Complex-Valued Periodic Stochastic Processes for Retrieving Unknown Atomic Structures

Gyoung S. Na[1,2,*] **and Chanyoung Park**[2,*]
[1]Korea Research Institute of Chemical Technology (KRICT), Republic of Korea
[2]Korea Advanced Institute of Science & Technology (KAIST), Republic of Korea
ngs0@krict.re.kr, cy.park@kaist.ac.kr

Abstract

Retrieving unknown atomic structures from observable analytical spectra or images remains a long-standing challenge across natural sciences. However, retrieval accuracy of existing retrieval methods on analytical data remains suboptimal because they have overlooked the underlying periodic quantum mechanical perturbations of non-equilibrium atomic structures behind analytical data. This paper proposes a Periodic Complex Stochastic Process (PCSP) that models such periodic perturbations and establishes theoretical backgrounds of periodic stochastic process in the complex-valued domain, including its sample diversity, process length, and periodicity. Finally, we develop a Complex-Valued Cross-modal Retrieval (CVCR) by integrating PCSP with cross-modal retrieval frameworks. CVCR outperformed existing cross-modal retrieval methods in cross-modal retrieval tasks of real-world analytical chemistry. Moreover, CVCR achieved state-of-the-art retrieval accuracy in zero-shot settings on 40 million of molecules.

1 Introduction

The central insight of current successes of artificial intelligence lies in learning the relations between physical quantities of matters (e.g., molecules and crystalline materials) and their atomic structures [11, 48]. However, obtaining ground-truth atomic structures of unknown matters is typically infeasible due to irreducible quantum mechanical uncertainty and limitations in measurement resolution [26, 4], which imposes fundamental limitations on artificial intelligence for science. In particular, the absence of reliable structural ground truth and the substantial experimental cost to obtain it constitutes a major bottleneck in closing the autonomous discovery loop of self-driving laboratories [20, 6].

Retrieving molecular structures from analytical observations requires capturing how a static structure gives rise to the measured response, such as infrared (IR) spectrum [26] and X-ray diffraction pattern [10]. The observations commonly reflect molecular vibrations rather than a direct snapshot of a fixed atomic arrangement, as illustrated in Fig. 1. A molecular structure supports multiple vibrational modes whose characteristic frequencies are determined by atomic masses and the local curvature of the molecular potential-energy surface. Under the harmonic approximation around an equilibrium structure, small molecular vibrations can be decomposed into normal modes, each corresponding to a periodic vibrational component [46, 33, 44]. Infrared spectra record how strongly the molecule absorbs light at frequencies associated with these vibrations [34, 26]. This connects a static molecular structure to its analytical observations through the periodic components of its dynamical response.

Despite recent advances in analytical instruments and methods, identifying atomic structures from analytical data remains a longstanding challenge [26]. Various data-driven methods have been investigated in machine learning and analytical chemistry [31, 1], and cross-modal retrieval methods in particular have demonstrated substantial accuracy [28]. However, their accuracy remains inherently suboptimal because existing retrieval methods attempt to learn a direct mapping between analytical data and the ground-truth equilibrium structure [19, 28], overlooking that analytical data is generated

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

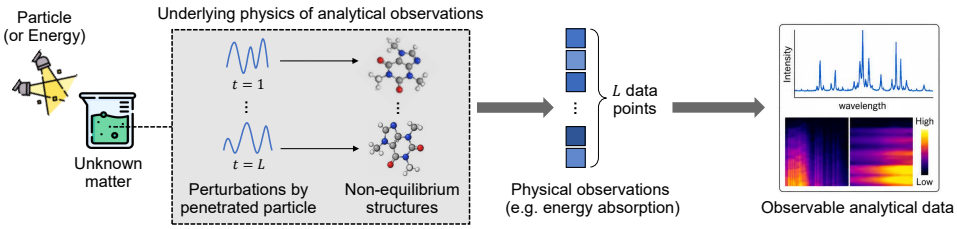

Figure 1: The overall analytical process to generate analytical data of unknown matters.

from a set of periodically perturbed atomic structures [26, 18] rather than an equilibrium structure in ideal closed and static environments.

Around an equilibrium molecular structure, small molecular vibrations can be decomposed under the harmonic approximation into normal modes, each exhibiting periodic motion with a characteristic frequency. This physical principle yields a motivation to incorporate periodicity into the stochastic generation of molecular representations, such that variations from a static structure are organized according to periodic rather than unconstrained stochastic dynamics. However, implementing a continuous and learnable stochastic process is not trivial, since generated samples $\mathbf{x}_t$ should satisfy the following periodicity for all $t \in \mathbb{N}$ as: $p(\mathbf{x}_t|\mathbf{x}_{t-1}, ..., \mathbf{x}_0) = p(\mathbf{x}_{t+L}|\mathbf{x}_{t+L-1}, ..., \mathbf{x}_0)$, where $L \in \mathbb{N}$ is the period of the stochastic process. However, enforcing the stochastic periodicity through explicit constraints in model training yields an unbounded number of constraints with respect to L, because the model should satisfy the stochastic periodicity for all pairs of timesteps in $\{(t, t + nL) \mid t \in \{0, 1, ..., L - 1\}, n \in \mathbb{N}\}$. Although several data-driven approaches to build periodic stochastic processes have been explored [51, 24], these methods have two fundamental limitations: (1) Their training algorithms require a large amount of training data covering multiple periods to learn the periodicity. (2) They are essentially interpolation models, and hence do not guarantee the periodicity in extrapolation regions.

We propose a Periodic Complex Stochastic Process (PCSP), which generates continuous and learnable periodic perturbations. PCSP enables a direct modeling of the periodically perturbed molecular representations behind observed analytical data, as described in Fig. 1. PCSP adopts the architectures of Gaussian AR(1) models for generative processes [21]. Subsequently, PCSP with complex normal distributions (Normal-PCSP) yields a transition probability based on complex normal distributions [12] with learnable means and variances. The generated sample $\mathbf{s}_t \in \mathbb{C}^d$ represents a d-dimensional perturbed embedding at t, and downstream retrieval methods of Normal-PCSP treat $\mathbf{s}_t$ as the embedding of the t-th perturbed atomic structures behind the input analytical data. Normal-PCSP defines the accumulated displacement $\prod_{i=1}^{t} \alpha_i$ as the Euler's formula [30]. In Section 3.2.1, we theoretically prove that a sequence of samples $\mathbf{s}_0, ..., \mathbf{s}_{L-1}, ...$ drawn from Normal-PCSP guarantee the periodicity.

Based on Normal-PCSP, we develop a Complex-Valued Cross-modal Retrieval (CVCR) that integrates Normal-PCSP into conventional cross-modal retrieval architectures. CVCR first calculates perturbed structure embeddings through Normal-PCSP and subsequently aligns these perturbed structure embeddings with the corresponding analytical data. In the inference step, CVCR identifies the atomic structures of unknown matters by comparing similarities between the embedding vectors of the analytical data and the perturbed embedding vectors of the atomic structures.

We evaluated the retrieval performance of CVCR on real-world analytical datasets, each comprising paired analytical data and corresponding ground-truth atomic structures. We used six benchmark datasets generated from widely used analytical methods, such as infrared spectroscopy [52, 25] and neutron scattering diffractometer [7]. In cross-modal retrieval tasks, CVCR substantially outperformed state-of-the-art methods by leveraging perturbed structure embeddings generated through Normal-PCSP. Notably, the retrieval accuracy of CVCR reached 98% in a task of identifying molecules from their Raman spectra. Moreover, CVCR improves the retrieval accuracy from 0.121 to 0.452 (273% improvement) in zero-shot retrieval settings over the state-of-the-art methods on a library of 40 million molecules. We further designed a fully realistic regression setting in which ground-truth atomic structures are unavailable and only analytical data are provided. In this setting, downstream prediction models using CVCR outperformed those based on existing cross-modal retrieval methods.

2 Related Work

Periodic Stochastic Processes. A function is called an almost periodic function [43] if $\forall \epsilon > 0, \exists L > 0$ such that every interval of L contains a shift τ satisfying: $||f(t+\tau) - f(t)|| < \epsilon$ for all $t \in \mathbb{R}$. There are several attempts to design periodic stochastic processes based on the almost periodic functions [16]. However, almost periodic stochastic processes are computationally expensive and sometimes assume unrealistic input domains [16]. In particular, periodic stochastic processes that guarantee the original periodicity [3] in general input domains has not yet been investigated.

Beyond stochastic processes, machine learning methods that learn mapping functions from system inputs to periodic outputs have also been studied [51, 24, 47]. However, their objectives are fundamentally different from PCSP: they learn input-to-output mappings rather than the underlying periodic stochastic process itself, and consequently inherit the two limitations discussed in Section 1.

Cross-Modal Retrieval Methods. Cross-modal retrieval methods aim to retrieve relevant answers whose modalities differ from that of the input query [37]. In natural science, cross-modal retrieval methods for retrieving molecular structures from their analytical spectra or images have received significant interests, as identifying atomic structures of unknown matters is a prerequisite task in many scientific applications, such as drug design [31], materials discovery [22], and biochemical analysis [40]. In chemical science, cross-modal retrieval methods, called spectra and molecule encoder network (SMEN) [19] and Vib2Mol [28], were proposed for identifying unknown molecules from their spectroscopy data. However, they still overlooked the underlying periodic stochastic processes behind analytical spectra, which leads them to suboptimal models. Although CellCLIP [27] was proposed for image-to-structure retrieval under cell imaging perturbations, its scope is protein-level perturbations and limited to image-to-structure modalities.

Graph Neural Networks for Molecular Dynamics. Molecular dynamics (MD) simulation aims to calculate inter-molecular dynamics based on quantum mechanical principles [14]. In machine learning and chemical science, various graph neural networks (GNNs) have been developed to learn inter-atomic interactions between target molecules and their environments [29, 5]. However, the objectives of MD simulation and PCSP are different: MD simulation and its GNN-based surrogate models calculate an equilibrium state under the molecule-level dynamics [50], whereas PCSP aims to simulate non-equilibrium states perturbed by inter-atomic interactions.

3 Method

3.1 Periodic Complex Stochastic Processes (PCSP)

PCSP is a complex-valued stochastic process that generates samples satisfying the stochastic periodicity. For a given initial value $\mathbf{s}_0$, it satisfies stochastic L-periodicity constraints in Eq. (1) for all $t \in \mathbb{N}$ to ensure the periodicity over the entire stochastic process.

$$p(\mathbf{s}_{t+L}|\mathbf{s}_0) = p(\mathbf{s}_t|\mathbf{s}_0), \tag{1}$$

where $\mathbf{s}_t \in \mathbb{C}^d$ is a complex random variable at t, and $L \in \mathbb{N}$ is the period of PCSP. The periodicity constraint in Eq. (1) imposes a challenging or sometimes infeasible objective to conventional time-series models. First, the constraint must hold for every $t \in \mathbb{N}$, so enforcing it through explicit constraints yields an unbounded number of constraints in model parameter optimization. Data-driven models that learn periodic patterns from empirical observations can address periodicity only within the range of their training data, and periodicity beyond training horizon is not guaranteed. In the next section, PCSP realizes a periodic stochastic process [12] that theoretically satisfies Eq. (1) for all $t \in \mathbb{N}$, without observations beyond the initial state $\mathbf{s}_0$.

3.2 PCSP with Complex Normal Distributions

Normal-PCSP is an implementation of PCSP based on the complex normal distribution $\mathcal{CN}$ [12]. Formally, Normal-PCSP decomposes the perturbed embedding at t into a deterministic mean trajectory $\boldsymbol{\mu}_t \in \mathbb{C}^d$ and an independent stochastic perturbation $\boldsymbol{\xi}_t$ as $\mathbf{s}_t = \boldsymbol{\mu}_t + \boldsymbol{\xi}_t$. Normal-PCSP implements a physically-motivated model in which $\boldsymbol{\mu}_t$ is the mean displacement of the equilibrium atomic structure $\mathbf{s}_0$ at vibrational phase t, and the stochastic perturbations $\boldsymbol{\xi}_t$ are superimposed on $\boldsymbol{\mu}_t$ at each t. Normal-PCSP models the mean trajectory as an AR(1) process, $\boldsymbol{\mu}_t = \alpha_t \boldsymbol{\mu}_{t-1}$ with $\boldsymbol{\mu}_0 = \mathbf{s}_0$ and a complex-valued transition coefficient $\alpha_t \in \mathbb{C}$, adopting the well-established AR(1) structure that underlies both molecular dynamics simulations [36] and the forward processes of diffusion generative models [21]. The stochastic perturbation follows a non-circular complex normal

distribution $\boldsymbol{\xi}_t \sim \mathcal{CN}(\mathbf{0}, \Gamma_t, C_t)$, where Γ_t is a covariance matrix and C_t is a complex symmetric pseudo-covariance matrix at t. Equivalently, the transition distribution of Normal-PCSP is given by $\mathbf{s}_t \sim \mathcal{CN}(\alpha_t \boldsymbol{\mu}_{t-1}, \Gamma_t, C_t)$.

Normal-PCSP defines the complex-valued transition coefficient α_t based on Euler's formula with a period-dependent coefficient $\delta = 2\pi/L$. Formally, α_t is given by $\alpha_t = e^{i\delta}$, and the AR(1)-based recursion yields $\boldsymbol{\mu}_t = \bar{\alpha}_t \mathbf{s}_0$ with $\bar{\alpha}_t = e^{i\delta t}$. Therefore, the distribution of $\mathbf{s}_t$ with respect to the initial sample $\mathbf{s}_0$ is given in closed form as:

$$\mathbf{s}_t \sim \mathcal{CN}\left(\bar{\alpha}_t \mathbf{s}_0, \Gamma_t, C_t\right), \tag{2}$$

where $\bar{\alpha}_t = e^{i\delta t}$ is a time-dependent accumulated transition coefficient, $\Gamma_t = \frac{\gamma^2}{2}(2 - \Re(\bar{\alpha}_t^2) - \Im(\bar{\alpha}_t^2))\mathbf{I}$, $C_t = \frac{\gamma^2}{2}(\Im(\bar{\alpha}_t^2) - \Re(\bar{\alpha}_t^2))\mathbf{I}$, and $\gamma^2 > 0$ is a pre-defined scaling parameter. Note that we assume $\mathbf{s}_0$ is an initial structure embedding generated by real-valued embedding networks, i.e., $\Re(\mathbf{s}_0) = \mathbf{s}_0$ and $\Im(\mathbf{s}_0) = 0$. The full derivation of Eq. (2) is provided in Appendix A.

To define valid probability distributions over the complex variables $\mathbf{s}_t$, we decompose the complex-valued variance into real and imaginary parts. For a complex normal random variable characterized by Γ_t and C_t, its real and imaginary components have covariances $(\Gamma_t + C_t)/2$ and $(\Gamma_t - C_t)/2$, respectively. Using $\bar{\alpha}_t = e^{i\delta t}$, Eq. 2 therefore yields the following real-valued normal distributions:

$$\Re(\mathbf{s}_t) \sim \mathcal{N}\left(\cos(\delta t)\mathbf{s}_0, \frac{\gamma^2}{2}(1 - \cos(2\delta t))\mathbf{I}\right), \quad \Im(\mathbf{s}_t) \sim \mathcal{N}\left(\sin(\delta t)\mathbf{s}_0, \frac{\gamma^2}{2}(1 - \sin(2\delta t))\mathbf{I}\right). \tag{3}$$

3.2.1 Stochastic Periodicity

Periodicity is the fundamental characteristics of the quantum mechanical perturbations [41], and Normal-PCSP is designed to generate periodic samples that reflect periodic perturbations grounded on the initial state $\mathbf{s}_0$. In this section, we mathematically validate the periodicity of Normal-PCSP and derive its hyperparameter estimators.

The periodicity of Normal-PCSP is defined in the *stochastic* sense rather than the deterministic pathwise sense. Due to the additive Gaussian noise in the transition distribution, the periodicity of an individual sample is not defined in deterministic sense. Instead, the induced distribution $p(\mathbf{s}_t)$ and its expectation $E[\mathbf{s}_t]$ generated through Normal-PCSP satisfy periodicity. This distinction motivates the term *periodic stochastic process*: the periodicity arises at the distributional level, while individual realizations remain stochastic.

Based on the periodicity in stochastic sense, we define stochastic periodicity with the period L under an error level $\zeta \geq 0$ as:

$$p\left(|\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L})| < \frac{\zeta}{2}\right) = \beta, \quad p\left(|\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L})| < \frac{\zeta}{2}\right) = \beta, \tag{4}$$

where $\beta \approx 1$ is a hyperparameter indicating the probability that the periodicity is satisfied at t. The periodicity criterion β, analogous to a confidence level, is typically chosen from $\{0.90, 0.95, 0.99\}$ in statistics. We evaluate sampling quality and retrieval accuracy of Normal-PCSP and its downstream retrieval methods for different values of β in Section 4.1 and Appendix H.

Mathematically, we can confirm the periodicity of the accumulated transition coefficient $\bar{\alpha}_t$. For a given period L, $\bar{\alpha}_t = e^{i\delta t}$ is equal to $\bar{\alpha}_{t+L}$ by the definition of Euler's formula as follows.

$$\bar{\alpha}_{t+L} = \cos(\delta(t+L)) + i\sin(\delta(t+L)) = \cos(\delta t + 2\pi) + i\sin(\delta t + 2\pi) = e^{i\delta t}. \tag{5}$$

Therefore, $\bar{\alpha}_t$ is periodic for the period L, and consequently the expected values of $\mathbf{s}_t$ is also periodic as: $E[\bar{\alpha}_t \mathbf{s}_0] = E[\bar{\alpha}_{t+L}\mathbf{s}_0] \Leftrightarrow E[\mathbf{s}_t] = E[\mathbf{s}_{t+L}]$ for all t.

3.2.2 Estimation of γ under β

The scaling parameter γ is a hyperparameter of Normal-PCSP, which controls the variance of the generated samples. However, there is a conflict between the periodicity and the sample variance because the periodicity essentially requires the almost same samples at t and $t + L$, i.e., small variances in the generated samples, as defined by the stochastic periodicity in Eq. (4). In this section, we develop a statistical estimator of γ satisfying the stochastic periodicity under the given β.

By the periodicity between $E[\mathbf{s}_t]$ and $E[\mathbf{s}_{t+L}]$, $\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L})$ and $\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L})$ follow normal distributions $\mathcal{N}(0, \sigma_R^2)$ and $\mathcal{N}(0, \sigma_I^2)$, respectively. The variances of the real and imaginary parts are calculated by $\sigma_R^2 = \gamma^2(1 - \cos(2\delta t))$ and $\sigma_I^2 = \gamma^2(1 - \sin(2\delta t))$, respectively.

Subsequently, the sample difference $D_R = |\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L})|$ of the real part follows a half-normal distribution [39], yielding tractable a probability density function (PDF) as $f_{D_R}(d_R; \sigma_R) = \frac{\sqrt{2}}{\sigma_R\sqrt{\pi}} \exp\left(-\frac{(\Re(\mathbf{s}_t)-\Re(\mathbf{s}_{t+L}))^2}{2\sigma_R^2}\right)$. Also, the sample different $D_I = |\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L})|$ of the imaginary part has the following PDF: $f_{D_I}(d_I; \sigma_I) = \frac{\sqrt{2}}{\sigma_I\sqrt{\pi}} \exp\left(-\frac{(\Im(\mathbf{s}_t)-\Im(\mathbf{s}_{t+L}))^2}{2\sigma_I^2}\right)$. Therefore, their cumulative distribution functions (CDFs) of D_R and D_I for a small positive real value $\zeta_h = \zeta/2$ are analytically given by:

$$F_{D_R}(\zeta_h) = \mathrm{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\cos 2\delta t)}}\right),\ F_{D_I}(\zeta_h) = \mathrm{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\sin 2\delta t)}}\right). \tag{6}$$

Based on these CDFs, an estimator for γ that satisfies the stochastic periodicity with a probability β under an error level ζ is derived as Eq. (7).

$$\gamma = \frac{\zeta_h}{\mathrm{erf}^{-1}(\beta)\max_t\{\sqrt{2(1-\phi(2\delta t))}\}}, \tag{7}$$

where ϕ is the cosine and sine functions for real and imaginary parts, respectively. Detailed derivation are provided in Appendix B. In the implementation of Normal-PCSP, we estimated γ under the conventional setting: $\zeta = 1/\sqrt{d}$ and $\beta = 0.99$.

3.2.3 Period and Sample Diversity

Naturally, the period L is equivalent to the process length in PCSP, because the samples at $t = L, L+1, ...$ are just repetitions of the previous samples at $t = 0, 1, ..., L-1$ under the periodicity. Simultaneously, L indicates the number of perturbed phases of the atomic structures in generating analytical data, i.e., the optimal value of L would be the number of perturbed phases captured in the analytical spectra or images. However, this information may be not feasible in most real-world applications. In this section, we develop an estimator for L that guarantees a certain minimum level of sample diversity under the periodicity criterion β.

After the generation process of Normal-PCSP, the real part of the final embedding vectors of target atomic structures are calculated by the generated samples $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_{L-1}$ as:

$$\Re(\mathbf{z})_i = \mathbf{w}_{0,i}\Re(\mathbf{s}_0)_i + \mathbf{w}_{1,i}\Re(\mathbf{s}_1)_i + \cdots + \mathbf{w}_{L-1,i}\Re(\mathbf{s}_{L-1})_i, \tag{8}$$

where $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{L-1}$ are pre-defined weight vectors based on problem definitions or learning from empirical observations. In this embedding process, we impose a constraint that $E[Var[\mathbf{z}_i]] \geq 1/d$ to ensure a minimum level of sample diversity under the given β. By the linearity of the normal random variables, the i-th variance of $\Re(\mathbf{z})$ is calculated by:

$$Var[\Re(\mathbf{z})_i] = \frac{\gamma^2}{2}\sum_{t=0}^{L-1}\mathbf{w}_{t,i}^2(1-\cos(2\delta t)). \tag{9}$$

In downstream models of Normal-PCSP, however, we treat $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{L-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$ as learnable variables. Thus, the variance in Eq. (9) eventually follows a generalized chi-squared distribution [8], and its expected value is given by $E[Var[\Re(\mathbf{z})_i]] = \frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right)$, as explained in Appendix C. From the given criterion $E[Var[\mathbf{z}_i]] \geq 1/d$ for maintaining sample diversity above a certain threshold, we decompose the criterion into $E[Var[\Re(\mathbf{z})_i]] \geq 1/2d$ and $E[Var[\Im(\mathbf{z})_i]] \geq 1/2d$, allocating the sample diversity equally to the real and imaginary parts. Hence, the following two inequalities to estimate the period L are derived:

$$\frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right) \geq \frac{1}{2d}, \quad \frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\sin(2\delta t)\right) \geq \frac{1}{2d}. \tag{10}$$

However, calculating an analytical solution for L satisfying Eq. (10) is not feasible because γ itself depends on L and derives an maximization problem over the L phase points, as derived in Eq. (7). In the following Sections 3.2.4–3.2.6, we propose an efficient decoupling method of γ and L to derive their analytical estimators and subsequent closed-form sampling procedure.

3.2.4 Decoupling of γ and L

As shown in Eq. (7), γ is itself a nonlinear function of L, as its value is determined by the maximum values of $\sqrt{2(1-\cos(2\delta t))}$ and $\sqrt{2(1-\sin(2\delta t))}$. However, we can rewrite γ as a variable independent of L under a mild condition of $L \in \{8n | n \in \mathbb{N}\}$. Numerically, for all $L \in \{8n \mid n \in \mathbb{N}\}$, π and

$\frac{3}{2}\pi$ always exist in $\{2\delta t \,|\, t = 0, 1, ..., L-1\}$, and thus the maximum values of $\sqrt{2(1-\cos(2\delta t))}$ and $\sqrt{2(1-\sin(2\delta t))}$ are fixed to the global maximum value $\sqrt{2^2} = 2$. Therefore, γ is decoupled from L under $L \in \{8n | n \in \mathbb{N}\}$, and a decoupled scaling parameter γ_{dcp} independent of L is calculated by fixing $\max_t\{\sqrt{2(1-cos(2\delta t)}\}$ and $\max_t\{\sqrt{2(1-sin(2\delta t)}\}$ to the global maximum value as:

$$\gamma_{dcp} = \frac{\zeta}{4\text{erf}^{-1}(\beta)}. \tag{11}$$

In the implementation of Normal-PCSP, we used γ_{dcp} instead of γ under $L \in \{8n \mid n \in \mathbb{N}\}$.

$L \in \{8n \mid n \in \mathbb{N}\}$ is a property of the discrete-time representation in Normal-PCSP, not a constraint or assumption on the physical period of the perturbation. L corresponds to the number of phases sampled along the continuous trajectory, determined by the analytical protocol (e.g., spectrometer resolution), while the true physical period can take arbitrary continuous values.

3.2.5 Estimation of L Under Decoupled γ

Decoupling of γ and L in Section 3.2.4 provides a starting point to derive an analytical estimator of L. The sums $\sum_{t=0}^{L-1}\cos(2\delta t)$ and $\sum_{t=0}^{L-1}\sin(2\delta t)$ in Eq. (10) vanish for all $L \geq 3$, as proved in Appendix C. Consequently, both inequalities reduce to $\frac{\gamma^2 L}{2} \geq \frac{1}{2d}$ under $L \in \{8n \mid n \in \mathbb{N}\}$. Thus, substituting γ_{dcp} into γ under $L \in \{8n \mid n \in \mathbb{N}\}$ yields the following inequality with respect to L:

$$L \geq \frac{1}{\gamma_{dcp}^2 d}. \tag{12}$$

Among the possible values, we set L as the minimum satisfying Eq. (12) to reduce the computational cost and the number of parameters of downstream models. Finally, our estimator for L is given by:

$$L_{dcp} = \min\left\{8n \mid n \in \mathbb{N}, 8n \geq 1/(\gamma_{dcp}^2 d)\right\}. \tag{13}$$

3.2.6 Closed-Form Sampling Procedure with γ_{dcp} and L_{dcp}

Based on the tractable distribution in Eq. (3), Normal-PCSP can directly samples $\mathbf{s}_t$ from the closed-form marginal distribution $p(\mathbf{s}_t|\mathbf{s}_0)$ without a time-consuming autoregressive process. Based on the reparameterization trick, $\Re(\mathbf{s}_t)$ and $\Im(\mathbf{s}_t)$ are individually generated through normal distributions as:

$$\Re(\mathbf{s}_t) = \cos(\delta t)\Re(\mathbf{s}_0) + \gamma_{dcp}\sqrt{\tfrac{1-\cos 2\delta t}{2}}\,\boldsymbol{\epsilon}_R, \quad \Im(\mathbf{s}_t) = \sin(\delta t)\Re(\mathbf{s}_0) + \gamma_{dcp}\sqrt{\tfrac{1-\sin 2\delta t}{2}}\,\boldsymbol{\epsilon}_I. \tag{14}$$

where $\boldsymbol{\epsilon}_R, \boldsymbol{\epsilon}_I \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$, and $\delta = 2\pi/L$. Normal-PCSP can parallelly generate samples from $p(\mathbf{s}_t|\mathbf{s}_0)$ through Eq. (14). We evaluated computational efficiency of Normal-PCSP in Appendix G.

3.2.7 Projection of Complex-Valued Representations

Since Normal-PCSP is performed in the complex-valued space, we need to project the complex-valued samples $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_{L-1}$ into the real-valued space to leverage them in real-valued downstream models. To this end, Normal-PCSP transforms the d-dimensional complex-valued samples $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_{L-1}$ into $(2d)$-dimensional real-valued vectors $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$ based on a vector concatenation as: $\tilde{\mathbf{z}}_t = \phi(\Re(\mathbf{s}_t) \oplus \Im(\mathbf{s}_t))$, where $\oplus$ is the vector concatenation operator, and ϕ is a learnable function.

The projected sample $\tilde{\mathbf{z}}_t$ is also periodic due to the following two lemmas: (1) The vector concatenation operator preserves the periodicity if two operands have the same periodicity. (2) A composite function $g \circ f$ of an arbitrary function g and a periodic function f is always periodic. In Section 4.1, we will empirically demonstrate the periodicity of the projected samples. Algorithm 1 in Appendix D shows the overall process of Normal-PCSP to generate $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$ from an initial embedding $\mathbf{s}_0$.

3.3 Cross-Modal Retrieval with Normal-PCSP

Based on Normal-PCSP, we developed a cross-modal retrieval method called complex-valued cross-modal retrieval (CVCR) for identifying atomic structures of unknown matters from their analytical spectra or images. CVCR employs the conventional architectures of the cross-modal retrieval methods, which consist of a query encoder $h_{\mathcal{Q}}$ and a target encoder $h_{\mathcal{A}}$ in different modalities. In CVCR, $h_{\mathcal{Q}}$ and $h_{\mathcal{A}}$ calculate d-dimensional latent embeddings $\mathbf{h}$ and $\mathbf{s}_0$ for the input analytical data and atomic structure, respectively. Then, the final structure embedding $\mathbf{z}$ derived from $\mathbf{s}_0$ is calculated based on the perturbed structure embeddings $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$ with learnable weights $\omega_0, \omega_1, ..., \omega_{L-1}$ as:

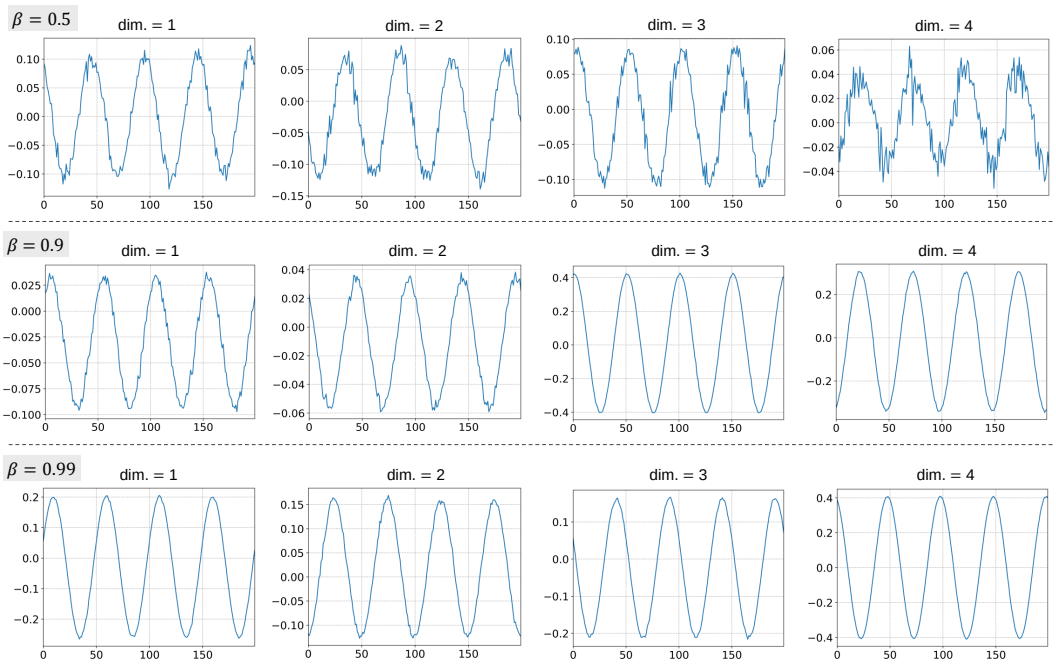

Figure 2: Visualization of 4-dimensional perturbed structure embeddings. X- and Y-axes are the timestep and the embedding value, respectively.

Table 1: Cross-modal retrieval tasks from six benchmark datasets.

Retrieval Task	Dataset	Input Query	Target Data	# of Samples
Sequence-to-graph information retrieval	QMe14S-IR	IR spectrum	Molecule	186,102
	QMe14S-RM	Raman spectrum	Molecule	186,102
	NIST	IR spectrum	Molecule	8,832
	OpenXRD	X-ray diffraction patterns	Crystalline material	872
Image-to-graph information retrieval	2DNMRGym	NMR scattering image	Molecule	11,825
	INS-MAT	Neutron scattering image	Crystalline material	9,928

$\mathbf{z} = \sum_{t=0}^{L-1} \omega_t \tilde{\mathbf{z}}_t$. Algorithm 2 in Appendix D presents the overall retrieval process of CVCR. An ablation study to assess the effectiveness of each component in CVCR is conducted in Appendix J.

We employ the InfoNCE loss [35] to train CVCR. However, we compare the similarity between $\mathbf{h}$ and $\mathbf{z}$, rather than between $\mathbf{h}$ and $\mathbf{s}_0$ as:

$$L_{align} = -\frac{1}{N}\sum_{i=1}^{N}\left\{\ln\left(\frac{e^{\tilde{u}_{ii}/\tau}}{\sum_{j=1}^{N} e^{\tilde{u}_{ij}/\tau}}\right) + \ln\left(\frac{e^{\tilde{u}_{ii}/\tau}}{\sum_{j=1}^{N} e^{\tilde{u}_{ji}/\tau}}\right)\right\}, \tag{15}$$

where $\tau \in \mathbb{R}^+$ is a temperature parameter, and $\tilde{u}_{ij} = \mathbf{h}_i \cdot \mathbf{z}_j$ is the similarity between the embeddings of the i-th query (analytical data) and j-th atomic structure.

4 Experiments

We conducted a numerical analysis to assess the periodicity of Normal-PCSP and further performed cross-modal molecular retrieval on real-world benchmark datasets to evaluate the final retrieval capabilities of CVCR. Table 1 summarizes cross-modal retrieval tasks from six benchmark datasets, and detailed descriptions of the retrieval tasks and benchmark datasets are provided in Appendix E. Implementation details and hyperparameter settings of Normal-PCSP and CVCR are provided in Appendix F. Source codes are available at `https://huggingface.co/KRICT-AI/PCSP_CVCR`.

4.1 Periodicity Analysis

For empirical assessment of the periodicity of Normal-PCSP, we generated 4-dimensional structure embeddings $\mathbf{z}$ through Normal-PCSP followed by CVCR. Fig. 2 visualizes the perturbed embeddings

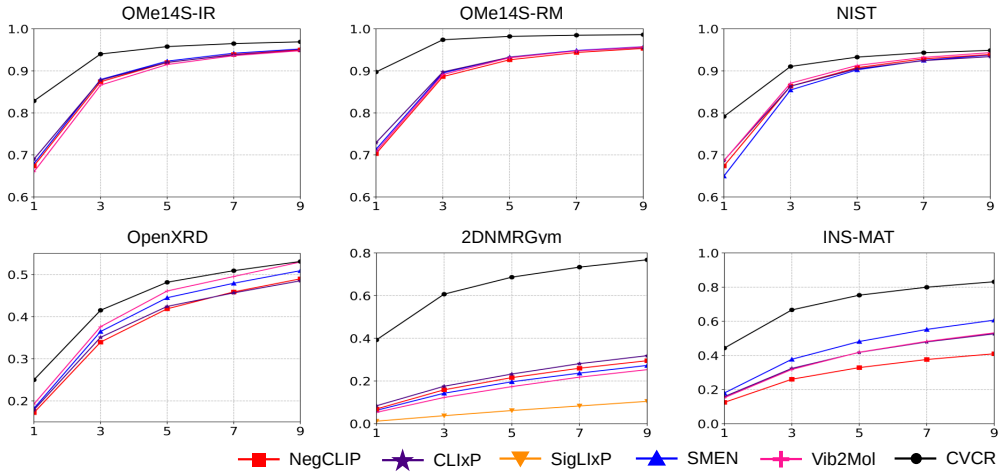

Figure 3: Recall@k of the competitor methods and CVCR on the six cross-modal retrieval datasets. X-axis: $k \in \{1, 3, 5, 7, 9\}$, Y-axis: measured recall@k.

Table 2: Zero-shot retrieval accuracy of the cross-modal retrieval competitors and CVCR.

Method	Recall@k			Tanimoto Similarity		
	$k=1$	$k=5$	$k=9$	$k=1$	$k=5$	$k=9$
SMEN	0.013 (0.011)	0.105 (0.017)	0.121 (0.006)	0.265 (0.014)	0.438 (0.021)	0.483 (0.012)
Vib2Mol	0.015 (0.013)	0.096 (0.008)	0.112 (0.008)	0.289 (0.013)	0.424 (0.017)	0.480 (0.009)
CVCR	**0.202 (0.008)**	**0.362 (0.006)**	**0.452 (0.005)**	**0.499 (0.015)**	**0.647 (0.013)**	**0.708 (0.010)**

of Normal-PCSP with $\beta \in \{0.5, 0.9, 0.99\}$. As expected, $\mathbf{z}$ generated with $\beta = 0.5$ showed a weak periodicity, whereas $\mathbf{z}$ generated with $\beta = 0.99$ exhibited a strong periodicity over the entire timesteps. For $\beta = 0.99$ enforcing a strong periodicity, the entire embedding $\mathbf{z}$ has the periodicity with the period L. As β increases from 0.5 to 0.99, the generated embeddings transition from weakly periodic to strongly periodic, consistent with the theoretical role of β as a periodicity criterion.

4.2 Cross-Modal Retrieval on Benchmark Datasets

We employed six benchmark datasets in analytical chemistry, covering sequence-to-graph retrieval (QMe14S-IR [52], QMe14S-RM [52], NIST [31], OpenXRD [45]) and image-to-graph retrieval (2DNMRGym [23], INS-MAT [7]). Dataset statistics and task definitions are summarized in Table 1. We compared retrieval accuracy of CVCR with five state-of-the-art cross-modal retrieval methods: NegCLIP [53], CLIxP [38], SigLIxP [38], SMEN [19], and Vib2Mol [28]. Note that SMEN and Vib2Mol are specially designed for retrieving unknown molecules from their analytical data. However, as CLOOME [40] was highly sensitive to hyperparameter settings and consequently failed on many benchmark datasets, we were not able to include it in our evaluations. The retrieval accuracy were measured by recall@k for $k \in \{1, 3, 5, 7, 9\}$ via 5-fold cross-validation.

Fig. 3 presents recall@k of the competitor methods and CVCR for each dataset. CVCR showed higher recall@k than those of the competitor methods for all datasets. Notably, CVCR outperformed the competitor methods for all k in the QMe14S-IR, QMe14S-RM, 2DNMRGym, and INS-MAT datasets. In particular, the recall@k of CVCR reached to 0.99 in the QMe14S-RM dataset, i.e., CVCR retrieved molecules from their Raman spectra with 98% accuracy.

CVCR achieved substantial accuracy improvements on the 2DNMRGym and INS-MAT datasets, and its recall@k approximately reached 0.8 at $k = 9$. Since many analytical images are directly generated through an aggregation of physical observations from perturbed atomic structures [26, 7], the retrieval process of CVCR, which is based on perturbed structure embeddings, was highly effective in capturing the underlying physics of the analytical images in the 2DNMRGym and INS-MAT datasets. Therefore, the significant accuracy improvements in the 2DNMRGym and INS-MAT datasets demonstrate the necessity of considering structural perturbations in data-driven analytical chemistry. We further note that the strong performance on the INS-MAT dataset, whose unit cells reach up to 43,512 Da and overlap with the molecular weight range typically associated with macromolecules, provides supporting evidence that CVCR generalizes beyond the small-molecule scale.

Table 3: Recall@k of CVCR for different autoregressive generative models.

Method	NIST			2DNMRGym		
	$k=1$	$k=5$	$k=9$	$k=1$	$k=5$	$k=9$
DDPM	0.698 (0.013)	0.785 (0.009)	0.913 (0.005)	0.252 (0.027)	0.513 (0.019)	0.686 (0.021)
DSB	0.714 (0.015)	0.895 (0.009)	0.922 (0.008)	0.239 (0.018)	0.520 (0.016)	0.683 (0.014)
Normal-PCSP	**0.792 (0.016)**	**0.929 (0.010)**	**0.948 (0.008)**	**0.393 (0.017)**	**0.638 (0.015)**	**0.768 (0.016)**

4.3 Data-Efficient Zero-Shot Retrieval

We trained CVCR and two cross-modal retrieval methods specialized for chemical data, SMEN and Vib2Mol, on the NIST dataset containing only 8,832 molecules. Then, we evaluated their zero-shot retrieval accuracy on an external PubChem subset of approximately 40 million molecules, presenting the majority of synthesizable compounds. Given the extremely large candidate space and zero-shot setting, exact-match top-k accuracy can be overly restrictive. For this reason, we additionally measured Tanimoto similarity, which better reflects chemical similarity between retrieved and ground-truth molecules. As shown in Table 2, despite training on only 0.022% of the library, CVCR still performed reasonably well in this large-scale zero-shot setting. Quantitatively, CVCR improved Recall@9 from 0.121 to 0.452 and Tanimoto similarity at $k = 9$ from 0.483 to 0.708.

4.4 Comparison with Generative Models and Augmentation Baselines

Although Normal-PCSP is architecturally similar to AR(1) diffusion processes [21, 42] and generates multiple perturbed embeddings from $\mathbf{s}_0$ akin to data augmentation, it is fundamentally distinct from both: it constructs periodic samples through intrinsic structures rather than data-dependent parameters or noise injection, directly modeling the underlying physics of analytical data.

Comparison with generative models. We compared recall@k of CVCR with denoising diffusion probabilistic model (DDPM) [21], Diffusion Schrödinger bridge (DSB) [42], and Normal-PCSP on the experimentally collected NIST and 2DNMRGym datasets. DDPM and DSB were trained to generate diverse L samples grounded on a given $\mathbf{s}_0$. As shown in Table 3, CVCR with Normal-PCSP outperformed its variants with DDPM and DSB, and these results demonstrate the importance to capture the underlying periodic perturbations behind analytical data in retrieving atomic structures, which generic generative models do not encode as a structural prior.

Comparison with data augmentation methods. PCSP is conceptually distinct from data augmentation: standard augmentation methods enlarge the training set by injecting generic perturbations at the data level, whereas PCSP enriches the embedding space itself. In this experiment, we further compared PCSP against three standard augmentation strategies on the NIST dataset: (i) Gaussian noise injection into the input spectra ($\mathcal{N}$-AUG-SIGNAL), (ii) Gaussian noise injection into the structure embedding ($\mathcal{N}$-AUG-STRUCT), and (iii) SMILES augmentation (SMILES-AUG) [2]. As shown in Table 4, PCSP consistently outperformed all baselines across all k. These accuracy gains reveal that PCSP is not merely a training-time data augmentation strategy but a representation enhancement mechanism applicable to both training and inference.

Table 4: Recall@k of PCSP and standard data augmentation baselines.

Method	$k=1$	$k=5$	$k=9$
$\mathcal{N}$-AUG-SIGNAL	0.718	0.899	0.925
$\mathcal{N}$-AUG-STRUCT	0.705	0.900	0.943
SMILES-AUG	0.675	0.898	0.941
PCSP	**0.792**	**0.929**	**0.948**

4.5 Robustness Analysis

We examine this from two complementary angles: (i) noise in the input analytical spectra, which directly affects the query embedding, and (ii) noise or limited expressiveness in the initial structure embedding $\mathbf{s}_0$, which seeds the entire stochastic process in Normal-PCSP. Table 5 summarizes both.

Robustness to input spectral noise. We injected additive Gaussian noise at SNR$= 15$ dB[1] into the IR and Raman spectra of the QMe14S-IR and QMe14S-RM datasets, corresponding to a substantial level of corruption relative to typical experimental noise. As shown in Table 5a, Recall@k degraded only marginally, and recall@9 stayed above 0.95 on both datasets despite the strong noise injection.

[1]We follow the standard definition of $\text{SNR}_{\text{dB}} = 10\log_{10}(P_{\text{signal}}/P_{\text{noise}})$, where P_{signal} is the average power of the clean signal. The noise variance is computed as $\sigma^2 = P_{\text{signal}}/10^{\text{SNR}/10}$, and the same convention is used for noise injection into $\mathbf{s}_0$.

Table 5: Robustness analysis of CVCR. (a) Retrieval accuracy under noised input spectra (SNR$=$ 15 dB). (b) Retrieval accuracy on the NIST dataset under weaker GNN backbones and noised $\mathbf{s}_0$.

(a) Recall@k on noised input spectra.

Dataset (Noise Setting)	$k=1$	$k=5$	$k=9$
QMe14S-IR (noisy)	0.785	0.939	0.957
QMe14S-IR (clean)	0.828	0.950	0.969
QMe14S-RM (noisy)	0.843	0.957	0.968
QMe14S-RM (clean)	0.898	0.970	0.986

(b) Recall@k of backbone GNNs under noised $\mathbf{s}_0$.

Backbone	Noise (SNR)	$k=1$	$k=5$	$k=9$
GAT	-10 dB	0.388	0.754	0.848
GAT	-5 dB	0.609	0.887	0.925
GAT	5 dB	0.751	0.927	0.950
MPNN	-10 dB	0.440	0.792	0.873
MPNN	-5 dB	0.651	0.902	0.935
MPNN	5 dB	0.784	0.929	0.952
AttentiveFP	clean	0.792	0.929	0.948

Table 6: R^2-scores of the downstream GNNs. NO RETRIEVAL is the baseline methods predicting targets directly from analytical data through a spectrum encoder [31] or an image encoder [13].

Dataset	NO RETRIEVAL	NegCLIP	CLIxP	SigLIxP	SMEN	Vib2Mol	CVCR
ESOL	0.613 (0.046)	0.848 (0.014)	0.903 (0.012)	0.895 (0.014)	0.875 (0.012)	0.908 (0.007)	**0.962 (0.007)**
PCQM4Mv2	0.682 (0.065)	0.863 (0.076)	0.972 (0.013)	0.965 (0.011)	0.972 (0.007)	0.973 (0.009)	**0.987 (0.012)**
MP	0.541 (0.031)	0.193 (0.036)	0.163 (0.050)	0.218 (0.021)	0.421 (0.013)	0.308 (0.011)	**0.619 (0.005)**

Robustness to the initial structure embedding. We test two perturbations on $\mathbf{s}_0$: (i) weaker GNN backbones (GAT and MPNN) producing less expressive embeddings, and (ii) Gaussian noise injected into $\mathbf{s}_0$ at SNR $\in \{-10, -5, 5\}$ dB (SNR<0 dB = noise dominates signal) on the NIST dataset. As shown in Table 5b, CVCR maintained recall@9 in $[0.84, 0.95]$. Rather than relying on a single static $\mathbf{s}_0$, Normal-PCSP aggregates over L phases in the complex domain, smoothing localized errors across the periodic trajectory, which acts as an implicit stochastic regularizer.

4.6 Regression on Unknown Matters

One of the fundamental challenges in artificial intelligence (AI) for science is that ground-truth atomic structures assumed to be available are not feasible in most real-world applications. We designed fully realistic regression settings where the ground-truth atomic structure is unknown and only analytical data is provided. We made joint chemical datasets by collecting intersected data in both an analytical dataset (NIST or INS-MAT) and a target regression dataset. Then, the cross-modal retrieval methods retrieve atomic structures on the joint chemical datasets. In the end, we train GNNs to predict molecular or material properties from retrieved atomic structures instead of ground-truth atomic structures. The detailed experimental setup and procedure are provided in Appendix K.

Three well-known chemical datasets were employed in this experiment: ESOL [9], PCQM4Mv2 [15], and Materials Project (MP) [17]. Brief descriptions about these datasets are provided in Appendix E. In the analytical datasets, 230, 1,471 and 9,675 matters were identified for the ESOL, PCQM4Mv2, and MP datasets, respectively. We used AttentiveFP [49] and CGCNN [48] to implement downstream GNNs for molecules and materials, respectively. As shown in Table 6, the downstream GNNs with CVCR achieved the highest R^2-scores for all datasets. Although the baseline method (NO RETRIEVAL) showed a high R^2-score on the MP dataset because neutron scattering patterns inherently include fundamental information about crystal structures, the downstream GNN with CVCR still outperformed this baseline method. These experiment results show the practical potential of CVCR in the complete real-world settings, which is the next challenge of AI for science.

5 Conclusion

Capturing the underlying physics induced by the perturbed atomic structures is essential for retrieving atomic structures from their analytical data. This paper first tackled the problem of considering the perturbed atomic structures in cross-modal retrieval on analytical data. Normal-PCSP extends real-valued stochastic processes into complex-valued stochastic processes to generate periodic structure embeddings. CVCR leverages Normal-PCSP to accurately retrieve atomic structures from analytical data and achieved state-of-the-art retrieval accuracy on benchmark datasets. Future work should extend CVCR into fully generative models, enabling the identification of atomic structures that remain unexplored by current scientific knowledge.

Acknowledgments

This work was supported by Korea Research Institute of Chemical Technology (KK2661-51), Korea Institute for Advancement of Technology funded by Ministry of Trade, Industry and Resource (2410016815), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25607361).

Competing Interests

The authors declare no competing interests.

References

[1] Marvin Alberts, Filippo Ficarra, and Teodoro Laino. Language model enabled structure prediction from infrared spectra of mixtures. In *AI for Accelerated Materials Design-NeurIPS 2025*, 2025.

[2] Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11(1):71, 2019.

[3] Alan F. Beardon. *Periodic Functions*, pages 83–96. Springer New York, New York, NY, 1997.

[4] Andrei A Bunaciu, Elena Gabriela UdriŞTioiu, and Hassan Y Aboul-Enein. X-ray diffraction: instrumentation and applications. *Critical reviews in analytical chemistry*, 45(4):289–299, 2015.

[5] Jian Chang and Shuze Zhu. Mgnn: Moment graph neural network for universal molecular potentials. *npj Computational Materials*, 11(1):55, 2025.

[6] Ming-Chiang Chang, Sebastian Ament, Maximilian Amsler, Duncan R Sutherland, Lan Zhou, John M Gregoire, Carla P Gomes, R Bruce van Dover, and Michael O Thompson. Probabilistic phase labeling and lattice refinement for autonomous materials research. *npj Computational Materials*, 11(1):148, 2025.

[7] Yongqiang Cheng, Matthew B Stone, and Anibal J Ramirez-Cuesta. A database of synthetic inelastic neutron scattering spectra from molecules and crystals. *Scientific Data*, 10(1):54, 2023.

[8] Robert B Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, pages 323–333, 1980.

[9] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

[10] J Epp. X-ray diffraction (xrd) techniques for materials characterization. In *Materials characterization using nondestructive evaluation (NDE) methods*, pages 81–124. Elsevier, 2016.

[11] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. Pmlr, 2017.

[12] Nathaniel R Goodman. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of mathematical statistics*, 34(1):152–177, 1963.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6): 1129–1143, 2018.

[15] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.

[16] Alexander Iksanov, Zakhar Kabluchko, and Alexander Marynych. Almost periodic stochastic processes with applications to analytic number theory. *arXiv preprint arXiv:2502.04969*, 2025.

[17] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[18] Cy M Jeffries, Jan Ilavsky, Anne Martel, Stephan Hinrichs, Andreas Meyer, Jan Skov Pedersen, Anna V Sokolova, and Dmitri I Svergun. Small-angle x-ray and neutron scattering. *Nature Reviews Methods Primers*, 1(1):70, 2021.

[19] Ganesh Chandan Kanakala, Bhuvanesh Sridharan, and U Deva Priyakumar. Spectra to structure: contrastive learning framework for library ranking and generating molecular structures for infrared spectra. *Digital Discovery*, 3(12):2417–2423, 2024.

[20] Martin A Karlsen and Andy S Anker. The importance of fit-for-purpose structure characterization in autonomous laboratories. *Nature Reviews Materials*, pages 1–3, 2026.

[21] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[22] Qingsi Lai, Fanjie Xu, Lin Yao, Zhifeng Gao, Siyuan Liu, Hongshuai Wang, Shuqi Lu, Di He, Liwei Wang, Linfeng Zhang, et al. End-to-end crystal structure prediction from powder x-ray diffraction. *Advanced Science*, 12(8):2410722, 2025.

[23] Yunrui Li, Hao Xu, and Pengyu Hong. 2dnmrgym: An annotated experimental dataset for atom-level molecular representation learning in 2d nmr via surrogate supervision. *arXiv preprint arXiv:2505.18181*, 2025.

[24] Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.

[25] Peter J Linstrom and William G Mallard. The nist chemistry webbook: A chemical data resource on the internet. *Journal of Chemical & Engineering Data*, 46(5):1059–1063, 2001.

[26] Ángela I López-Lorente and Boris Mizaikoff. Mid-infrared spectroscopy for protein analysis: potential and challenges. *Analytical and bioanalytical chemistry*, 408(11):2875–2889, 2016.

[27] MingYu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP - learning perturbation effects in cell painting via text-guided contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] Xinyu Lu, Hao Ma, Hui Li, Jia Li, Yuqiang Li, Tong Zhu, Guokun Liu, and Bin Ren. Vib2mol: from vibrational spectra to molecular structures-a versatile deep learning model, 2025. URL `https://arxiv.org/abs/2503.07014`.

[29] Debasis Maji, Atish Ghosh, Debaditya Barman, and Pranab Sarkar. Accelerating molecular dynamics with a graph neural network: A scalable approach through e (q) c-gnn. *The Journal of Physical Chemistry Letters*, 16(9):2254–2264, 2025.

[30] Martin Moskowitz. *A course in complex analysis in one variable*. World Scientific Publishing Company, 2002.

[31] Gyoung S Na and Yecheol Rho. Explainable machine learning for characterizing unknown molecular structures in infrared spectra. *Analytical Chemistry*, 97(38):20869–20878, 2025.

[32] Maho Nakata and Tomomi Shimazaki. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of chemical information and modeling*, 57(6):1300–1308, 2017.

[33] Johannes Neugebauer, Markus Reiher, Carsten Kind, and Bernd A Hess. Quantum chemical calculation of vibrational spectra of large molecules—raman and ir spectra for buckminsterfullerene. *Journal of computational chemistry*, 23(9):895–910, 2002.

[34] Lily M Ng and Reiko Simmons. Infrared spectroscopy. *Analytical chemistry*, 71(12):343–350, 1999.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[36] Michael Plainer, Hao Wu, Leon Klein, Stephan Günnemann, and Frank Noé. Consistent sampling and simulation: Molecular dynamics with energy-based diffusion models. *Advances in Neural Information Processing Systems*, 38:24460–24505, 2026.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[38] Karsten Roth, Zeynep Akata, Dima Damen, Ivana Balazevic, and Olivier J Hénaff. Context-aware multimodal pretraining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4267–4279, 2025.

[39] Christian Röver, Ralf Bender, Sofia Dias, Christopher H Schmid, Heinz Schmidli, Sibylle Sturtz, Sebastian Weber, and Tim Friede. On weakly informative prior distributions for the heterogeneity parameter in bayesian random-effects meta-analysis. *Research Synthesis Methods*, 12(4):448–474, 2021.

[40] Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023.

[41] Ramamurti Shankar. *Principles of quantum mechanics*. Springer Science & Business Media, 2012.

[42] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

[43] Mikhail Aleksandrovich Shubin. Almost periodic functions and partial differential operators. *Russian Mathematical Surveys*, 33(2):1, 1978.

[44] Kirill Speranskiy and Maria Kurnikova. Accurate theoretical prediction of vibrational frequencies in an inhomogeneous dynamic environment: A case study of a glutamate molecule in water solution and in a protein-bound form. *The Journal of chemical physics*, 121(3):1516–1524, 2004.

[45] Ali Vosoughi, Ayoub Shahnazari, Yufeng Xi, Zeliang Zhang, Griffin Hess, Chenliang Xu, and Niaz Abdolrahim. Openxrd: A comprehensive benchmark and enhancement framework for llm/mllm xrd question answering. *arXiv preprint arXiv:2507.09155*, 2025.

[46] E Bright Wilson Jr. The normal modes and frequencies of vibration of the regular plane hexagon model of the benzene molecule. *Physical Review*, 45(10):706, 1934.

[47] Qiang Wu, Gechang Yao, Zhixi Feng, and Yang Shuyuan. Peri-midformer: Periodic pyramid transformer for time series analysis. *Advances in Neural Information Processing Systems*, 37: 13035–13073, 2024.

[48] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[49] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

[50] Takuma Yagasaki and Shinji Saito. Molecular dynamics simulation of nonlinear spectroscopies of intermolecular motions in liquid water. *Accounts of chemical research*, 42(9):1250–1258, 2009.

[51] Chunwei Yang, Xiaoxu Chen, Lijun Sun, Hongyu Yang, and Yuankai Wu. Enhancing representation learning for periodic time series with floss: A frequency domain regularization approach. *arXiv preprint arXiv:2308.01011*, 2023.

[52] Mingzhi Yuan, Zihan Zou, Yi Luo, Jun Jiang, and Wei Hu. Qme14s: A comprehensive and efficient spectral data set for small organic molecules. *The Journal of Physical Chemistry Letters*, 16(16):3972–3979, 2025.

[53] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

A Sampling from Closed-Form Marginal Complex Normal Distributions

Normal-PCSP defines the perturbed embedding at phase t as

$$\mathbf{s}_t = \boldsymbol{\mu}_t + \boldsymbol{\xi}_t, \qquad \boldsymbol{\mu} t = \alpha_t \boldsymbol{\mu}_{t-1}, \quad \boldsymbol{\mu}_0 = \mathbf{s}_0, \qquad \boldsymbol{\xi}_t \sim \mathcal{CN}(\mathbf{0}, \Gamma_t, C_t), \tag{16}$$

where $\mathbf{s}_0 \in \mathbb{R}^d$ is the initial structure embedding and $\alpha_t = e^{i\delta}$ with $\delta = 2\pi/L$. Unrolling the recursion of the mean trajectory gives

$$\boldsymbol{\mu}_t = \alpha_t \alpha_{t-1} \cdots \alpha_1 \boldsymbol{\mu}_0 = e^{i\delta t} \mathbf{s}_0 = \bar{\alpha}_t \mathbf{s}_0 \tag{17}$$

where $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k = e^{i\delta t}$ is the accumulated transition coefficient. Since $\boldsymbol{\mu}_t$ is deterministic and the complex normal family is closed under translation by a constant, adding $\boldsymbol{\mu}_t$ to $\boldsymbol{\xi}_t$ shifts the mean while leaving the covariance and pseudo-covariance unchanged. Therefore,

$$\mathbf{s}_t = \bar{\alpha}_t \mathbf{s}_0 + \boldsymbol{\xi}_t \sim \mathcal{CN}\big(\bar{\alpha}_t \mathbf{s}_0,\ \Gamma_t,\ C_t\big), \tag{18}$$

which is Eq. (2). This complex-valued distribution is well defined for all t, since $\Gamma_t + C_t = \gamma^2(1 - \cos(2\delta t))\mathbf{I} \succeq \mathbf{0}$ and $\Gamma_t - C_t = \gamma^2(1 - \sin(2\delta t))\mathbf{I} \succeq \mathbf{0}$ imply $\Gamma_t \succeq |C_t|$.

B Maximizing CDFs of the Half-Normal Distributions for Estimating γ

For a given periodicity criterion β and error level ζ, the inequalities in Eq. (4) indicate that CDFs of D_R and D_I should greater than β for all $t \in \{0, 1, ..., L-1\}$ as follows.

$$\text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\cos(2\delta t))}}\right) \geq \beta, \tag{19}$$

$$\text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\sin(2\delta t))}}\right) \geq \beta, \tag{20}$$

where $\zeta_h = \zeta/2$. For the real part, Eq. (19) can be rewritten with respect to the minimum value over t as:

$$\min_{t\in\{0,1,\ldots,L-1\}} \left\{ \text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\cos(2\delta t))}}\right)\right\} = \beta \tag{21}$$

Since the error function (erf) is increasing, the erf is minimized when $\sqrt{2(1-\cos(2\delta t))}$ is maximized, i.e., we can convert the problem in Eq. (21) into the following maximization problem as:

$$\text{erf}\left(\frac{\zeta_h}{\gamma \max_t \left\{\sqrt{2(1-\cos(2\delta t))}\right\}}\right) = \beta \tag{22}$$

By taking the inverse of erf, γ for the real part is calculated as follows.

$$\gamma = \frac{\zeta_h}{\text{erf}^{-1}(\beta) \max_t \left\{\sqrt{2(1-\cos(2\delta t))}\right\}} \tag{23}$$

Similarly, we can calculate γ for the imaginary part as:

$$\gamma = \frac{\zeta_h}{\text{erf}^{-1}(\beta) \max_t \left\{\sqrt{2(1-\sin(2\delta t))}\right\}}$$

Finally, we can derive an estimator for γ under the given β as follows.

$$\gamma = \begin{cases} \frac{\zeta_h}{\text{erf}^{-1}(\beta) \max_t \left\{\sqrt{2(1-\cos(2\delta t))}\right\}}, & \text{for real} \\ \frac{\zeta_h}{\text{erf}^{-1}(\beta) \max_t \left\{\sqrt{2(1-\sin(2\delta t))}\right\}}, & \text{for imaginary.} \end{cases} \tag{24}$$

C Sample Variances and Estimation of L

By the definition in Eq. (3), the variance of the i-th real part of the generated sample is given by:

$$Var(\Re(\mathbf{s}_t)_i) = \frac{\gamma^2}{2}(1 - \cos(2\delta t)). \tag{25}$$

Subsequently, for given weights $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{L-1}$, the variance of the i-th weighted latent feature of the real part is calculated by the linearity of the variance as:

$$\begin{aligned} Var(\Re(\mathbf{z})_i) &= Var(\mathbf{w}_{0,i}\Re(\mathbf{s}_0)_i + \cdots + \mathbf{w}_{L-1,i}\Re(\mathbf{s}_{L-1})_i) \\ &= \frac{\gamma^2}{2}\sum_{t=0}^{L-1}\mathbf{w}_{t,i}^2(1 - \cos(2\delta t)) \end{aligned} \tag{26}$$

However, we assume $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{L-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$ during the training process of CVCR. Thus, we can define a random variable $Y = Var(\Re(\mathbf{z})_i)$ as a weighted sum of independent noncentral chi-square variables, which are defined as:

$$\frac{\gamma^2}{2}(1 - \cos(2\delta t))\mathbf{w}_{t,i}^2 \sim \mathcal{X}_1^{'2} \tag{27}$$

Therefore, $Var(\Re(\mathbf{z})_i)$ eventually follows a generalized chi-squared distribution [8], and its expected value is calculated by:

$$E[Var[\Re(\mathbf{z})_i]] = \sum_{t=0}^{L-1}\frac{\gamma^2}{2}(1 - \cos(2\delta t)) \tag{28}$$

$$= \frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right). \tag{29}$$

In the estimation procedure for L, our criterion for the sample diversity is $E[Var[\mathbf{z}_i]] \geq 1/d$ to ensure a minimum level of the unit variance of the latent embeddings. Allocating this budget equally to the real and imaginary parts, the period L should satisfy $E[Var[\Re(\mathbf{z})_i]] \geq 1/2d$ and $E[Var[\Im(\mathbf{z})_i]] \geq 1/2d$. Hence, the period (= process length) L for the real part should satisfy the following inequality.

$$\frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right) \geq \frac{1}{2d}. \tag{30}$$

Similarly, the following inequality is derived for the imaginary part.

$$\frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\sin(2\delta t)\right) \geq \frac{1}{2d}. \tag{31}$$

However, since $\delta = 2\pi/L$, this sum is the real part of a geometric series:

$$\sum_{t=0}^{L-1} e^{i\cdot 2\delta t} = \sum_{t=0}^{L-1} e^{i\cdot 4\pi t/L} = \frac{1 - e^{i\cdot 4\pi}}{1 - e^{i\cdot 4\pi/L}} = 0, \tag{32}$$

where the numerator vanishes because $e^{i4\pi} = 1$, and the denominator is nonzero whenever $e^{i4\pi/L} \neq 1$, i.e., $L \geq 3$. Taking the real and imaginary parts of Eq. (32) yields

$$\sum_{t=0}^{L-1}\cos(2\delta t) = 0, \qquad \sum_{t=0}^{L-1}\sin(2\delta t) = 0 \qquad \text{for all } L \geq 3. \tag{33}$$

Hence, for $L \geq 3$, Eqs. (30) and (31) are simplified into $L \geq 1/(\gamma^2 d)$. Under $L \in \{8n \mid n \in \mathbb{N}\}$, we can derive an inequality to calculate an analytical solution to estimate L by substituting γ_{dcp} into γ as follows.

$$L \geq \frac{1}{\gamma_{dcp}^2 d} \tag{34}$$

Algorithm 1 The Generative Process of Normal-PCSP

Input: initial structure embedding $\mathbf{s}_0$,
periodicity criterion $\beta \in \{0.9, 0.95, 0.99\}$.
Output: perturbed structure embeddings $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$.
$\gamma_{dcp} = \frac{\zeta}{4\text{erf}^{-1}(\beta)}$
$L_{dcp} = \min\left\{8n \mid n \in \mathbb{N}, 8n \geq \frac{1}{\gamma_{dcp}^2 d}\right\}$
$\delta = \frac{2\pi}{L_{dcp}}$
for $t = 0$ **to** $L-1$ **do**
  $\Re(\mathbf{s}_t) = \cos(\delta t)\Re(\mathbf{s}_0) + \gamma_{dcp}\sqrt{\frac{1-\cos(2\delta t)}{2}}\boldsymbol{\epsilon}_R$
  $\Im(\mathbf{s}_t) = \sin(\delta t)\Re(\mathbf{s}_0) + \gamma_{dcp}\sqrt{\frac{1-\sin(2\delta t)}{2}}\boldsymbol{\epsilon}_I$
  $\tilde{\mathbf{z}}_t = \phi(\Re(\mathbf{s}_t) \oplus \Im(\mathbf{s}_t))$
end for
Return: $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$

Algorithm 2 Information Retrieval Process of CVCR

Input: analytical data $\mathbf{x}$,
periodicity criterion $\beta \in \{0.9, 0.95, 0.99\}$.
Output: index of the retrieved atomic structure j^*.
for $\mathcal{A} \in \mathcal{D}_{\mathcal{A}}$ **do**
  $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1} = \texttt{Normal-PCSP}(h_{\mathcal{A}}(\mathcal{A}), \beta)$
  $\mathbf{z} = \sum_{t=0}^{L-1} \omega_t \tilde{\mathbf{z}}_t$
  $\mathcal{S} = \mathcal{S} + [h_{\mathcal{Q}}(\mathbf{x}) \cdot \mathbf{z}]$ // Append the calculated similarity.
end for
Return: argmax $\mathcal{S}$

D Algorithmic Descriptions of Normal-PCSP and CVCR

Algorithm 1 shows the generative process of Normal-PCSP to sample periodic embeddings grounded on a given initial structure embedding $\mathbf{s}_0$. Algorithm 2 presents the information retrieval process of CVCR based on the perturbed structure embeddings $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$ generated through Normal-PCSP.

E Datasets

In the experiments, six analytical and three molecular regression datasets were used to assess the retrieval and prediction capabilities of the proposed methods. Brief descriptions of the six analytical datasets are as follows.

- **QMe14S-IR** [52]: It contains 186,102 molecules and their IR spectra calculated through a high-throughput quantum mechanical simulation. The QMe14S-IR dataset covers diverse small organic molecules featuring 14 elements (H, B, C, N, O, F, Al, Si, P, S, Cl, As, Se, and Br) and 47 functional groups. The optimized molecular structures and their IR spectra were calculated at a B3LYP/TZVP level.
- **QMe14S-RM** [52]: Similarly, the QMe14S-RM dataset contains 186,102 small organic molecules and their Raman spectra calculated through a high-throughput quantum mechanical simulation at a B3LYP/TZVP level.
- **NIST** [31]: It is an analytical dataset containing 8,832 gas-phase molecules and their IR spectra collected experimentally, which was extracted from the original NIST database [25]. While the NIST dataset provides a smaller number of molecules than the QMe14S-RM datasets, it includes a broader and more diverse set of molecules and their experimental IR spectra.
- **OpenXRD** [45]: The OpenXRD is a subset of a public database in analytical materials science. It contains 872 crystalline materials and experimentally measured X-ray diffraction patterns of them. The crystalline materials in the OpenXRD dataset cover diverse range of crystal structures of synthesized materials.

- **2DNMRGym** [23]: It is a collection of 11,825 molecules and experimentally collected nuclear magnetic resonance (NMR) results formatted by 2D images. The 2D-NMR images exhibit substantially different characteristics from well-known benchmark image data: sparse, discrete, and heterogeneous.
- **INS-MAT** [7]: This dataset provides 9,928 crystalline materials and their inelastic neutron scattering (INS) images, which are used to study the vibrational dynamics in a material. The 2D-INS images were calculated by a first principle calculation method under the following configurations: 0-150meV energy range, 1.5meV energy resolution, and 0.05 Å^{-1} bin size. Notably, the crystalline materials in INS-MAT span a wide range of structural complexity: the average molecular weight of unit cells is 4,865 ($\pm$ 3,806) Da, with the largest reaching 43,512 Da, and the dataset covers diverse elements with atomic numbers from 1 to 100. This range substantially exceeds the typical scale of small organic molecules (commonly under 1,000 Da) and overlaps with the macromolecular regime, providing a meaningful testbed for evaluating scalability beyond small molecules.

Descriptions of the three molecular and materials regression datasets are as follows.

- **ESOL** [9]: It is a well-known benchmark molecular dataset in cheminformatics and physical chemistry. This dataset provides experimentally observed 1,128 molecules and their aqueous solubilities. We use the ESOL dataset to evaluate the prediction capabilities of GNNs on retrieved molecular structures.
- **PCQM4Mv2** [15]: The PCQM4Mv2 dataset is a large benchmark dataset originally curated under the PubChemQC project [32]. It contains 3,378,606 and 147,037 molecules for training and test of machine learning models, respectively. The target HOMO-LUMO gaps of the molecules were calculated by density functional theory (DFT) on the PubChemQC.
- **MP** [17]: Materials Project (MP) is one of the most well-known database in materials science, which aims to accelerate data-driven materials discovery in battery, photovoltaics, and thermoelectric materials. We collected 155,353 crystalline materials and their band gaps calculated by DFT calculations from the original MP database.

F Implementations

Normal-PCSP and CVCR were implemented based on Python 3.12, PyTorch 2.9.0, and CUDA 13.0. The latent dimension of the query and structure embeddings were fixed to 512 for all benchmark datasets. We followed a conventional setting of the temperature parameter τ and fixed τ^{-1} to 14.3. For the competitor methods, we used the hyperparameters provided in their original papers [53, 38, 19, 28]. The structure encoders in the retrieval methods were implemented by AttentiveFP [49] and CGCNN [48] for molecules and crystalline materials, respectively. All source codes and experiment scripts of this work are publicly available at `https://huggingface.co/KRICT-AI/PCSP_CVCR`. All experiments were conducted on a machine with A100 64GB GPU.

G Computational Efficiency of the Sampling Procedure

As described in Section 3.2.6, Normal-PCSP can parallelly generate perturbed samples $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_{L-1}$ from the closed-form marginal distribution $p(\mathbf{s}_t|\mathbf{s}_0)$ without a time-consuming autoregressive process. In this section, we assess the entire computational efficiency of CVCR, which is a downstream information retrieval method of Normal-PCSP. Table 7 presents execution time of the competitor information retrieval methods and CVCR in seconds. The execution time was measured for a single training epoch in the training step, whereas the execution time of the inference step was measured over the entire retrieval dataset.

Table 7: Execution time of the competitor information retrieval methods and CVCR in seconds.

Dataset	Training Step						Inference Step					
	NegCLIP	CLIxP	SigLIxP	SMEN	Vib2Mol	CVCR	NegCLIP	CLIxP	SigLIxP	SMEN	Vib2Mol	CVCR
QMe14S-IR	41.965	43.636	45.611	41.776	42.913	44.452	3.209	3.571	5.525	3.619	3.591	3.760
QMe14S-RM	42.030	43.988	45.668	41.741	43.060	44.514	3.574	3.497	6.023	3.560	3.589	3.601
NIST	2.024	2.136	2.211	2.054	2.082	2.137	0.184	0.183	0.186	0.265	0.186	0.191
OpenXRD	1.917	1.924	1.937	1.902	1.878	1.860	0.270	0.277	0.267	0.281	0.245	0.244
2DNMRGym	15.262	15.915	15.581	15.540	15.843	16.004	1.626	1.635	1.722	2.141	1.627	1.690
INS-MAT	12.518	12.632	12.525	12.346	12.942	12.337	1.363	1.360	1.356	1.390	1.375	1.405

As shown in Table 7, CVCR with Normal-PCSP showed similar execution time with the competitor methods in both the training and inference steps. These results demonstrate that the additional computational overhead from the sampling procedure of Normal-PCSP is negligible in the information retrieval processes. Therefore, the computational costs of Normal-PCSP and CVCR are still practical in real-world applications.

H Analysis on the Periodicity Criterion

The periodicity criterion β is the fundamental criterion of Normal-PCSP to determine γ and L through the hyperparameter estimators γ_{dcp} and L_{dcp} in Eqs. (11) and (13). Furthermore, in CVCR, β can be interpreted as an importance of the periodicity because β controls the probability that the generated perturbed structure embeddings satisfy the periodicity. In this section, we measured the retrieval accuracy of CVCR for different values of $\beta \in \{0.90, 0.95, 0.99\}$ to assess the robustness and retrieval accuracy of CVCR in the cross-modal retrieval tasks under the different periodicity criteria, i.e. the different criteria of modeling constraints.

Table 8: Recall@1 of CVCR for different values of β.

Dataset	$\beta = 0.90$	$\beta = 0.95$	$\beta = 0.99$
QMe14S-IR	0.825 (0.012)	0.822 (0.009)	0.828 (0.005)
QMe14S-RM	0.857 (0.006)	0.852 (0.005)	0.898 (0.006)
NIST	0.790 (0.016)	0.788 (0.021)	0.792 (0.016)
OpenXRD	0.233 (0.117)	0.202 (0.101)	0.252 (0.027)
2DNMRGym	0.331 (0.024)	0.332 (0.039)	0.393 (0.017)
INS-MAT	0.301 (0.156)	0.323 (0.183)	0.433 (0.051)

Table 8 shows recall@1 of CVCR implemented with $\beta = 0.90, 0.95$, and 0.99. In the experiments, CVCR was robust to β on the QMe14S-IR, QMe14S-RM, NIST, and OpenXRD datasets, which contains input queries generated from optical spectrometers that aggregating perturbed atomic structures are relatively less important. However, a strict periodicity criterion $\beta = 0.99$ was crucial on the 2DNMRGym and INS-MAT datasets because the structure embeddings generated through the periodic perturbations were directly related to the input analytical data.

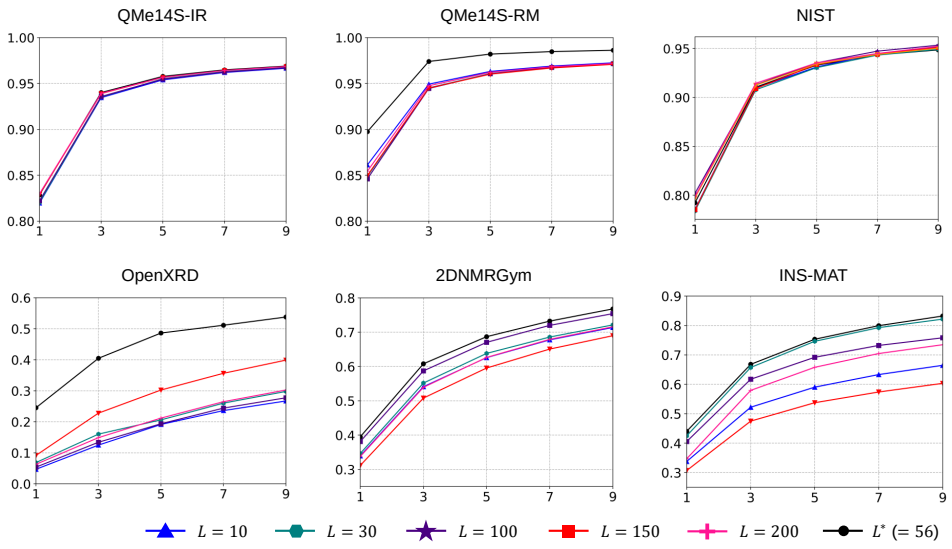

Figure 4: Recall@k of CVCR with Normal-PCSPs for different values of L.

I Retrieval Accuracy for Different Values of L

The period L is one of the essential parameters of Normal-PCSP, and it eventually determines the entire behaviors and retrieval accuracy of the downstream CVCR. In this experiment, we measured

Recall@k of CVCR with Normal-PCSPs for different values of $L \in \{10, 30, 100, 150, 200\}$. Fig. 4 shows the measured Recall@k for different values of L. We plotted Recall@k of the proposed CVCR with the estimator $L_{dcp}(= 56)$ as the black line. Although CVCR was robust to L on the QMe14S-IR and NIST datasets, its retrieval accuracy was somewhat different to the values of L on the OpenXRD, 2DNMRGym, and INS-MAT datasets. Nonetheless, the proposed estimator L_{dcp} led CVCR the highest retrieval accuracy regardless of the benchmark datasets. In particular, the accuracy improvements by L_{dcp} were on the OpenXRD dataset, which is a small experimental dataset containing experimental outliers and noises. These experimental results in Fig. 4 demonstrate the robustness of our estimator L_{dcp} in real-world cross-modal retrieval tasks.

J Ablation Study

We conducted an ablation study to assess the effectiveness of each module in CVCR in cross-modal retrieval tasks. We generated three ablation models by removing each module from the complete CVCR as follows. (1) w/o PCSP: removing the Normal-PCSP ($\approx$ SMEN). (2) w/o γ_{dcp}: removing the γ estimator in Eq. (11) and setting γ to 0.01 manually. (3) w/o L_{dcp}: removing the L estimator in Eq. (13) and setting L to 30 manually.

Table 9: Ablation study results of CVCR in cross-modal retrieval tasks.

Dataset	Recall@1				Recall@9			
	w/o PCSP	w/o γ_{dcp}	w/o L_{dcp}	CVCR	w/o PCSP	w/o γ_{dcp}	w/o L_{dcp}	CVCR
QMe14S-IR	0.683 (0.003)	0.821 (0.004)	0.820 (0.005)	0.828 (0.005)	0.923 (0.002)	0.960 (0.001)	0.967 (0.001)	0.969 (0.001)
QMe14S-RM	0.715 (0.006)	0.893 (0.005)	0.849 (0.007)	0.898 (0.006)	0.935 (0.002)	0.986 (0.001)	0.971 (0.003)	0.986 (0.001)
NIST	0.677 (0.018)	0.792 (0.015)	0.785 (0.010)	0.792 (0.016)	0.913 (0.004)	0.945 (0.005)	0.946 (0.003)	0.948 (0.008)
OpenXRD	0.185 (0.031)	0.178 (0.102)	0.075 (0.035)	0.252 (0.027)	0.488 (0.073)	0.439 (0.197)	0.291 (0.061)	0.533 (0.051)
2DNMRGym	0.103 (0.039)	0.340 (0.042)	0.346 (0.035)	0.393 (0.017)	0.228 (0.047)	0.716 (0.044)	0.722 (0.043)	0.768 (0.016)
INS-MAT	0.197 (0.053)	0.314 (0.151)	0.424 (0.043)	0.433 (0.051)	0.605 (0.041)	0.647 (0.142)	0.812 (0.035)	0.833 (0.036)

Table 9 presents recall@1 and recall@9 of three ablation models on the six benchmark datasets. We mainly observed three results in the ablation study. (1) As shown in the recall values of w/o PCSP (without PCSP), the retrieval accuracy was significantly degraded, and this result strongly supports the necessity of the perturbed structure embeddings in cross-modal retrieval with analytical observations. (2) As shown in the ablation models w/o γ_{dcp} and w/o L_{dcp}, the retrieval accuracy of CVCR was sensitive to the hyperparameters of Normal-PCSP (γ and L), but CVCR with the hyperparameters estimated by Eqs. (11) and (13) showed the highest retrieval accuracy regardless of the types of the input queries and target matters. (3) As shown in the retrieval accuracy on the OpenXRD dataset, the hyperparameter estimators γ_{dcp} and L_{dcp} were robust even on the small dataset.

K Experiment Setup of Retrieval-Based Molecular Regression

Existing machine learning models on chemical data typically assume that ground-truth atomic structures of unknown matters are known and available in the training and inference processes [11, 48]. However, obtaining atomic structures of unknown matters in real-world nature is expensive and sometimes infeasible due to the irreducible quantum mechanical uncertainty in measuring the atomic structures. Therefore, the typical assumption on the availability of ground-truth atomic structures is not valid in most real-world scientific applications.

In this experiment, we assess the availability of existing machine learning methods in a complete real-world regression setting where input ground-truth atomic structures are not given. We generated a joint chemical dataset from an analytical dataset (NIST or INS-MAT) and a target regression dataset, as illustrated in Fig. 5. The overall process to generate a joint chemical dataset consists of the following two steps: (1) Find the data in both the training analytical dataset and the target regression dataset based on its chemical identifier (e.g., SMILES and MP-ID). (2) Collect the matched data and construct a joint chemical dataset for training a cross-modal retrieval model and its downstream GNN. The same procedure is applied to the test analytical dataset and the target regression dataset to generate a joint chemical dataset for evaluating GNNs.

The NIST and INS-MAT datasets were used for regression tasks on molecules and crystalline materials, respectively. The target regression datasets for molecular and materials property predictions are presented in Appendix E. Table 6 shows the R^2-scores of the downstream GNNs trained with the atomic structures retrieved by each retrieval method on the training joint chemical dataset.

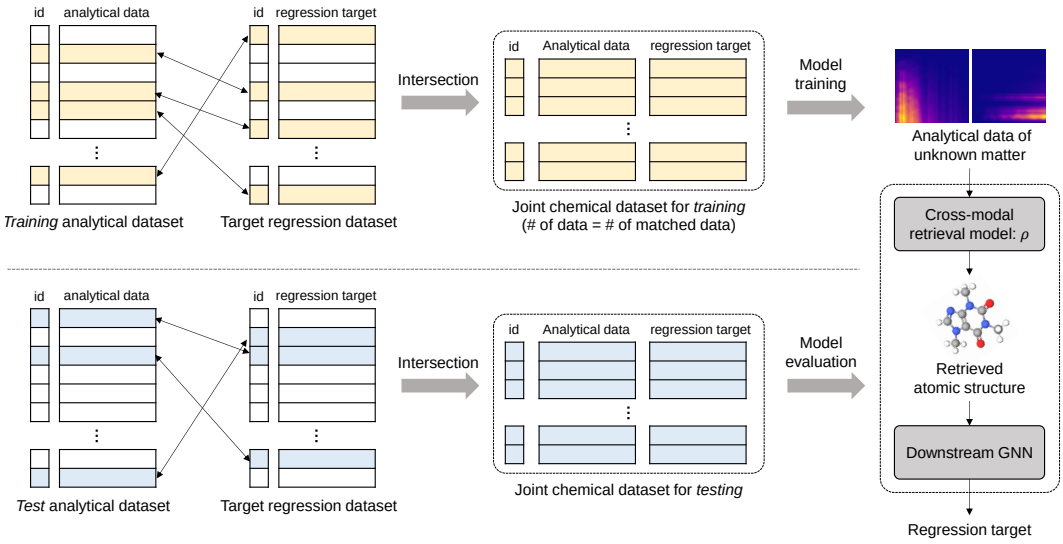

Figure 5: The overall processes to generate the joint chemical datasets and molecular regression in the complete real-world settings. The yellow rows indicate the matched data in both the training analytical dataset and the target regression dataset, and a set of the matched data compose the joint chemical dataset for training the cross-modal retrieval model and the downstream GNN. Similarly, the blue rows indicate the data matched from the test analytical dataset, and the matched data compose the test dataset for model evaluations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clarified in the abstract that capturing the underlying physics of the analytical data is essential, particularly the aggregation of non-equilibrium atomic structures induced by quantum mechanical perturbations.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed future work of the proposed method in the Conclusion section.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: This paper aims to present theoretical backgrounds of complex-valued periodic stochastic processes with a complete proof. Essential proofs are presented in the main text, and the full set of supplementary assumptions and proofs are provided in Appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all information regarding the benchmark datasets, competing methods, hyperparameter settings, and task specifications in the main text and Appendix. We have also released the source code and experimental scripts for the proposed method in the Experiments section.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data sources are presented in the main text. All implementation codes and experiment scripts are provided at `https://huggingface.co/KRICT-AI/PCSP_CVCR`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: A detailed description of the experimental setup and evaluation protocol is provided in the Experiment section and Appendix F.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We reported the standard deviations measured by the repetitions of the evaluation processes.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: A detailed description of computing resources is provided in Appendix F.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. To the best of our knowledge, the research conducted in this paper conforms, in every respect, with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper investigates the fundamental limitations of current AI for science in real-world scientific applications and demonstrates the effectiveness of the proposed method in fully realistic settings.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The datasets and models used in this work are limited to scientific regression tasks and do not involve sensitive, personal, or generative capabilities that could reasonably enable harmful misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes. The creators and original owners of the assets used in this paper are properly credited through citations.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The source code and implementation detail of the proposed method are documented with descriptions of their structure in the manuscript.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable. This paper does not involve crowdsourcing experiments or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve study participants or human-subject research.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.