Supporting Information

Generative Adversarial Networks for Crystal Structure Prediction

Sungwon Kim[1,†], Juhwan Noh[1,†], Geun Ho Gu[1], Alán Aspuru-Guzik[2,3,4], Yousung Jung[1,*]

[1]Department of Chemical and Biomolecular Engineering, KAIST, 291 Daehak-ro, Daejeon 34141, South Korea

[2]Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M55S 3H6, Canada

[3]Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1M1, Canada

[4]Canadian Institute for Advanced Research (CIFAR) Lebovic Fellow, Toronto, Ontario M5S 1M1, Canada

† These authors contribute equally to this work.

*E-mail: ysjn@kaist.ac.kr

Table of contents

S1. Representation details

S1.1 Specific representation of Mg-Mn-O system and rescaling fractional coordinates

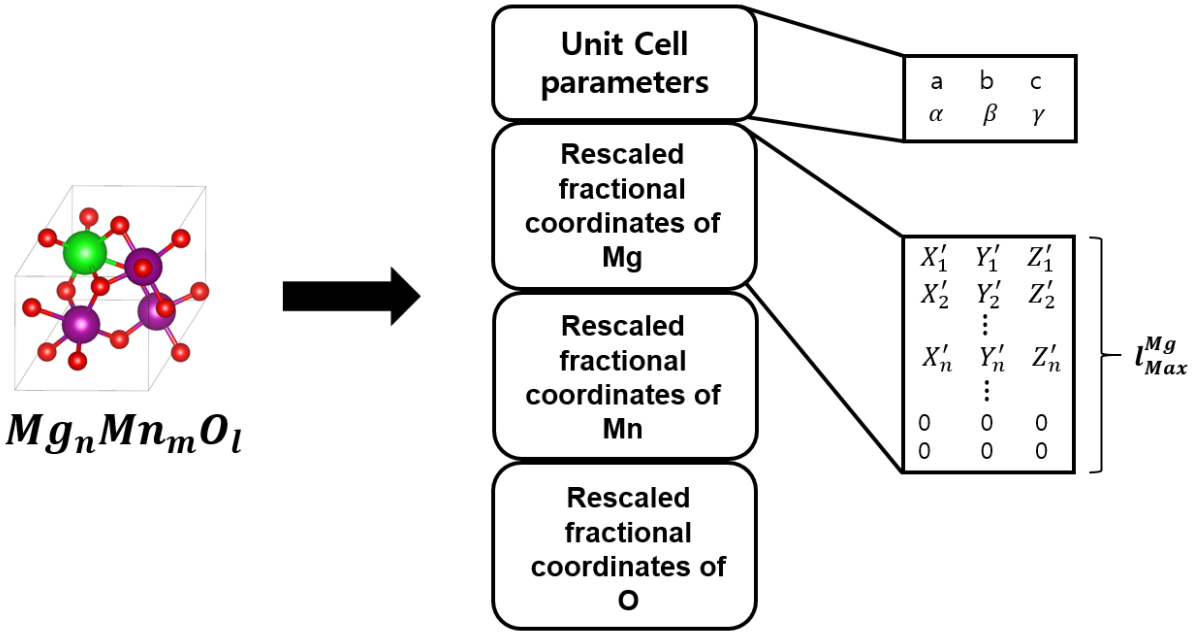


Figure S1. Specific representation of Mg-Mn-O structure.

The representation is composed of unit cell parameters and the sets of rescaled fractional coordinates of atoms. Each set of rescaled fractional coordinates represents an element. If the number of elements increases, the number of sets of coordinates also increases. The length of each part of the coordinates is fixed to $l^{element}_{Max}$ and the rows where atom does not exist are filled with zero-paddings

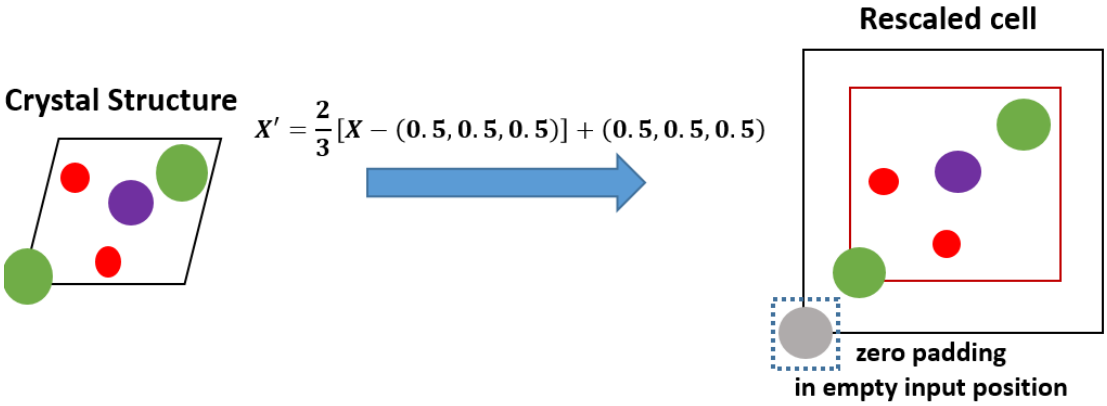


Figure S2. Scheme of rescaling fractional coordinates. To distinguish real atoms located at (0,0,0) from zero-paddings which were used to fix the shape of the representation, we rescaled

the fractional coordinates.

The first two rows of the 2D-representation contain the lengths of the unit cell edge and angles between them, and the fractional coordinates of atoms are listed below. To fix the shape of the representation, we set the maximum number of atoms in each element, $l^{\mathrm{A}}_{\mathrm{Max}}$ and zero-padding (0, 0, 0) is used if the number of atoms with specific element type A is lower than $l^{\mathrm{A}}_{\mathrm{Max}}$. In the Mg-Mn-O system used here, for example, $l^{\mathrm{Mg}}_{\mathrm{Max}}$ is 8, $l^{\mathrm{Mn}}_{\mathrm{Max}}$ is 8 and $l^{\mathrm{O}}_{\mathrm{Max}}$ is 12. (See **Figure S1**) In addition, as shown in the **Figure S2**, we apply the rescaling operation to fractional coordinates of the point atom, $\boldsymbol{P}(X, Y, Z)$ to distinguish the zero paddings from the atoms located at (0, 0, 0) position.

$$\boldsymbol{P'} = \frac{2}{3}[\boldsymbol{P} - (0.5, 0.5, 0.5)] + (0.5, 0.5, 0.5)$$

where $\boldsymbol{P'}$ is the rescaled fractional coordinates used in our representation. Since our representation only requires the atomic coordinates and cell information, it requires almost no preparation and memory cost to store the raw input data that can be compared with the 3D voxel representations necessitating substantial memory space to store grid data.

S1.2 Data augmentation

- **Supercell operation**

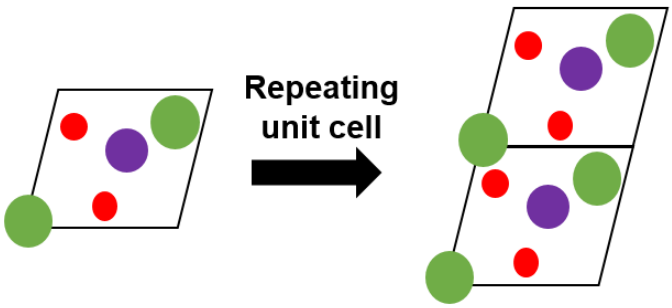
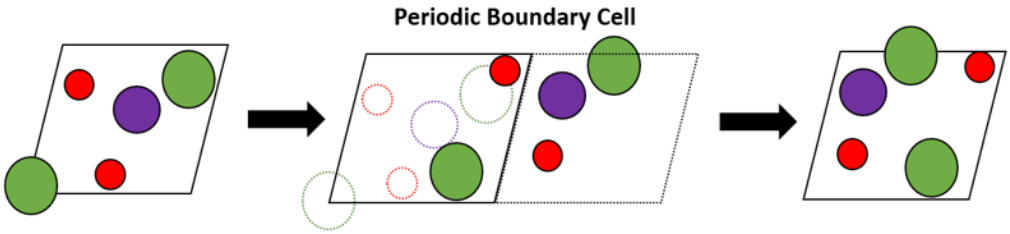


Figure S3. Scheme of supercell operation We made super cells by repeating unit cells twice in x-y axis, y-z axis, and x-z axis respectively where the number of atoms in the cell does not exceed the maximum number of atoms our representation.

- **Translation operation**



$$P' = P + \Delta P$$

$$\Delta P = \lambda * (l_x, l_y, l_z) \ , (0< \lambda<1, \ l = cell\ length)$$

Figure S4. Scheme of translation operation We applied translation operation on structure data by moving atoms in unit cell by random distances which are smaller than cell length.

- **Rotating operation**

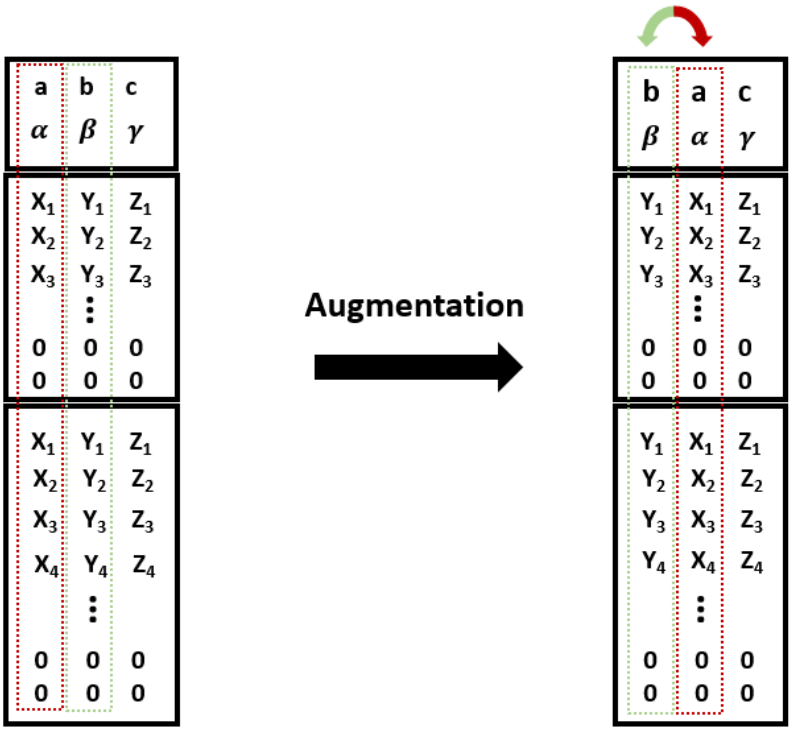


Figure S5. Scheme of rotating operation We applied rotational operation on structure data by swapping two axes (column).

S2. Model details

S2.1 Architecture of Composition-Conditioned Crystal GAN

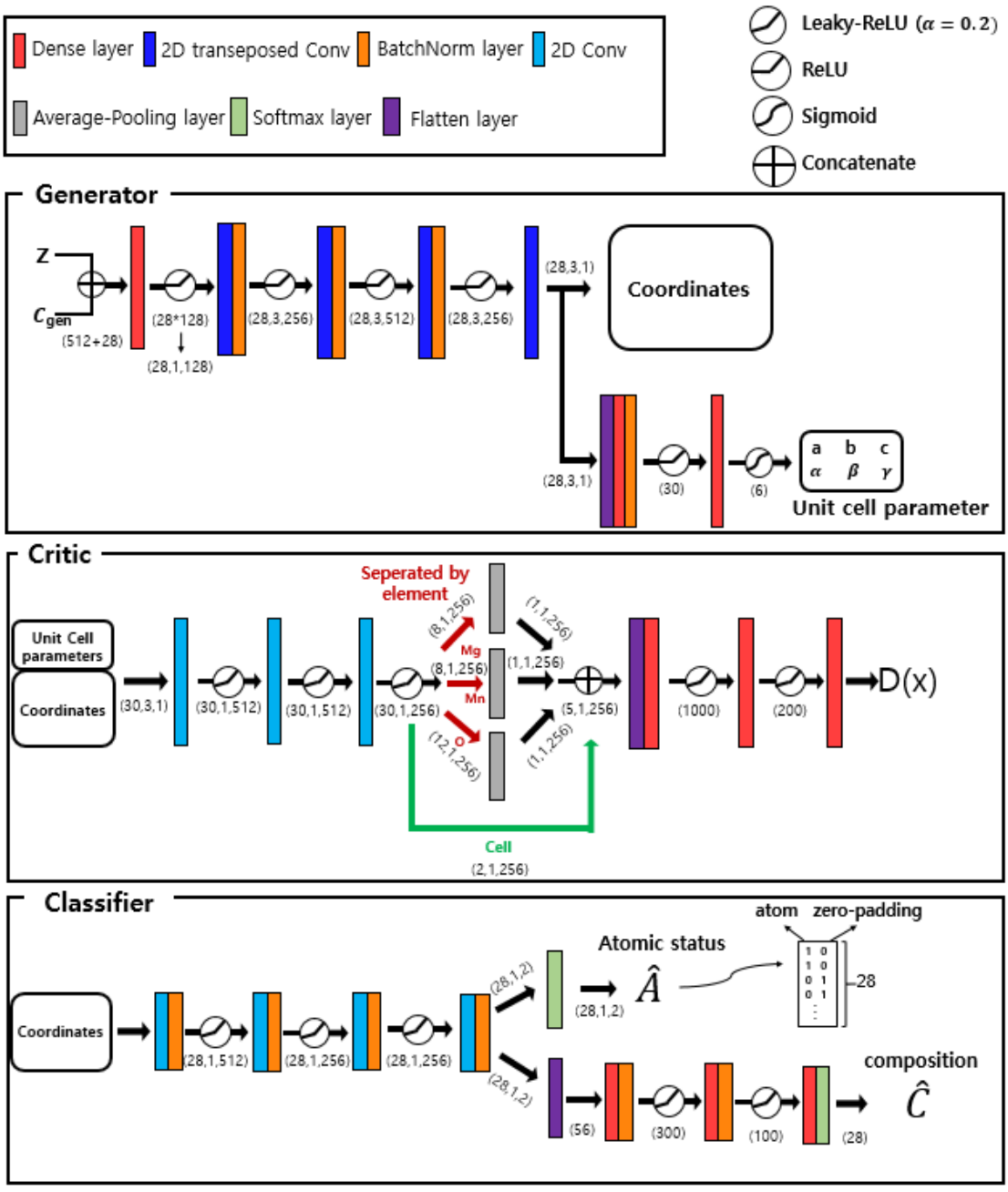


Figure S6. Architecture of Composition-Conditioned Crystal GAN

Table S1. Hyperparameters of Composition-Conditioned Crystal GAN

Hyperparameter		Value
Mini batch size		32
Adam optimizer	Learning rate	0.0001
	β_1	0.5
	β_2	0.999
Coefficient of Gradient penalty , λ		10
Coefficient of generated data in $L_{class-comp}$, λ_1		1
Coefficient of generated data in $L_{class-atom}$, λ_2		1
Coefficient of Composition Classification, λ_c		0.3

S2.2 Loss function of Composition-Conditioned Crystal GAN

In this work, we implemented variant of GAN called WGAN (Wasserstein GAN[1]) WGAN overcomes the shortcoming of GAN such as unstable training or mode collapse by using the Wasserstein distance between real and generated data distributions as the loss function. This loss function train the generator to create materials that are similar to the real materials. In addition, for improved training of WGAN, Gulrajani et al.[2] proposed additional term in loss function of WGAN, gradient penalty term. This regularizer term has enabled more stable training of WGAN. In detail, the loss function is

$$L_{WGAN} = \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)] + \lambda \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2]$$

Where D is the critic function, $P_{\hat{x}}$ is sampling uniformly along straight lines between pairs of points sampled from the real data distribution, P_r and the generator distribution, P_g, and λ is penalty coefficient, set to be 10. In order to train the generator to create materials with target property, the generator is trained together with the classifier with a loss function

$$L_{class-comp} = CE\left(C_{real}, \hat{C}_{real}\right) + \lambda_1 CE(C_{gen}, \hat{C}_{gen})$$

$$L_{class-atom} = CE\left(A_{real}, \hat{A}_{real}\right) + \lambda_2 CE(A_{gen}, \hat{A}_{gen})$$

$$L_{class} = L_{class-atom} + \lambda_c L_{class-comp}$$

$$CE(t, x) = -\sum_{i}^{C} t_i \log(x_i)$$

where CE is cross entropy, x_i is ith value of output value for classifier function, C is the number of classes and t_i is ith target value. C_{real}, $\hat{C}_{real}$, C_{gen}, $\hat{C}_{gen}$ are true (no hat) and predicted (hat) property value of real (real subscript) and generated (gen subscript) materials. λ_1, λ_2, and λ_c are generator coefficient in composition, generator coefficient in atomic state, and composition coefficient respectively.

S3. Learning curve of the Composition-Conditioned Crystal GAN

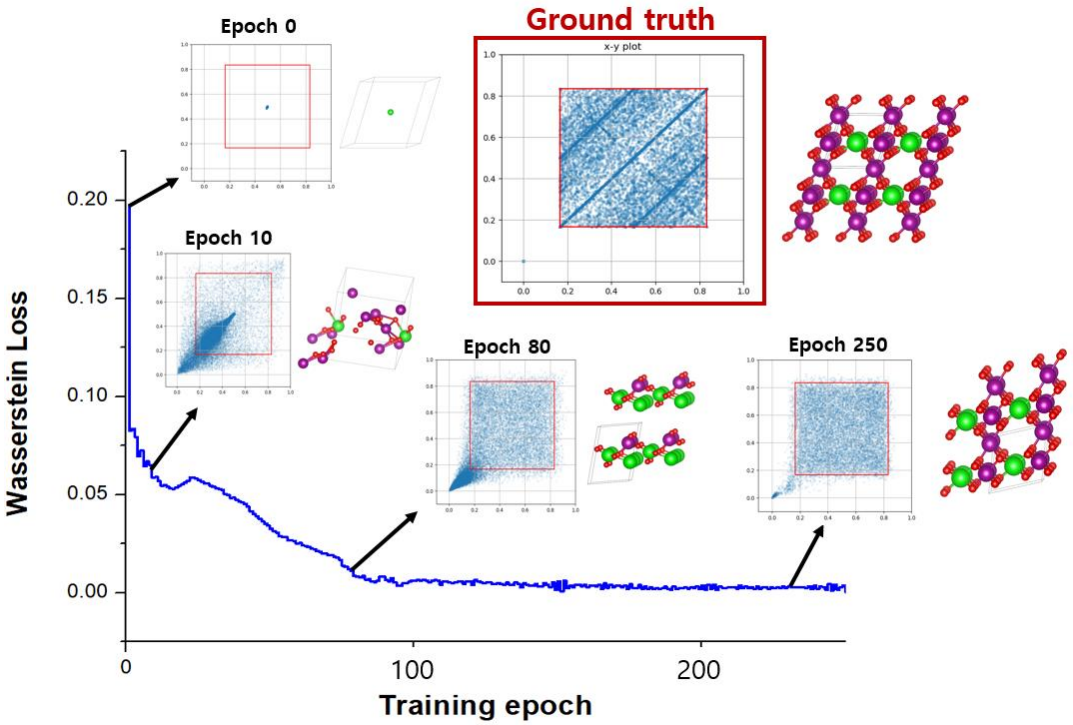


Figure S7. The learning curve of Composition-Conditioned Crystal GAN. The Wasserstein loss as a function of the training epoch. The inset figures are scattered plot where scaled fractional coordinates of all atoms in generated structures are plotted. For clarity, only X-Y coordinates of atom positions are shown. The example of a generated structure according to the epoch is on the right side of the scatter plot. As the training progresses, generated atomic positions are becoming more and more similar to the ground truth representing the training dataset**.**

The learning process of our model is described in **Figure S7**, demonstrating that the Wasserstein distance converges to zero as training progresses and the generator can generate the data similar to real data. To help visualize it, we included the inset figures for scatter plot of all atoms' positions in randomly selected 1000 structures generated via Composition-Conditioned Crystal GAN in the corresponding training epoch. In the ground truth indicating training dataset, all atoms in the structures of training dataset are located in the red box except zero paddings because the coordinates were rescaled. In the early stages of training, atoms were bound together and located in erroneous coordinates, however, as training progresses and Wasserstein distance converges to zero, the atomic and zero padding coordinates become

reasonable with the correct composition. The examples of converted to Mg-Mn-O structure were randomly selected from the generated structures in each corresponding training epoch and is located in the right side of each scatter plot.

S4. Computational details

S4.1 DFT calculations for the generated VO materials

For the comparison with iMatGen, we performed identical VASP[3] calculation method with Noh et al.[4] Also, phase diagram and energy above convex hull were calculated by the method described in methodology section of Noh et al[4] For all generated structures, we performed spin polarized GGA+U[5,6] calculations, with the same U parameter for V used in the Materials Projects database.[7] We relaxed both atomic positions and cell parameters using conjugate gradient descent method with convergence criteria of $1.0e^{-5}$ for energy and 0.05 eV/Å for force with 500 eV cut off energy. To compare the phase stability among the generated structures for all generated materials we first used sparse reciprocal lattice grid with a grid spacing of 0.5 $Å^{-1}$. The formation energy (or formation enthalpy, E_f) was calculated using $E_{V_xO_y} - (xE_V - yE_O)/(x+y)$ (in eV/atom). Then, for a smaller set of materials that satisfy energy stability ($E_{hull} \leq 0.2$ eV/atom), we refined the formation energy calculations using a denser reciprocal lattice grid with grid spacing of 0.25 $Å^{-1}$.

S4.2 DFT calculations for the generated MgMnO materials

For the comparison with high-throughput screening method, we performed identical VASP calculation method with Noh et al.[8] Also, phase diagram and energy above convex hull were calculated by the method described in methodology section of Noh et al.[8] For all generated structures, we performed spin polarized PBE+*U*[5,6] calculations and PAW[9]–PBE pseudopotentials as implemented in the *ab initio* package, VASP, and we used 3.9 as U-value for Mn taken from Materials Project.[7] We relaxed both atomic positions and unit cell parameters using conjugate gradient descent method with convergence criteria of 1.0e-5 for

energy and 0.05 for $eV/Å$ force with 500 cut off energy. To compare the phase stability among the generated structures for all generated materials, Brillouin zone is used with k-point densities at or larger than 500 k-points per atoms using the *Pymatgen*[10] package. Duplicates for the converged structures are removed using the *StructureMatcher* function implemented in *Pymatgen* package. After that, we performed the latter computations with dense k-space (i.e. Brillouin zone with k-point densities at or larger than 1000 k-points per atoms using the *Pymatgen* package).

S4.3 Band gap calculations

We performed HSE[11] hybrid DFT functional implemented in VASP[3] with a mixing parameter of 0.2. For computational efficiency, a uniform reduction factor for the q-point grid of the exact exchange potential is applied (NKRED = 2) with gamma centered even number k-points (with a k-point densities at or larger than 1000 k-points per atoms).

S5. Detailed statistics of the result

S5.1 Materials generation statistics in Mg-Mn-O system

Table S2. Numerical statistic of Mg-Mn-O polymorphs generated from Composition-Conditioned Crystal GAN with compositions in MP.

Composition	E above hull≤200meV	E above hull≤80meV
MgMnO3	3	-
MgMn2O4	35	8
MgMn2O5	17	-
MgMn3O6	15	2
MgMn3O7	42	-
MgMn4O6	52	7
MgMn4O8	51	1
Mg2Mn2O5	24	4
Mg2Mn3O6	95	5
Mg2Mn3O8	4	-
Mg6MnO8	30	8

Table S3. Numerical statistic of Mg-Mn-O polymorphs generated from Composition-Conditioned Crystal GAN with compositions not in MP.

Composition	E above hull≤200meV	E above hull≤80meV
MgMn3O4	37	13
MgMn3O5	39	7
MgMn4O5	29	3
MgMn4O7	26	3
MgMn5O6	31	11
Mg2MnO3	12	3
Mg2MnO4	13	2
Mg2Mn3O5	67	8
Mg2Mn4O7	84	5
Mg3MnO5	22	1
Mg3Mn2O5	28	3
Mg3Mn2O6	68	10
Mg3Mn2O7	22	-
Mg3Mn3O7	74	6
Mg4MnO5	12	3
Mg4MnO6	15	4

Mg4Mn2O6	55	15
Mg4Mn2O7	63	4
Mg5MnO6	30	10
Mg5MnO7	26	2

S5.2 Structure comparison

We first note that out of 28 Mg-Mn-O phases identified as promising photoanode materials, 14 phases correspond to new composition not included in database, meaning that the newly found structures are completely new and cannot be obtained from simple structural distortions. For the remaining 14 phases in 8 existing compositions, we estimated the similarity of the structures using the local structure order parameters with the 'Structure Matcher' function implemented in Pymatgen python package. (see **Figure S8**)

In **Figure S8**, we listed the 14 newly identified Mg-Mn-O structures with the most similar structures in our database based on the calculated dissimilarity values cited above. Indeed, we find that 5 of 14 identified structures (red boxed) are classified as the same structures to those in our database. For the other 9 structures, however, the generated structures seem to have very different structural motifs with large structural dissimilarity values compared those in the database. Thus, we believe the 23 out of 28 identified structures correspond to either completely new (with new compositions) or very different (with large dissimilarity values) structural motifs that are not in the databases.

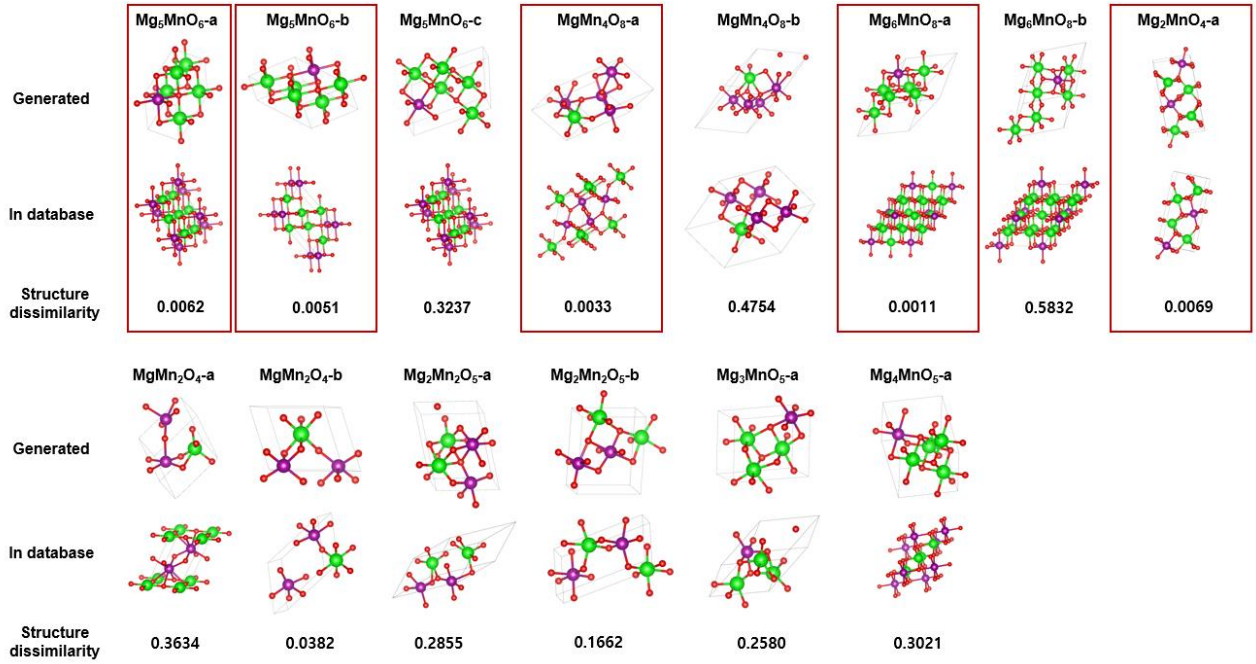


Figure S8. The result of comparison of 14 generated structures considered as promising photoanode materials and database structures having same chemical composition. In first row, 14 generated structures are located and in second row, the database structures most similar with corresponding generated structure are located. Structure dissimilarity in third row was value calculated between two structures in same column.

S5.3 Examples of promising photoanode materials

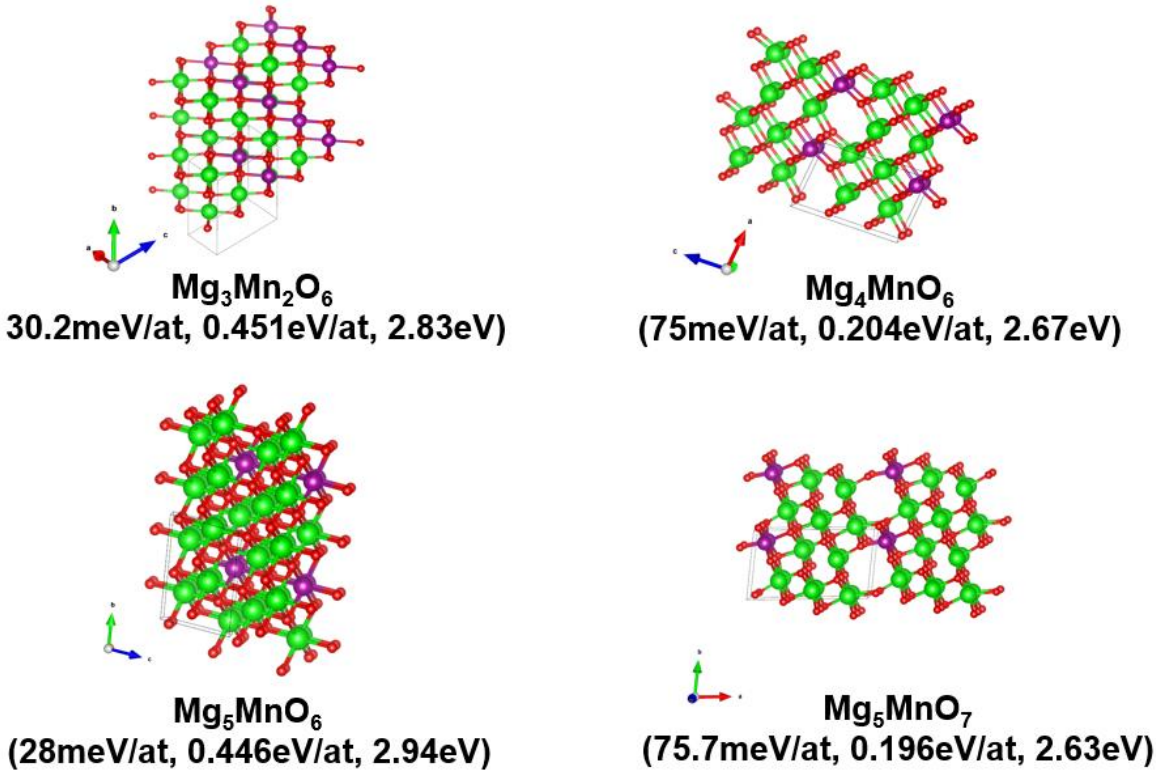


Figure S9. Examples of promising photoanode materials with composition not in MP. For each structure, the energy above the convex hull, pourbaix stability, and bandgap energy are also shown below the structure.

S6. V.O system

S6.1 V.O dataset

In V-O system, l^{V}_{Max} is 8 and l^{O}_{Max} is 12. V-O initial dataset is constructed by using the elemental substitution for the binary compounds existing in the Material Project (MP) database, and removed duplicates. From this, we get a total of 1396 unique structures with 86 compositions for the V-O system. After data augmentation with the same way in Mg-Mn-O system, totaling 86,000 for V-O training dataset were constructed.

S6.2 V.O results

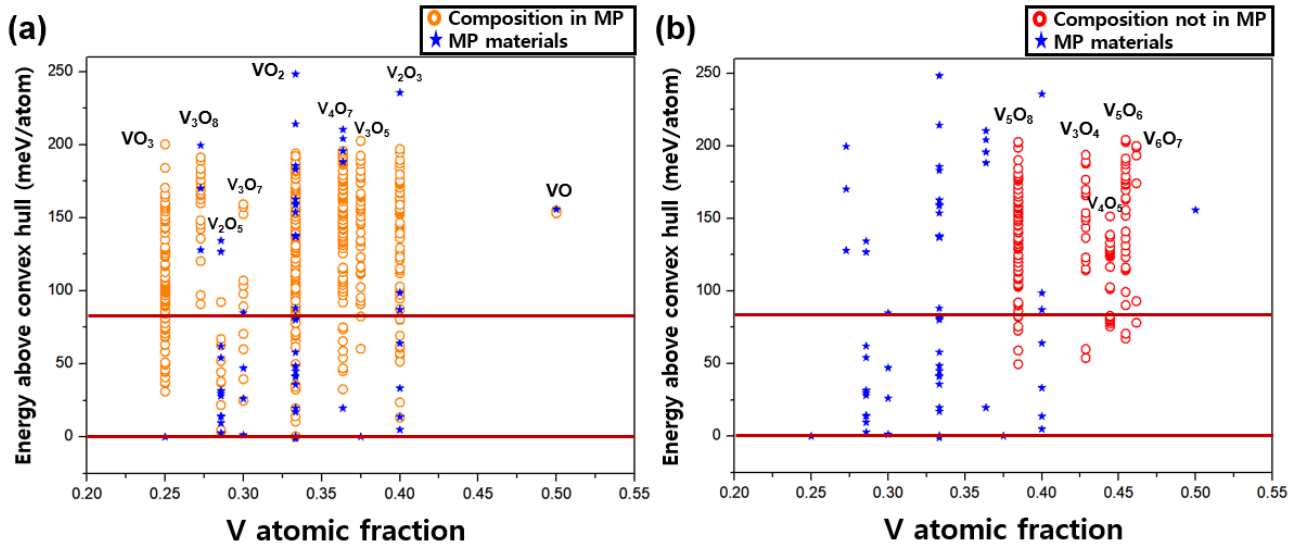


Figure S10. DFT calculated formation energies for generated VO polymorphs. (a) DFT calculated formation energies for the generated VO polymorphs with composition existing in MP (b) DFT calculated formation energies for the generated VO polymorphs with composition not in MP. Orange circles and Red circles are generated materials with composition existing in MP and not in MP respectively. Blue stars correspond to the materials of the V-O database in MP. The horizontal red lines represent 80meV/atom and 0meV/atom respectively.

To validate that our model can generate stable and realistic polymorphs of VxOy with given composition condition, we generated VxOy compounds of several known compositions in MP (see **Figure S10**), and screened them with their formation energy. We sampled structures

according to these compositions and selected the structures only if the given composition condition (C_{gen}) is exactly reconstructed by the classifier ($\hat{C}_{gen}$). Also, we removed the structures if atoms are too close to each other. From this post process, 300 structures in each composition thus totaling 5400 structures in 18 compositions were sampled, and further DFT calculations are performed to compute stability (i.e. the energy above the convex hull, E_{hull}) of the generated materials as shown in **Figure S10**. And for energy comparison with materials actually existing in MP, the formation energies of them were also plotted in **Figure S10**. Here, three main results are noteworthy:

(1) As shown in **Figure S10a**, 562 of unique and entirely new structures are predicted as theoretically metastable (i.e. $E_{\mathrm{hull}} \leq 200$ meV/atom) among the total 4800 structures indicating that the proposed Composition-Conditioned Crystal GAN can effectively generate stable and new materials. Furthermore, considering that 80 percent of the experimentally known identified sulfides and oxides were within this criterion[12] 91 unique structures are predicted as potentially synthesizable (i.e. $E_{\mathrm{hull}} \leq 80$ meV/atom).

(2) Among 53 V-O polymorphs of MP which we compared with our generated results energetically in **Figure S10a** (blue stars), 15 materials have the same number of atoms in unit cell as the materials generated by our model. Although we randomly generated new materials using the trained model, 7 of 15 V-O polymorphs of MP with compositions which we selected to generate polymorphs are successfully rediscovered. Here, we expect the latter success ratio to increase further if we increase the number of samples.

(3) From comparison to the generated materials from iMatGen[4] (see **Figure S10b**), 40 percent of previous discovery was generated within the range of $E_{\text{hull}} \leq 200$ meV/atom by our framework (see **Figure S10b**). It is notable that for V_3O_4 our model generates more stable materials than the most stable previous discovery from iMatGen. In addition, in the case of V_6O_7, while iMatGen does not find stable materials in the range of $E_{\text{hull}} \leq 80$ meV/atom, our model can successfully generate new and stable polymorphs showing the effectiveness of our model to explore unseen chemical space.

Table S4. Comparison with the results of iMatGen[4] Each value in the table is the number of structures. Considering structures with E_{hull} less than 200 meV/atom, 13 of 33 structures generated via iMatGen are composed in structures generated via Composition-Conditioned Crystal GAN**.**

Composition	Proposed Crystal GAN model	iMatGen	Identical structure
V_3O_4	**20**	**10**	**3**
V_4O_5	**3**	**3**	**1**
V_5O_6	**32**	**8**	**3**
V_5O_8	**75**	**10**	**4**
V_6O_7	**6**	**2**	**2**

S7. The effect of data augmentation in generative model.

To clarify that the data augmentation is actually affect the critic network to indirectly learn invariance under symmetric operations, we implemented additional experiment. To identify the effect of data augmentation and the minimum number of required data augmentation, we trained addition four independent models by changing the number of applied data augmentation as following:

-'No Aug': No augmentation (1,253 data points),

-'Aug1000': 1,000 structures per composition (112,000 data points total),

-'Aug2000': 2,000 structures per composition (224,000 data points total),

-'Aug3000': 3,000 structures per composition (336,000 data points total)

Next, we made a separate test dataset for 25 compositions. For each composition, there is 1 unique structure and 999 structures are augmented by applying the same procedure in the manuscript, yielding a total of 25,000 data points (25 x 1,000 data for each composition). To quantify the effect of data augmentation, we computed the difference of the output of the critic network between the base structure and the augmented structures. Here, the model would learn the invariance under the symmetric operations if the computed difference values are close to zero as shown in below figure. To be specific, in 'No Aug' case, there are relatively large deviation from the base structure (i.e. red line) compared to the other augmented models because the critic network cannot learn additional information from augmented structures. Notably, for the other 3 cases (i.e. 'Aug1000', 'Aug2000' and 'Aug3000'), the distribution is centered at zero (i.e. red line) indicating that the model can learn invariance under the symmetric operations from the augmented structures. Therefore,

we can roughly estimate that augmenting 1000 data points per composition (in this work) would be sufficient.

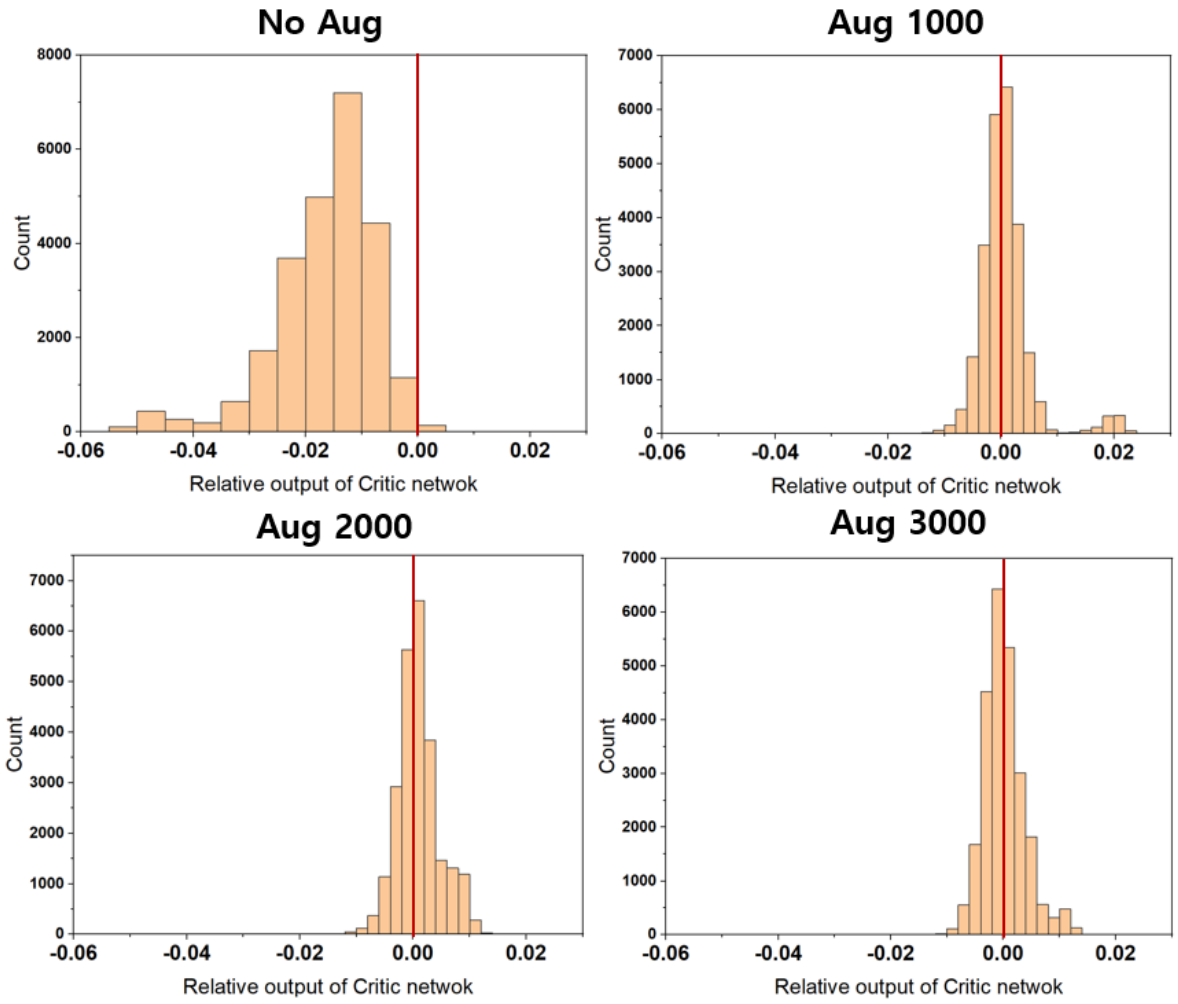


Figure S11. Distribution of the difference of the output of critic network between the base structure and augmented structures.

References

(1) Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875* **2017**.

(2) Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. C. Improved training of wasserstein gans. **2017**;5767.

(3) Kresse, G.; Furthmüller, J. Software VASP, vienna (1999). *Phys. Rev. B* **1996,** *54* (11), 169.

(4) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019,** *1* (5), 1370.

(5) Anisimov, V. I.; Aryasetiawan, F.; Lichtenstein, A. I. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+ U method. *J. Phys. Condens. Matter* **1997,** *9* (4), 767 %@ 0953.

(6) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996,** *77* (18), 3865.

(7) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013,** *1* (1), 011002.

(8) Noh, J.; Kim, S.; ho Gu, G.; Shinde, A.; Zhou, L.; Gregoire, J. M.; Jung, Y. Unveiling new stable manganese based photoanode materials via theoretical high-throughput screening and experiments. *ChemComm.* **2019,** *55* (89), 13418.

(9) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B.* **1994,** *50* (24), 17953.

(10) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013,** *68*, 314.

(11) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003,** *118* (18), 8207.

(12) Singh, A. K.; Montoya, J. H.; Gregoire, J. M.; Persson, K. A. Robust and synthesizable photocatalysts for CO 2 reduction: a data-driven materials discovery. *Nat. Commun.* **2019,** *10* (1), 443.