Electronic Supplementary Material (ESI) for Materials Horizons.

Supplementary materials

Machine-learning-assisted search for functional materials over extended chemical space

Vadim Korolev,*[a,b] Artem Mitrofanov [a,b], Artem Eliseev [b,c] and Valery Tkachenko [a]

[a] Science Data Software, LLC, 14909 Forest Landing Circle, Rockville, Maryland 20850, USA.
[b] Department of Chemistry, Lomonosov Moscow State University, Leninskie gory 1/3, Moscow 119991, Russia.
[c] Department of Materials Science, Lomonosov Moscow State University, Leninskie gory 1/73, Moscow 119991, Russia.

* Address correspondence to: korolev@colloid.chem.msu.ru

Theoretical Details

Structure representation via concatenated latent spaces

We use generative neural networks with complex architecture known as variational autoencoders[1] (VAEs) to convert the sparse representation of chemical structures (stoichiometric composition and XRD pattern) to denser one, i.e., with lower dimension. VAE consists of two neural networks: encoder and decoder. Encoder $Q(z|x)$ takes input data x (normalized stoichiometric composition or XRD pattern) and returns its dense representation z (point in the "latent space"). For clarity, initial 100-dimensional vectors of stoichiometric composition are converted to 24-dimensional latent space; 1601-dimensional vectors of XRD patterns are converted to 64-dimensional latent space. The decoder takes z as input and returns image x^* of initial data x. Intrinsic parameters of encoder and decoder (its weights and biases) are optimized to reproduce initial data by its image. "Training" VAEs is aimed at minimizing loss function represented for single structure as follows:

$$l_i = -E\left[\log D(x_i|z)\right] + D_{KL}\left[Q(z|x_i)||p(z)\right]$$

where E — expected negative log-likelihood, D_{KL} — the Kullback-Leibler divergence between the encoder's distribution $Q(z|x)$ and standard normal distribution $p(z)$. Total loss for a set of structures equals to the sum of absolute values of l_i. E is a "reconstruction loss" forcing the decoded data x^* to match initial inputs x. Value of D_{KL} reflects the similarity of encoder's distribution to normal distribution $p(z)$. This regularization term distinguishes VAEs from "vanilla" autoencoders that are not suitable for sampling new points from latent space.

Architectures of compositional and structural VAEs implemented with Keras package are presented in Fig. S1. The architecture of structural VAE was inspired by the structure of variational convolutional autoencoder for stellar spectra analysis (from astroNN package[2]). Hyperparameters of both models were optimized with the Hyperopt framework[3]. Permissible ranges and optimal values are presented in Table S3.

To verify trained VAEs, we inspected a set of structures from the Materials Project database that were not using during VAEs training.

Target property optimization via concatenated latent spaces

Taken into account all previous observations (pairs (x_i, y_i), where x_i is a point in latent space and y_i is a corresponding value of an optimized variable), sequential model-based optimization (SMBO) methods provide new points in configurational space that potentially satisfy target functionality, i.e. minimum/maximum value of an optimized variable[4]. Tree-structured Parzen estimator (TPE) strategy models $p(x|y)$ and $p(y)$, whereas most other SMBO algorithms, e.g., Gaussian-process (GP) based, model $p(y|x)$ directly. Tree-based SMBO methods often outperform not only random search but also GP based methods, especially if search space contains many conditional/categorical hyperparameters. Moreover, the runtime of each iteration of TPE optimization can scale linearly in configurational space, whereas the runtime of each iteration of GP optimization scales cubically[4].

The hyperparameter space is described using uniform distributed variables while original latent variables are normally distributed. TPE approach suggests the following replacement: uniform → truncated Gaussian mixture. The posterior distribution $p(x|y)$ in modified configuration space defines through two non-parametric densities as follows:

$$p(x|y) = \begin{cases} l(x), & if \ \ y < y^* \\ g(x), & if \ \ y \geq y^* \end{cases}$$

where y^* is usually defined as a fixed quantile of the observed losses. TPE optimization algorithm may be summarized as a two-step iterative procedure:

- Evaluate the loss function in the point of the configuration space where ratio $\frac{l(x)}{g(x)}$ reaches its maximum;
- (Re-)calculate $l(x)$ and $g(x)$ densities.

We illustrate the optimization performance of two above-mentioned Bayesian approaches in comparison with a random search on several 2-dimensional artificial test functions. TPE and GP algorithms are implemented with the Hyperopt framework[3] and the scikit-optimize library (https://doi.org/10.5281/zenodo.1207017), correspondingly. Benchmarking results are presented in Fig. S8. While both optimization strategies demonstrate comparable convergence, TPE seems preferable due to its better scalability.

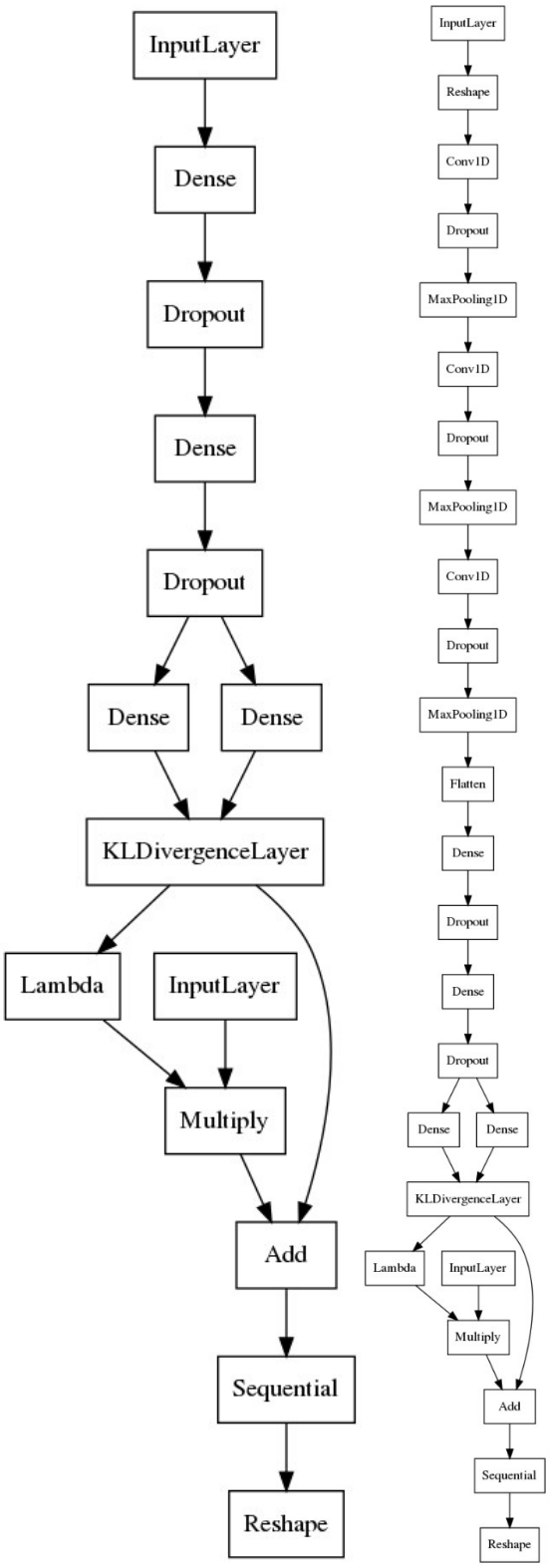


Fig. S1. Architectures of compositional (left) and structural (right) VAEs.

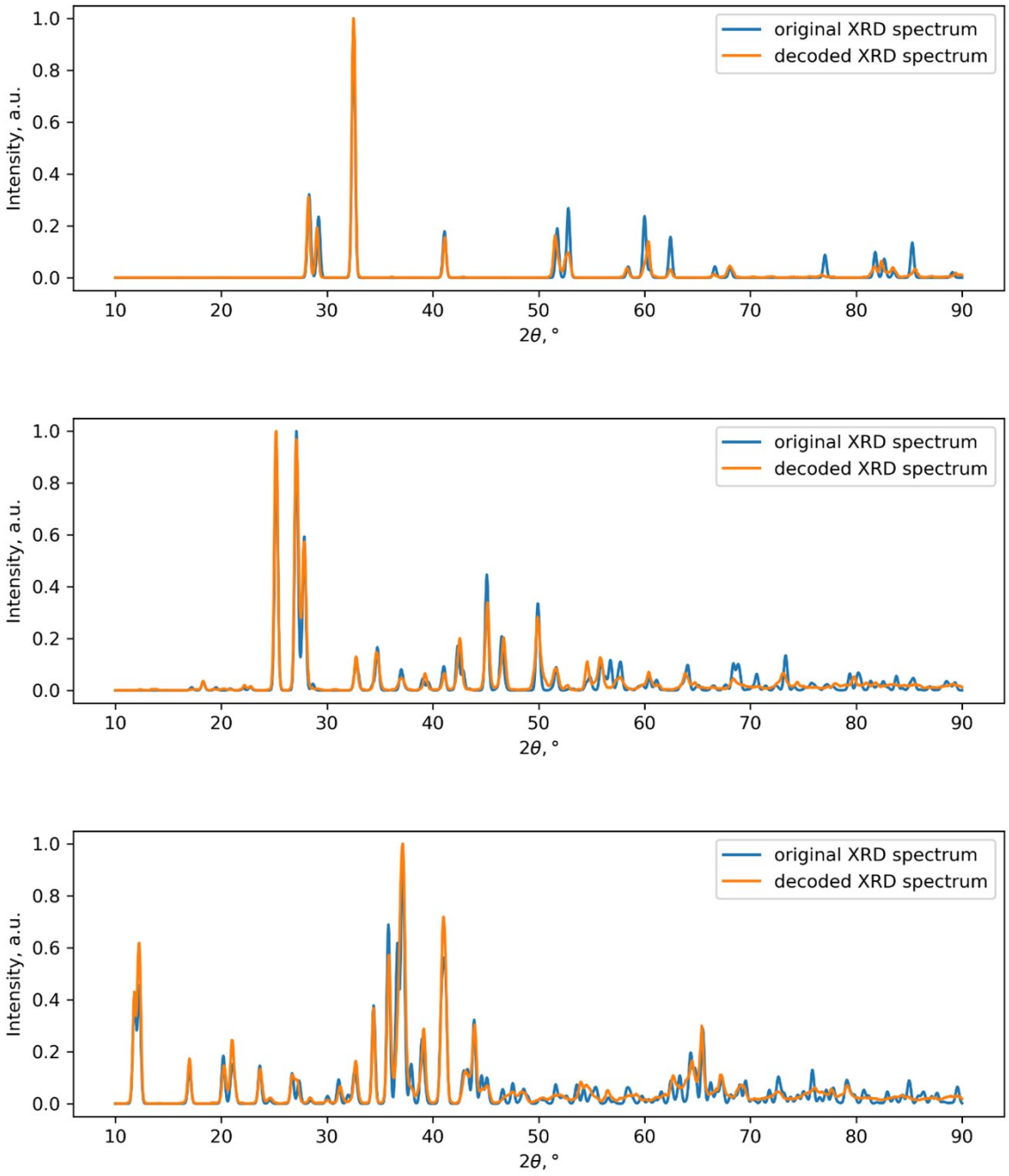


Fig. S2. Structural VAE verification. Minor set of structures from the Materials Project database was not used to train VAEs. We test the validity of trained models via visual inspection of original XRD spectra and its decoded image.

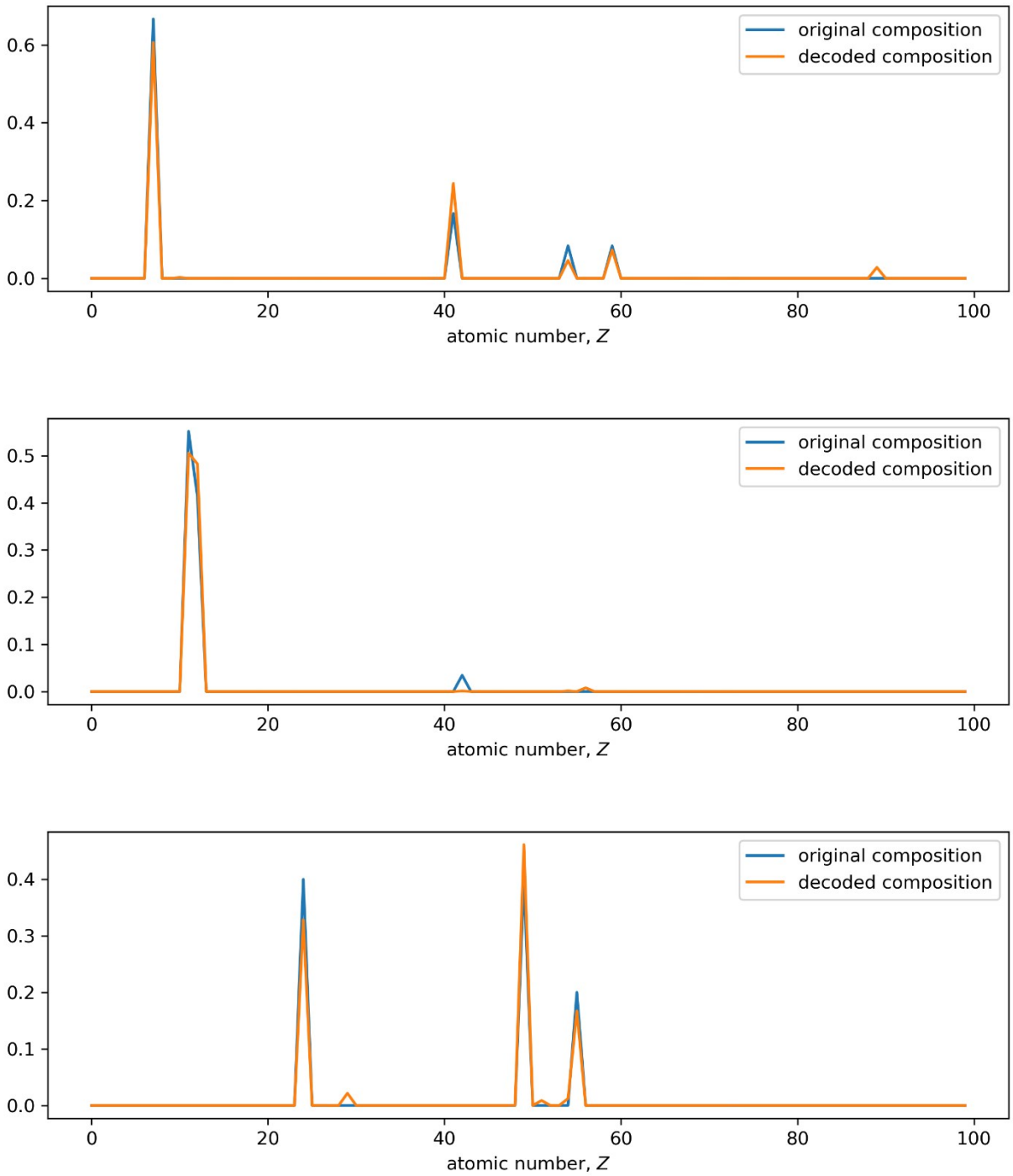


Fig. S3. Compositional VAE verification. Minor set of structures from the Materials Project database was not used to train VAEs. We test the validity of trained models via visual inspection of original compositions and its decoded image.

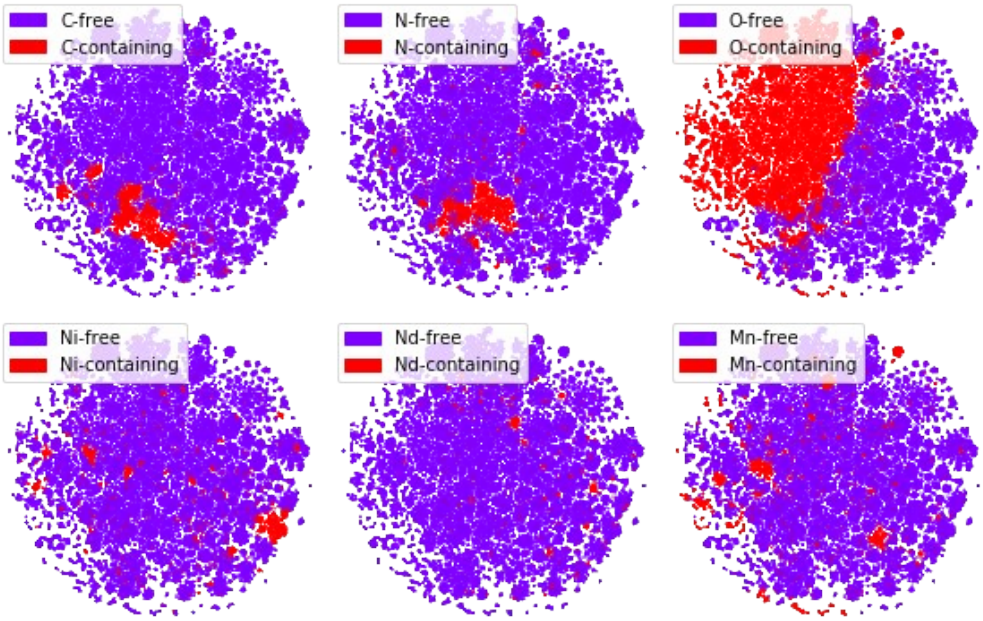


Fig. S4. Two-dimensional t-distributed stochastic neighbour embedding (t-SNE) projection of compositional latent (sub)space variables of structures from the Materials Project database. Structures contained corresponding element are marked in red. All considered elements (except for oxygen) form several distinct clusters that correspond to different chemical classes of compounds.

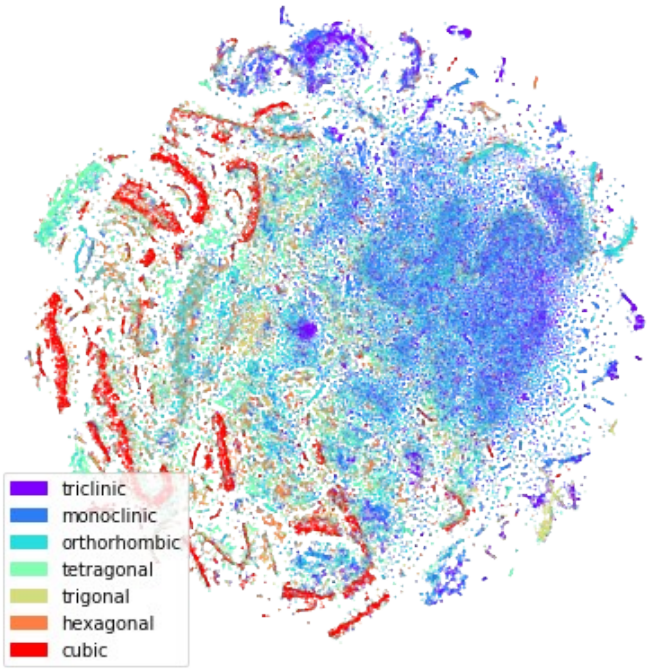


Fig. S5. Two-dimensional t-distributed stochastic neighbour embedding (t-SNE) projection of structural latent (sub)space variables of structures from the Materials Project database. All structures are marked in accordance with its crystal symmetry. Structures with trigonal, hexagonal, and cubic crystal symmetry form well-resolved clusters. Structures with triclinic, monoclinic, orthorhombic, and tetragonal crystal symmetry form poorly separable groups (at least in two-dimensional projection). As in the case of chemical latent (sub)space (structures contained specific chemical element), each of crystal symmetry forms several distinct clusters due to the structural diversity of corresponding space groups.

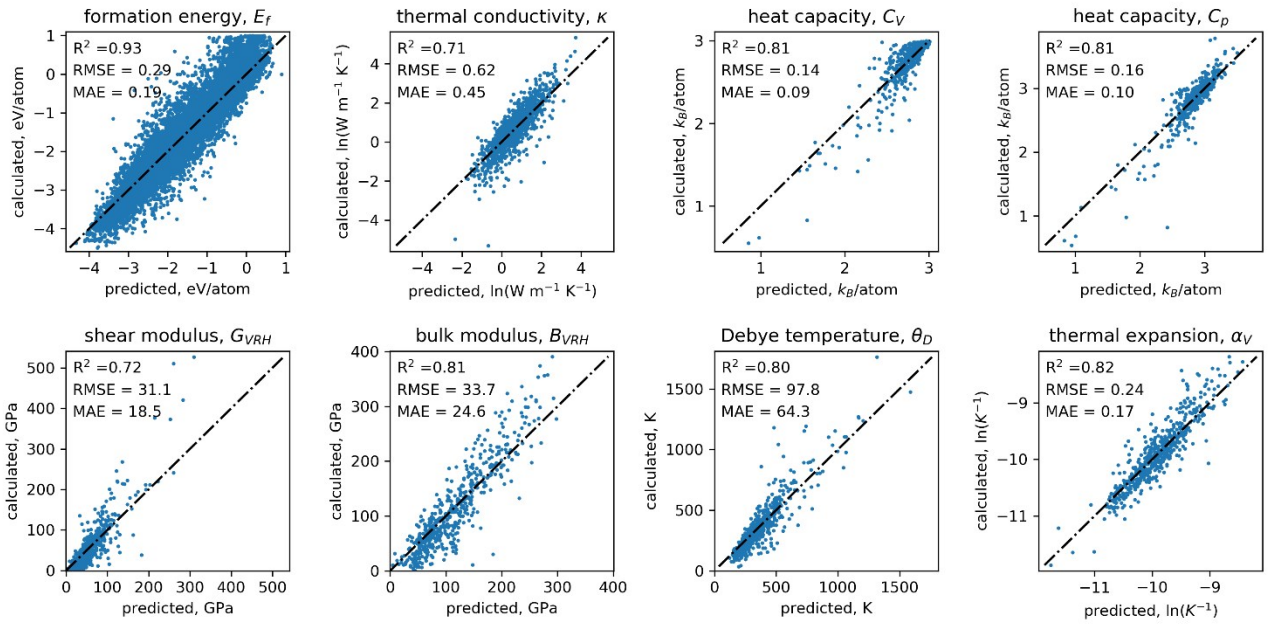


Fig. S6. Predicted vs. calculated values of physicochemical properties used to verify proposed structure representation (regression tasks).

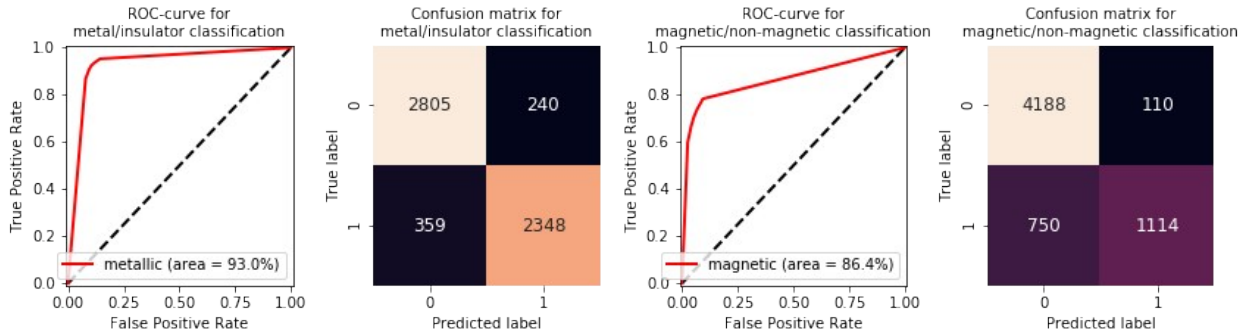
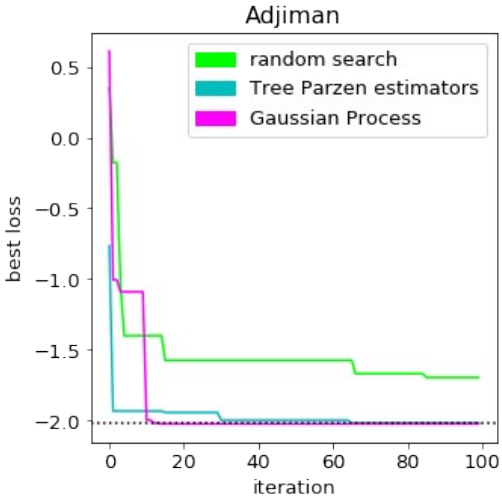
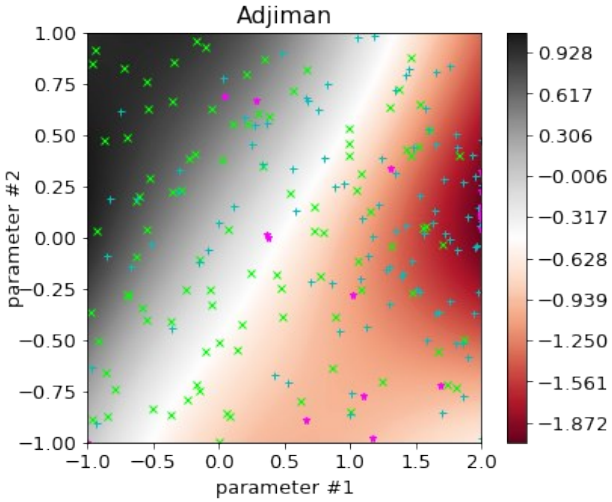
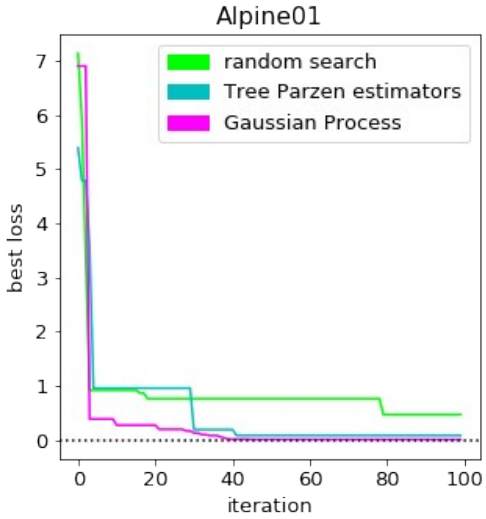
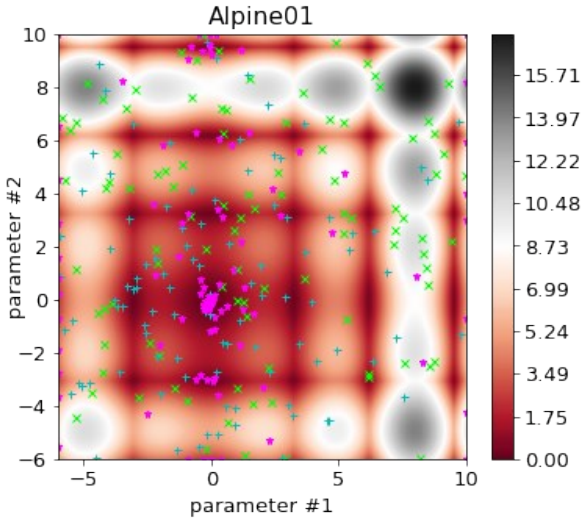


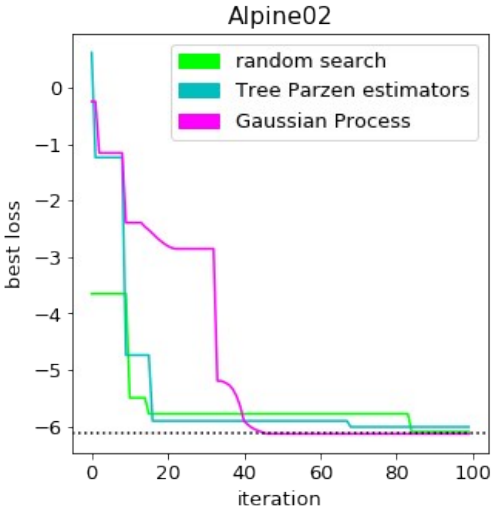
Fig. S7. Predicted vs. calculated values of physicochemical properties used to verify proposed structure representation (classification tasks).

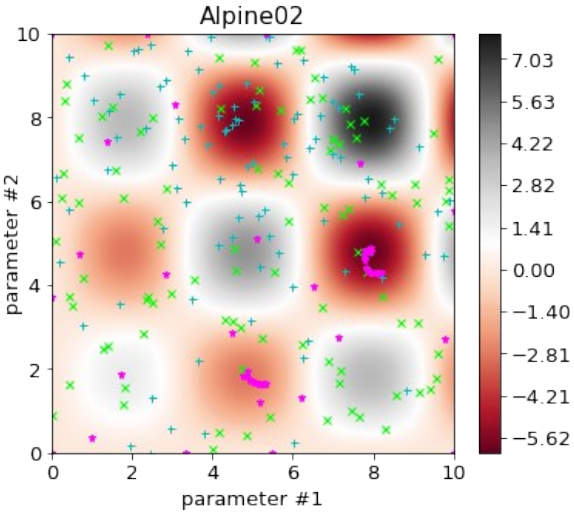

Adjiman
random search
Tree Parzen estimators
Gaussian Process
best loss
iteration


Adjiman
parameter #2
parameter #1


Alpine01
random search
Tree Parzen estimators
Gaussian Process
best loss
iteration


Alpine01
parameter #2
parameter #1


Alpine02
random search
Tree Parzen estimators
Gaussian Process
best loss
iteration


Alpine02
parameter #2
parameter #1

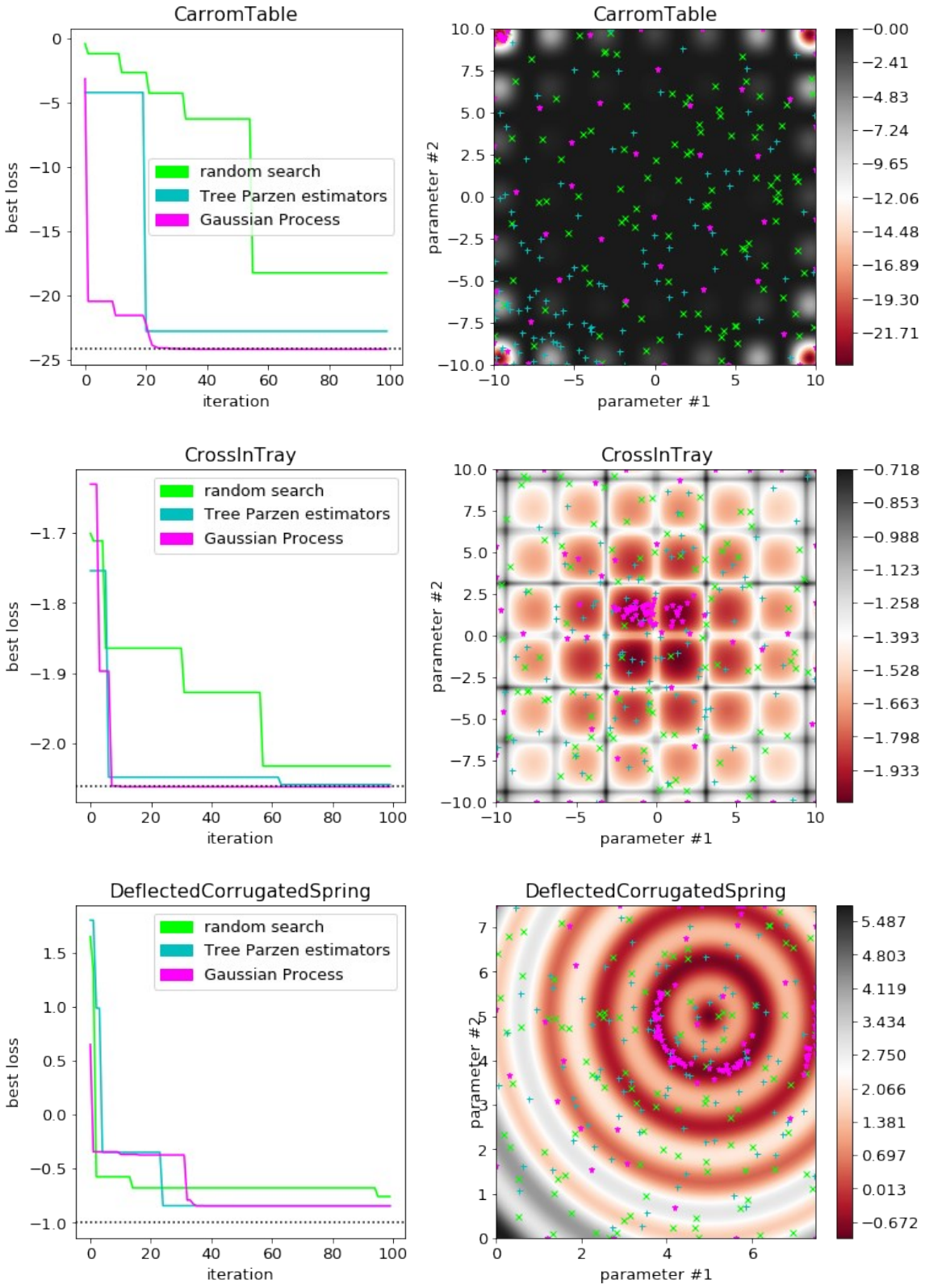


Fig. S8. Progress of optimization runs for 2-dimensional test functions with three strategies: random search, Gaussian process, and Tree-Structured Parzen Estimators.

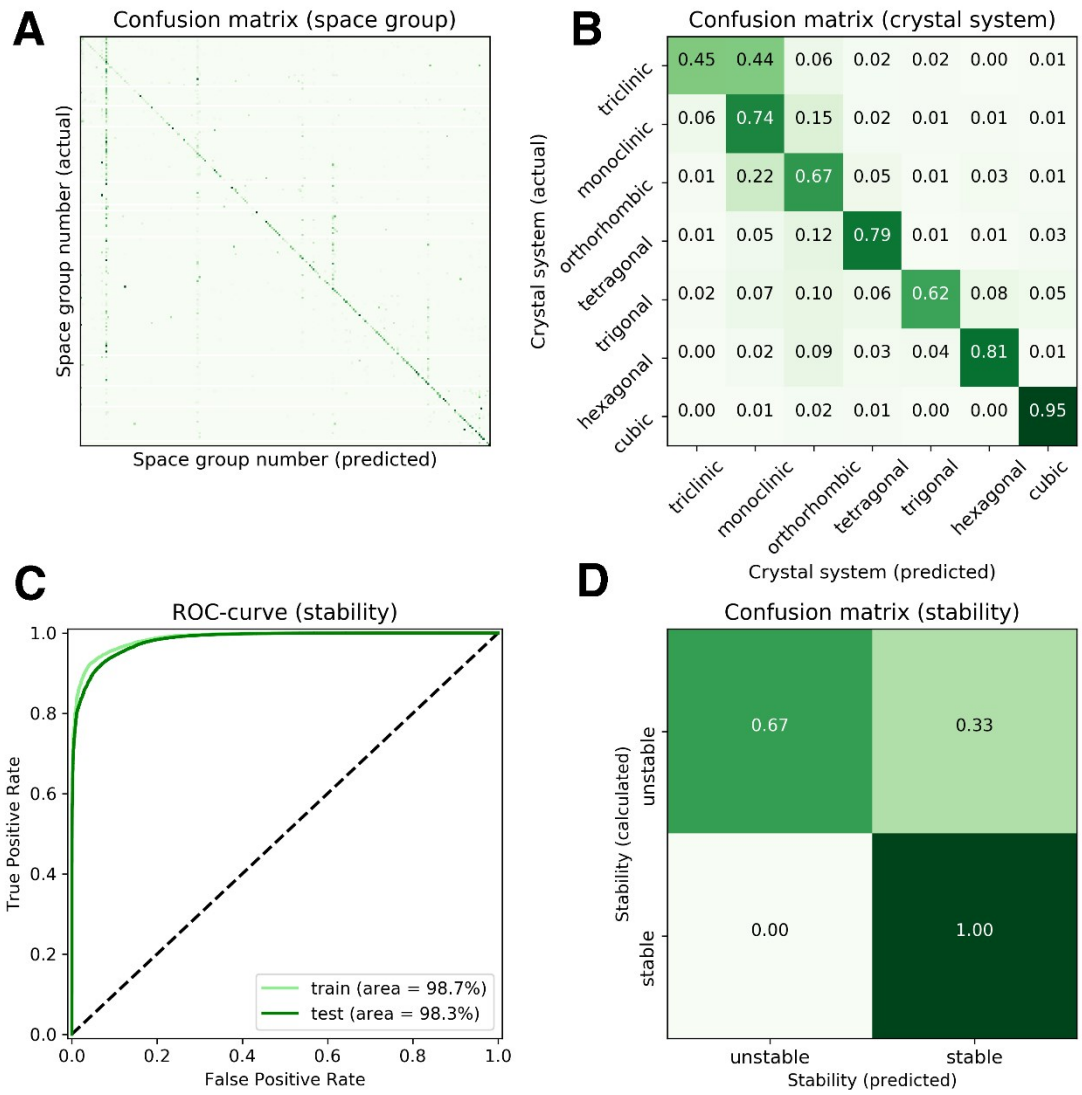


Fig. S9. Performance metrics of filters. (A) The confusion matrix for the space group classification. Since the number of classes exceeds two hundred, this representation looks sparse and unrepresentative. (B) The confusion matrix for the crystal system classification. We do not build a separate predictive model, output signals of the space group classifier are summed up following which crystal system a given space group belongs to. The average values of accuracy for space groups and crystal system predictions are 0.54 and 0.72, respectively. These values are significantly lower than those presented in the original study[5]. On the other hand, a similar accuracy of around 54% was obtained on experimental data[6]. The presented performance metrics refer to diffraction patterns that have passed through a convolutional autoencoder used to construct the structural subspace. (C) Receiver operating characteristic (ROC) curve and (D) confusion matrix for stability classifier. The model shows excellent predictive power with the area under the ROC curve at 0.98 and accuracy at 0.96 on an external test set.

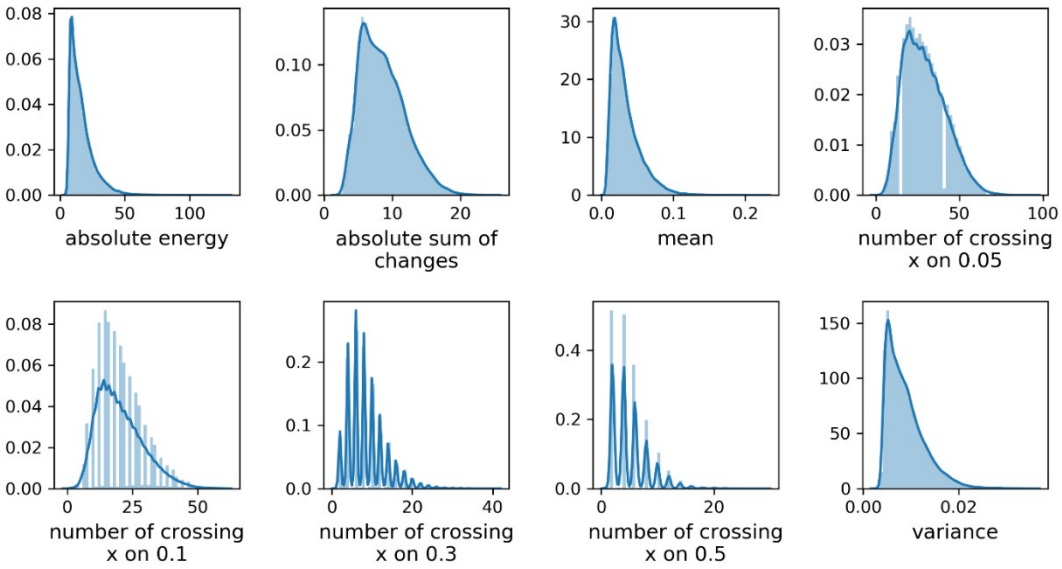


Fig. S10. The distributions of "time series" characteristics for structures from the Materials Project Database.

Table S1. Ranges and optimal values of hyperparameters used to build best-performance models.

Model	Hyperparameter	Distribution type	Range	Optimal value
Structural VAE	filters_1	integers	8…16	16
	filters_2	integers	16…32	17
	filters_3	integers	32…64	37
	kernel_size	integers	2…5	4
	pool_size	integers	2…5	2
	units_1	integers	512…1024	693
	units_2	integers	256…512	403
	dropout	uniform	[0, 0.06]	2.04×10^{-2}
	regularizer_l1	log-uniform	$[10^{-7}, 3\times10^{-5}]$	1.46×10^{-6}
	regularizer_l2	log-uniform	$[10^{-7}, 3\times10^{-5}]$	6.42×10^{-7}
	kernel_initializer	categories	[lecun_normal, he_normal, he_uniform]	lecun_normal
	activation	categories	[tanh, selu, elu, relu]	selu
	optimizer	categories	[sgd, rmsprop, adagrad, adadelta, adam, adamax, nadam]	adamax
Compositional VAE	units_1	integers	48…96	95
	units_2	integers	20…64	50
	dropout	uniform	[0, 0.1]	4.49×10^{-6}
	regularizer_l1	log-uniform	$[10^{-7}, 3\times10^{-5}]$	1.94×10^{-7}
	regularizer_l2	log-uniform	$[10^{-7}, 3\times10^{-5}]$	1.09×10^{-6}
	kernel_initializer	categories	[lecun_normal, he_normal, he_uniform]	he_normal
	activation	categories	[tanh, selu, elu, relu]	selu
	optimizer	categories	[sgd, rmsprop, adagrad, adadelta, adam, adamax, nadam]	nadam

Table S2. Ranges and optimal values of hyperparameters used to build best-performance models.

Model	**Hyperparameter**	**Distribution type**	**Range**	**Optimal value**
Space group classifier	filters_1	integers	8…128	104
	filters_2	integers	8…128	51
	filters_3	integers	8…128	117
	strides_1	integers	1…3	2
	strides_2	integers	1…3	1
	strides_3	integers	1…3	2
	kernel_size_1	integers	4…80	34
	kernel_size_2	integers	8…40	37
	kernel_size_3	integers	8…20	13
	units_1	integers	256…2500	2047
	units_2	integers	128…1600	943
	pool_size	integers	2…4	3
	conv_dropout	uniform	[0, 0.85]	0.498
	conn_dropout	uniform	[0, 0.75]	0.675
	kernel_initializer	categories	[lecun_normal, he_normal]	lecun_normal
	activation	categories	[tanh, selu, elu, relu]	relu
	optimizer	categories	[adagrad, adadelta, adam, adamax]	adamax
Formation energy predictor	units	integers	8…256	171
	dropout_rate	uniform	[0, 0.75]	0.347
	kernel_initializer	categories	[lecun_normal, he_normal]	lecun_normal
	activation	categories	[tanh, selu, elu, relu]	elu
	optimizer	categories	[adagrad, adadelta, adam, adamax]	adam

Bulk/shear modulus predictor	max_depth	integers	3…11	6
	alpha	log-uniform	$[10^{-4}, 10^{2}]$	1.91
	gamma	log-uniform	$[10^{-4}, 1]$	0.783
	lambda	log-uniform	[1, 10]	9.01
	eta	log-uniform	$[10^{-3}, 10^{-1}]$	2.12×10^{-2}
	subsample	uniform	[0.3, 1.0]	0.846
	colsample_bytree	uniform	[0.2, 1.0]	0.888
	colsample_bylevel	uniform	[0.2, 1.0]	0.334

Table S3. Summary of performance (regression tasks). Corresponding XGBoost models were trained on concatenated latent spaces.

endpoint	number of structures	source of data	performance metrics (this study)	performance metrics (benchmarking)
formation energy, eV/atom	130998	Materials Project[7]	$R^2 = 0.93$ MAE = 0.19 RMSE = 0.29	MAE = 0.03[8]
thermal conductivity at 600 K, ln(W/m×K)	5511	AFLOWLIB[9]	$R^2 = 0.71$ MAE = 0.45 RMSE = 0.62	
heat capacity at constant volume, k_B/atom	2721	AFLOWLIB[9]	$R^2 = 0.81$ MAE = 0.09 RMSE = 0.14	$R^2 = 0.95$[10] MAE = 0.04[10] RMSE = 0.07[10]
heat capacity at constant pressure, k_B/atom	2721	AFLOWLIB[9]	$R^2 = 0.81$ MAE = 0.10 RMSE = 0.16	$R^2 = 0.95$[10] MAE = 0.05[10] RMSE = 0.09[10]
bulk modulus, GPa	2721	AFLOWLIB[9]	$R^2 = 0.81$ MAE = 24.6 RMSE = 33.7	$R^2 = 0.97$[10] MAE = 8.7[10] RMSE = 14.3[10]
shear modulus, GPa	2721	AFLOWLIB[9]	$R^2 = 0.81$ MAE = 18.5 RMSE = 31.1	$R^2 = 0.88$[10] MAE = 10.6[10] RMSE = 18.4[10]
Debye temperature, K	2721	AFLOWLIB[9]	$R^2 = 0.80$ MAE = 64.3 RMSE = 97.8	$R^2 = 0.95$[10] MAE = 35.9[10] RMSE = 57.0[10]
coefficient of thermal expansion, ln(1/K)	2721	AFLOWLIB[9]	$R^2 = 0.82$ MAE = 0.17 RMSE = 0.24	$R^2 = 0.91$ (1/K)[10]

Table S4. Summary of performance (classification tasks). Corresponding XGBoost models were trained on concatenated latent spaces.

endpoint	**number of structures**	**source of data**	**performance metrics (this study)**	**performance metrics (benchmarking)**
band gap, eV (cutoff 0 eV)	28737	AFLOWLIB[9]	AUC ROC = 0.93 accuracy = 0.90 F1 score = 0.87	AUC ROC = 0.98[10] accuracy = 0.93[10]
magnetic moment, μ_B (cutoff 0 μ_B)	30787	JARVIS-DFT[11]	AUC ROC = 0.86 accuracy = 0.86 F1 score = 0.72	AUC ROC = 0.96[12]

References

1 D. P. Kingma and M. Welling, *arXiv Prepr. arXiv1312.6114*.

2 H. W. Leung and J. Bovy, *Mon. Not. R. Astron. Soc.*, 2018, **483**, 3255–3277.

3 J. Bergstra, D. Yamins and D. D. Cox, *Proc. 30th Int. Conf. Mach. Learn.*, 2013, 115–123.

4 J. S. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, in *Advances in neural information processing systems*, 2011, pp. 2546–2554.

5 W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin and K.-S. Sohn, *IUCrJ*, 2017, **4**, 486–494.

6 P. M. Vecsei, K. Choo, J. Chang and T. Neupert, *Phys. Rev. B*, 2019, **99**, 245120.

7 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

8 C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain and G. Ceder, *npj Comput. Mater.*, 2020, **6**, 1–11.

9 S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo and O. Levy, *Comput. Mater. Sci.*, 2012, **58**, 227–235.

10 O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, *Nat. Commun.*, 2017, **8**, 1–12.

11 K. Choudhary, I. Kalish, R. Beams and F. Tavazza, *Sci. Rep.*, 2017, **7**, 1–16.

12 K. Choudhary, B. DeCost and F. Tavazza, *Phys. Rev. Mater.*, 2018, **2**, 1–8.