ill-converged crystals, the full database contains 817,133 materials.

To pre-train the GNN model, we used batchsize of 128 and 512 for 2D and 3D materials, respectively, with the Adam optimizer[7]. Each model was trained with 60% of the data, validated with 20% of the data, and the best-performing model in the validation set was selected. The remaining 20% of the data was used to test the generalization ability of the selected model. The pretrain model was trained for 120 epochs in a multi-learning rate style, with a learning rate of 0.01 for the first 100 epochs and a learning rate of 0.001 for the last 20 epochs. The performance of the model in the test set and the mean square error (MAE) of the validation set during training are shown in Fig. S5.

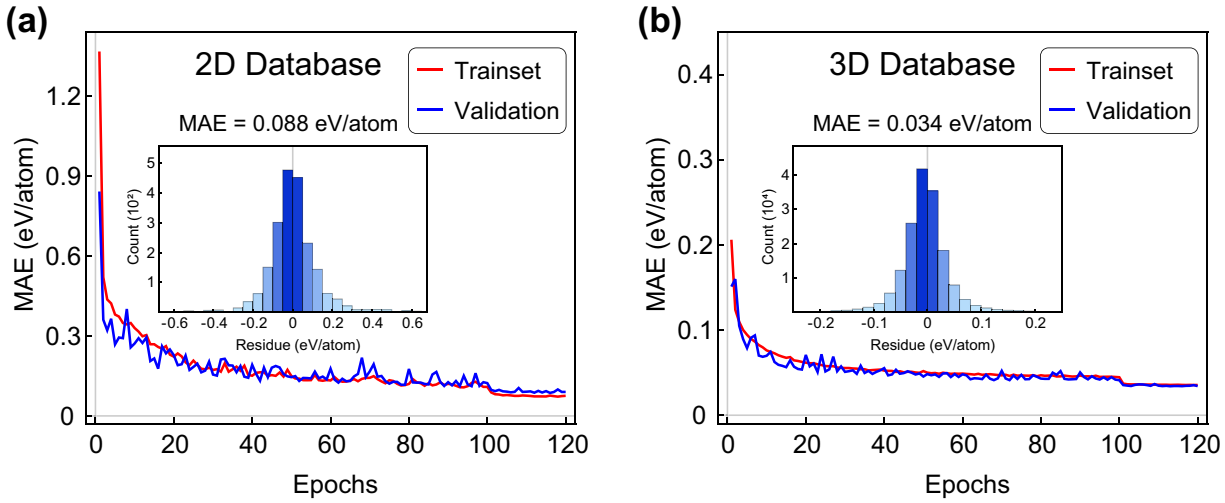


Fig. S5. The MAE of the GNN model during training and the MAE of the best-performing model on the test set for (a) 2D materials and (b) 3D materials.

## REFERENCES


[1]T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).

[2]K. Choudhary, K. F. Garrity, A. C. E. Reid, *et al.*, npj Comput. Mater. **6**, 173 (2020).

[3]S. Haastrup, M. Strange, M. Pandey, *et al.*, 2D Mater. **5**, 042002 (2018).

[4]J. Zhou, L. Shen, M. D. Costa, *et al.*, Sci. Data **6**, 86 (2019).

[5]J. E. Saal, S. Kirklin, M. Aykol, *et al.*, JOM **65**, 1501 (2013).

[6]M. G. Brik and A. M. Srivastava, J. Am. Ceram. Soc. **95**, 1454 (2012).

[7]D. P. Kingma and J. Ba, in *International Conference on Learning Representations* (2015).