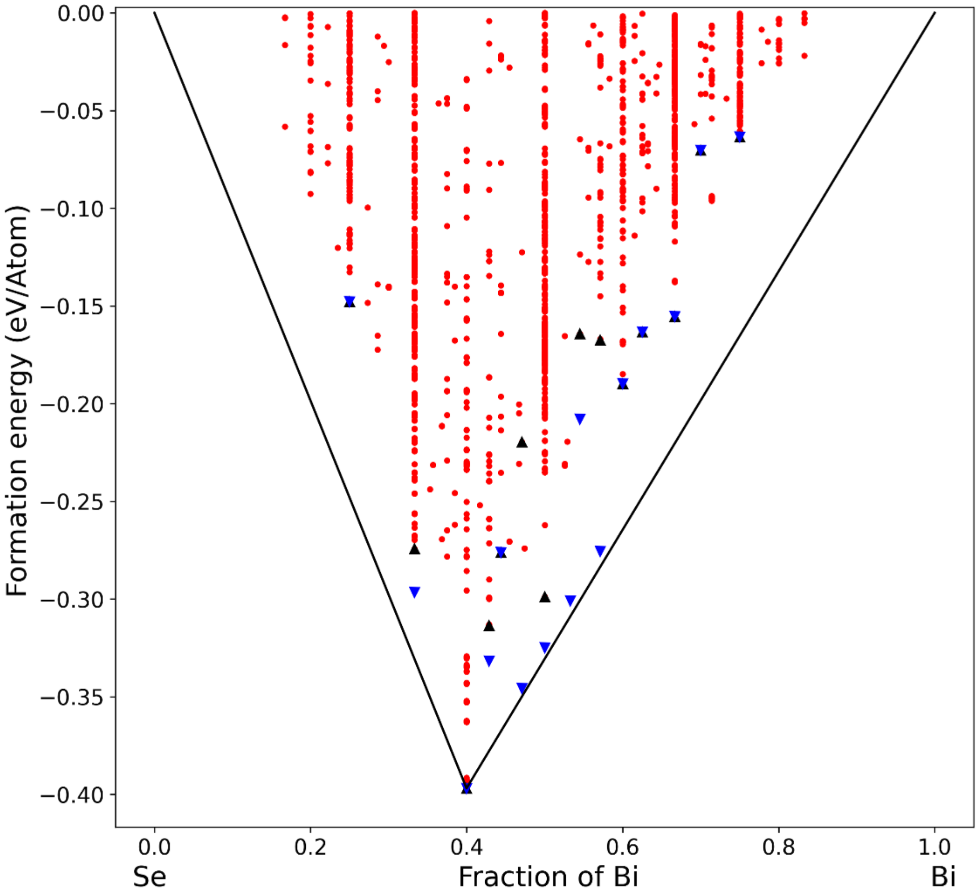
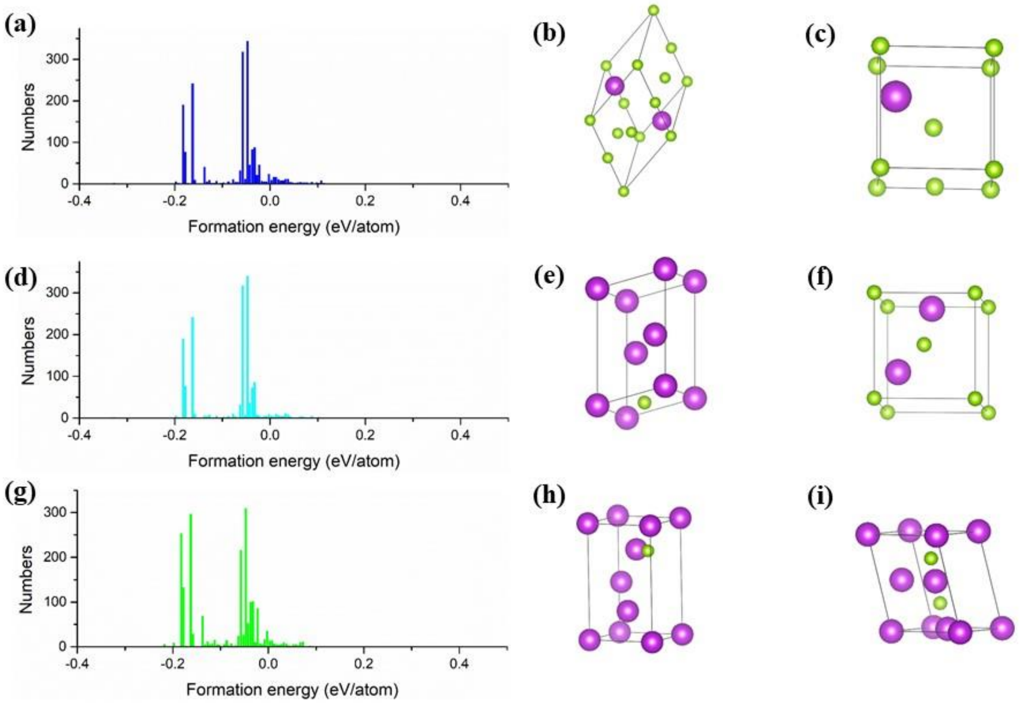
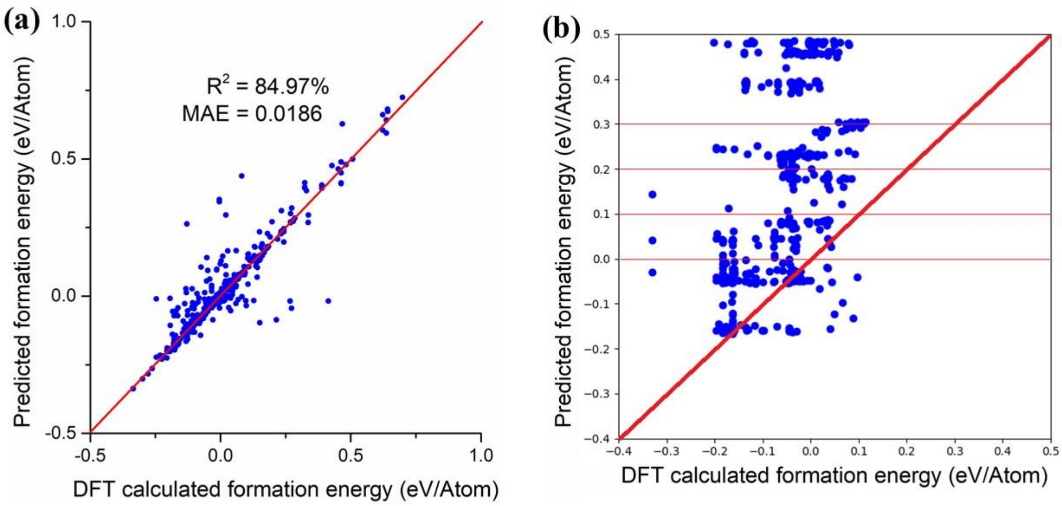
Supplementary Figures



Supplementary Figure 1. Convex hull of Bi-Se system, where the black line represents the convex hull, red points are the structures, blue triangles are the experimentally achievable phases, black triangles are the corresponding structures in the MP database.



Supplementary Figure 2. Distribution and structures of generated data. (a) Distribution of formation energy of DCGAN generated data, where the size of bin is 5 meV; (b) crystal structure of generated Bi_2Se_4; (c) crystal structure of generated Bi_1Se_3 ; (d) Distribution of formation energy of DCGAN + constraint generated data, where the size of bin is 5 meV; (e) crystal structure of generated Bi_3Se_1; (f) crystal structure of generated Bi_2Se_2; (g) Distribution of formation energy of CCDCGAN generated data, where the size of bin is 5 meV; (h) crystal structure of generated Bi_4Se_1; (i) crystal structure of generated Bi_4Se_2.



Supplementary Figure 3. Comparison of formation energy. (a) DFT calculated formation energy vs predicted formation energy for test set of original database; (b) Predicted formation energy of generated structures by DCGAN before structure relaxation vs their corresponding formation energy after relaxation.

Supplementary Tables

Supplementary Table 1. Design of sites autoencoder.

Model type	Layer type	Stride	Activation	Padding	Input size	Output size
Encoder	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	64,64,64	32,32,32
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	32,32,32	16,16,16
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	16,16,16	8,8,8
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	8,8,8	4,4,4
	3D Convolution	1,1,1	tanh	VALID	4,4,4	200
Decoder	3D Convolution		LeakyRELU(0.2)	VALID	200	4,4,4
	3D Convolution		LeakyRELU(0.2)	SAME	4,4,4	8,8,8
	3D Convolution		LeakyRELU(0.2)	SAME	8,8,8	16,16,16
	3D Convolution		LeakyRELU(0.2)	SAME	16,16,16	32,32,32
	3D Convolution		sigmoid	SAME	32,32,32	64,64,64

Supplementary Table 2. Design of lattice autoencoder.

Model type	Layer type	Stride	Activation	Padding	Input size	Output size
Encoder	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	32,32,32	16,16,16
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	16,16,16	8,8,8
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	8,8,8	4,4,4
	3D Convolution	1,1,1	tanh	VALID	4,4,4	200
Decoder	3D Convolution		LeakyRELU(0.2)	VALID	200	4,4,4
	3D Convolution		LeakyRELU(0.2)	SAME	4,4,4	8,8,8
	3D Convolution		LeakyRELU(0.2)	SAME	8,8,8	16,16,16
	3D Convolution		sigmoid	SAME	16,16,16	32,32,32

Supplementary Table 3. Design of generator.

Layer type	Stride	Activation	Padding	Input size	Output size
Fully connected		RELU		200	7,7,128
2D Convolution		RELU	SAME	7,7,128	14,14,128
BatchNormallization(0.8)				14,14,128	14,14,128
2D Convolution		RELU	SAME	14,14,128	28,28,128
2D Convolution		RELU	SAME	28,28,128	28,28,64
BatchNormallization(0.8)				28,28,64	28,28,64
2D Convolution		tanh	SAME	28,28,64	28,28,1

Supplementary Table 4. Design of discriminator.

Layer type	Stride	Activation	Padding	Input size	Output size
2D Convolution	2,2	LeakyRELU(0.2)	SAME	28,28,1	14,14,32
Dropout(0.25)				14,14,32	14,14,32
2D Convolution	2,2	LeakyRELU(0.2)	PADDING	14,14,32	8,8,64
BatchNormallization(0.8)				8,8,64	8,8,64
Dropout(0.25)				8,8,64	8,8,64
2D Convolution	2,2	LeakyRELU(0.2)	SAME	8,8,64	4,4,128
BatchNormallization(0.8)				4,4,128	4,4,128
Dropout(0.25)				4,4,128	4,4,128
2D Convolution	1,1	LeakyRELU(0.2)	SAME	4,4,128	4,4,256
BatchNormallization(0.8)				4,4,256	4,4,256
Dropout(0.25)				4,4,256	4,4,256
Flatten				4,4,256	4096
Fully connected		sgimoid		4096	1

Supplementary Table 5. Design of constraint model.

Layer type	Stride	Activation	Padding	Input size	Output size
2D Convolution	2,2	LeakyRELU(0.2)	SAME	28,28,1	14,14,32
Dropout(0.25)				14,14,32	14,14,32
2D Convolution	2,2	LeakyRELU(0.2)	PADDING	14,14,32	8,8,64
BatchNormallization(0.8)				8,8,64	8,8,64
Dropout(0.25)				8,8,64	8,8,64
2D Convolution	2,2	LeakyRELU(0.2)	SAME	8,8,64	4,4,128
BatchNormallization(0.8)				4,4,128	4,4,128
Dropout(0.25)				4,4,128	4,4,128
2D Convolution	1,1	LeakyRELU(0.2)	SAME	4,4,128	4,4,256
BatchNormallization(0.8)				4,4,256	4,4,256
Dropout(0.25)				4,4,256	4,4,256
Flatten				4,4,256	4096
Fully connected		LeakyRELU(0.2)		4096	1024
Fully connected		LeakyRELU(0.2)		1024	256
Fully connected		LeakyRELU(0.2)		256	256
Fully connected		LeakyRELU(0.2)		256	256
Fully connected		LeakyRELU(0.2)		256	64
Fully connected		LeakyRELU(0.2)		64	1

Table Supplementary 6. Generated distinct structures.

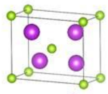
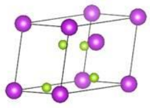
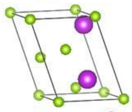
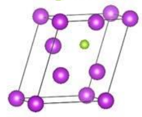
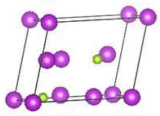
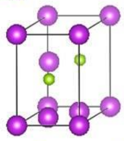
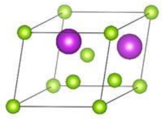
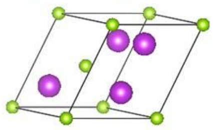
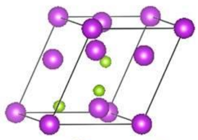
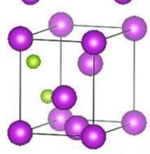
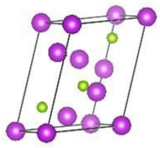
ID	Formula	Formation energy (eV/atom)	Convex hull (eV/atom)	Structure	Space group #
DCGAN Epoch_11731	Bi_4Se_2	-0.1391	0.0798		141
DCGAN Epoch_2368	Bi_2Se_4	-0.2481	0.0803		12
DCGAN Epoch_1651	Bi_2Se_4	-0.2447	0.0837		2
DCGAN Epoch_1343	Bi_5Se_1	-0.021	0.0893		8
DCGAN Epoch_876	Bi_6Se_2	-0.066	0.0981		12
DCGAN Epoch_7598	Bi_4Se_2	-0.1201	0.0988		8
CCDCGAN Epoch_315	Bi_2Se_4	-0.2871	0.0413		2
CCDCGAN Epoch_8688	Bi_4Se_2	-0.1394	0.0795		141
CCDCGAN Epoch_12310	Bi_6Se_3	-0.1314	0.0876		1
CCDCGAN Epoch_4803	Bi_4Se_2	-0.1258	0.0931		10
CCDCGAN Epoch_9982	Bi_6Se_3	-0.1214	0.0975		1

Table Supplementary 7. Compilation of the experimentally achievable (EXP) phases and the corresponding generated phases (CCDCGAN) of the binary Bi-Se system. The fifth column shows the formation energy of the structures with corresponding compositions included in the training. [1]

Phase	Composition % of Bi	Formation energy (EXP) (eV/atom)	Distance to the convex hull (EXP) (eV/atom)	Formation energy (training) (eV/atom)	Formation energy (CCDCGAN) (eV/atom)
$BiSe_3$	0.25	-0.15	0.10	-0.15	-0.15
$BiSe_2$	0.333	-0.30	0.03	-0.27	-0.30
Bi_2Se_3	0.4	-0.39	0	-0.39	-0.39
Bi_3Se_4	0.429	-0.33	0.06	-0.31	-0.33
Bi_4Se_5	0.444	-0.28	0.09	-0.28	-0.28
Bi_8Se_9	0.471	-0.35	0.01	-0.22	-0.22
BiSe	0.5	-0.32	0.01	-0.30	-0.30
Bi_8Se_7	0.533	-0.30	0.01	None	-0.02
Bi_6Se_5	0.545	-0.21	0.13	-0.16	-0.21
Bi_4Se_3	0.571	-0.28	0.01	-0.17	-0.26
Bi_3Se_2	0.6	-0.19	0.07	-0.19	-0.19
Bi_5Se_3	0.625	-0.16	0.08	-0.16	-0.16
Bi_2Se	0.667	-0.16	0.08	-0.16	-0.16
Bi_7Se_3	0.7	-0.07	0.13	-0.07	-0.07
Bi_3Se	0.75	-0.07	0.10	-0.07	-0.07

Supplementary Discussion

Batch training is conducted for both sites autoencoder and lattice autoencoder, the batch size is 822 and 311 respectively, and the epoch used in training are 100 and 200 respectively. Exact model designs are listed in Supplementary Table 1 and 2. The generated vectors are resized into a 28×28 2d graphs where the rest are padding zeros.

Batch training is conducted for DCGAN model, batch size is 128 and total epoch is 500000. Detailed model design is listed in Supplementary Table 3 and 4. The loss function used in the GAN model is "binary_crossentropy", generator loss is to minimize the entropy between 1 and generated structures, while the discriminator loss is to minimize the entropy between 1 and real structures.

Batch training is conducted for constraint model, batch size is 128 and total epoch is 2000. Detailed model design is listed in Supplementary Table 5.

The performance of constraint model is in Supplementary Figure 3 (a). In the DCGAN + constraint model, the selection criterion is 0.3 eV/atom. The selection is calculated for two reasons: first, the MAE of the model is not exactly 0, so cutting at 0 eV/atom is highly likely to drop some reasonable structures; second, DFT is highly likely to reduce the formation energy of generated structures through relaxation as Supplementary Figure 3 (b) demonstrates, so we select this number to balance these two factors. And the effects of constraint in the latent space are demonstrated in movie file Supplementary Movie 1.

All parameters and designs are the same as DCGAN, except the loss of generator, it is the combination of "binary_crossentropy" and the average of generated structures predicted by the constraint.

Supplementary Reference

1. *ASM handbook*. (ASM International, 1990).