INPUT CONSISTENCY REGULARIZATION FOR MODELING NEGATIVE THERMAL EXPANSION CHARACTERISTICS OF $A_2M_3O_{12}$ FAMILY OF COMPOUNDS

A PREPRINT

Natalia V. Kireeva
Laboratory of novel physicochemical problems
Frumkin Institute of Physical Chemistry and Electrochemistry RAS
Leninsky Prospect, 31, Moscow, 119071, Russia
kireeva@phyche.ac.ru

Aslan Yu. Tsivadze
Laboratory of novel physicochemical problems
Frumkin Institute of Physical Chemistry and Electrochemistry RAS
Leninsky Prospect, 31, Moscow, 119071, Russia
tsiv@phyche.ac.ru

August 27, 2024

ABSTRACT

The prediction confidence is one of the goals of any machine learning-based study with no respect if this is distinguished as the aim of the study or is the associated desired concomitant. The possibility do not import the additional error into the pre-experimental estimation of the studied characteristics should be the goal of machine learning-based approaches in any case limited in their accuracy by the precision and confidence of the experimental data. In this study, we consider the approaches related to the input consistency regularization which may provide with the required enhancement in the prediction confidence and are able to recoup the part of the experimental error associated with obtaining the data using the methods of different precision. The methodology of regularization of input data consistency is considered in relation to the problem of the predictive modeling of the functional characteristics of $A_2M_3O_{12}$ family of ceramics with negative thermal expansion (NTE) property. The methodological part of this study includes several problems. The Hessian-based analysis of the loss function landscape was considered as the criterion of the generalizability and model performance. The continuity of the property change as a function of the data description coupled with the p-values for the experiment-prediction output were considered as the auxiliary criteria concerned with the input consistency regularization.

Keywords input consistency regularization · diffusion probabilistic models · $A_2M_3O_{12}$ · Hessian function

1 Introduction

The accuracy of the machine learning-based methods is directly concerned with the quality and quantity of the experimental data in use. Therefore, do not introduce the error exceeding the experimental one is of importance for any machine learning-based approach. In this study, we consider the approaches related to the input consistency regularization which may provide with the enhancement in the prediction accuracy and confidence and are able to recoup the part of the experimental data discrepancy associated with obtaining the data using the methods or measured with different precision.
According to the theory, homogeneity of the chemical potential is a condition for thermodynamic equilibrium of the

system being studied [58] and, therefore, the regularized values of the observed variables should be considered as corresponding to the thermodynamic equilibrium while the input data before the regularization may contain the data, which are characterized by the deviations from the system "equilibrium". Accepting this analogy allows us to consider the experimental data for some process/objects obtained from different sources as the thermodynamic system that should obey the thermodynamics laws. Henri Poincare noted "Physics cannot do without mathematics: it furnishes the only language it can speak". Thus, importantly and par for the course, this problem can be performed by using the diffusion equations that are based on the branches of mathematics appearing the major approaches for many physical and chemical problems.

We would, first of all, refer to the works in which authors have provided with a broad mathematics ground to the diffusion probabilistic models [28, 74, 66, 55, 19, 92, 54, 15]. One of the studies in machine learning most directly relating the problem of input consistency regularization with the thermodynamics is the work by Sohl-Dickstein et al. [76] introducing the concept of non-equilibrium thermodynamics into the deep unsupervised learning namely for accepting the experimental data as the object determined by the thermodynamic laws. It should be emphasized that using the thermodynamics principles in statistical learning in general is widely explored methodology. Among the approaches most closely related to the considered family of diffusion probabilistic models [13] one should mention using the concept of Helmholtz free energy of the system involved in a variety of machine learning approaches including the pioneering works imparting to diffusion probabilistic models [37]. Another example is using the notion of the quasi-equilibrium and low-temperature approximation characterizing the minima in the study involving a diffusion theory for Deep Learning (DL) dynamics [89]. The loss landscape is discussed in a context of the minimum energy paths that are located in the flat regions separated by the saddle points of high-loss areas of the landscape.

One of the possibilities to improve the performance of the machine learning methods which is discussed in the literature is to partially and selectively transform the data in the input consistency regularization procedure to be robust against the occasional and non-systematic deviations from the result pretending to be considered as the "true" value[60]. Several families of machine learning methods aimed at involving the special techniques to enhance the predictive performance are known among them are the approaches related to diffusion probabilistic models [76, 81, 40, 57, 62] and that operate with the so-called privileged information [3, 83, 84]. Recently, the generative diffusion model was applied for the optimization of the surface structure [69] and for modelling the impedance spectra [20]. The DPMs are widely explored in the nonequilibrium physics in biology [26] and one can expect the growth of the related researches based on DPMs in this field. In the same comprehensive review authors overview the problem of symmetry constraints that include the translational, rotational and permutation invariance or equivariance. They argue that the physical prior knowledge integrated into the data by augmenting or synthesizing it expectedly improve the general model performance. Related studies for drug design are also known [43, 33, 56, 73]. Using the same concept of input data regularization was demonstrated its efficiency in combination [45, 17] with dropout [36] and other regularization techniques [2, 44, 4, 60, 86, 39], latent variable models [7] and unsupervised feature transformation-based [24]. The adversarial attacks [90], direction can be considered as a particular case of using the privileged information in modeling. In physics-informed machine learning, partial differential equations (PDEs) and ordinary differential equations (ODEs) are involved in use as a regularization terms in the loss function or to construct the neural architecture (Neural ODE) [34].The combined approach of using the variational inference in the ensemble with the generative model firstly was reported in some earlier works including the Helmholtz Machine (HM) [22], Wake-Sleep algorithm (WS) [37], Annealed Importance Sampling [61], Reverse Annealing Importance Sampling Estimator (RAISE) [12], Deep Autoregressive Networks [32] and some others with very effective upgrowth in the direction of the family of diffusion probabilistic models. Deep Gaussian Processes (DGP) are well-known methodology that, however, has not been considered earlier as the input consistency regularization technique as this property is realized at the model level and was not discussed in the considered context earlier. However, this study, in contrast to our earlier work [50], DGP reveal rather the sensitivity to the input data discussed earlier in the context of the strict requirements to the number of the introduced inducing points [75] than the propensity of this approach to compensate the complexity of the data pattern. The problem of this complexity can be considered as the direct consequence of the chosen data description, the general character of which is concerned with another problem of a reliable assessment of the success in model development. Using the Hessian-based methodology is of interest as the way to elucidate the relationship between the behaviour of the loss function as a function of several factors and the generalizability of the model. One can draw upon the consideration in parallel from the two distinct points of view: the first one is the data-based one when the data invariance to local perturbations can serve as an evidence on the robust decision, while, on the other hand, the Hessian-based analysis of the loss function [65, 71, 70, 31] infers on the importance of the flatness of the landscape in the weight space in general without the presence of the negative eigenvalues in the Hessian spectra. In the former case, one of the first studies dates back to the work of Caruana et al. [14] where it was shown that using the extra features for multitask learning is more useful as inputs if these features have low noise but which become more useful as the outputs as their noise increases. Thus, the features most resistant to the adversarial attacks are beneficial for input data while the features allowing the variance in their values are of importance as an additional target thus enhancing the robustness of the model. This may serve as one of the first insights on the problem of input consistency regularization. The models demonstrating a little or no change in the loss function in the most of the directions except those of eigenvectors that correspond to the large eigenvalues of the Hessian [71] can appear the objects of further enhancement due to the identification of these directions. In [41] it was shown that the generalization of deep networks is related to the curvature of the loss function. The complementary observations were made in [39], where authors discuss the role of the weights distribution and that of their noisy replicas in model generalizability (it is worth to note that the notion of the probabilistic neural networks was introduced in this study). Defining the curvature of the landscape as a function of the spectrum is well-motivated due to its invariance to the rotations of the landscape [90]. Therefore, one may assess the obtained results in input consistency regularization using the Hessian spectra focusing the attention on the values of the eigenvalues and their quantity. If the landscape of the loss function is flat in all of the directions one may infer on the high confidence in the results obtained.

It is worth to note that using the Hessian as a second order characteristic of the loss function has already shown the efficiency in optimization problem. It is well-known that any, even the most advanced methods, may provide with rather poor results in a case of the inaccurate configurational parameters setup. Thus, in [59] it was used to predict the largest useful batch size by estimating the introduced parameter called the gradient noise scale. The batch size impact on the Hessian-described loss function was the subject of attention in [47]. The presence of the negative eigenvalues evidences that the algorithm does not find the local minimum yet and this latter observation can be of importance since the minimal or no changes in the loss function itself does not confirm finding the minimum [71]. In [64] authors have distinguished three aspects of the Hessian-based SGD training: *(i)* generalization performance, *(ii)* optimization and *(iii)* a landscape of the loss function. The latter one can be related to the problem of data description complexity. In [93], authors have shown that training landscape is easy to optimize even when there is no clear notion of generalization. The over-parameterization as a result of the model architecture leads to the enhanced flatness of the landscape and, therefore, the investigation of the problem of data description complexity is of primary importance.

In this study, we consider the methodology of input consistency regularization in relation to the important problem, the predictive modeling of the functional characteristics of $A_2M_3O_{12}$ class of solid oxides with negative thermal expansion (NTE) property. The coefficient of linear thermal expansion and monoclinic to orthorombic phase transition temperature are considered. NTE as a response to the thermal stress is a rare physical property of solids. In most of the cases, this property is associated with a ferroelastic phase transition as a result of the appearance of the spontaneous stress in the structure with increasing temperature combined with the lattice degrees of freedom allowing such a transition. The structure of NTE solid oxides of the $A_2M_3O_{12}$ family (Figure 1) is quite flexible, but with the expected limitations in the composition. The cation A can be represented by trivalent rare-earth elements Ln^{3+}, Sc^{3+}, Y^{3+}, Al^{3+}, Fe^{3+}, Cr^{3+} and In^{3+}, by a combination of two- and four- valence elements (Hf^{4+} and Mg^{2+}, Zr^{4+} and Mg^{2+} allowing the substitution by other elements like Zn^{2+}, Ga^{3+}, Mn^{2+})), or by a combination of three- and four- valence elements by introducing V^{5+} or P^{5+} as M cation thus preserving the electroneutrality condition. The M cations in this structure are represented by W^{6+}, Mo^{6+}, V^{5+} and P^{5+} that ensures the target structural features are preserved. The structure can be represented as the 2D slabs of corner sharing polyhedra parallel to the *ac* plane, which are stacked together. Each individual slab can be considered as being formed by the cross-linking via oxygen atoms of the AO_6 and WO_4 groups [25]. In this arrangement a single oxygen ion links the individual slabs thus preserving the structure of the "layered" type. In the case, when a cation A is represented by the cations of different oxidation states, the cations occupy each atomic position not randomly, but in an ordered manner [63].

The Conditional Score-Based Diffusion Models for Imputation [82] was involved as the input consistence regularization approach for the data analysis. The consistency regularized data were used in modelling the negative thermal expansion property studied within one particular class of solid oxides by using different machine learning approaches including DGP performing the data consistency regularization at the model level. The statistical parameters of modelling demonstrate the enhanced performance of about 30 % in the predictive error decrease. Deep Autoencoding Gaussian Mixture Model has been involved to analyze the impact of the input consistency regularization on the changes in the data between the obtained "equilibrium" state and the nominal data [95]. The redistribution of the energy and reconstruction error of the method responsible for the anomalies detection allows one to infer on the additional feature of the input consistency regularization procedure to distinguish the input data containing the novelty.

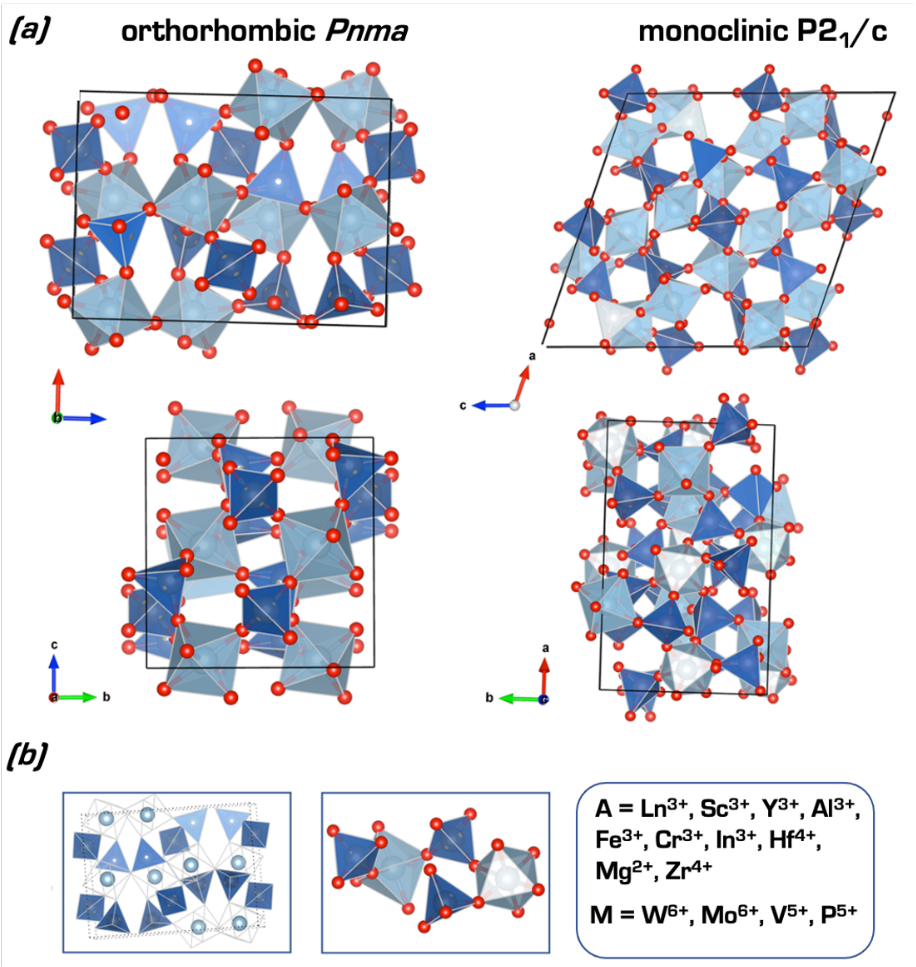

Figure 1: *(a)* Structure of $Dy_2(WO_4)_3$ compound: orthorhombic (*Pnma*) (NTE) and monoclinic ($P2_1/c$) (PTE) polymorphs; *(b)* diffusion path, lantern and list of cations could be accomodated by structure type (orthorhombic).

2 Materials and methods

2.1 Methods

2.1.1 Diffusion Probabilistic Models

In classical thermodynamics, Clausius defines a reversible process as a slowly varying process wherein successive states of this process differ by infinitesimals from the equilibrium system states. Diffusion probabilistic models directly implement this principle in this precise formulation in practice by using the large variety of approaches. Diffusion Probabilistic Model (DPM) is a parameterized Markov chain trained using the variational inference. In all the works authors consider the stationary distributions of time-homogeneous processes, where the data are measurably embedded in a space (usually Euclidean). Important statement for this methodology is that if the Markov chain is ergodic, then its stationary distribution will define the joint distribution between the random variables, even if the conditionals are not consistent with it. This, in particular, allows to work with the complex multimodal distributions satisfying some requirements/conditions [8]. In the special cases the group of the DPMs which use the Riemannian metric is used. The consecutive steps of introducing slowly increasing noise according to a predetermined rule, followed by removing such a noise using the backward steps in the case of sufficient training performance, determined, in particular, by a limited amount of Gaussian noise [10], guarantee the obtaining of samples of increased reliability at the sampling stage after a finite time. The linear drift part is considered as self-adjoint and uniformly positive while the diffusion term is assumed to match a tractable Gaussian distribution.
Let $p_\Theta(\mathrm{x}_0)$ denote the unknown distribution of a dataset consisting of D-dimensional i.i.d. samples. In diffusion probabilistic models, the problem can be formulated as a learning a distribution $\mathrm{p}_\Theta(\mathrm{x}_0)$ that approximates target $\mathrm{q}(\mathrm{x}_0)$

distribution and allows to sample from by diffusing *p(x)* towards the noise distribution followed by the reverse process. Diffusion probabilistic models (DPMs) are the latent variable models that can be represented in the form:

$$p_\Theta(x_0) = \int p_\Theta(x_{0:T})dx_{1:T}, \tag{1}$$

where

$$p_\Theta(x_{0:T}) := p_\Theta(x_T)\prod_{t=1}^{T} p_\Theta^{(t)}(x_{t-1}|x_0) \tag{2}$$

here x_1, ..., x_T are latent variables in the same sample space as x_0. The parameters Θ are learned to fit the data distribution *q(x_0)* by maximizing the variational lower bound, where *q($x_{1:T}|x_0$)* is some inference distribution over the latent variables:

$$\max_\Theta E_{q(x_0)}[logp_{\Theta(x_0)}] \leq \max_\Theta E_{q(x_0,x_1,\ldots,x_T)}[logp_{\Theta(x_{0:T})} - logq(x_{1:T}|x_0)] \tag{3}$$

The problem can be formulated as follows:
dX(t)=(-AX(t)+f(X(t)))dt+σ(X(t))dW(t)
X(0) = x $\in$ E,
where E - is a state space of X measurably embedded in a Hilbert space H, *-AX(t)+f(X(t)))dt* corresponds to a drift term, *σ(X(t))dW(t)* corresponds to a diffusion term, for which W(t)$_{t\geq 0}$ is a Wiener process. The solution of this stochastic differential equation is a diffusion process (or Feller process) X(t)$_{t\in[0..T]}$, with the imposed time constraints that is important condition for the problem formulation [28]. The resulting prior distribution at t = T supposed to be analytically tractable and is considered as a starting point for the reverse process.
The known weakness of this family of methods follows from the thermodynamic sense that the changes in these systems have to be of such a constrained character to guarantee the reversibility of the stochastic process and therefore the large number of the steps is required. In [30] authors have stated the notion of the z-reversibility for the case of a finite state space extending to the arbitrary state space introducing the balanced partitions concept and using the Markov transition kernel to better control the transition from t to $t+1$ thus providing with the very convenient way to avoid the extensive estimates of the conditional probabilities for the complete stochastic sequence. The similar methodology was implemented in several diffusion probabilistic models without, however, the detailed discussion at the level of the theory of the stochastic processes. The time-reversal process is used to generate the refined samples by denoising the distribution. The following SDE equation corresponds to this denoising procedure: *dx = [f(x,t) - g(t)$^2\nabla_x$logp$_t$(x)]dt = g(t)d$\bar{w}$*. Here, we provide with the score based method aimed at the learning of the time-dependent score function ∇_xlogp$_t$(x). In a process of learning the loss function is minimized, the one of the commonly used is the following:

$$J_{SM}(\Theta;\lambda(\cdot)) := \frac{1}{2}\int_0^T Ep_t(x)[\lambda(t)||\nabla logp_t(x) - s\Theta(x,t)||_2^2]dt \tag{4}$$

where λ:[0,T] is a positive weighting function. In practice, this parameter is defined as $\propto \frac{1}{E}[||\nabla_{x(t)} logp_{0t}(x(t|x(0))))||_2^2]$.
The transition kernel *logp$_{0t}$(x(t)|x(0))* is the object to elucidation.
The scheme in Figure 2 shows the main groups of methods related to the family of DPMs.
Afterwards the approach of Sohl-Dickstein et al. [76] the numerous methods were proposed [81, 40, 57, 90, 62, 11, 72].
In Denoising Diffusion Probabilistic Models [40] authors fix the forward process variances β_t to constants: Σ_Θ(x$_t$, t) = σ_t^2I.
$p_\Theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\ \mu_\Theta(x_t,\ t),\ \sigma_t^2 I)$
Thus, the approximate posterior q has no learnable parameters. Second, to represent the mean μ_Θ(x$_t$, t) authors propose a specific parameterization: a model that predicts $\tilde{\mu}_t$, the forward process posterior mean.
In Denoising Diffusion Implicit Models [77] authors propose the algorithm based on the non-Markovian inference process replacing the original one. Authors implement the forward process defined not on all the latent variables x$_{1:T}$, but on a subset of the latent variables. This simplification was suggested as a result of observation that DPM objective depends only on the marginals *q($x_t|x_0$)* rather than on the joint q(x$_{1:T}$|x$_0$) distribution. The obtained results demonstrate that such a reducing does not affect the performance of the obtained solution. In improved denoising diffusion models[62] authors demonstrate that learning the variances of the reverse diffusion process allows sampling with an order of magnitude fewer forward passes while a negligible difference in sample quality. In [57] Pseudo-Numerical Diffusion Methods were used to solve ordinary differential equations (ODEs). The improvement of optimization was demonstrated in both the convergence rate and the solution accuracy by means of the junction of Runge-Kotta method with the linear multi-step method. This extension has the benefits over its counterparts in a way how the decision is obtained. The optimization process consists of two steps of different intervals: first is based on Runge-Kutta method (RK) for faster convergence of the algorithm while the second one is the pseudo-linear multi-step method (pseudo-LMS) related to PNDM approaches, which is intended to find the optimization minimum by using more fine steps based on the second derivative of the coordinate estimates.

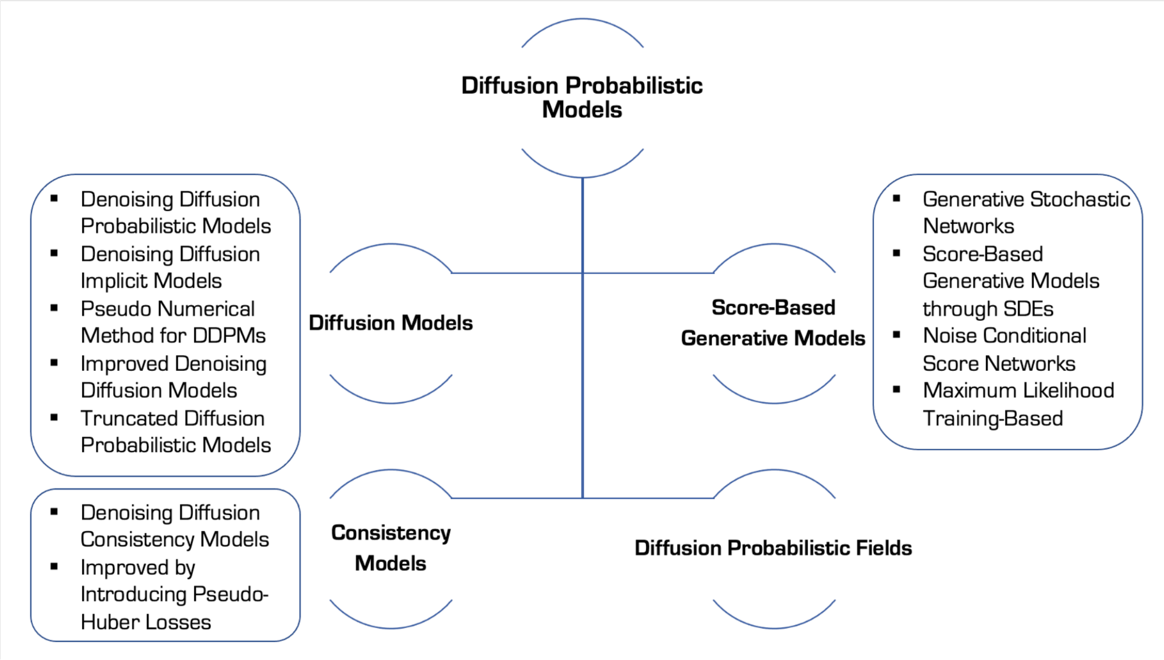
Diffusion Probabilistic Models
Denoising Diffusion Probabilistic Models
Denoising Diffusion Implicit Models
Pseudo Numerical Method for DDPMs
Improved Denoising Diffusion Models
Truncated Diffusion Probabilistic Models
Diffusion Models
Score-Based Generative Models
Generative Stochastic Networks
Score-Based Generative Models through SDEs
Noise Conditional Score Networks
Maximum Likelihood Training-Based
Denoising Diffusion Consistency Models
Improved by Introducing Pseudo-Huber Losses
Consistency Models
Diffusion Probabilistic Fields

Figure 2: Scheme representing the main groups of methods related to Diffusion Probabilistic Models family.

The linear multi-step method satisfies the following condition:
$\mathrm{x}_{t+\delta}$=x_t+$\frac{\delta}{24}(55f_t)$-$59f_{t-\delta}$+$37f_{t-2\delta}$-$9f_{t-3\delta}), f_t = f(\mathrm{x}_t, \mathrm{t})$
In [94] proposed method is based on truncating the forward diffusion chain and learning the distribution at the corresponding time coordinate to start the reverse diffusion process. Thus, the process of a noise introduction is performed up to the certain threshold corresponding to the state, where the initial distribution is not completely blurred by the noise introduction and, therefore, the information on the original data is largely preserved.
Score-based generative models is a type of the diffusion probabilistic models. For continuous state spaces, the DPM training objective implicitly computes scores at each noise scale. Generative Stochastic Networks (GSNs) [9] is one of the first works on DPMs. Given an original data point x of *P(x)*, GSN obtain a corrupted $\tilde{x}$ by sampling from some "corruption" distribution (some analogue of the noise-injected distribution) followed by training the network to reconstruct the point. The basic idea behind is that such a reconstructed distribution $P(x|\tilde{x})$ is evidently easier to estimate compared to the learning the whole data distribution while it should guarantees the same quality of solution. Besides, the work is highly interesting by the theoretical analysis of the diffusion probabilistic models in general. One of the important observations emphasized by authors is concerned with the artificial replacement of the methodology of unsupervised density estimation by the problem more similar to the supervised learning due to the assigned role of the loss function. It is well-known that this could principally enhance the performance of the models obtained. Another important aspect considered in this study to pay attention is that the reconstructed distribution $P(x|\tilde{x})$ must be of the same complexity with *P(x)*: to model the complex *P(x)* we need to provide both $P(x|\tilde{x})$ and corrupted distribition $\mathrm{C}(\tilde{x}|\mathrm{x})$ of similar complexity. Another important study using the same principle is Score-Based Generative Modeling through stochastic differential equations [81]. SDE depends only on the time-dependent gradient field and, thus, one can use the numerical SDE solvers to generate samples. In [81] authors propose to improve score-based generative modeling by perturbing the data using various levels of noise and simultaneously estimating scores corresponding to all noise levels by training a single conditional score network. In [80] authors have shown that using the intentionally introduced weighting function λ(t), the combination of score matching losses $\mathrm{J}_{SM}(\Theta;\lambda(\cdot))$ can be considered as an upper bound on Kullback-Leibler divergence D_{KL}, and therefore, serve as an efficient proxy for maximum likelihood training.
Recently, Consistency Models (CMs) [79, 78] were proposed as a stand-alone type of DPMs. In these models authors propose to learn the probability flows (PFs), ordinary diffusion equations, to be able to map any point at any time to the trajectory-defined starting point.
$dx_t = [\mu(x_t, t)\text{-}1/2\ \sigma(t)^2\nabla log\ p_t(x_t)]dt$
Authors adopt the setting from [46] with zero mean (μ(x, t) = 0) and variance σ(t) = $\sqrt{2t}$. By using a generative model to sample the point from the data distribution the score of the model is obtained via score matching procedure $\mathrm{s}_\phi(\mathrm{x,t})\approx$ ∇log p_t(x) followed by using this score model to obtain the estimate of the PFs as follows:
$\mathrm{dx}_t/\mathrm{dt}$ = -t $\mathrm{s}_\phi(x,t)$

This equation is called the empirical PF ODE. As the SDE is designed to match a tractable Gaussian distribution π(x), one can sample from the final point of the solution trajectory $\hat{x}_t \sim \pi = \mathcal{N}(0, T^2 I)$ to initialize the PF ODE and to reverse toward the origin with any numerical solvers. Improved modification [78] uses metric learning (Pseudo-Huber metric function). In Neural Diffusion Models (NDMs) [6], which are related by authors to the modifications of [77]), authors propose to induce the non-linear transformations for the data that extends the refinement capability of DPMs. Authors also noticed the parallels of DPMs with the variational autoencoders, where the latent variables are inferred using scaling of data points and injecting of Gaussian noise. We would like also noticed the alternative parallels with the Convolutional Neural Networks (CNNs) with, however, the opposite to CNNs principle directed rather to deconvolute-convolute operations.

The separate direction of using DPMs for the data of a complex geometry was distinguished by the fast growth of the approaches involving the Riemannian manifold concept [13, 29, 23]. Another interesting representative of the DPMs are Autoregressive Diffusion Models [42].

The inherent property of the diffusion probabilistic models, namely, the operational time may be any measure allowing to preserve the ergodicity in the function approximating the data distrubution. As such a measure one can consider the temperature, volume, or such variables as a depth of penetration, force and any suggesting the clear physical or chemical meaning while does not dispense with the ergodicity condition [27]. The rationale from the stochastic resetting [68] (Figure 3) supports the hypothesis on achieving the enhancement in both the performance of learning the DPMs as well as, more generally, the input consistency regularization. The regularized values of the observed variables can be considered as corresponding to the thermodynamic equilibrium while the input data before the regularization may contain the data characterizing by the deviations from the system equilibrium. This observation is also in agreement with the observations made by Caruana et al. [14].

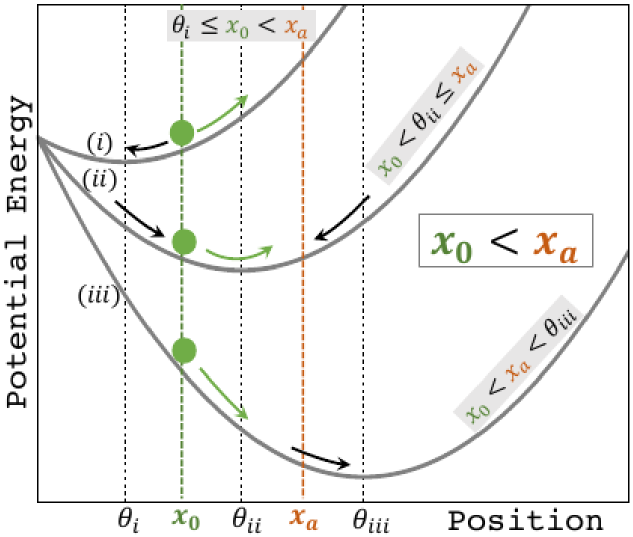

Figure 3: Stochastic resetting [68] as a decision aimed at the search of global minimum of energy. According to the theory, homogeneity of the chemical potential is a condition for thermodynamic equilibrium of the system being studied [58] and, therefore, the regularized values of the observed variables should be considered as corresponding to the thermodynamic equilibrium while the input data before the regularization should be considered as included those with the deviations from the system equilibrium. The figure is reproduced with permission from [68] APS 2022.

Conditional Score-Based Diffusion models [82].

This approach was developed for processing the time series data while can work with other types of the data, and the conditional probability-based implementation of the diffusion model allows one to exploit the information on the correlations between the observed values. This method directly learns the conditional distribution in contrast to other diffusion models for imputing the missing data in time series.

Let us denote the input data as $X = \{X_{1:K,1:L}\} \in R^{KxL}$. The observation mask $M = \{m_{1:K,1:L}\} \in 0{,}1^{KxL}$ is used for realizing the capability to use this approach as a data imputation technique. The time steps denoted as $S = \{s_{1:L}\}$. Given samples x_0 one can distinguish the data with completely or partially missing data. The imputation targets or the data afforded for the evaluation x_0^{ta} are generated by using the conditional distribution based on the observed data samples x^{co}. The aim is to estimate the true conditional data distribution $q(x_0^{ta}|x_0^{co})$ by using the distribution approximated

by the model $p_{\Theta}(x_0^{ta}|x_0^{co})$. This approach differs from other DPMs in the extending the parameterization of DPM to the conditional case defining the conditional denoising function ε_{Θ}:($\mathrm{X}^{ta} \times R|\mathrm{X}^{co}$) $\rightarrow X^{ta}$, which takes conditional observations x^{co} as the input. The following parameterization is performed: $\mu_{\Theta}(\mathrm{x}_t^{ta}|\mathrm{x}_0^{co}) = \mu^{DPM}(x_t^{ta}, t, \varepsilon(x_t^{ta}, t|\mathrm{x}_0^{co}))$, $\sigma\Theta(x_t^{ta}, t)|\mathrm{x}_0^{co}) = \sigma^{DPM}(\mathrm{x}_t^{ta}, t)$. With the function$\varepsilon_{\Theta}$ and data x_0 the data generation is performed in a reverse diffusion process by sampling the data using the following conditional denoising function and the loss function:
$\min_{\Theta}\mathscr{L}(\Theta)$:=$\min_{\Theta}E_{x_0 \sim q(x_0, \varepsilon \sim \mathscr{N})(0,I),t}||(\varepsilon\text{-}\varepsilon_{\Theta}(\mathrm{x}_t^{ta}, t|x_0^{co}))||_2^2$.
We have used two schemes of noise introduction: *(i)* the progressively introduced noise according to the following pattern (α $[\mathrm{t}]^{0.5}$)*X_{observ} + $\beta[\mathrm{t}]^{0.5}$ * noise (where the α = 1-β and β is constrained (β_{start} = 0.0001, β_{start} = 0.05 using the following defined constraints ${\beta_{start}}^{0.9}$ and ${\beta_{end}}^{0.9}$) and *(ii)* introducing the noise according to the following equation $\int_0^1 \sqrt{x - x^2}dx$. In some allied fields the term band-pass filtering is applied for describing the cases of the noise constrained in a specific way. The noise magnitude and the schedule is of primary importance since the correct estimation of the lower and upper bounds of the distribution defines the diffusion process [30]. This allows one to avoid the necessity to use the special techniques for generative models (preserving the local character of generator, locally compact character of the topological space and guarantying the twice differentiability of the functions essentially determined by the knowledge of *P(t, x, Γ)* and the initial probability distribution μ*(Γ = Pr{X(0) $\in$ Γ}))* as well as to the diffusion model itself. Thus, the problem of the choice of the noise amplitude and schedule can be formulated as achieving the coarsest topology that makes functions continuous. From the other hand, the noise term can be linked to all the factors which were not accounted by the model directly and, therefore, the amplitude of the noise is responsible for the DPM performance.

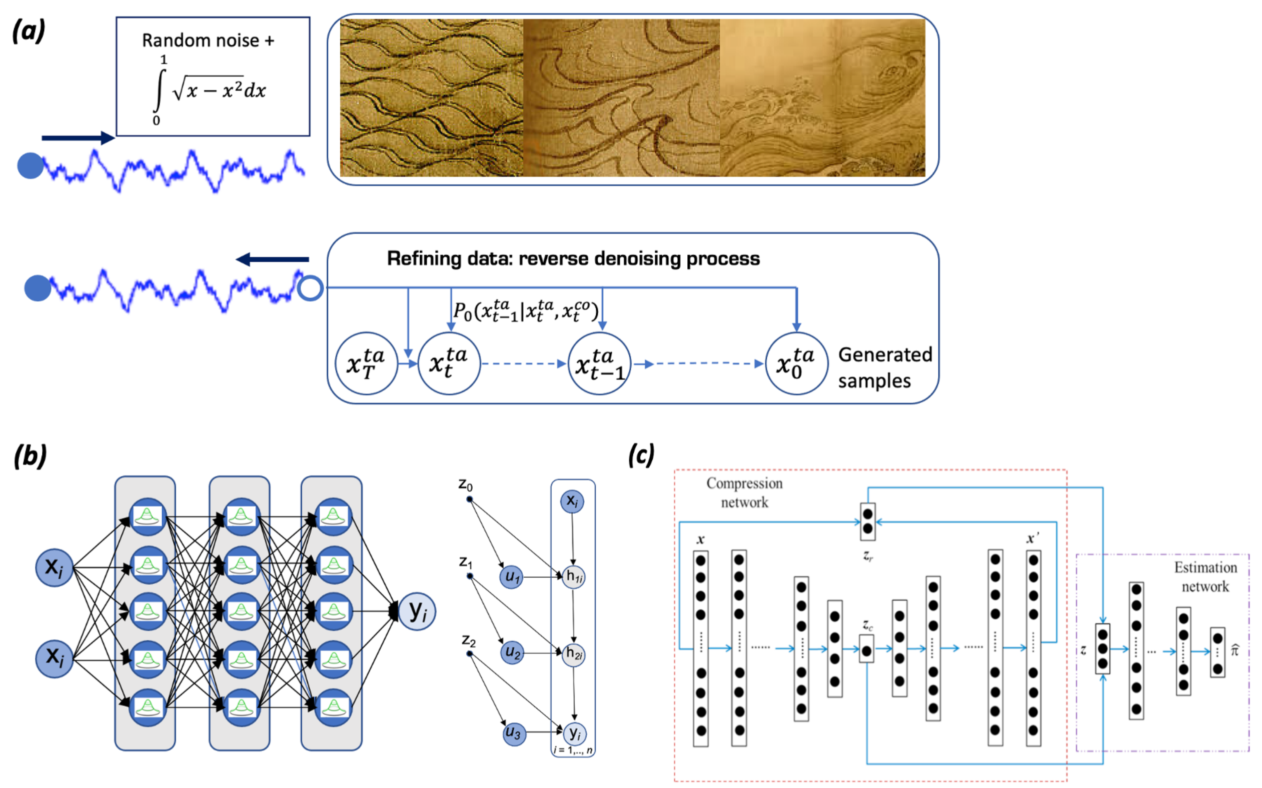

Figure 4: Scheme of machine learning methods used in this study: *(a)* Deep Gaussian Processes (DGP), *(b)* Conditional Score-based Diffusion models for Imputation (CSDI), *(b)* Deep Autoencoder Gaussian Mixture Model (DAGGM) (reprinted from [95]). Part of this Figure reproduces the works of Ma Yuang, a Chinese painter of the Song dynasty.

2.1.2 Regression methods

Deep Gaussian Processes [21]. DGP are a deep belief network based on Gaussian process method. The DGPs are organized in the hierarchy in a way similar to the neural networks, where the inputs of one layer are the outputs of the previous layer. A single layer model is equivalent to the individual GP or the GP latent variable model (GP-LVM). The benefits of this methodology is two-fold: *(i)* deep models are known to have the advantages over the shallow methods

as they are able to extract more complex relationships from the processed data and *(ii)* the proposed method allows one to perform some sort of the feature weighting that significantly improves the stability of the models for the data of different complexity. Principally, authors have realized the variational inference to marginalize the latent variables in the hierarchy. The probabilistic nature of the algorithm additionally contributes to the model stability that otherwise can be sensitive to the outlier objects.
The proposed architecture can be represented as a graphical model with the three kinds of nodes: *(i)* the leaf nodes $Y \in \mathbb{R}^{N \times D}$, where D is the number of Gaussian process priors, *(ii)* the intermediate latent spaces in amount of the hidden layers in the architecture $X_h \in \mathbb{R}^{N \times Q_h}$, where $h = 1,..., H$ (H is the number of hidden layers) and *(iii)* the parent latent node Z that can be unobserved in the architectures of the special purposes $Z = X_H \in \mathbb{R}^{N \times Q_h}$.
For the architectures containing only two hidden units, the generative process takes the form:
$y_{nd} = f^y_d(x_n) + \varepsilon^y_{nd}$, where $d = 1,..., D$; $x_n \in \mathbb{R}^Q$
$x_{nq} = f^x_q(z_n) + \varepsilon^x_{nq}$, where $q = 1,..., Q$; $z_n \in \mathbb{R}^{Q_z}$.
where f^y and f^x are the input and the output, respectively. Authors have proposed two features: at the first step, they have introduced the automatic relevance determination (ARD) covariance functions for the GPs:
$K(x_i, x_j) = \sigma^2_{ard} e^{-1/2 \sum_{q=1}^{Q} w_q (x_{i,q} - x_{j,q})^2}$
To avoid the difficulties for Bayesian treatment of the introduced nonlinearities, which are, however, of principal importance for the weighting, the special pseudo-inputs are introduced that are known as the inducing points. Their number K is defined in the configuration of the model. The location of these points is optimized by the likelihood function. The properly found locations correspond to the most confident data. The cases, when there are difficulties with finding these locations if are observed require some additional operations with input data as this discrepancy is of critical character for finding the robust solution [75]. This allows to realize some sort of the surrogate function defining the performance of the models developed during the training procedure.
Probabilistic Backpropagation Bayesian Neural Networks (PBP) [35]. Given data $\mathcal{D} = \{\mathrm{x}_n, \mathrm{y}_n\}^N_{n=1}$, where $\mathrm{x}_n \in \mathbb{R}^{\mathcal{D}}$ and corresponding scalar variables $\mathrm{y}_n \in \mathbb{R}$, $\mathrm{y}_n = f(\mathrm{x}_n; \mathscr{W}) + \varepsilon_n$, where $f(\cdot; \mathscr{W})$ is the output value with weights given by $\mathscr{W}$ and noise variables ε_n.
The likelihood of dependent variable given weights $\mathscr{W}$ and the noise precision γ is defined as following:

$$p(y|\mathscr{W}, X, \gamma) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(x_n; \mathscr{W}), \gamma^{-1}) \tag{5}$$

PBP does not use the point estimates of the weights during the training instead the set of the Gaussians is generated:

$$p(\mathscr{W}|\lambda) = \prod_{l=1}^{L} \prod_{i=1}^{V_l} \prod_{j=1}^{V_{l-1}+1} \mathcal{N}(w_{ij,l} | 0, \lambda^{-1}) \tag{6}$$

where $w_{ij,l}$ is the weights and λ is a precision parameter. The λ value is a tunable parameter.
The posterior distribution for the parameters $\mathscr{W}$, γ and λ can then be obtained according to Bayes' rule. The output predictions are performed using predictive posterior distribution:

$$p(y_{target} | x_{target}, \mathcal{D}) = \int p(y_{target} | x_{target}, \mathscr{W}, \gamma) p(\mathscr{W}, \gamma, \delta | \mathcal{D}) d\gamma d\lambda d\mathscr{W} \tag{7}$$

where $p(\mathrm{y}_{target} | \mathrm{x}_{target}, \mathscr{W}, \gamma) = \mathcal{N}(\mathrm{y}_{target} | f(\mathrm{x}_{target}), \gamma)$. At the end of the forward stage, PBP computes the logarithm of the marginal probability of the dependent variable. At the stage of backward propagation, the network propagates the gradient of this quantity with respect to the means and the variances of the approximate Gaussian posterior, which in turn are used to update the corresponding values of the means and the variances of the posterior approximation of the Gaussians. The weights are updated as follows:

$$s(w) = Z^{-1} f(w) \mathcal{N}(w | m, v) \tag{8}$$

where Z is the normalization constant. The updated values for the means and the variances are obtained using the gradient of the logarithm of the normalization constant Z:

$$m^{new} = m + v \frac{\partial logZ}{\partial m} \tag{9}$$

$$v^{new} = v - v^2 \left[\left(\frac{\partial logZ}{\partial m} \right)^2 - 2 \left(\frac{\partial logZ}{\partial v} \right) \right] \tag{10}$$

Gradient Tree Boosting (XGBoost) from XGBoost package [18] was used in this study for the regression problem. XGBoost regressor was used in this study with the following parameters: max depth=3, learning rate=0.1, number of estimators = 100 and booster 'gbtree'.

2.1.3 Novelty/anomalies detection methods

Deep Autoencoding Gaussian Mixture Model [95]. In its original form, the autoencoders that are nowadays at the root of one of the most popular family of statistical learning methods rose to prominence in the late 80's [5] and gained traction during the last decades simultaneously with onrush of the methodology of deep learning [38]. The method involved in this study, Deep Autoencoding Gaussian Mixture Model (DAGGM), is a hybrid of variational deep autoencoder and Gaussian Mixture Model thus providing with additional powerful capabilities.

DAGGM consists of two components: the compression network and the estimation network. The compression network functionalizes in an enforced manner by outputting the data of two types: *(i)* the reduced low-dimensional representation learned by a deep autoencoder and *(ii)* the data derived using the reconstruction error evaluated by the decoder component. The estimation networks uses both types of information to evaluate the likelihood/energy by using Gaussian Mixture Model (GMM). Given the low-dimensional representation, the estimation network outputs the probability density of the data by using GMM. During the training with some initialized mixture component distribution ϕ, the mixture means μ, and the mixture covariance Σ, the network estimates the parameters without using procedures as Expectation-Maximization (EM) by means of multilayer neural network to predict the mixture membership for each of the samples.

$$p = MLN(z; \Theta_m) \tag{11}$$

$$\gamma = softmax(p) \tag{12}$$

where γ is a K-dimensional vector for the soft mixture-component membership prediction, K is a number of Gaussians and p is the output of a multilayer network parameterized by Θ_m.

2.1.4 Metric learning

Metric Learning for Kernel Regression (MLKR) was involved in our study [88]. Despite the impressive performance demonstrated for the data description-property landscape, we were not successful in obtaining the stable 10 fold-cross-validation results and thus in our study this method was used only for the first purpose. However, we assume that the Mahalanobis distance used in the supervised version of the metric learning, Large Margin Nearest Neighbors method [87] would be very efficient as was shown in one of our earlier work [52].

2.2 Descriptors

The parameters used in the study as the descriptors can be related to several categories: ***(i)*** a composition of the compounds, ***(ii)*** atomic characteristics (Shannon ionic radii, electronegativity (Pauling), work function and effective nuclear charge) [1], ***(iii)*** synthesis time and temperature.

The atomic characteristics are well-known and successfully applied in materials informatics as a simple materials fingerprints [85, 16]. Among the atomic characteristics involved in this study the choice of Shannon ionic radii and electronegativity (Pauling) was expected due to the numerous discussions in the literature.

In recent years the descriptors related to the process of synthesis were efficiently used in materials informatics [48, 67, 91, 53, 49]. The same descriptors set has shown good results on the same data set[51].

2.3 Performance evaluation

Statistical coefficients for regression problem. Predictive performance of developed regression models was assessed using the ten-fold external cross-validation (10-CV) procedure where the entire data set was split into ten non-overlapping pairs of the training and test sets of compounds. The models were obtained using the training set followed by their validation using the corresponding test set. The determination coefficient R^2 and root mean square error ($RMSE$) and p-value were used as the statistical parameters of the models:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_{(pred,i)} - Y_{(exp,i)})^2}{\sum_{i=1}^{N}(Y_{(exp,i)} - \bar{Y}_{exp})^2} \tag{13}$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{1}{N}(Y_{(pred,i)} - Y_{(exp,i)})^2} \tag{14}$$

Here, $\mathrm{Y}_{(exp,i)}$ and $\mathrm{Y}_{(pred,i)}$ are, respectively, experimental and predicted values of the modeling property, N is the number of data points, $\bar{Y}_{exp}$ is the average value of the experimental property.

The initial data set was reshuffled 100 times, followed by model development. The obtained results were averaged through the averageng the predicted values. The p-value was used for characterizing the statistical significance of the results obtained.

3 Results and Analysis

3.1 Regression problem: enhancement of predictive accuracy, a landscape descriptors-property, sharpness score

If we assume the descriptors reasonably describing the data this means the continuity in the data description–property landscape. In this study, the analysis of the data description–property landscape has allowed to draw the important conclusions on the impact of the chosen configuration of the input consistency regularization on the model performance. Figure 5 shows the impact of consistency regularization procedure on the property landscape: in contract to metric learning the metrics of space is preserved, two different diffusion models (conventional stepwise noise introduction vs. introduction of the noise in a mid-course manner according to the equation $\int_0^1 \sqrt{x-x^2}$) provide with the weak transformation of the landscape.

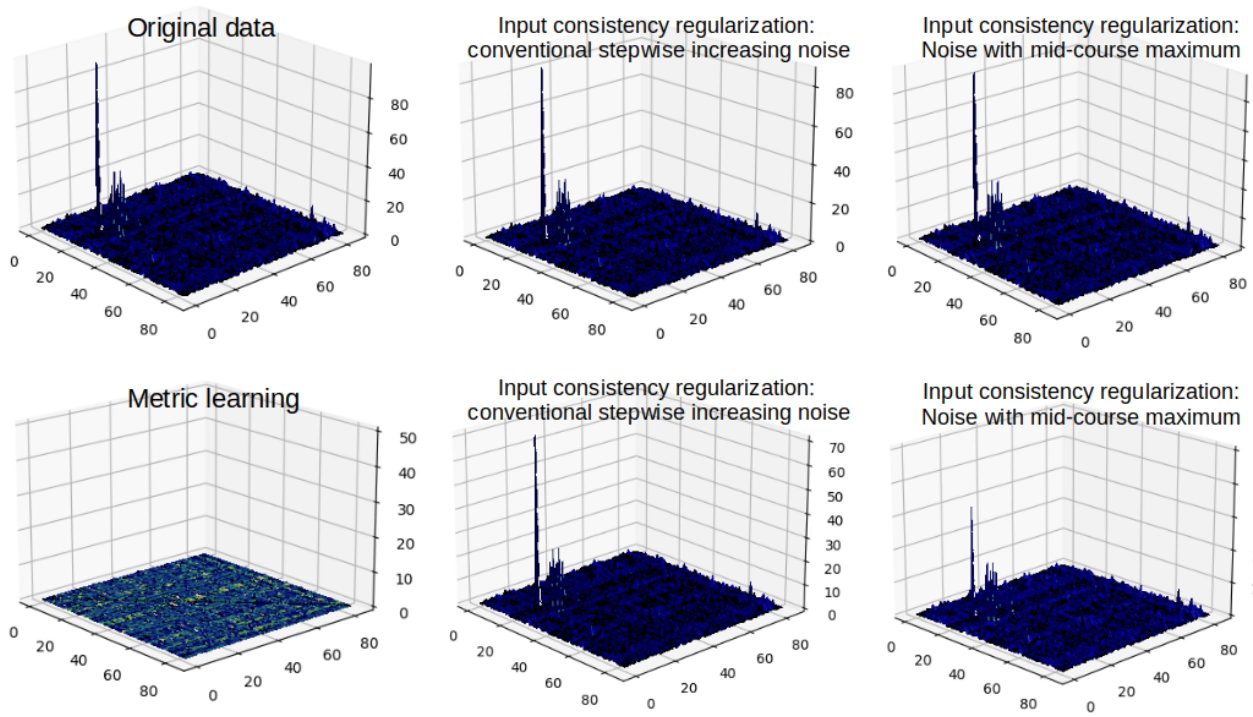

Figure 5: Impact of consistency regularization procedure on the property landscape: in contract to metric learning the metrics of space is preserved, two different diffusion models (conventional stepwise noise introduction vs. introduction of the noise in a mid-course manner according to the equation $\int_0^1 \sqrt{x-x^2}$) provide with the distinct transformation of the landscape.

The decrease in the cliff's magnitude is observed as a result of the input consistency regularization. The comparison with the statistical coefficients characterizing the model performance before and after the input consistency regularization shows the direct relationship between the magnitude of the discontinuities of the landscape and the model performance. From the Figure 5 one can see that the positions of the peaks remain the same. The known conditions on the local compactness of the noise magnitudes allows one to discuss the existence of the certain noise level providing with the enhancement of the data characteristics. This is in an agreement with the thermodynamics theory which introduces the clear notions of the reversibility of the processes occurred. Thus, increasing the noise magnitude above the certain threshold have resulted in the opposite trend in the statistical characteristics of the obtained models. Another conclusion can be made from the results obtained is that the magnitude of the noise has more severe impact compared with the noise type used by DPMs. This observation, however, requires more detailed verification involving the data of different

complexity and to the significantly extended set of DPMs.
The statistical parmeters of the models are the following: the R^2 value varies in the range of 0.78 to 0.80 for XGBoost, 0.74 to 0.75 for PBP and 0.71 to 0.72 for DGP as a function of the noise magnitude introduced in the models. The prediction error ($RMSE$) decreases from 1.266 to 0.88 for XGBoost, from 1.421 to 0.97 for PBP and from 1.606 to 0.993 for DGP. The error values varies in the reasonable range that confirms the overall stability of the models regardless using the input consistency regularization or not.

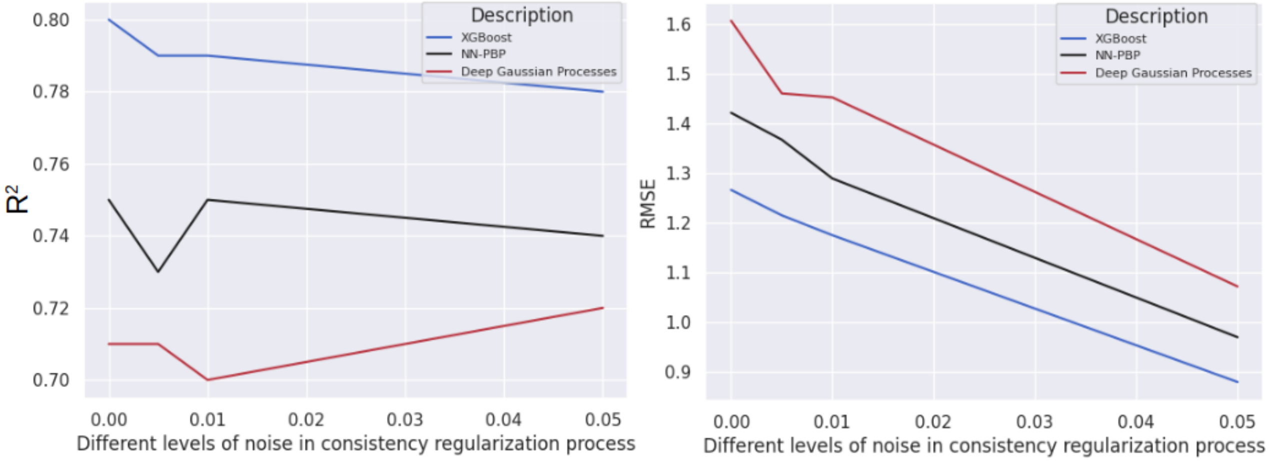

Figure 6: Statistical performance of XGBoost, PBP and DGP models: determination coefficient R^2 and $RMSE$ as a function of the noise introduced in the diffusion model.

Table 1: Statistical parameters for modeling the linear coefficient of thermal expansion α_L: determination coefficient R^2 and root mean square error ($RMSE$) with averaged (the number of models are shown in the brackets) std. errors for ten-fold cross-validation.

β or ω	N	**DGP** Avg. R^2 with std. error and p-value	**DGP** Avg. $RMSE$ and std. error	**XGBoost** Avg. R^2 with std. error and p-value	**XGBoost** Avg. $RMSE$ and std. error	**PBP** Avg. R^2 with std. error and p-value	**PBP** Avg. $RMSE$ and std. error
w/o	84	0.71±0.01, 3.5e-23 (10)	1.606±0.044 (10)	0.80±0.02, 1.24e-27 (100)	1.266±0.045 (100)	0.75±0.04, 1.6e-26 (100)	1.421±0.105(100)
0.005	84	0.71±0.02, 6.7e-22 (10)	1.460±0.056 (10)	0.79±0.02, 9.42e-30 (100)	1.215±0.045 (100)	0.73±0.05, 4.3e-25 (100)	1.367±0.124(100)
0.01	84	0.70±0.01, 3.0e-23 (10)	1.452±0.038 (10)	0.79±0.02, 6.0e-30 (100)	1.175±0.064 (100)	0.75±0.07, 2.2e-26 (100)	1.289±0.087(100)
0.05	84	0.72±0.008, 8.0e-24 (10)	1.072±0.014 (10)	0.78±0.02, 2.9e-29 (100)	0.880±0.036 (100)	0.74±0.03, 1.8e-25 (100)	0.970±0.073(100)
ω = 0.005	84	0.71±0.02, 8.8e-24 (10)	0.993±0.031 (10)	0.79±0.02, 2.95e-29 (100)	0.867±0.037 (100)	0.74±0.05, 8.6e-26 (100)	0.970±0.086(100)

One may see the almost insignificant variations in the determination coefficients R^2 of the models which do not exceed the estimated standard error value (Table 1) while the $RMSE$ values of the model predictions were decreased significantly. Importantly, the $RMSE$ decrease has a synchronous character between the involved methods. This can be considered as the evidence of the impact of the input consistency regularization procedure and is mathematically satisfied.

3.2 Hessian-based analysis of the loss function landscape as a criterion of model generalizability

Hessian of the loss function show the similar flat character for both types of the diffusion models (Figure 7). The landscape data description-property demonstrates decreased magnitude of the cliffs observed. This corresponds to the slightly more smooth pattern of MSE values for the second diffusion model. One can also note the decreasing MSE values with a time-step t. During the denoising procedure one can distinguish the local outliers in the MSE values for the second DPM. This, however, does not affect the final accuracy of the models. One of the most important observations were found is that the increase in the model performance appears the result of the increased difference in the denoised and the noised values of the property value. This means that the certain, very limited level of the deviations from the nominal data are desired, one may hypothesize that the enhancement expectations can be related and screened by using the Hessian of the difference in the denoised and the noised values of the property. The latter is the important conclusion that allows to use the Hessian-based analysis for determining the optimal values of the time-step thus providing with the early stop or, in more conventional terminology, the optimal time point of the truncation of the diffusion process with no respect what data or what the noise injection scheme is involved. One may mention the more prominent character of the discussed discrepancy for the second DPM which is presumably concerned with the feature of the denoising procedure that corresponds to the denoising of the first DPM while the noise introduction differs.

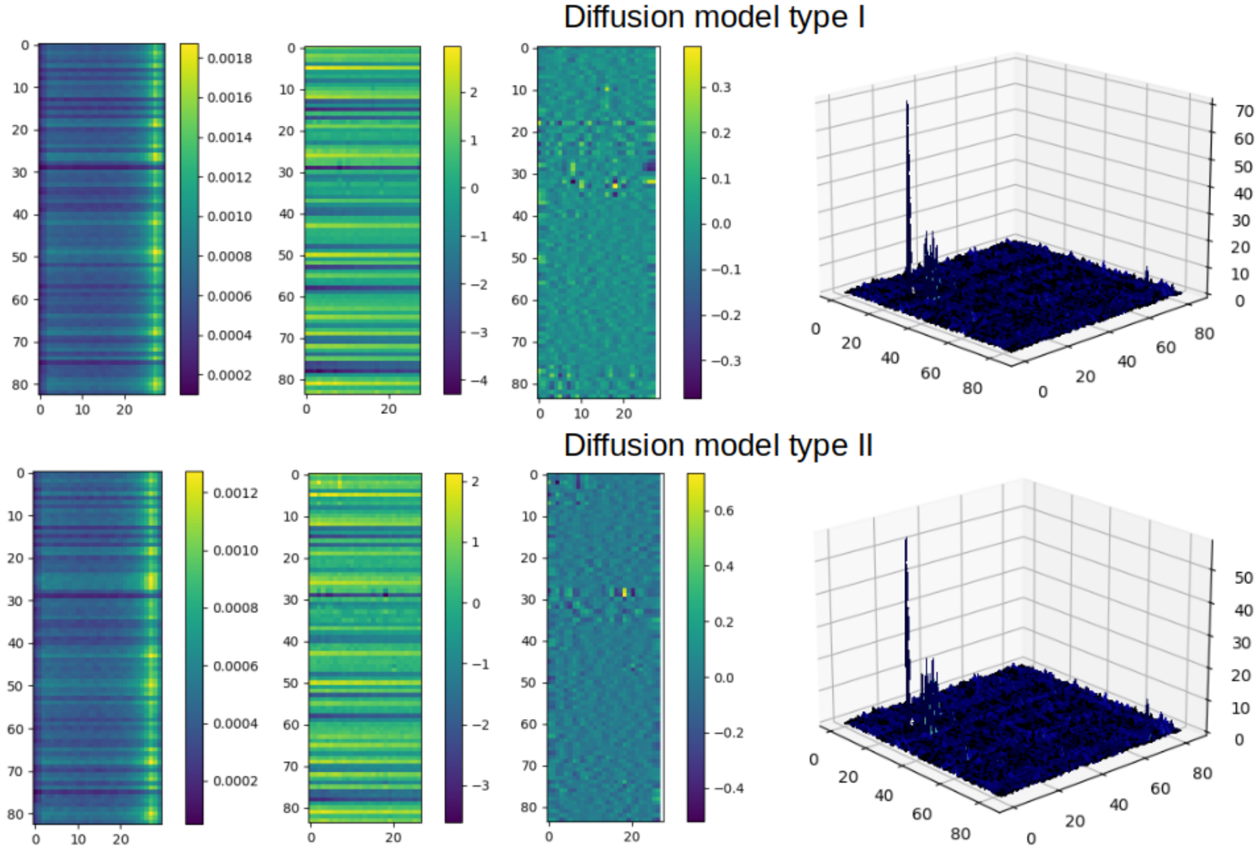

Figure 7: Hessian-based analysis of the results: (left) Hessian-based analysis of the difference in the denoised and the noised values of the linear thermal expansion coefficient α_L averaged over the descriptor values (CSDI), MSE function as a function of the time-step t, Hessian-based analysis of the experimental vs. predicted values (DGP) of the linear thermal expansion coefficient as a function of the time-step t α_L, (right) the corresponding plots of the data description-property landscape (DGP model with beta=0.05, omega = 0.1).

3.3 DAGGM-based analysis of the energy and reconstruction error values

In order to evaluate the data rearrangement invoked by the input consistency regularization the novelty detection method DAGGM was applied. Figure 8 shows the reconstruction error (RE) and the energy values (E) for the data obtained from the both types of diffusion models. One may see that the overall character of the data distribution according to these characteristics was preserved. However, it is worth noting that some observations (compounds) from the bins of high ER and E values were recognized as belonging to the bins with the lesser or higher outlierness which means that the input consistency regularization procedure can be considered as the additional pre-processing step for novelty detection. This study does not perform such an analysis due to the data diversity, however, this finding can effectively enhance the precision of the outlier/novelty detection approaches. The limited changes in the distribution is in agreement with the observations obtained from the data description-property landscape highlighting the difference between the input consistency regularization and metric learning. In the former case the metric remains preserved whereas in the latter case the metric is learned that corresponds to the more significant and data-specific procedure. Normally, some limited decrease in the outlierness of the data is expected as a result of the ICR procedure and, therefore, the opposite trend can be considered for more careful analysis for presence of the outliers/novelty in the data.

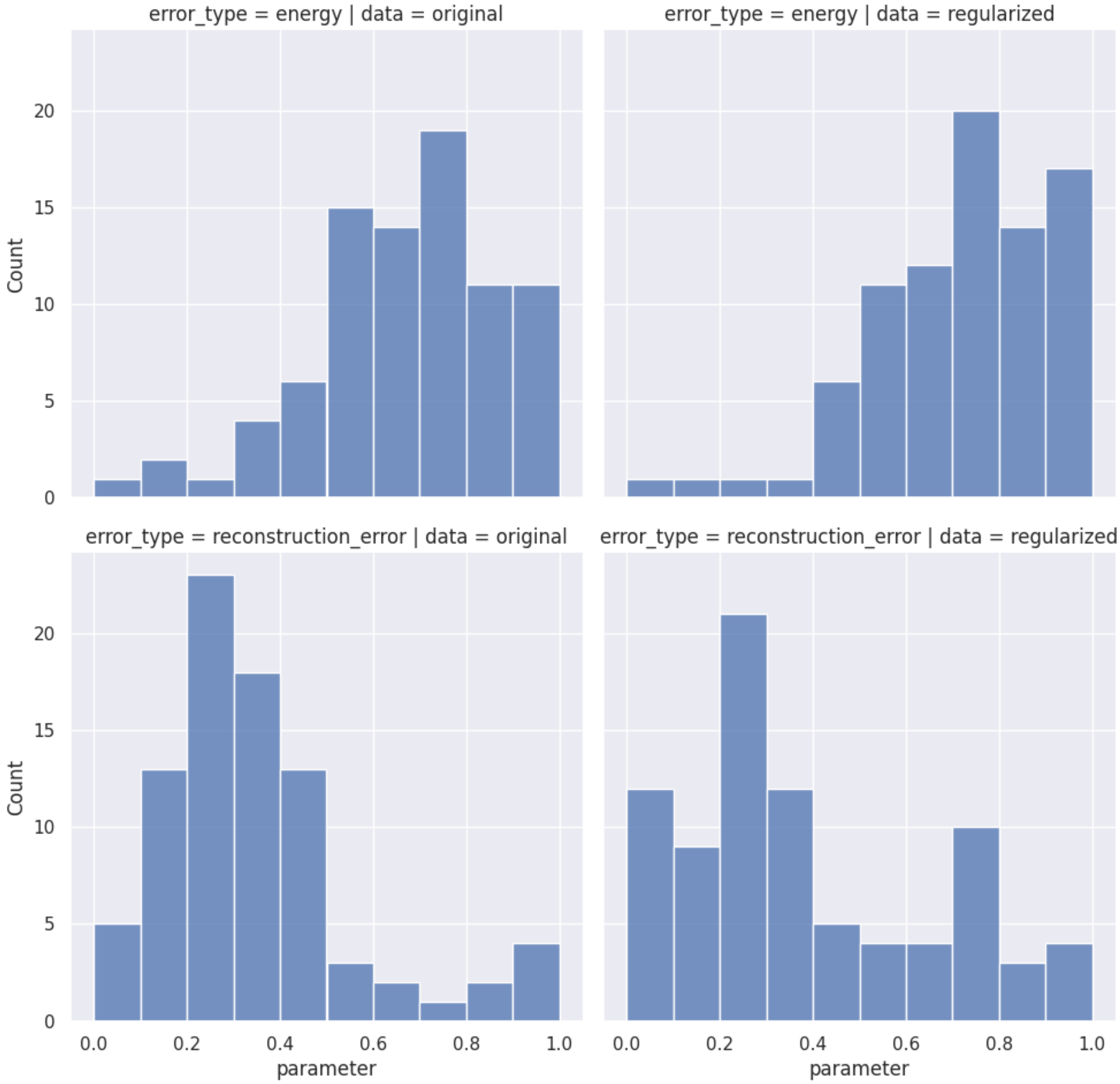

Figure 8: Distribution of energy and reconstruction error values before and after input consistency regularization.

4 Conclusions

This study demonstrates the performance of the input consistence regularization based on diffusion probabilistic models aimed at the refining of the experimental data for the regularization of the experimental data on linear thermal expansion coefficient obtained by high-temperature XPRD measurements. The results obtained in this study have shown the decrease in the predictive error of about 30 % for all the machine learning methods used in these studies. The Hessian-based analysis of the loss function has confirmed the stability of the obtained solution. One of the important results obtained in this study allows one to recommend the Hessian-based analysis for determining the optimal values of the number of time-step thus providing with the early stop or, in more conventional terminology, the optimal time point of the truncation of the diffusion process with no respect what data or the noise introduction scheme are involved. The similar character of the dependence of the RMSE error values as a function of the noise introduced in DPM observed for all machine learning methods used in this study allows to conclude on the reasonable approximation of the data density distribution by the diffusion probabilistic models. New methodology with the controlled noise amplitude has shown promise. The metric learning is recommended as the promising way to enhance the DPMs performance.

5 Conflict of interest

Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6 Acknowledgements

N.K. thanks Dr. Andrei Kazennov for the discussion of the methodology used in this study.

References

[1] Database of properties of chemical elements.

[2] M. Abbas, J. Kivinen, and T. Raiko. Understanding regularization by virtual adversarial training ladder networks and others. 05 2016.

[3] Y. S. Abu-Mostafa. Learning from hints in neural networks. *Journal of Complexity*, 6(2):192–198, 1990.

[4] P. Bachman, O. Alsharif, and D. Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 4, 12 2014.

[5] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[6] G. Bartosh, D. Vetrov, and C. A. Naesseth. Neural diffusion models, 2024.

[7] P. Becker, O. Arenz, and G. Neumann. Expected information maximization: Using the i-projection for mixture density estimation, 01 2020.

[8] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. *Advances in Neural Information Processing Systems*, 05 2013.

[9] Y. Bengio, Éric Thibodeau-Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop, 2014.

[10] S. Bernstein. *Sobranie Sochinenii. (Collected Works, 4 vols.).* Gostehizdat, Moscow-Leningrad, 1964.

[11] D. Berthelot, A. Autef, J. Lin, D. A. Yap, S. Zhai, S. Hu, D. Zheng, W. Talbott, and E. Gu. Tract: Denoising diffusion models with transitive closure time-distillation, 2023.

[12] Y. Burda, R. Grosse, and R. Salakhutdinov. Accurate and conservative estimates of mrf log-likelihood using reverse annealing. 12 2014.

[13] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 36:2814–2830, 2022.

[14] R. Caruana and V. de Sa. Promoting poor features to supervisors: Some inputs work better as outputs. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[15] S. Cfrrai. Ergodicity for stochastic reaction-diffusion systems with polynomial coefficients. *Stochastics and Stochastic Reports*, 67(1-2):17–51, 1999.

[16] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, and S. P. Ong. A critical review of machine learning of energy materials. *Advanced Energy Materials*, 10(8):1903242, 2020.

[17] J. Chen, Z. Wu, J. Zhang, and F. Li. Mutual information-based dropout: Learning deep relevant feature representation architectures. *Neurocomputing*, 361:173–184, 2019.

[18] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[19] P.-L. Chow and R. Z. Khasminskii. Stationary solutions of nonlinear stochastic evolution equations 1. *Stochastic Analysis and Applications*, 15(5):671–699, 1997.

[20] W. Clarke, G. Richardson, and P. Cameron. Understanding the full zoo of perovskite solar cell impedance spectra with the standard drift-diffusion model. *Advanced Energy Materials*, n/a(n/a):2400955.

[21] A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

[22] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural computation*, 7:889–904, Sep 1995.

[23] V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. Teh, and A. Doucet. Riemannian score-based generative modeling, Feb. 2022.

[24] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[25] J. Evans, T. Mary, and A. Sleight. Negative thermal expansion in sc2(wo4)3. *Journal of Solid State Chemistry*, 137(1):148–160, 1998.

[26] X. Fang, K. Kruse, T. Lu, and J. Wang. Nonequilibrium physics in biology. *Rev. Mod. Phys.*, 91:045004, Dec 2019.

[27] W. Feller. On the theory of stochastic processes, with particular reference to applications. 1949.

[28] W. Feller. Diffusion processes in one dimension. *Transactions of the American Mathematical Society*, 77:1–31, 1954.

[29] N. Fishman, L. Klarner, V. D. Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion models for constrained domains. *ArXiv*, abs/2304.05364, 2023.

[30] M. Gallegos and G. Ritter. Balanced partitions for markov chains. *Results in Mathematics*, 37, 05 2000.

[31] B. Ghorbani, S. Krishnan, and Y. Xiao. An investigation into neural net optimization via hessian eigenvalue density. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 09–15 Jun 2019.

[32] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra. Deep autoregressive networks. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1242–1250, Bejing, China, 22–24 Jun 2014. PMLR.

[33] J. Guan, W. W. Qian, X. Peng, Y. Su, J. Peng, and J. Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction, 2023.

[34] Z. Hao, S. Liu, Y. Zhang, C. Ying, Y. Feng, H. Su, and J. Zhu. Physics-informed machine learning: A survey on problems, methods and applications, 11 2022.

[35] J. M. Hernández-Lobato and R. P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks, 2015.

[36] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint*, arXiv, 07 2012.

[37] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[38] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[39] G. E. Hinton and D. van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Annual Conference Computational Learning Theory*, 1993.

[40] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020.

[41] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9:1–42, 02 1997.

[42] E. Hoogeboom, A. Gritsenko, J. Bastings, B. Poole, R. Berg, and T. Salimans. Autoregressive diffusion models, 10 2021.

[43] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling. Equivariant diffusion for molecule generation in 3d, 2022.

[44] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama. Learning discrete representations via information maximizing self-augmented training, 2017.

[45] P.-A. Kamienny, K. Arulkumaran, F. Behbahani, W. Boehmer, and S. Whiteson. Privileged information dropout in reinforcement learning, 2020.

[46] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models, 2022.

[47] N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. 09 2016.

[48] E. Kim, K. Huang, S. Jegelka, and E. Olivetti. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational Materials*, 3(1):53, 2017.

[49] N. Kireeva, V. S. Pervov, and A. Y. Tsivadze. Machine learning-based evaluation of functional characteristics of li-rich layered oxide cathode materials using the data of xps and xrd spectra. *Computational Materials Science*, 231:112591, 2024.

[50] N. Kireeva and A. Y. Tsivadze. Novelty detection in the design of synthesis of garnet-structured solid electrolytes. *Journal of Solid State Chemistry*, 334:124669, 2024.

[51] N. Kireeva and A. Y. Tsivadze. Oxide ceramics of a2m3o12 family with negative and close-to-zero thermal expansion coefficients: Machine learning-based modeling of functional characteristics. *Journal of Alloys and Compounds*, 990:174356, 2024.

[52] N. V. Kireeva, S. I. Ovchinnikova, S. L. Kuznetsov, A. M. Kazennov, and A. Y. Tsivadze. Impact of distance-based metric learning on classification and visualization model performance and structure-activity landscapes. *Journal of Computer-Aided Molecular Design*, 28(2):61–73, 2014.

[53] N. V. Kireeva, A. Y. Tsivadze, and V. S. Pervov. Modeling ionic conductivity and activation energy in garnet-structured solid electrolytes: The role of composition, grain boundaries and processing. *Solid State Ionics*, 399:116293, 2023.

[54] G. Leha and G. Ritter. Lyapunov-type conditions for stationary distributions of diffusion processes on hilbert spaces. *Stochastics and Stochastic Reports*, 48(3-4):195–225, 1994.

[55] G. Leha and G. Ritter. Lyapunov functions and stationary distributions of stochastic evolution equations. *Stochastic Analysis and Applications - STOCHASTIC ANAL APPL*, 21:763–799, 02 2007.

[56] H. Lin, Y. Huang, O. Zhang, L. Wu, S. Li, Z. Chen, and S. Z. Li. Functional-group-based diffusion for pocket-specific molecule generation and elaboration, 2024.

[57] L. Liu, Y. Ren, Z. Lin, and Z. Zhao. Pseudo numerical methods for diffusion models on manifolds, 02 2022.

[58] J. Maier. On the correlation of macroscopic and microscopic rate constants in solid state chemistry. *Solid State Ionics*, 112(3):197–228, 1998.

[59] S. McCandlish, J. Kaplan, D. Amodei, and O. Team. An empirical model of large-batch training, 12 2018.

[60] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 04 2017.

[61] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[62] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[63] A. Omote, S. Yotsuhashi, Y. Zenitani, and Y. Yamada. High ion conductivity in mghf(wo4)3 solids with ordered structure: 1-d alignments of mg2+ and hf4+ ions. *Journal of the American Ceramic Society*, 94(8):2285–2288, 2011.

[64] V. Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size, 2019.

[65] B. A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994.

[66] F. Petit. Time reversal and reflected diffusions. *Stochastic Processes and their Applications*, 69(1):25–53, 1997.

[67] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533:73 EP –, 05 2016.

[68] S. Ray. Expediting feller process with stochastic resetting. *Phys. Rev. E*, 106:034133, Sep 2022.

[69] N. Ronne, A. Aspuru-Guzik, and B. Hammer. Generative diffusion model for surface structure discovery, 2024.

[70] L. Sagun, L. Bottou, and Y. LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond, 2017.

[71] L. Sagun, U. Evci, V. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parametrized neural networks. 06 2017.

[72] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models, 2022.

[73] A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lio, C. Gomes, M. Welling, M. Bronstein, and B. Correia. Structure-based drug design with equivariant diffusion models, 10 2022.

[74] A. V. Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961.

[75] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

[76] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[77] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2022.

[78] Y. Song and P. Dhariwal. Improved techniques for training consistency models, 2023.

[79] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models, 2023.

[80] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models, 2021.

[81] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2021.

[82] Y. Tashiro, J. Song, Y. Song, and S. Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Neural Information Processing Systems*, 2021.

[83] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. Advances in Neural Networks Research: IJCNN2009.

[84] V. Vapnik, A. Vashist, and N. Pavlovitch. Learning using hidden information (learning with teacher). In *2009 International Joint Conference on Neural Networks*, pages 3188–3195, 2009.

[85] P. Villars, J. DAAMS, Y. SHIKATA, K. Rajan, and S. Iwata. A new approach to describe elemental-property parameters. 01 2008.

[86] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation, 05 2018.

[87] K. Weinberger, J. Blitzer, and L. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification*, volume 10. 01 2006.

[88] K. Weinberger and G. Tesauro. Metric learning for kernel regression. *Journal of Machine Learning Research - Proceedings Track*, 2:612–619, 01 2007.

[89] Z. Xie, I. Sato, and M. Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *ArXiv*, abs/2002.03495, 2020.

[90] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. 02 2018.

[91] S. R. Young, A. Maksov, M. Ziatdinov, Y. Cao, M. Burch, J. Balachandran, L. Li, S. Somnath, R. M. Patton, S. V. Kalinin, and R. K. Vasudevan. Data mining for better material synthesis: The case of pulsed laser deposition of complex oxides. *Journal of Applied Physics*, 123(11):115303, 2018.

[92] J. Zabczyk. Structural properties and limit behaviour of linear stochastic systems in hilbert spaces. *Banach Center Publications*, 14(1):591–609, 1985.

[93] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization, 2017.

[94] H. Zheng, P. He, W. Chen, and M. Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders, 2023.

[95] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. ki Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.