Supplementary file: Physics Guided Deep Learning for Generative Design of Crystal Materials with Symmetry Constraints

Yong Zhao, Edirisuriya M. Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu*, Jianjun Hu*

1 Dataset Curation

We select the material data from three databases: MP [2], ICSD [1], OQMD [3]. The selection criteria are described following:

1. Ternary materials with only three base atom sites (a.k.a. one element is allowed to have only one base atom site);
2. Only keep materials that do not contain elements in Lanthanoid and Actinoid;
3. Ternary materials whose space group has more than 400 materials totally in three databases;
4. Ternary materials in OQMD whose fractional coordinates does not all belong to the set $[0.0, 0.25, 0.5, 0.75]$ since materials with fractional coordinates all falling in that set dominate the database [4].

In total, 42072 materials are selected and 20 space groups are found in those materials following above criteria. The statistics of materials in each space group is shown in Supplementary Table 1.

Supplementary Table 1: 20 space groups and their frequency in dataset **MIO**.

SG	SG Id	# samples	SG	SG Id	# samples
$P4/mmm$	123	1180	$Immm$	71	4679
$Fm\bar{3}m$	225	3716	$Cmcm$	63	1004
$I4_1/amd$	141	588	$I\bar{4}2d$	122	749
$Pm\bar{3}m$	221	1462	$R\bar{3}$	148	1969
$F\bar{4}3m$	216	898	$I4/mmm$	139	6162
$P6_3/mmc$	194	5599	$Fd\bar{3}m$	227	3292
$P\bar{3}m1$	164	1191	$Pnma$	62	2527
$P6/mmm$	191	2214	$R\bar{3}m$	166	1479
$I4/mcm$	140	433	$P6_3mc$	186	692
$R\bar{3}c$	167	1246	$P4/nmm$	129	992

Supplementary Table 2: 20 space groups and their frequency in dataset **TST**.

SG	SG Id	# samples	SG	SG Id	# samples
$P4/mmm$	123	317	$Immm$	71	59
$Fm\bar{3}m$	225	675	$Cmcm$	63	507
$I4_1/amd$	141	168	$I\bar{4}2d$	122	482
$P4/nmm$	129	719	$R\bar{3}$	148	374
$F\bar{4}3m$	216	60	$I4/mmm$	139	768
$P6_3/mmc$	194	1713	$Fd\bar{3}m$	227	239
$P\bar{3}m1$	164	674	$Pnma$	62	1386
$P6/mmm$	191	281	$R\bar{3}m$	166	576
$I4/mcm$	140	81	$P6_3mc$	186	151
$R\bar{3}c$	167	211	$Pm\bar{3}m$	221	0

We use first, second, and four criteria above to select materials in new released OQMD and the distribution of materials in 20 space groups is shown in Supplementary Table 2. 9441 materials are chosen and space group $Pm\bar{3}m$ does not have any new released materials.

2 Model Details

2.1 Implementation Hyperparameters for training PGCGM

Supplementary Table 3 shows the hyper-parameters in **PGCGM**. We use $\lambda_1 = 1$ and $\lambda_2 = 1$ for atom distance losses. We use *Property distribution* to select best atom dist bound ϕs combination and then using best ϕs, we optimize the best base coordinates and average full coordinates loss coefficients λ_1 and λ_2. Supplementary Table 4 and Supplementary Table 5 show the performance with different settings. We use 9 different combinations of ϕs and the best average *Property distribution* is achieved when ϕs are $(0.3, 7.5, 0.15, 7.5)$ as shown in 4. With best ϕs, we add coordinates based losses for the generator and the best λ_1 and λ_2 are 0.001 and 0.01 averagely as shown in Supplementary Table 5.

Supplementary Table 3: Hyper-parameters for training.

Hyper-parameters		Values
Adam optimizer	learning rate	0.0002
	β_1	0.5
	β_2	0.5
batch size		8192
gradient penalty coefficient		10
# of iterations of D per G iteration		5
low bound for inter atom dist ϕ_{inter}^{lower}		0.3
upper bound for inter atom dist ϕ_{inter}^{upper}		7.5
low bound for intra atom dist ϕ_{intra}^{lower}		0.15
upper bound for intra atom dist ϕ_{intra}^{upper}		7.5
inter dist loss coefficient λ_1		1.0
intra dist loss coefficient λ_2		1.0
base coord diff loss coefficient λ_3		0.001
avg. full coord loss coefficient λ_4		0.1

Supplementary Table 4: Choose the best ϕs (ϕ_{inter}^{lower}, ϕ_{inter}^{upper}, ϕ_{intra}^{lower}, ϕ_{intra}^{upper}) when adding dist losses.

ϕs	minD	maxD	density	avg.
(0.3, 7.8, 0.0009, 30.5)	0.220	0.846	1.481	0.849
(0.3, 12.5, 0.15, 25.0)	0.256	1.703	1.770	1.243
(0.3, 15.0, 0.15, 25.0)	0.228	1.879	2.139	1.415
(0.3, 7.5, 0.15, 12.5)	0.401	0.834	0.548	0.594
(0.3, 7.5, 0.15, 20.0)	0.301	1.000	1.176	0.826
(0.3, 7.5, 0.15, 7.5)	**0.354**	**0.512**	**0.757**	**0.541**
(0.3, 2.75, 0.15, 2.75)	0.573	2.157	3.214	1.981
(0.3, 5.0, 0.15, 5.0)	0.424	0.590	0.721	0.578
(0.3, 2.0, 0.15, 2.0)	0.728	2.322	3.848	2.299

Supplementary Table 5: Choose the best λ_1 and λ_2 when adding coordinates based losses.

(λ_1, λ_2)	minD	maxD	density	avg.
(0.001, 0.0001)	0.301	0.594	0.993	0.629
(0.001, 0.001)	0.258	1.103	0.823	0.728
(0.0001, 0.0001)	0.299	1.346	1.206	0.950
(0.0001, 0.001)	0.367	0.770	0.728	0.622
(0.01, 0.001)	0.337	1.032	1.440	0.936
(0.01, 0.01)	0.203	1.147	2.17	1.173
(0.001, 0.01)	**0.308**	**0.504**	**0.689**	**0.500**
(0.01, 0.0001)	0.251	0.991	0.942	0.728
(0.1, 0.1)	0.159	1.359	2.918	1.479

2.2 Model Structures

Supplementary Table 6 and 7 show the detailed architectures of discriminator and generator.

Supplementary Table 6: Discriminator configuration. **Mat** is the input material representations with shape of $3 \times 8 \times 8$. $\mathbf{SymOp}$ is the zero-padded symmetric operation matrix for space groups of materials. The 2D convolutional layer parameters are denoted as "C2D-<number of channels>-<receptive field size>". The fully connected layer parameters are denotes as "FC-<number of neurons>". The concatenation is denoted as "CAT-<number of neurons>". We use *LeakyReLU* as the activation function after each layer except for the last layer. The negative slope for it is 0.2.

Discriminator Configuration	
Mat-$3 \times 8 \times 8$	
C2D-16-2	
C2D-32-2	
C2D-64-2	
C2D-96-2	$\mathbf{SymOp}$-$192 \times 4 \times 4$
C2D-128-2	C2D-64-2
C2D-192-2	C2D-128-2
C2D-256-2	C2D-256-2
CAT-512	
FC-265	
FC-1	

Supplementary Table 7: Generator configuration. $\mathbf{SymOp}$ is the zero-padded symmetric operation matrix for space groups of materials. $\mathbf{Z}$ is the random noise with shape of 128 and it shared by two branches for generating unit cell length $\mathbf{P}^*$ and three set of base atom sites $(\mathbf{B}^0_{fake}, \mathbf{B}^1_{fake}, \mathbf{B}^2_{fake})$. The 2D convolutional layer parameters are denoted as "C2D-<number of channels>-<receptive field size>". The 2D deconvolutional layer parameters are denoted as "TC2D-<number of channels>-<receptive field size>". The fully connected layer parameters are denotes as "FC-<number of neurons>". The concatenation is denoted as "CAT-<number of neurons>". We use batch normalization and *ReLU* after each layer except for the last layers of two branches. They are followed by a *Tanh* activation to generate lengths and atom coordinates.

Generator Configuration		
		$\mathbf{ElemProp}$-23×3
$\mathbf{SymOp}$-$192 \times 4 \times 4$		C1D-64-2
C2D-64-2		C1D-128-2
C2D-128-2	$\mathbf{Z}$-128	flatten
C2D-256-2	FC-256	FC-256
CAT-512		CAT-512
FC-128		TC2D-1024-2
FC-64		TC2D-512-2
FC-32		TC2D-256-1
FC-16		TC2D-128-1
FC-3		TC2D-64-1
output: $\mathbf{P}^*$-3		TC2D-3-1
		output: $\mathbf{B}$ - $3 \times 3 \times 3$

3 DFT configuration

The structures were optimized by density functional theory (DFT) that were carried out with Vienna ab initio simulation package (VASP). The structure optimization convergence criteria of force and energy are 10^{-4} eV/A and 10^{-7} eV, respectively. VASP runs were performed with full degree of freedom in terms of allowing the atomic coordinates, lattice size, lattice constant, and lattice shape to change to reach the convergence criteria of force and energy in the structure optimization process. The Perdew–Burke–Ernzerhof (PBE) of the generalized gradient approximation (GGA) was used for exchange–correlation functional. The kinetic energy cutoff was set to be 520 eV for the electronic wavefunction having a plane wave basis set which was obtained using the projector augmented-wave method. The Monkhorst–pack k-mesh grids selected to sample the Brillouin zone in the calculations were determined depending on the lattice constants. The product of the number of k-meshes in

one direction and the lattice constant (measured in Angstrom) in the same direction is roughly set as 60, which is dense enough for structure optimization.

References

[1] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta. Crystallogr. B Struct. Sci.*, 58(3):364–369, 2002.

[2] Anubhav Jain et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1):011002, 2013.

[3] Scott Kirklin et al. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *NPJ Comput. Mater.*, 1(1):1–15, 2015.

[4] Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.