Supplementary

Inverse design of crystals using generalized invertible crystallographic representation

S1 Latent space of trained VAE

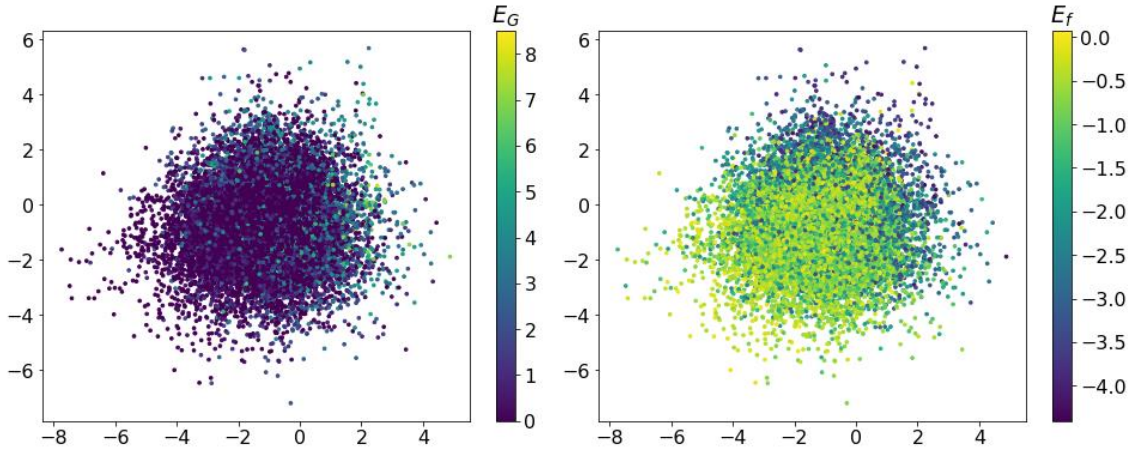
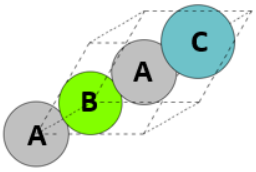


Figure S1 The first two dimensions of latent space (256 dimensions in total). Different colors represent different materials properties.

Figure S1 shows the first two latent distribution with different materials properties. By jointly training the VAE and gated regression, the latent space becomes an organized and continuous crystal representation with different material properties. Inverse design of new crystals leverages this structured latent space by sampling in regions with desired properties.

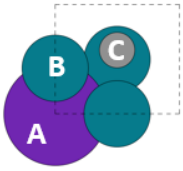
S2 14 generated crystals that are not in the database

Class 1: A_2BC-type



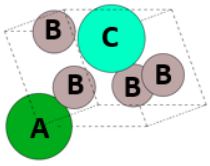
A	B	C
Ag	Mg	Nb
Ru	Li	Zr
Sr	Cu	Pb
Al	Ac	Os

Class 2: AB_3C-type



A	B	C
Rb	Rh	C
Ce	Ge	Fe

Class 3: AB_4C-type



A	B	C
Lu	Al	Dy
Pm	Al	Ho
La	Al	Dy

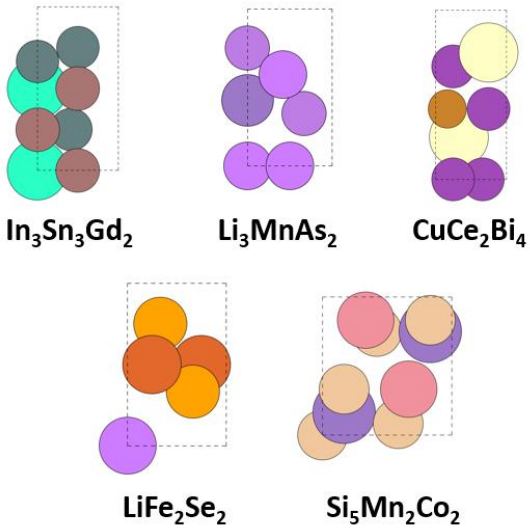


Figure S2 14 generated crystals that don't exist in the dataset

Figure S2 shows the generated crystal structure of 14 unique crystals that are generated by the VAE. There are 8 different crystal structure with >30 different elements in those 14 crystals. This shows the VAE can generate crystals accessing a wide range of structures and chemistries.

S3 DFT validation of generated crystals with different Ef values

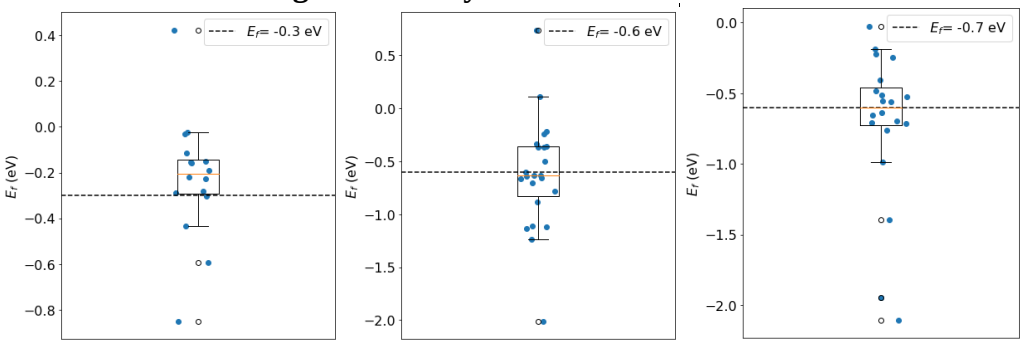


Figure S3 DFT calculated material property for 80 generated crystals with different E_f values

Figure S3 shows 3 boxplots of generated crystals with different E_f values. 20 out of those 80 compounds achieve targeted E_f value within an error of 0.1 eV.

S4 Dissimilarity value of 97 generated crystals

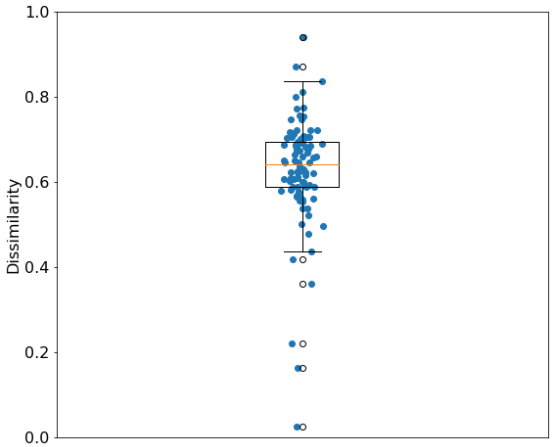


Figure S4 dissimilarity value of 97 generated compounds compared to all training point

We use dissimilarity value to assess the structural uniqueness of those generated crystals for E_f validation. The dissimilarity value is the vector distance between two structures based on local coordination information from all sites in the two structures[1]. Figure S5 shows the minimum dissimilarity value compared to every crystal in the training dataset. The median dissimilarity value is 0.64 with a number of compounds have a dissimilarity value above 0.75 which is the cut off dissimilarity value in Materials Project website.

S5 Distribution of Ef and Eg in the training dataset

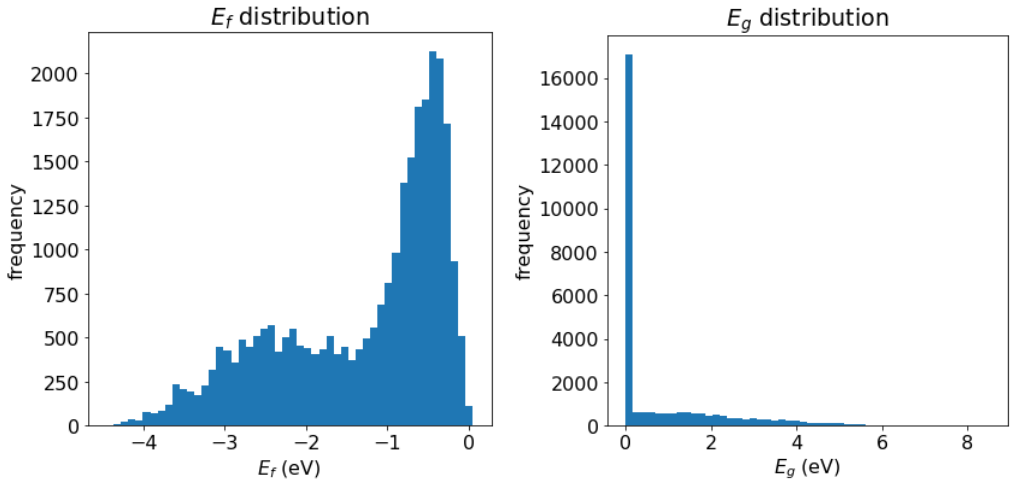


Figure S5 Histogram of E_f and E_g in the compounds used in training the VAE to generate crystals with targeted properties (E_f < 1.5 eV and E_g =1.5eV).

As shown in Figure S5, the number of crystals that satisfy the targeted properties (E_f < 1.5 eV and E_g =1.5eV) are quite low. Especially for the bandgap (E_g), more than 60% of compounds in the training dataset has E_g of zero. The probability of random sampling crystals that meet the property requirement is less than 1% while our method reports a success rate of 37.5% (6 out of 16).

S6 Machine learning model details

The structure of 1D convolution variational autoencoder is described in this section. The encoder consists of 3 1D convolutional layers with kernel size {5,3,3}, strides {2,2,1} and the number of channels {32,64,128}. Batch normalization and Leaky-ReLU activation with parameters of 0.2 are used between convolutional layers. After the output of the convolutional layers is flattened and connected to 2 Dense layers {1024, 256}. The latent space has 256 dimensions. The decoder has a mirrored structure with the same kernel sizes, strides and number of channels. The gated regression consists of 2 dense layers with {128,32} with sigmoid activation. For training of VAE, we use RMSprop optimizer with a batch size of 128 and learning rate of 0.008.

Reference:

[1] N. E. Zimmermann and A. Jain, "Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity," *RSC Advances*, vol. 10, no. 10, pp. 6063-6081, 2020.