PCVAE: A Physics-informed Neural Network for Determining the Symmetry and Geometry of Crystals

Ke Liu*†, Shangde Gao*, Kaifan Yang*, and Yuqiang Han*†‡
*College of Computer Science and Technology, Zhejiang University, China
†ZJU-Hangzhou Global Scientific and Technological Innovation Center, China
{lk2017,gaosde,yangkaifan,hyq2015}@zju.edu.cn

***Abstract*—The symmetry and geometry of a crystal fundamentally determine its various physical and chemical properties. However, crystal structure prediction, including space group determination and crystal structure optimization, remains an ongoing challenge because traditional DFT-based approaches are time- and computational-intensive even for one specific set of material, not to mention structure prediction of massive materials. This paper determines the geometric structure of massive crystals solely from chemical formulae from scratch. In addition, due to various phases or changing environmental conditions, different pressure and temperature for example, there could be multiple crystal structures corresponding to one chemical formula (MS4OF), which has been overlooked or poorly addressed in previous research. Hereby, we propose a Physics-informed Conditional Variational Auto Encoder (PCVAE) to encode possible symmetry and geometry distribution as well as various phases of a crystal with Gaussian distributions. PCVAE achieves a new state-of-the-art in crystal structure prediction. Extensive experiments demonstrate the strong predictive power of PCVAE. The code and datasets are available at https://github.com/zjuKeLiu/PCVAE.**

I. Introduction

Crystal structure plays a vital important role in determining physical and chemical properties. However, crystal structure prediction (CSP) remains to be an open challenge [1]–[3]. The structure of one crystal material consists of a unit cell and the collection of symmetry operations of the unit cell, i.e., space group g. Specifically, the unit cell is defined as a parallelepiped, which can be determined by six lattice constants, including the lengths of three cell edges, (a, b, c) and the angles between them, i.e., (α, β, γ).

The symmetry of a crystal is defined by the space group. Crystals have a total of 230 different space groups, which can be grouped into 14 Bravais lattice types and 7 lattice systems [4]. The mapping relations exist among the space groups, Bravais lattices, and crystal systems, where the space group is the most fine-grained category among them. Once the space group is given, the bravais lattice and crystal system are determined. Hence, CSP seeks to predict the six lattice constants and the space group from a given chemical formula.

Traditional computational approaches, such as molecular dynamics (MD) and density functional theory (DFT), are arduous and computationally expensive. They calculate the energy of randomly generated crystal structure candidates and optimize the structures [5].

‡Corresponding author: Yuqiang Han.

TABLE I
Part of phases of $Li_{18}Mn_4Co_{10}O_{32}$ in the same Bravais lattice type, Monoclinic(C). α and γ for all the phases are 90°

space group	a(Å)	b(Å)	c(Å)	β(°)
8	**5.1223**	2.8688	39.1435	91.1565
8	**38.9833**	2.8714	5.1407	92.6744
12	5.2723	**2.8439**	39.2574	90.3422
12	10.1598	**11.6879**	5.1606	109.8335
12	19.6491	5.8773	**5.0044**	100.7293
8	5.2135	2.8815	**39.4486**	90.8430
12	5.2245	2.8728	38.8203	**90.0608**
5	10.2583	11.6423	5.9102	**124.7122**

The majority of current machine learning-based crystal structure prediction algorithms have poor generalization performance since they are restricted to specific crystal families, e.g., the same mole ratio, the same space group [6]–[11].Thus the datasets in these works are often small. Random hold-out or cross-validation methods are frequently used to train and evaluate these models. Despite the remarkable precision of these works, the models cannot be generalized to other materials.

Besides, the **MS4OF** problem, i.e. Multiple Structures exist For One chemical Formula, has been overlooked or poorly addressed in previous works [12]. For example, the stable structures are of Ce_4Se_8 and Si_8O_{16} scattered throughout more than 10 Bravais lattice types. $Li_{18}Mn_4Co_{10}O_{32}$ and $Li_8Mn_2Co_4O_{14}$ have more than 120 stable structures in one Bravais lattice type. Previous machine learning models for determining the geometry of crystals can only predict one unique set of lattice constants, because the input features are the same for different structures of one chemical formula with different crystal structures [5], [13].

MS4OF problem cannot be ignored because the lattice constants vary over a wide range. Take the chemical formula $Li_{18}Mn_4Co_{10}O_{32}$ for example, part of its stable structures are shown in Table I. The lattice constants, a, b, c, β vary in [5.1223, 38.9833], [2.8439, 11.6879], [5.0044, 39.4486], and [90.0608, 124.7122] respectively. [5], [13] predict one set of lattice constants with prior knowledge of space group, i.e. different sets of lattice constants are predicted for the same chemical formula with different space groups. However, Multiple stable structures exist even in the same space group. The MS4OF problem remains to be unsolved. To the best of our knowledge, our model is the first work in tackling

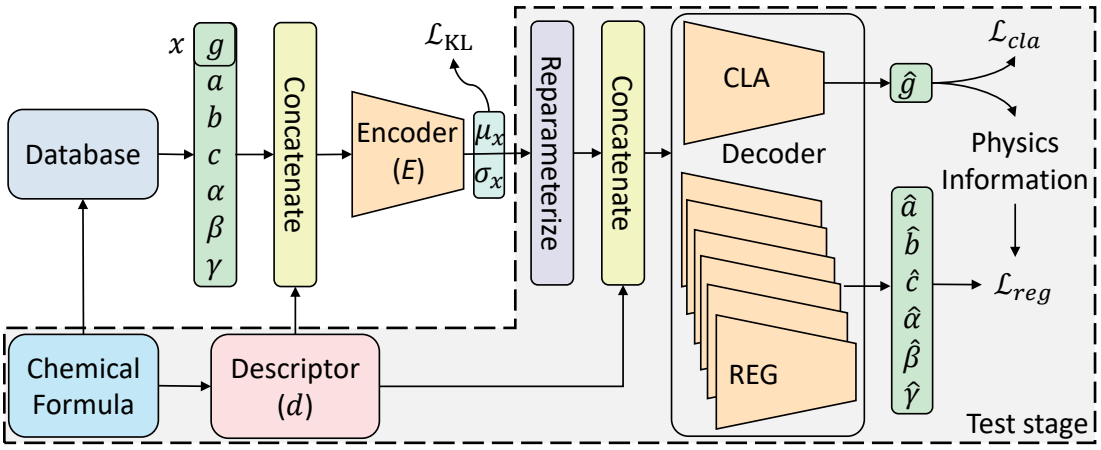

Fig. 1. Illustration of PCVAE. The grey box denotes the workflow of test stage. Best viewed in color. Database indicates the database where lattice constants and space group, x, can be obtained. Descriptor is the features calculated through the composition of chemical formula. CLA and REG are neural networks for space group classification and lattice constant regression respectively. x is the ground truth and $\hat{}$ denotes the prediction.

this problem by encoding the multiple stable structures into a Gaussian distribution.

In this work, focusing on space group classification and lattice constant regression, we propose a **P**hysics-informed **C**onditional **V**ariational **A**uto **E**ncoder **(PCVAE)**. Compared to previous work, PCVAE has achieved much better accuracy in determining the symmetry and geometry of crystals. PCVAE makes good use of the physics information and encodes the multiple stable structures into a Gaussian distribution, resulting in substantially higher accuracy in determining crystal symmetry and geometry.

The main contributions of our work can be summarized as follows:

- We propose a generic Physics-informed Conditional Variational Auto Encoder (PCVAE), in which the MS4OF problem can be solved by encoding multiple stable structures into a Gaussian distribution.
- We construct a dataset for the MS4OF problem, in which at least two structures exist for one chemical formula.
- We introduce physics knowledge to our model. With physics information, PCVAE is generalized to all materials with any mole ratio and predicts the crystal structure without any prior knowledge about the space group.
- Comprehensive experiments are conducted on standard benchmark datasets. The empirical results demonstrate that PCVAE achieves a new state-of-the-art in predicting both the symmetry and geometry of crystal structure.

The rest of this paper is organized as follows. Section II reviews the related works on crystal structure prediction and possible solutions to MS4OF. Section III introduces our PCVAE, including the descriptors and physics information in detail. In section IV, we report our carefully designed experiments and experimental results. We draw conclusions and look ahead to future work in section V. To promote the development of AI for science, we will release our code and dataset upon acceptance.

II. Related Work

a) Crystal Structure Prediction: With the development of machine learning, a variety of machine learning methods have been applied to crystal structure prediction, including support vector machines (SVM) [14]–[16], neural networks (NN) [5], [9], random forests (RF) [13], and so on [17]. Almost all the works focus on a specific family of materials, which makes it difficult to generalize these models to other materials. CRYSPNet built a neural network-based model for each Bravais lattice and achieved a state-of-the-art in CSP [5]. MLatticeABC proposed a series of new features and built random forest models for each lattice constant of each Bravais lattice, achieving a state-of-the-art in crystal constants prediction [13]. Both CRYSPNet and MLatticeABC took the Bravais lattice as prior knowledge and can not solve the MS4OF problem.

b) Generative Models for CSP: Few previous works can be used to tackle this MS4OF challenge. IMatgen, which is not controllable, utilized Variational Auto Encoder to discover stable new materials [18]. Generative Adversarial Networks (GAN) were proposed to predict the crystal structure by [17], [19]. Williams *et al.* focused on generating the cubic inorganic perovskites structure. All these generative models take the microscopic material images or 3D point clouds from expensive experiments as input. In addition, they focus on generating stable crystal structures without the control of chemical formulae instead of predicting the structure of specific materials.

c) Conditional Variational Auto Encoder (CVAE): GSNN encodes conditions and features respectively and gets the two distributions close [20]. CVAE was proposed to model complex structured output representations, effectively perform probabilistic inference and make diverse predictions [21]. CVAE is widely used in controllable generation tasks. A CVAE-based model for molecular conformation generation

TABLE II
DESCRIPTORS USED IN PCVAE

Magpie descriptors	Atomic number, Mendeleev number, Atomic weight, Melting temperature, Periodic Table row & column, The number of Valence e^- in each Orbital (s, p, d, f, total), The number of unfilled e^- in each orbital (s, p, d, f, total), Ground state band gap energy, Covalent radius, Ground state magnetic moment.
Statistic descriptors	Maximum of the atomic numbers, Minimum of the atomic numbers, Average (all atoms divided by the number of element types) of the atomic numbers, Variance of the atomic numbers, Total atom number.
Additional descriptors	Stoichiometry p-norm (p=0,2,3,5,7), Elemental fraction, Fraction of electrons in each orbital, Band center, Ion Property (possible to form ionic compound, ionic charge)

was proposed in [22]. CVAE was also used to determine the geometry of molecular [23].

III. METHODOLOGY

To tackle the MS4OF problem that multiple stable structures exist for one chemical formula, we introduce CVAE and physics information to determine the symmetry and geometry of crystals. The overall PCVAE is illustrated in Fig. 1. In this section, the model will be explained in detail, including the descriptors, workflow, loss functions, and physics information.

A. Workflow

Fig. 1 shows the workflow of PCVAE. The training is performed as following: (1) For a chemical formula, the ground truth x, i.e., lattice constants and space group is obtained through the database. (2) The descriptor d is calculated according to the statistics of the components and Materials Agnostic Platform for Informatics and Exploration (Magpie) with only the chemical formula. (3) Then descriptor d is concatenated with the ground truth x and encoder into the Gaussian distribution through Encoder E. (4) Reparameterization trick is performed on μ_x and σ_x from the Encoder to get the latent vector z. (5) z is concatenated with descriptor d and input into the Decoder, which consists of a CLA module and 6 REG modules for space group prediction and lattice constants prediction respectively. (6) The predicted space group $\hat{g}$ and lattice constants $\hat{x}$ are obtained through the Decoder. $\mathcal{L}_{KL}$, $\mathcal{L}_{cla}$ and $\mathcal{L}_{reg}$ denotes the KL divergence loss of μ_x and μ_d, cross-entropy loss for space group classification, and the mean square error for lattice constants regression. The losses are integrated through physics information which is described in section III-C.

The modules in grey box in Fig. 1 shows the workflow for prediction. Given any chemical formula, descriptor is calculated and concatenated with the latent vector z, which is randomly sampled from a standard Gaussian distribution. Then Decoder outputs one set of lattice constants and space group for the crystal.

B. Descriptors

For a crystal material with only the chemical formula, the descriptors d can be calculated based on the properties of its constituent atoms without any prior knowledge about the crystal geometry. Following [13], the descriptors in our PCVAE mainly consist of the Magpie descriptors [24], statistic descriptors like the number of elements and element fractions, and additional descriptors like stoichiometry p-norm, valence orbitals, ion properties. The whole set of descriptors is listed in Table II.

C. Physics Information

The physics knowledge includes two parts: *(a)* The mapping of crystal system, Bravais lattice and space group.With the mapping, we can get the Bravais lattice type of a crystal via only predicting the space group. While the crystal system directly determines the constraints of lattice constants and space group defines the symmetry of crystals. *(b)* The constraints of lattice constants in Table III. Since the lengths of three cell edges and the angles between them are constrained in a specific Bravais lattice, the modules for space group classification (CLA) and lattice constants regression (REG) are strongly related to each other. The physics knowledge is integrated into PCVAE through the coefficient p, where $p = 0$ for lattice constants that are fixed or can be determined by other lattice constants, otherwise $p = 1$.

Specifically, Space group is obtained through the CLA prediction in testing and the given ground truth in training in PCVAE. Then the crystal system can be looked up by space group via the mapping relationship. Finally the coefficient p and the constraints of lattice constants can be looked up by crystal system via Table III.

D. Physics-informed Conditional Variational Auto Encoder

Corresponding to one chemical formula with descriptors d, there can be multiple stable crystal structures $X = \{x_1, x_2, ...x_n\}$. One set of crystal structure consists of space group g and the lengths of three cell edges $\{a, b, c\}$ and the angles between them $\{\alpha, \beta, \gamma\}$, i.e. $x = \{g, a, b, c, \alpha, \beta, \gamma\}$. The distribution of crystal structure x is modeled conditioning on descriptors d in PCVAE, i.e. $P(x|d)$.

TABLE III
PHYSICS INFORMATION

Crystal system	Constraints on lattice constants Edge	Angle (rad)	p
Triclinic	-	-	[1,1,1,1,1,1]
Monoclinic	$a \neq c$	$\alpha = \gamma = \frac{\pi}{2}, \beta \neq \frac{\pi}{2}$	[1,1,1,0,0,1]
Orthorhombic	$a \neq b \neq c$	$\alpha = \beta = \gamma = \frac{\pi}{2}$	[1,1,1,0,0,0]
Tetragonal	$a = b \neq c$	$\alpha = \beta = \gamma = \frac{\pi}{2}$	[1,0,1,0,0,0]
Rhombohedral	$a = b \neq c$	$\alpha = \beta = \frac{\pi}{2}, \gamma = \frac{2\pi}{3}$	[1,0,1,0,0,0]
Hexagonal	$a = b$	$\alpha = \beta = \frac{\pi}{2}, \gamma = \frac{2\pi}{3}$	[1,0,1,0,0,0]
Cubic	$a = b = c$	$\alpha = \beta = \gamma = \frac{\pi}{2}$	[1,0,0,0,0,0]

A latent variable z is introduced to model the multiple stable structures of the same chemical formula in this model. A prior distribution of latent variable $E_\psi(z|d)$ and a group of decoders $D_\theta(x|z)$ are used to capture the conditional distribution of x given d. z is obtained through the Encoder $E_\phi(z|x,d)$. ψ and θ denote the module parameters. The decoder consists of one classification module for space group prediction and six regression modules for lattice constant prediction. All the encoders and decoders are jointly trained to maximize the evidence lower bound (ELBO) of data log-likelihood as (1) [21], [22].

$$\begin{aligned}\log P(x|d) \geq & \mathbb{E}_{z\sim E(z|x,d)}[\log p_\theta(x|z,d)] \\ & - D_{KL}[E_\phi(z|x,d)||E_\psi(z|d)] \\ = & \mathcal{L}_{recon} - \mathcal{L}_{KL}\end{aligned} \tag{1}$$

The two terms in (1) can be explained as the negative reconstruction error $\mathcal{L}_{recon}$ and latent space prior regularizer $\mathcal{L}_{KL}$ respectively. $\mathcal{L}_{KL}$ can be calculated as (2), where J, μ, and σ denote the dimension of Gaussian function, mean and variance respectively [25].

In PCVAE, the latent vector z is regarded as external condition features. Each chemical formula indicates a distinct compound of material. Due to various external environmental conditions, multiple stable structures exist. The structure of a crystal is almost unique if the external condition is given. One chemical formula with a specific structure actually represents one set of external conditions, which is encoded into a standard Gaussian distribution i.e., $E_\phi(z|d) \sim \mathcal{N}(0, \mathbf{I})$. Therefore, the $\mathcal{L}_{KL}$ in PCVAE is modified into (3).

$$\mathcal{L}_{KL}(E_\phi||E_\psi) = \log\frac{\sigma_d}{\sigma_x} + \frac{\sigma_x^2 + (\mu_x - \mu_d)^2}{2\sigma_d^2} - \frac{1}{2} \tag{2}$$

$$\mathcal{L}_{KL}(E_\phi||E_\psi) = -\log\sigma_x + \frac{\mu_x^2 + \sigma_x^2 - 1}{2} \tag{3}$$

The reconstruction loss is divided into regression loss $\mathcal{L}_{reg}$ and classification loss $\mathcal{L}_{cla}$ as (4) and (5), where N is the number of materials, $r = \{a, b, c, \alpha, \beta, \gamma\}$ denotes the lattice constants, $t_i(g)$ is the indicator variable for the true space group label of the i-th materials ($t_i(g)$ is 1 for $g = g_{true}$ and 0 for all other g), D and K denote the numbers of lattice constants and space groups, $y_{g,i}$ is the predicted probability of space group g for the i-th chemical formula, and $p_{ij}(g)$ is determined by the space group and the physics information [25].

$$\mathcal{L}_{reg} = \frac{1}{N}\frac{1}{D}\sum_{i=1}^{N}\sum_{j=1}^{D} p_{ij}(\hat{g})\sqrt{(r_{ij} - \hat{r}_{ij})^2} \tag{4}$$

$$\mathcal{L}_{cla} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{K} t_i(g)\log(y_{g,i}) \tag{5}$$

The workflow of PCVAE is shown in Fig. 1. In practice, crystal structure x and chemical formula d are concatenated and input into the Encoder E_ϕ, which then gives the prediction of mean, μ_x, and variance, σ_x i.e. the distribution of possible external environmental conditions. The Reparameterization trick is then applied to obtain the Latent vector z, as (6), where ϵ_1 is auxiliary noise variable, and $\epsilon \sim \mathcal{N}(0, 1)$. Finally, by concatenating z and d and feeding them into the CLA and REG modules, the space group $\hat{g}$ and crystal structure are obtained.

$$z = \mu_d + \sigma_d\epsilon \tag{6}$$

During prediction, z is obtained by randomly sampling from the Gaussian distribution $\mathcal{N}(0, 1)$. Since the more stable structures exist under common external conditions, truncation is employed as (7), where $\overline{z}$ indicates the mean latent vector. Truncation is a common trick used in the inference of state-of-the-art GANs to improve the synthesis quality [26].

$$z' = \beta z + (1 - \beta)\overline{z} \tag{7}$$

IV. EXPERIMENTS & DISCUSSION

In this section, we present the experimental results and discussion in details. Both two datasets are randomly split into 80% for training and the remaining 20% for testing.

A. Experimental Setup

1) Dataset: We construct two datasets from the Material Project [27], 125,278 crystal structures in ***Whole dataset*** and 33,160 crystal structures in ***Multi-structure dataset***. Following the setup of MLatticeABC [13], each entry of both datasets consists of space group, lattice constants, and 249 descriptors listed in Table II.

All the materials can be grouped into 14 Bravais lattice types. In ***Whole dataset***, only 8.31% chemical formulae have more than one stable structure. However, almost all the materials have at least two stable structures in the real world [4]. Hence, we construct a more practical dataset, ***Multi-structure dataset***, in which more than one stable structures exist for one chemical formula.

2) Compared Approaches: We mainly compare our proposal with several previous crystal structure prediction methods. For **SVM** [14]–[16], **NN** [9], and **RF** [28], we train one model for each lattice constant. For **CRYSPNet** [5] and **MLatticeABC** [13], we follow the setups in their paper. Besides, we also perform experiments with **VAE** [21] and **GSNN** [20] which are suitable but never used to tackle the MS4OF problem.

TABLE IV
R^2 SCORES (THE COEFFICIENT OF DETERMINATION) OF LATTICE CONSTANT PREDICTION ON *Whole dataset*.

	a	b	c	α	β	γ
SVM	0.4672 ± 0.0051	0.3535 ± 0.0066	0.3264 ± 0.0015	0.0231 ± 0.0037	-0.0821 ± 0.0125	-0.0993 ± 0.0326
RF	0.4274 ± 0.0072	0.4179 ± 0.0418	0.4306 ± 0.0050	0.0159 ± 0.0129	0.0110 ± 0.0082	0.0103 ± 0.0324
NN	0.3364 ± 0.0242	0.3258 ± 0.0213	0.1732 ± 0.0517	-1.7682 ± 0.1942	-0.3447 ± 0.0143	-0.2279 ± 0.0205
MLatticeABC	0.4353 ± 0.0149	0.4357 ± 0.0098	0.4463 ± 0.0106	-0.0159 ± 0.0293	-0.1098 ± 0.0010	-0.0329 ± 0.0204
CRYSPNet	0.3766 ± 0.0151	0.3759 ± 0.0383	0.2378 ± 0.0131	-1.4034 ± 0.0181	-0.3501 ± 0.0119	-0.1733 ± 0.0699
GSNN + **PI**	0.4436 ± 0.0135	0.4603 ± 0.0235	0.2921 ± 0.0573	0.0025 ± 0.0153	0.2651 ± 0.0183	0.7689 ± 0.0235
VAE + **PI**	0.5832 ± 0.0157	0.6053 ± 0.0176	0.6287 ± 0.0645	0.0538 ± 0.0115	0.3328 ± 0.0186	0.7856 ± 0.0539
PCVAE	**0.7905** ± 0.0124	**0.6895** ± 0.0053	**0.8124** ± 0.0114	**0.0647** ± 0.0110	**0.4093** ± 0.0177	**0.8065** ± 0.0306

3) Evaluation Metrics: We employ the coefficient of determination R^2 score as (8), where n, x_i, $\hat{x}_i$, and $\overline{x}$ denote the number of data, ground truth, prediction, and the mean of ground truth respectively, to evaluate lattice constants regression. Accuracy is used to evaluate the and space group classification. Following previous works [22], Matching (MAT) score is used to get the prediction and ground truth matched as (9), where $\mathbb{S}_g(\hat{x})$ and $\mathbb{S}_r(x)$ denote the generated and the reference structure sets respectively. Both of the evaluation metrics are calculated based on the minimal MAT, which is commonly used in evaluating the generative models. MAT is used for evaluating PCVAE since multiple structures exist for one chemical formula and multiple structures are predicted from PCVAE. The ground truth and predicted structures should be matched.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{8}$$

$$MAT(\mathbb{S}_g(\hat{x}), \mathbb{S}_r(x)) = \frac{1}{|\mathbb{S}_r|}\sum_{x\in\mathbb{S}_r}\min_{\hat{x}\in\mathbb{S}_g}(\mathcal{L}_{reg}(\hat{x},x)+\mathcal{L}_{cla}(\hat{x},x)) \tag{9}$$

B. Implementation Details

This section gives implementation details of PCVAE for better reproducibility, including the model structure, hyperparameters and hardware.

a) Model Structure: Two-layer neural networks are used to implement the encoder E_ϕ. The first layer contains 512 neurons, while the second layer consists of two parts, each with 64 neurons, one for encoding the mean μ and the other for encoding the variation σ. The architecture of the six REG neural network-based modules is the same, but the parameters are not shared. The REG consists of 6 layers, with 256, 512, 256, 128, 64, and 1 neurons in each layer. The 6 layers of CLA neural networks have 256, 512, 256, 128, 64, and 230 neurons, respectively. In both CLA and REG modules, $Droptout$ is appended to the last three layers. $ReLu$ is the activation function for all modules.

b) Hyperparameters Setting: Following MLatticeABC [13], 249 descriptors are adapted in PCVAE. The number of training epochs is 500 and the batch size is 128. The learning rate is 0.001 and the weight decay is set to 0.0001, starting from the 100 epochs. The dropout rate is 0.1. Adam optimizer is applied. β is 0.9.

c) Hardware Environment: All our experiments are conducted on a computing cluster with GPUs of NVIDIA® GeForce® RTX 2080 Ti 11GB and CPUs of Intel® Xeon® Gold 6139 CPU @ 2.30GHz.

C. Result and Discussion

This section gives the experimental results on both *Whole dataset* and *Multi-structure dataset* as well as a discussion of the results.

1) Results on Whole Dataset: The R^2 score results for lattice constants prediction on the *Whole dataset* are shown in Table IV, where PI denotes the physics information knowledge. For a fair comparison, all the VAE-based models are trained and tested with physics information. PCVAE exceeds the performance of all previous methods on all the lattice constants prediction. The average MAT scores for GSNN+PI, VAE+PI, and PCVAE on *Whole dataset* are 6.1638, 4.1057, and 3.5169 respectively. PCVAE achieves a better MAT score than other VAE-based methods which means that the predicted crystal structures from PCVAE covers more existing crystal structures.

The neural network-based model, CRYSPNet works better than NN because the MS4OF problem is alleviated. CRYSPNet builds models for each lattice constant of each Bravais lattice type. Therefore, CRYSPNet can predict multiple structures for one chemical formula within different Bravais lattices. However, the MC4OF problem is still not solved since multiple structures also exist in the one Bravais lattice type or even space group.

In addition, almost all the VAE-based models improve the R^2 scores substantially, because multiple stable structures are encoded into the Gaussian distribution, which solves the MC4OF problem thoroughly. Although the R^2 scores are improved substantially, they are still not high for the prediction of lattice constants, α, β, and γ. This is because the distributions are still skewed, which is already improved from the original distribution through the physics information. The distributions are shown in Fig. 4 and Fig. 3.

2) Results on Multi-structure dataset: The results of lattice constants regression on the more challenging but practical *Multi-structure dataset* are shown in Table V. The average MAT scores for GSNN+PI, VAE+PI, and PCVAE on *Multi-structure dataset* are 6.1923, 4.4050, and 3.1023 respectively, which also demonstrates that the predicted crystal structures from PCVAE covers more existing crystal structures. The difference in R^2 scores between the VAE-based models and other methods is

TABLE V
R^2 SCORES (THE COEFFICIENT OF DETERMINATION) OF LATTICE CONSTANT PREDICTION ON *Multi-structure dataset*.

	a	b	c	α	β	γ
SVM	0.3448 ± 0.0109	0.2920 ± 0.0303	0.3075 ± 0.0163	0.0712 ± 0.0136	-0.0402 ± 0.0128	0.0778 ± 0.0095
RF	0.3577 ± 0.0258	0.3083 ± 0.0818	0.6404 ± 0.0292	-0.0140 ± 0.0065	-0.1097 ± 0.0245	0.0879 ± 0.0404
NN	0.2702 ± 0.0263	0.2822 ± 0.0456	0.5267 ± 0.0682	-1.2732 ± 0.0167	-0.3177 ± 0.0454	-0.1265 ± 0.0417
MLatticeABC	0.3770 ± 0.0258	0.3342 ± 0.0393	0.6749 ± 0.0404	-0.0112 ± 0.0054	-0.0708 ± 0.0336	0.0683 ± 0.0537
CRYSPNet	0.3145 ± 0.0119	0.3055 ± 0.0321	0.4643 ± 0.0115	-1.0878 ± 0.0032	-0.2867 ± 0.0678	-0.0534 ± 0.0266
GSNN + **PI**	0.3488 ± 0.0192	0.3804 ± 0.0271	0.6000 ± 0.0231	0.1134 ± 0.0241	0.1097 ± 0.0312	0.5951 ± 0.0321
VAE + **PI**	0.6259 ± 0.0218	0.6120 ± 0.0241	0.8612 ± 0.0274	0.1864 ± 0.0252	0.3120 ± 0.0263	0.6705 ± 0.0284
PCVAE	**0.7858** ± 0.0172	**0.8244** ± 0.0193	**0.8812** ± 0.0163	**0.2856** ± 0.0221	**0.5523** ± 0.1930	**0.7359** ± 0.0294

TABLE VI
CLASSIFICATION ACCURACY FOR SPACE GROUP AND BRAVAIS LATTICE ON *Multi-structure dataset*.

	Bravais lattice type			space group		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
SVM	0.1505 ± 0.0045	0.2824 ± 0.0067	0.4243 ± 0.0086	0.1464 ± 0.0023	0.2866 ± 0.0033	0.4244 ± 0.0042
RF	0.4298 ± 0.0095	0.6239 ± 0.0104	0.7271 ± 0.0093	0.3227 ± 0.0041	0.4781 ± 0.0054	0.5467 ± 0.0048
NN	0.2415 ± 0.0143	0.3269 ± 0.0346	0.4060 ± 0.0507	0.1449 ± 0.0023	0.1611 ± 0.0012	0.2447 ± 0.0068
MLatticeABC	0.4737 ± 0.0116	0.6783 ± 0.0142	0.7693 ± 0.0113	0.3605 ± 0.0104	0.5228 ± 0.0099	0.5913 ± 0.0069
CRYSPNet	0.5420 ± 0.0049	0.7060 ± 0.0040	0.8010 ± 0.0032	0.4322 ± 0.0089	0.5630 ± 0.0016	0.6472 ± 0.0058
GSNN+**PI**	0.2795 ± 0.0020	0.4447 ± 0.0019	0.5089 ± 0.0020	0.1859 ± 0.0021	0.2928 ± 0.0018	0.4304 ± 0.0019
VAE+**PI**	0.7280 ± 0.0024	0.8626 ± 0.0021	0.9226 ± 0.0016	**0.6573** ± 0.0012	0.7302 ± 0.0017	0.8290 ± 0.0014
PCVAE	**0.8542** ± 0.0018	**0.9563** ± 0.0024	**0.9837** ± 0.0014	0.6546 ± 0.0018	**0.8011** ± 0.0024	**0.8695** ± 0.0014

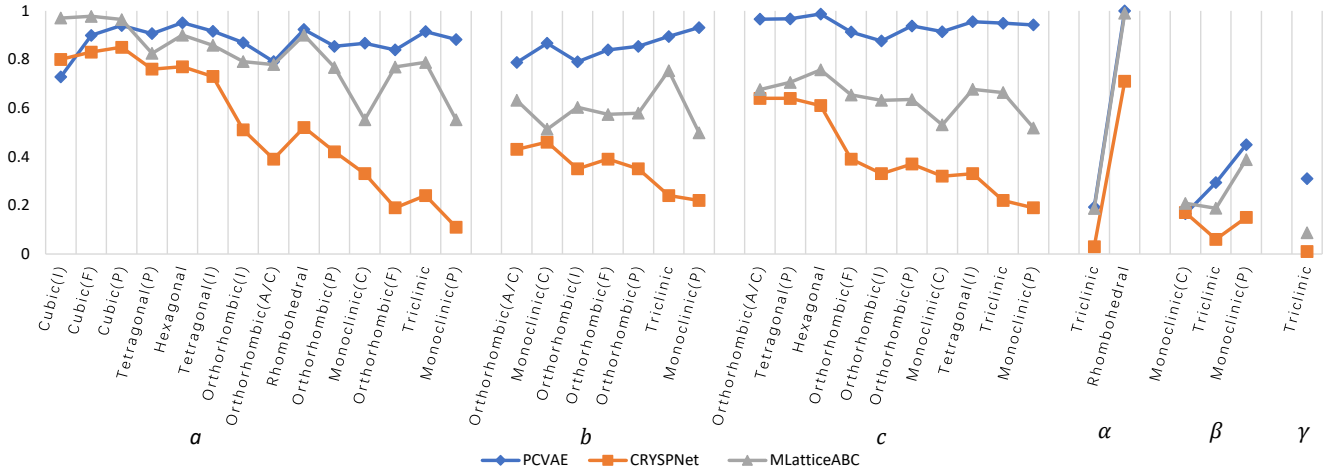

Fig. 2. Comparison of R^2 score between PCVAE and CRYSPNet. Horizontal and vertical axes indicate the Bravais lattice type and R^2 score respectively. The blue and orange lines denote the R^2 score of PCVAE and CRYSPNet respectively.

more significant compared to the results on *Whole dataset*. This is because the chemical formulae in *Multi-structure dataset* have at least two stable structures which make the MC4OF problem more pronounced.

The classification accuracies on *Multi-structure dataset* are shown in Table VI. PCVAE also outperforms other methods on both space group and Bravais lattice classification. CRYSPNet works better than other traditional methods since it takes the task as multi-label classification.

PCVAE outperforms other methods in our more practical task that determing the crystal structure without any prior knowledge about its space group. We also conduct experiments with the common setting in CRYSPNet and MLatticeABC that predicting crystal structures with the space group given. Thus the constraints can be obtained through physics information in PCVAE. Fig. 2 shows the comparison of R^2 scores of PCVAE, CRYSPNet and MLatticeABC on each Bravais lattice. PCVAE still outperforms other models on most tasks. MLatticeABC shows comparable predictive power because, for one chemical formula, fewer stable structures exist with the same Bravais lattice type. CRYSPNet performs worse than MLatticeABC because the material descriptors used in MLatticeABC are well-chosen [13].

TABLE VII
ABLATION STUDY: R^2 SCORES OF LATTICE CONSTANTS AND SPACE GROUP CLASSIFICATION ACCURACY ON *Phase dataset*

	Lattice constants						Space group		
	a	b	c	α	β	γ	Top-1	Top-2	Top-3
VAE	0.5534	0.4570	0.8136	-0.9937	-0.2028	-0.0318	0.6220	0.7103	0.7972
VAE+PI	0.6259	0.6120	0.8612	0.1864	0.3120	0.6705	0.6573	0.7302	0.8290
CVAE*	0.5885	0.5197	0.8603	-0.6077	-0.0636	0.0933	0.6488	0.7850	0.8429
PCVAE	0.7858	0.8244	0.8812	0.2856	0.5523	0.7359	0.6546	0.8011	0.8695

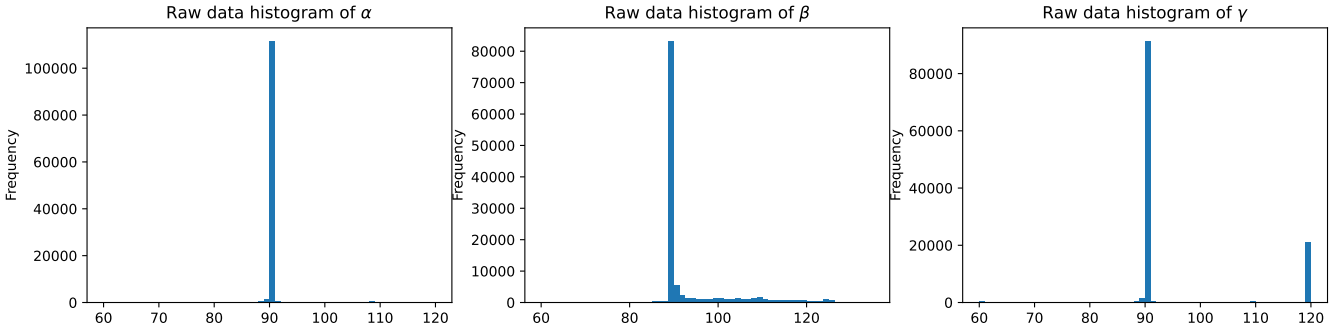

Fig. 3. Histogram of the lattice constant distribution. Horizontal and vertical axes indicate the length/angle and frequency respectively.

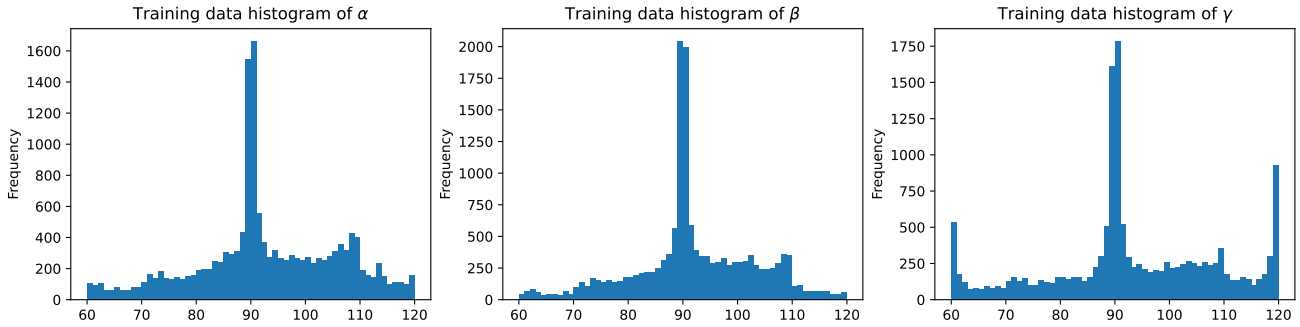

Fig. 4. Histogram of the lattice constant distribution in training data. The data that is not used due to physics information constraints is removed. Horizontal and vertical axes indicate the length/angle and frequency respectively.

D. Ablation Study

We conduct ablation experiments to verify that physics information works on both VAE and PCVAE. The results are shown in Table VII. VAE and CVAE denote the model without Physics information added. With physics information, Both R^2 scores and classification accuracies are improved, especially the R^2 scores of the three angles α, β, and γ. The physics information works by changing the distribution of lattice constants actually used in PCVAE. Physics information is integrated in PCVAE with the coefficient p on loss, therefore, the losses are ignored where $p = 0$, i.e., the data is not actually used in PCVAE. Fig. 3 shows the lattice constants distribution of the raw data from Material Project [27]. The distributions of α, β, and γ are extremely skewed. Fig. 4 shows that with physics information, the data for model learning is more evenly distributed, especially for α, β, and γ compared with Fig. 3. The more even distribution makes it easier for models to learn not only in the three most difficult angle regression tasks but also in the regression for three edge lengths, a, b, and c, and the classification for the space group. With the physics information, the intrinsic connections between the lattice constants are captured in our PCVAE.

E. Case Study

In addition, take the chemical formula, $Li_7Mn_2Co_3O_{12}$ with the most number of stable structures (164) for example. We predict 164 sets of space group and lattice constants. MAT is used to match the predicted crystal structures and ground truth. Then the evaluation metrics are calculated. Top-1, Top-2, and Top-3 Bravais lattice classification accuracies achieve 78.53%, 96.32%, and 100% respectively. The R^2 scores of lattice constant regression are 0.7035, 0.6083, 0.7394, 0.2395, 0.2063, and 0.3240 respectively. Both classification and regression achieve high accuracy. For other previous methods, only one set of lattice constants and space group can be predicted with one given chemical formula. These experiments show that PCVAE performs quite well in determining the geometry of crystals.

V. Conclusion

In this study, we propose a physics-informed conditional variational auto encoder (PCVAE). To tackle the MS4OF problem, we firstly propose the conditional variational auto encoder (CVAE) to encode crystal structure variations with a Gaussian distribution. Furthermore, since the pieces of the decoder output are physically inter-constrained, we introduce physics information on top of the CVAE. Comprehensive experiments are conducted and the empirical experimental results demonstrate that PCVAE achieves a new state-of-the-art in crystal structure prediction. Additional experiments on the more challenging *Multi-structure dataset* further prove the efficacy of PCVAE. With an additional dataset of correspondence among lattice constants and environmental conditions such as pressures and temperatures, we would obtain a physically conditional generative model, which will be done in our future work. We expect that Multi-structure dataset and PCVAE could inspire the data mining community to make more impact on data mining for Physical science.

References

[1] J. Graser, S. K. Kauwe, and T. D. Sparks, "Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons," *Chemistry of Materials*, vol. 30, no. 11, pp. 3601–3612, 2018. I

[2] A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, "Structure prediction drives materials discovery," *Nature Reviews Materials*, vol. 4, no. 5, pp. 331–348, 2019. I

[3] K. Liu, K. Yang, J. Zhang, and R. Xu, "S2snet: A pretrained neural network for superconductivity discovery," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 5101–5107, aI for Good. [Online]. Available: https://doi.org/10.24963/ijcai.2022/708 I

[4] C. Kittel, P. McEuen, and P. McEuen, *Introduction to solid state physics*. Wiley New York, 1996, vol. 8. I, IV-A1

[5] V. Stanev, A. Kusne, I. Takeuchi *et al.*, "Cryspnet: Machine learning tool for crystal structure predictions," *Bulletin of the American Physical Society*, 2021. I, I, II-0a, II-0a, IV-A2

[6] J. Briones, M. C. Guinto, and C. M. Pelicano, "Accelerated lattice constant prediction of perovskite materials (abx3, a2bb' o6) using partial least squares and principal component regression methods," *Materials Letters*, vol. 298, p. 130040, 2021. I

[7] L. Jiang, J. Guo, H. Liu, M. Zhu, X. Zhou, P. Wu, and C. Li, "Prediction of lattice constant in cubic perovskites," *Journal of Physics and Chemistry of Solids*, vol. 67, no. 7, pp. 1531–1536, 2006. I

[8] M. W. Lufaso and P. M. Woodward, "Prediction of the crystal structures of perovskites using the software program spuds," *Acta Crystallographica Section B: Structural Science*, vol. 57, no. 6, pp. 725–738, 2001. I

[9] A. Majid, A. Khan, G. Javed, and A. M. Mirza, "Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression," *Computational materials science*, vol. 50, no. 2, pp. 363–372, 2010. I, II-0a, IV-A2

[10] R. Ubic, "Revised method for the prediction of lattice constants in cubic and pseudocubic perovskites," *Journal of the American Ceramic Society*, vol. 90, no. 10, pp. 3326–3330, 2007. I

[11] A. Verma and V. Jindal, "Lattice constant of cubic perovskites," *Journal of alloys and compounds*, vol. 485, no. 1-2, pp. 514–518, 2009. I

[12] M. Modell and R. C. Reid, "Thermodynamics and its applications. prentic-hall," *Englewood Cliffs, NJ, USA*, 1974. I

[13] Y. Li, W. Yang, R. Dong, and J. Hu, "Mlatticeabc: generic lattice constant prediction of crystal materials using machine learning," *ACS omega*, vol. 6, no. 17, pp. 11 585–11 594, 2021. I, I, II-0a, II-0a, III-B, IV-A1, IV-A2, IV-B0b, IV-C2

[14] S. G. Javed, A. Khan, A. Majid, A. M. Mirza, and J. Bashir, "Lattice constant prediction of orthorhombic abo3 perovskites using support vector machines," *Computational materials science*, vol. 39, no. 3, pp. 627–634, 2007. II-0a, IV-A2

[15] T. O. Owolabi, "Extreme learning machine and swarm-based support vector regression methods for predicting crystal lattice parameters of pseudo-cubic/cubic perovskites," *Journal of Applied Physics*, vol. 127, no. 24, p. 245107, 2020. II-0a, IV-A2

[16] T. O. Owolabi, K. O. Akande, and S. O. Olatunji, "Estimation of superconducting transition temperature t c for superconductors of the doped mgb 2 system from the crystal lattice parameters using support vector regression," *Journal of Superconductivity and Novel Magnetism*, vol. 28, no. 1, pp. 75–81, 2015. II-0a, IV-A2

[17] L. Williams, A. Mukherjee, and K. Rajan, "Deep learning based prediction of perovskite lattice parameters from hirshfeld surface fingerprints," *The Journal of Physical Chemistry Letters*, vol. 11, no. 17, pp. 7462–7468, 2020. II-0a, II-0b

[18] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, and Y. Jung, "Inverse design of solid-state materials via a continuous representation," *Matter*, vol. 1, no. 5, pp. 1370–1384, 2019. II-0b

[19] B. Kim, S. Lee, and J. Kim, "Inverse design of porous materials using artificial neural networks," *Science Advances*, vol. 6, no. 1, p. eaax9324, 2020. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.aax9324 II-0b

[20] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in neural information processing systems*, vol. 25, pp. 2843–2851, 2012. II-0c, IV-A2

[21] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, pp. 3483–3491, 2015. II-0c, III-D, IV-A2

[22] M. Xu, W. Wang, S. Luo, C. Shi, Y. Bengio, R. Gomez-Bombarelli, and J. Tang, "An end-to-end framework for molecular conformation generation via bilevel programming," *arXiv preprint arXiv:2105.07246*, 2021. II-0c, III-D, IV-A3

[23] G. N. C. Simm and J. M. Hernández-Lobato, "A Generative Model for Molecular Distance Geometry," *arXiv e-prints*, p. arXiv:1909.11459, Sep. 2019. II-0c

[24] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, no. 1, pp. 1–7, 2016. III-B

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. III-D, III-D

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410. III-D

[27] 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, "The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013. [Online]. Available: http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi IV-A1, IV-D

[28] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282. IV-A2