Periodic Complex Stochastic Processes for Retrieving Atomic Structures of Unknown Matters

Gyoung S. Na [1 2] **Chanyoung Park** [2]

Abstract

Retrieving unknown atomic structures from observable analytical spectra or images remains a long-standing challenge in natural sciences. Although cross-modal retrieval methods achieved some notable successes in identifying atomic structures from the corresponding analytical data, their accuracy remains suboptimal because they have overlooked the underlying nature of analytical data: most analytical observations essentially reflect aggregations of multiple structural phases induced by periodic quantum mechanical perturbations. This paper proposes a periodic complex stochastic process (PCSP) that models such periodic perturbations and establishes theoretical backgrounds of PCSP, including its sample diversity, process length, and periodicity. Finally, for accurate atomic structure identifications on analytical data, we develop a complex-valued cross-modal retrieval (CVCR) by integrating PCSP with cross-modal retrieval frameworks. CVCR achieves state-of-the-art accuracy in cross-modal retrieval tasks of real-world analytical chemistry.

1. Introduction

Machine learning has substantially accelerated scientific discovery across physical and chemical sciences, enabling efficient data-driven molecular generation (Cornet et al., 2024; Chang & Ye, 2024), drug discovery (Alakhdar et al., 2024), and materials design (Zeni et al., 2025). The central insight of these successes lies in learning the relations between physical quantities of matters (e.g., molecules and crystalline materials) and their atomic structures (Gilmer et al., 2017; Xie & Grossman, 2018). However, obtaining ground-truth atomic structures of unknown matters is typically infeasible due to irreducible quantum mechanical uncertainty and limitations in measurement resolution (López-Lorente & Mizaikoff, 2016; Bunaciu et al., 2015). This lack of ground-truth atomic structures fundamentally constraints the practical potential of existing machine learning methods, because ground-truth atomic structures assumed to be accessible as model inputs are actually unavailable in most real-world applications (Delaney, 2004; Zhuo et al., 2018).

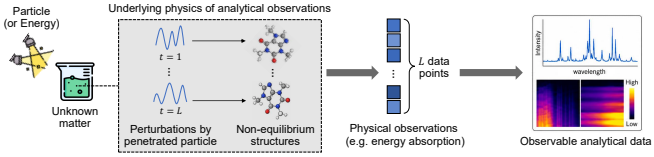

Figure 1. The overall analytical process based on quantum mechanical perturbations to generate analytical data of unknown matters.

Analytical chemistry is a research field to identify atomic structures of unknown matters from their analytical data (e.g., infrared spectra (López-Lorente & Mizaikoff, 2016) and neutron scattering images (Jeffries et al., 2021)). In analytical chemistry, most of analytical data are obtained by penetrating particles or energies into unknown matters, and the penetrated particles and energies generate quantum mechanical perturbations governed by the time-dependent Schrödinger equation based on periodic wave functions (Shankar, 2012; De Raedt, 1987) as:

$$H|\psi(t)\rangle = i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle, \tag{1}$$

where H is the Hamiltonian of the system, $|\psi(t)\rangle$ denotes the perturbed state at time t, and $\hbar$ is the reduced Planck constant. Multiple structural phases by quantum mechanical perturbations are captured by chemical instruments and subsequently aggregated to generate analytical spectra or images (López-Lorente & Mizaikoff, 2016; Jeffries et al., 2021). Therefore, as illustrated in Fig. 1, the underlying nature of the observed analytical data is originated from the perturbed states rather than a single equilibrium state.

Despite recent substantial advances in analytical instruments and methods, identifying atomic structures of unknown matters from their analytical data remains an open problem in natural sciences, because many of physical principles for

*Equal contribution [1]Korea Research Institute of Chemical Technology (KRICT) [2]Korea Advanced Institute of Science & Technology (KAIST). Correspondence to: Gyoung S. Na <ngs0@krict.re.kr>, Chanyoung Park <cy.park@kaist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

characterizing atomic structures from spectral and scattering observations are still not fully understood (López-Lorente & Mizaikoff, 2016). For more accurate identification of atomic structures, various data-driven methods have been investigated in both machine learning and analytical chemistry (Na & Rho, 2025; Alberts et al., 2025). In particular, cross-modal retrieval methods have demonstrated substantial accuracy in identifying atomic structures of unknown matters from their analytical data (Lu et al., 2025b). However, their accuracy in identifying atomic structures of unknown matters remains inherently suboptimal because they overlooked the underlying nature of analytical data. As illustrated in Fig. 1, analytical data is essentially generated from a set of periodically perturbed atomic structures (López-Lorente & Mizaikoff, 2016; Jeffries et al., 2021), i.e., the underlying nature of the analytical data is originated from perturbed atomic structures rather than a equilibrium structure in ideal closed and static environments. Nevertheless, existing information retrieval methods attempted to learn a direct mapping between analytical data and the ground-truth equilibrium structure (Kanakala et al., 2024; Lu et al., 2025b), which ultimately leads to suboptimal retrieval performance.

Physically, as shown in Eq. (1), quantum mechanical perturbations in atomic structures can be decomposed into time-dependent periodic functions. Hence, capturing the underlying periodic stochastic processes that generate perturbed atomic structures is crucial in atomic structure retrieval tasks. However, implementing a continuous and learnable stochastic process that generates periodic samples is not trivial, since generated samples $\mathbf{x}_t$ should satisfy the following constraint for all $t \in \mathbb{N}$ as:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, ..., \mathbf{x}_0) = p(\mathbf{x}_{t+L}|\mathbf{x}_{t+L-1}, ..., \mathbf{x}_0), \quad (2)$$

where $L \in \mathbb{N}$ is the period of the stochastic process. However, in the training problems, the number of constraints in Eq. (2) grows exponentially with respect to L, because the model should satisfy Eq. (2) 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 studied in machine learning (Yang et al., 2023; Lin et al., 2024), 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 processes (PCSP) to generate continuous and learnable periodic perturbations. PCSP adopts the architectures of Gaussian AR(1) models for generative processes (Kingma et al., 2021), and PCSP with complex normal distributions (Normal-PCSP) parameterizes its transition probability based on complex normal distributions with learnable means and variances as:

$$p(\mathbf{s}_t) = \mathcal{CN}(\mathbf{s}_t; \alpha_t \mathbf{s}_{t-1}, \gamma^2(1-\alpha_t^2)\mathbf{I}), \quad (3)$$

where $\mathcal{CN}$ denotes a complex normal distribution (Goodman, 1963), $\bar{\alpha}_t \in \mathbb{C}$ is a complex-valued coefficient of the stochastic process, and $\gamma^2 \in \mathbb{R}^+ \setminus \{0\}$ is a scaling parameter. 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 t-th perturbed atomic structures behind the input analytical data. Normal-PCSP defines the accumulated coefficients $\prod_{i=1}^{t} \alpha_i$ as the Euler's formula (Moskowitz, 2002). In Section 3.2.1, we prove that a sequence of samples $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_{L-1}, ...$ drawn from Normal-PCSP are strictly periodic. Furthermore, in Section 4.1, we demonstrate that the periodicity of the generated samples is preserved for nonlinear transformations.

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 for the stored atomic structures through Normal-PCSP. Subsequently, it 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 with the embedding vectors of the perturbed structures.

We evaluated the retrieval accuracy and computational efficiency of CVCR on real-world analytical datasets, each consisting of pairs of analytical data and its corresponding ground-truth atomic structure. We employed six benchmark datasets generated from widely used analytical methods, such as infrared spectroscopy (Yuan et al., 2025; Linstrom & Mallard, 2001) and neutron scattering diffractometer (Cheng et al., 2023). In cross-modal retrieval tasks, CVCR significantly outperformed state-of-the-art cross-modal retrieval methods by leveraging Normal-PCSP. In particular, the retrieval accuracy of CVCR reached 99% in a task of identifying molecules from their Raman spectra. Furthermore, we considered the complete real-world regression setting in which input ground-truth atomic structures are not available and only their analytical data is provided. In this setting, downstream prediction models with CVCR outperformed those with existing cross-modal retrieval methods.

2. Related Work

2.1. Periodic Stochastic Processes

A function is called an almost periodic function (Shubin, 1978) if $\forall \epsilon > 0, \exists L > 0$ such that every interval of L contains a shift τ satisfying:

$$||f(t+\tau) - f(t)|| < \epsilon, \forall t \in \mathbb{R}. \quad (4)$$

There are several attempts to design periodic stochastic processes based on the almost periodic functions (Iksanov et al., 2025). However, almost periodic stochastic processes are computationally expensive and sometimes assume unrealistic domain of input variables (Iksanov et al., 2025). In particular, periodic stochastic processes that guarantee the original periodicity (Beardon, 1997) in general input domains has not yet been investigated.

Rather than explicitly implementing periodic stochastic processes behind scientific systems, machine learning methods to learn mapping functions between system inputs and periodic outputs have been mainly studied (Yang et al., 2023; Lin et al., 2024; Wu et al., 2024). However, despite the notable prediction accuracy of existing methods, they still have two fundamental limitations due to their objective of learning mapping functions rather than learning the underlying periodic stochastic processes: (1) They require extensive training data covering multiple periods of the target systems to learn their periodic behaviors. (2) They are essentially interpolation models, i.e., periodicity of their outputs is not guaranteed in the extrapolation regions. Furthermore, the objectives of existing methods and PCSP are fundamentally different: existing methods learn a mapping from observed inputs to periodic outputs, whereas PCSP explicitly models a stochastic process that generates periodic outputs.

2.2. Cross-Modal Retrieval Methods

Cross-modal retrieval methods aim to retrieve relevant answers whose modalities differ from that of the input query (Radford et al., 2021). 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 (Na & Rho, 2025), materials discovery (Lai et al., 2025), and biochemical analysis (Sanchez-Fernandez et al., 2023). In chemical science, cross-modal retrieval methods, called spectra and molecule encoder network (SMEN) (Kanakala et al., 2024) and Vib2Mol (Lu et al., 2025b), 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 (Lu et al., 2025a) was proposed for image-to-structure retrieval under cell imaging perturbations, its scope is protein-level perturbations and limited to image-to-structure modalities.

2.3. Graph Neural Networks for Molecular Dynamics

Molecular dynamics (MD) simulation aims to calculate inter-molecular dynamics based on quantum mechanical principles (Hollingsworth & Dror, 2018). 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 (Maji et al., 2025; Chang & Zhu, 2025). 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 (Yagasaki & Saito, 2009), whereas PCSP aims to simulate non-equilibrium states perturbed by inter-atomic interactions within a closed molecular (or crystalline) system.

3. Method

3.1. Periodic Complex Stochastic Processes (PCSP)

PCSP is a complex-valued stochastic process that generates samples satisfying the periodicity. PCSP adopts the architecture of well-established time-dependent Gaussian AR(1) models, and thus the transition probability of PCSP is defined as $p(\mathbf{s}_t|\mathbf{s}_{t-1})$. However, PCSP should satisfy L periodicity constraints in Eq. (5) to ensure the periodicity over the entire stochastic process.

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = p(\mathbf{s}_{t+L}|\mathbf{s}_{t+L-1}), \forall t = \{0, 1, ..., L-1\}, \quad (5)$$

where $\mathbf{s}_t \in \mathbb{C}^d$ is a d-dimensional complex random variable at t, and $L \in \mathbb{N}$ is the period of PCSP. Therefore, we can define PCSP as a complex-valued stochastic processes that strictly follows the transition probability in Eq. (5).

3.2. PCSP with Complex Normal Distributions

Normal-PCSP is one of the specific implementations of PCSP and defines its transition probability based on a complex normal distribution $\mathcal{CN}$ (Goodman, 1963). Formally, Normal-PCSP models the stochastic process such that $\mathbf{s}_t$ follows a parameterized complex normal distribution as:

$$\mathbf{s}_t \sim \mathcal{CN}(\alpha_t \mathbf{s}_{t-1}, \gamma^2(1-\alpha_t^2)\mathbf{I}), \quad (6)$$

where $\gamma \in \mathbb{R}^+$ is a scaling parameter of the transition, and $\alpha_t \in \mathbb{C}$ is a complex-valued transition coefficient at t.

Normal-PCSP defines α_t based on Euler's formula with a period-dependent coefficient $\delta = 2\pi/L$. Formally, α_t is defined as a time-independent complex number for a given period as $\alpha_t = e^{i\delta}$. By the definition of $e^{i\delta}$, we can rewrite the distribution of $\mathbf{s}_t$ with respect to the initial sample $\mathbf{s}_0 \in \mathbb{R}^d$ based on the reparameterization trick as:

$$\mathbf{s}_t \sim \mathcal{CN}\left(\bar{\alpha}_t \mathbf{s}_0, \gamma^2(1-\bar{\alpha}_t)\mathbf{I}\right), \quad (7)$$

where $\bar{\alpha}_t = e^{i\delta t}$ is a time-dependent accumulated transition coefficient. Note that we assume $\mathbf{s}_0$ is an initial structure embedding generated by real-valued embedding networks, i.e., the real part $\Re(\mathbf{s}_0) = \mathbf{s}_0$ and the imaginary part $\Im(\mathbf{s}_0) = 0$. The full derivation of Eq. (7) is provided in Appendix A.

To define valid probability distributions for the complex variables $\mathbf{s}_t$, we decompose the complex-valued variance into real and imaginary parts. By the definition of complex normal distribution, decomposed real and imaginary parts of $\mathbf{s}_t$ each follow real-valued normal distributions as follows.

$$\Re(\mathbf{s}_t) \sim \mathcal{N}\left(\cos(\delta t)\mathbf{s}_0, \frac{\gamma^2}{2}(1-\cos(2\delta t))\mathbf{I}\right), \quad (8)$$

$$\Im(\mathbf{s}_t) \sim \mathcal{N}\left(\sin(\delta t)\mathbf{s}_0, \frac{\gamma^2}{2}(1-\sin(2\delta t))\mathbf{I}\right). \quad (9)$$

3.2.1. Stochastic Periodicity

Periodicity is the fundamental characteristics of the quantum mechanical perturbations (Shankar, 2012), and Normal-PCSP is designed to generate periodic samples that reflect periodic perturbations on the initial state $\mathbf{s}_0$. We mathematically validate the periodicity of Normal-PCSP and derive its hyperparameter estimators.

First, we 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.

$$\begin{aligned}\bar{\alpha}_{t+L} &= \cos(\delta(t+L)) + i\sin(\delta(t+L)) \\ &= \cos(\delta t + 2\pi) + i\sin(\delta t + 2\pi) = e^{i\delta t}. \end{aligned} \quad (10)$$

Therefore, $\bar{\alpha}_t$ is periodic for the period L, and consequently the expected values of $\mathbf{s}_t$ is also periodic as:

$$E[\bar{\alpha}_t \mathbf{s}_0] = E[\bar{\alpha}_{t+L}\mathbf{s}_0] \Leftrightarrow E[\mathbf{s}_t] = E[\mathbf{s}_{t+L}]. \quad (11)$$

Then, we prove the periodicity of the generated sample $\mathbf{s}_t$ based on the periodic characteristics of $\bar{\alpha}_t$ and $E[\mathbf{s}_t]$. By the definition of the periodicity with error level $\zeta \in \mathbb{R}$, stochastic periodicity between $\mathbf{s}_t$ and $\mathbf{s}_{t+L}$ can be defined as the following two equations.

$$p\left(|\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L})| < \frac{\zeta}{2}\right) = \beta, \quad (12)$$

$$p\left(|\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L})| < \frac{\zeta}{2}\right) = \beta, \quad (13)$$

where $\zeta \in \mathbb{R}^+$ is an error level, and $\beta \approx 1$ 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 for different values of β in Section 4.1 and Appendix H.

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 shown in Eqs. (12) and (13). 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 $E[\mathbf{s}_{t+L}]$, $\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L})$ follows a normal distribution $\mathcal{N}(0, \sigma_R^2)$, where σ_R^2 is the variance of the real part calculated by:

$$\sigma_R^2 = \gamma^2(1-\cos(2\delta t)). \quad (14)$$

Similarly, $\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L})$ follows a normal distribution $\mathcal{N}(0, \sigma_I^2)$ with the variance of the imaginary part as:

$$\sigma_I^2 = \gamma^2(1-\sin(2\delta t)). \quad (15)$$

Thus, the sample difference $D_R = |\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L})|$ follows a half-normal distribution (Röver et al., 2021), which a probability density function (PDF) is given by:

$$f_{D_R}(d_R; \sigma_R) = \frac{\sqrt{2}}{\sigma_R\sqrt{\pi}} \exp\left(-\frac{(\Re(\mathbf{s}_t) - \Re(\mathbf{s}_{t+L}))^2}{2\sigma_R^2}\right). \quad (16)$$

Also, the sample different $D_I = |\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L})|$ of the imaginary part has the following PDF.

$$f_{D_I}(d_I; \sigma_I) = \frac{\sqrt{2}}{\sigma_I\sqrt{\pi}} \exp\left(-\frac{(\Im(\mathbf{s}_t) - \Im(\mathbf{s}_{t+L}))^2}{2\sigma_I^2}\right). \quad (17)$$

Therefore, their cumulative distribution functions (CDFs) of D_R and D_I for a small positive real value $\zeta_h = \zeta/2$ are given as Eqs. (18) and (19), respectively.

$$F_{D_R}(\zeta_h) = \text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\cos(2\delta t))}}\right), \quad (18)$$

$$F_{D_I}(\zeta_h) = \text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\sin(2\delta t))}}\right). \quad (19)$$

Based on the CDFs, we can derive an estimator for γ as Eq (20), so that Normal-PCSP satisfies the stochastic periodicity with a probability β under an error level ζ. Detailed explanations are provided in Appendix B. In the implementation of Normal-PCSP, we estimated γ under $\zeta = 1/\sqrt{d}$ and $\beta = 0.99$.

$$\gamma = \begin{cases} \frac{\zeta_h}{2\text{erf}^{-1}(\beta)\max_t\{\sqrt{2(1-\cos(2\delta t))}\}}, & \text{for real} \\ \frac{\zeta_h}{2\text{erf}^{-1}(\beta)\max_t\{\sqrt{2(1-\sin(2\delta t))}\}}, & \text{for imaginary.} \end{cases} \quad (20)$$

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 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:

$$\Re(\mathbf{z})_i = \mathbf{w}_{0,i}\Re(\mathbf{s}_0)_i + \mathbf{w}_{1,i}\Re(\mathbf{s}_1)_i + \cdots + \mathbf{w}_{L-1,i}\Re(\mathbf{s}_{L-1})_i, \quad (21)$$

where $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{L-1}$ are pre-defined weight vectors based on problem definitions or 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:

$$Var[\Re(\mathbf{z})_i] = \frac{\gamma^2}{2}\sum_{t=0}^{L-1}\mathbf{w}_{t,i}^2(1-\cos(2\delta t)). \quad (22)$$

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. (22) eventually follows a generalized chi-squared distribution (Davies, 1980), as explained in Appendix C, and its expected value is given by:

$$E[Var[\Re(\mathbf{z})_i]] = \frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right). \quad (23)$$

By the diversity criterion $E[Var[\mathbf{z}_i]] \geq 1/d$, an inequality Eq. (24) to estimate the period L is derived as follows.

$$\frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right) \geq \frac{1}{2d} \quad (24)$$

Similarly, we can calculate $E[Var[\Im(\mathbf{z})_i]]$ and derive an inequality to estimate L as follows.

$$\frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\sin(2\delta t)\right) \geq \frac{1}{2d}. \quad (25)$$

However, calculating an analytical solution for L, which satisfies Eqs. (24) and (25), is not feasible because γ is itself a nonlinear function of δ determined by L, as derived in Eq. (20). In Section 3.2.4, we propose an efficient decoupling method of γ and L to calculate an analytical solution.

3.2.4. Decoupling of γ and L

As shown in Eq. (20), 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 $L \in \{4n | n \in \mathbb{N}\}$. Numerically, for all $L \in \{4n \mid n \in \mathbb{N}\}$, π and $\frac{3}{2}\pi$ always exist in $\{2\delta t \mid t = 0, 1, ..., L-1\}$, and thus the maximum values of $\sqrt{2(1-\cos(2\delta t))}$ and $\sqrt{2(1-\sin(2\delta t))}$ for t are fixed to the global maximum value $\sqrt{2^2} = 2$. Therefore, γ is decoupled from L under the constraint $L \in \{4n | n \in \mathbb{N}\}$, 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:

$$\gamma_{dcp} = \frac{\zeta}{4\text{erf}^{-1}(\beta)}. \quad (26)$$

In the implementation of Normal-PCSP, we use γ_{dcp} instead of γ under the mild condition of $L \in \{4n \mid n \in \mathbb{N}\}$.

3.2.5. Estimation of L

By substituting γ_{dcp} into γ under the condition $L \in \{4n \mid n \in \mathbb{N}\}$, we can simplify the inequalities Eqs. (24) and (25) into the following inequality with respect to L as:

$$L \geq \frac{1}{\gamma_{dcp}^2 d}. \quad (27)$$

Among the possible values of L, we set L as the minimum value satisfying Eq. (27) to reduce the computational cost and the number of parameters of downstream models. Finally, our estimator for L under γ_{dcp} is defined as:

$$L_{dcp} = \min\left\{4n \mid n \in \mathbb{N}, 4n \geq \frac{1}{\gamma_{dcp}^2 d}\right\}. \quad (28)$$

3.2.6. Sampling Procedure with γ_{dcp} and L_{dcp}

By Eqs. (8) and (9), 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. $\Re(\mathbf{s}_t)$ and $\Im(\mathbf{s}_t)$ are individually generated through normal distributions as:

$$\Re(\mathbf{s}_t) = \cos(\delta t)\Re(\mathbf{s}_0) + \gamma_{dcp}\sqrt{\frac{1-\cos(2\delta t)}{2}}\boldsymbol{\epsilon}_R, \quad (29)$$

$$\Im(\mathbf{s}_t) = \sin(\delta t)\Im(\mathbf{s}_0) + \gamma_{dcp}\sqrt{\frac{1-\sin(2\delta t)}{2}}\boldsymbol{\epsilon}_I, \quad (30)$$

where $\boldsymbol{\epsilon}_R, \boldsymbol{\epsilon}_I \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$, and $\delta = 2\pi/L_{dcp}$. By Eqs (29) and (30), Normal-PCSP can parallelly generate perturbed samples from $p(\mathbf{s}_t|\mathbf{s}_0)$ for an arbitrary t. We evaluate computational efficiency of Normal-PCSP in Appendix G.

3.2.7. Projection

Since Normal-PCSP is performed in the complex-valued space, we need to project the complex-valued samples $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{L-1}$ 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 $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{L-1}$ 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:

$$\tilde{\mathbf{z}}_t = \phi(\Re(\mathbf{s}_t) \oplus \Im(\mathbf{s}_t)), \tag{31}$$

where $\oplus$ is the vector concatenation operator, and ϕ can be a learnable nonlinear 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 generated samples. Algorithm 1 shows the overall process of Normal-PCSP to generate $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{L-1}$ for an initial embedding $\mathbf{s}_0$.

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

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 analysis 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 $\mathbf{q}$ 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 $\bar{\mathbf{z}}_0, \bar{\mathbf{z}}_1, ..., \bar{\mathbf{z}}_{L-1}$ with learnable weights $\omega_0, \omega_1, ..., \omega_{L-1}$ as:

$$\mathbf{z} = \sum_{t=0}^{L-1} \omega_t \tilde{\mathbf{z}}_t. \tag{32}$$

An ablation study to assess the effectiveness of each component in CVCR is conducted in Appendix F.

We employ the InfoNCE loss (Oord et al., 2018) to optimize model parameters of CVCR. However, we compare the similarity between $\mathbf{h}$ and $\mathbf{z}$, rather than between $\mathbf{h}$ and $\mathbf{s}_0$ as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left\{\ln\left(\frac{e^{\tilde{u}_{ii}/\tau}}{\sum_{j=1}^{N} e^{\tilde{u}_{ij}/\tau}}\right) + \ln\left(\frac{e^{\tilde{u}_{ii}/\tau}}{\sum_{j=1}^{N} e^{\tilde{u}_{ji}/\tau}}\right)\right\}, \tag{33}$$

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 experiments on real-world benchmark datasets to evaluate the retrieval capabilities of CVCR. Implementation details and hyperparameter settings of Normal-PCSP and CVCR are provided in Appendix E. All experiments were conducted on a machine with A100 64GB GPU. Source codes are publicly available at https://huggingface.co/KRICT-AI/PCSP_CVCR.

4.1. Periodicity Analysis

In this section, we empirically assess the periodicity of Normal-PCSP. We generated 4-dimensional perturbed structure embeddings $\mathbf{z}$ by Eq. (32). Fig. 2 visualizes $\mathbf{z}$ for different values of β in $\{0.5, 0.9, 0.99\}$. As expected, $\mathbf{z}$ generated by the Normal-PCSP 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 $\beta = 0.99$, since $\mathbf{z}$ has the same period across all latent dimensions, the entire embedding $\mathbf{z}$ has the periodicity with the period L.

4.2. Cross-Modal Retrieval on Benchmark Datasets

We employed six benchmark datasets in analytical chemistry, covering sequence-to-graph retrieval (QMe14S-IR (Yuan et al., 2025), QMe14S-RM (Yuan et al., 2025), NIST (Na & Rho, 2025), OpenXRD (Vosoughi et al., 2025)) and image-to-graph retrieval (2DNMRGym (Li et al., 2025), INS-MAT (Cheng et al., 2023)). 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 (Yuksekgonul et al., 2023), CLIxP (Roth et al., 2025), SigLIxP (Roth et al., 2025), SMEN (Kanakala et al., 2024), and Vib2Mol (Lu et al., 2025b). Note that SMEN and Vib2Mol are specially designed for retrieving unknown molecules from their analytical data. 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

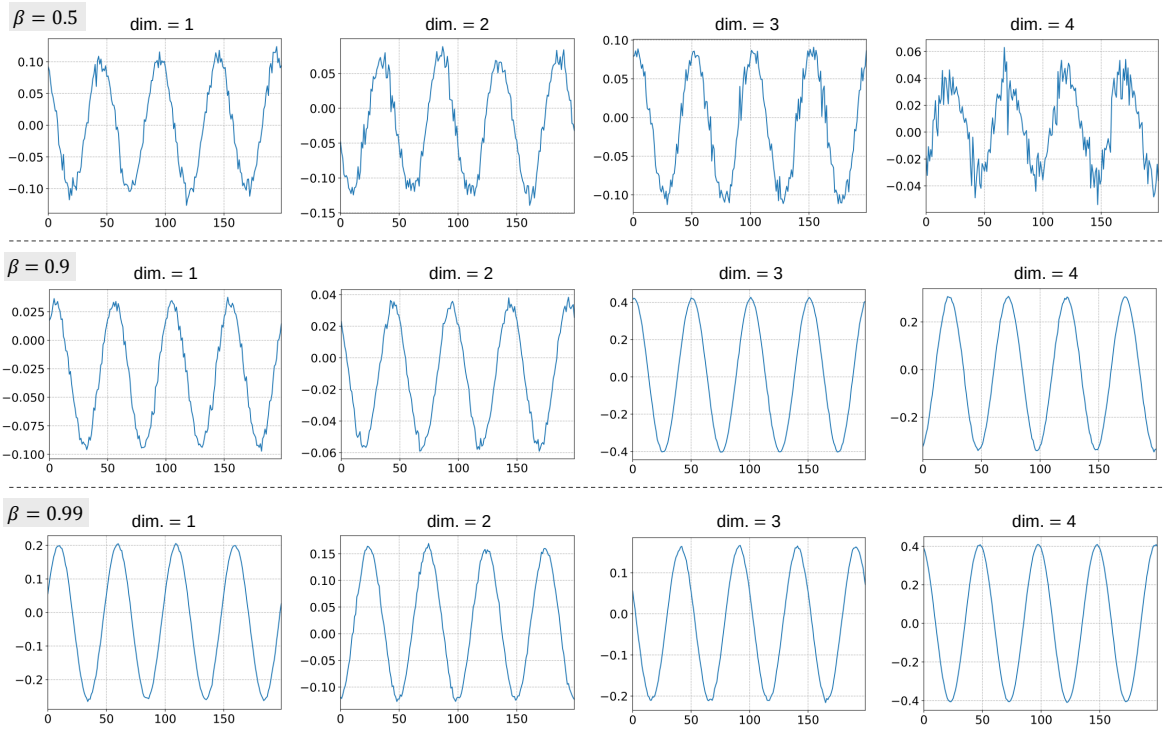

Figure 2. Visualization of 4-dimensional perturbed structure embeddings $\mathbf{z}$ for different values of β in $\{0.5, 0.9, 0.99\}$. X- and Y-axes are the timestep of the stochastic processes and the embedding value, respectively. The subfigures in each row visualize each dimension of $\mathbf{z}$.

Table 1. Two cross-modal retrieval tasks derived from six benchmark datasets in physical and chemical applications.

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

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 99% accuracy.

CVCR achieved significant accuracy improvements on the 2DNMRGym and INS-MAT datasets, and its recall@k reached 0.8 at $k = 9$. Since many analytical images are directly generated by aggregating physical observations of perturbed atomic states of unknown matters (López-Lorente & Mizaikoff, 2016; Cheng et al., 2023), the information 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.

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 7,066 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 0.018% of the library, CVCR still per-

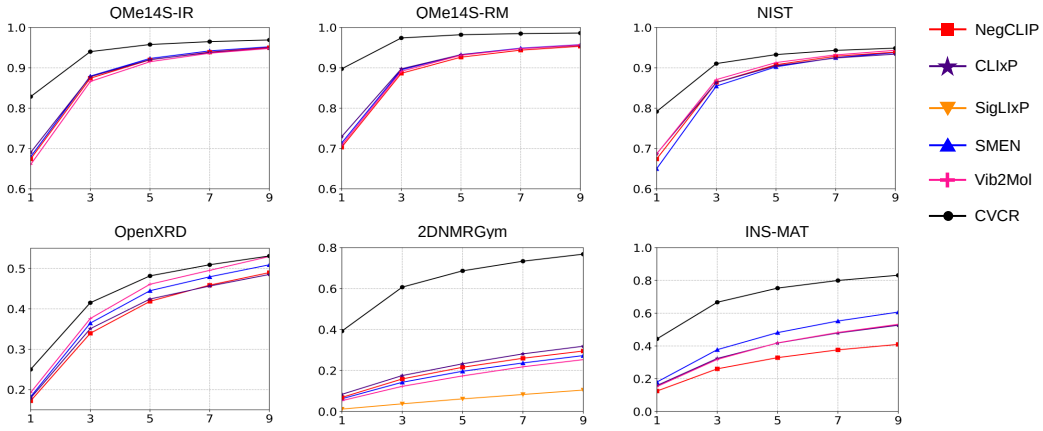

Figure 3. Recall@k of the competitor methods and CVCR. X-axis: $k \in \{1, 3, 5, 7, 9\}$, Y-axis: measured recall@k.

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)**

formed 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. Robustness Analysis

A natural concern with representation-level stochastic perturbation is whether the resulting retrieval performance degrades under noisy inputs. 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.

Robustness to Input Spectral Noise. In this experiment, we follow the standard definition $\text{SNR}_{\text{dB}} = 10\log_{10}(P_{\text{signal}}/P_{\text{noise}})$, 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 $\mathbf{s}_0$. We injected additive Gaussian noise at SNR$=$ 15 dB 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 3, Recall@k degraded only marginally, and recall@9 stayed above 0.95 on both datasets despite the strong noise injection.

Robustness to Initial Embeddings. 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 4, 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 — acting as an implicit stochastic regularizer.

Table 3. 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

Table 4. 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	0.791	0.929	0.948

4.5. Comparison with Augmentation Baselines

Although Normal-PCSP is architecturally similar to AR(1) diffusion processes (Kingma et al., 2021; Shi et al., 2023) 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.

Table 5. 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)**	**0.921 (0.010)**	**0.949 (0.008)**	**0.393 (0.017)**	**0.638 (0.015)**	**0.768 (0.016)**

Table 6. Recall@k of PCSP and standard data augmentation baselines on the NIST dataset.

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**

We compared recall@k of CVCR with denoising diffusion probabilistic model (DDPM) (Kingma et al., 2021), Diffusion Schrödinger bridge (DSB) (Shi et al., 2023), 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 5, 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.

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) (Arús-Pous et al., 2019). As shown in Table 6, PCSP outperformed the baselines, and these gains reflect a representation enhancement mechanism applicable to both training and inference rather than a generic regularization effect.

4.6. Regression in Complete Real-World Settings

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. In this experiment, we simulated the complete real-world regression setting 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 J.

Table 7. R^2-scores of downstream GNNs. None is the baseline predicting targets directly from analytical data through a spectrum encoder (Na & Rho, 2025) or an image encoder (He et al., 2016).

Retrieval Method	ESOL	PCQM4Mv2	MP
None	0.613 (0.046)	0.682 (0.065)	0.541 (0.031)
NegCLIP	0.848 (0.014)	0.863 (0.076)	0.193 (0.036)
CLIxP	0.903 (0.012)	0.972 (0.013)	0.163 (0.050)
SigLIxP	0.895 (0.014)	0.965 (0.011)	0.218 (0.021)
SMEN	0.875 (0.012)	0.972 (0.007)	0.421 (0.013)
Vib2Mol	0.908 (0.007)	0.973 (0.009)	0.308 (0.011)
CVCR	**0.962 (0.007)**	**0.987 (0.012)**	**0.619 (0.005)**

Three well-known chemical datasets were employed in this experiment: ESOL (Delaney, 2004), PCQM4Mv2 (Hu et al., 2021), and Materials Project (MP) (Jain et al., 2013). Brief descriptions about these datasets are provided in Appendix D. In the analytical datasets, 230, 1,471 and 9,675 matters were identified for the ESOL, PCQM4Mv2, and MP datasets, respectively. We used AttentiveFP (Xiong et al., 2019) and CGCNN (Xie & Grossman, 2018) to implement downstream GNNs for molecules and materials, respectively. Table 7 shows the R^2-scores of the downstream GNNs, which were trained on the datasets generated by each cross-modal retrieval method. The downstream GNNs with CVCR achieved the highest R^2-scores for all datasets, as CVCR accurately retrieved atomic structures. Although the baseline method (None) showed a high R^2-score in the MP dataset because X-ray diffraction 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 nature 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.

Impact Statements

In real-world scientific application, most AI-driven approaches for scientific discovery will ultimately face a fundamental limitation: molecular (or crystal) structures are typically unavailable because obtaining them can be time-consuming, costly, or infeasible in routine experiments. CVCR based on Normal-PCSP offers an efficient way to mitigate this fundamental limitation by enabling cross-modal retrieval to determine atomic structures from their analytical data. However, CVCR may retrieve incorrect atomic structures, and this limitation could cause serious issues in chemical applications that directly affect human health or the environment. Therefore, in high-stakes chemical applications, the atomic structures retrieved by CVCR should be re-checked and validated by qualified chemical experts before being used for downstream decisions or experiments.

Acknowledgments

This research was supported by Korea Research Institute of Chemical Technology (No. KK2661-51). This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-02304967, AI Star Fellowship(KAIST)), and by National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (RS-2022-NR068758).

References

Alakhdar, A., Poczos, B., and Washburn, N. Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, 64(19):7238–7256, 2024.

Alberts, M., Ficarra, F., and Laino, T. Language model enabled structure prediction from infrared spectra of mixtures. In *AI for Accelerated Materials Design-NeurIPS 2025*, 2025.

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., and Engkvist, O. Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11(1):71, 2019.

Beardon, A. F. *Periodic Functions*, pp. 83–96. Springer New York, New York, NY, 1997.

Bunaciu, A. A., UdriŞTioiu, E. G., and Aboul-Enein, H. Y. X-ray diffraction: instrumentation and applications. *Critical reviews in analytical chemistry*, 45(4):289–299, 2015.

Chang, J. and Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15(1):2323, 2024.

Chang, J. and Zhu, S. Mgnn: Moment graph neural network for universal molecular potentials. *npj Computational Materials*, 11(1):55, 2025.

Cheng, Y., Stone, M. B., and Ramirez-Cuesta, A. J. A database of synthetic inelastic neutron scattering spectra from molecules and crystals. *Scientific Data*, 10(1):54, 2023.

Cornet, F., Bartosh, G., Schmidt, M., and Andersson Naesseth, C. Equivariant neural diffusion for molecule generation. *Advances in Neural Information Processing Systems*, 37:49429–49460, 2024.

Davies, R. B. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, pp. 323–333, 1980.

De Raedt, H. Product formula algorithms for solving the time dependent schrödinger equation. *Computer Physics Reports*, 7(1):1–72, 1987.

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

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hollingsworth, S. A. and Dror, R. O. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

Hu, W., Fey, M., Ren, H., Nakata, M., Dong, Y., and Leskovec, J. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.

Iksanov, A., Kabluchko, Z., and Marynych, A. Almost periodic stochastic processes with applications to analytic number theory. *arXiv preprint arXiv:2502.04969*, 2025.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Jeffries, C. M., Ilavsky, J., Martel, A., Hinrichs, S., Meyer, A., Pedersen, J. S., Sokolova, A. V., and Svergun, D. I. Small-angle x-ray and neutron scattering. *Nature Reviews Methods Primers*, 1(1):70, 2021.

Kanakala, G. C., Sridharan, B., and Priyakumar, U. D. Spectra to structure: contrastive learning framework for library ranking and generating molecular structures for infrared spectra. *Digital Discovery*, 3(12):2417–2423, 2024.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Lai, Q., Xu, F., Yao, L., Gao, Z., Liu, S., Wang, H., Lu, S., He, D., Wang, L., Zhang, L., et al. End-to-end crystal structure prediction from powder x-ray diffraction. *Advanced Science*, 12(8):2410722, 2025.

Li, Y., Xu, H., and Hong, P. 2dnmrgym: An annotated experimental dataset for atom-level molecular representation learning in 2d nmr via surrogate supervision. *arXiv preprint arXiv:2505.18181*, 2025.

Lin, S., Lin, W., Hu, X., Wu, W., Mo, R., and Zhong, H. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.

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

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

Lu, M., Weinberger, E., Kim, C., and Lee, S.-I. CellCLIP - learning perturbation effects in cell painting via text-guided contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Lu, X., Ma, H., Li, H., Li, J., Li, Y., Zhu, T., Liu, G., and Ren, B. Vib2mol: from vibrational spectra to molecular structures-a versatile deep learning model, 2025b. URL https://arxiv.org/abs/2503.07014.

Maji, D., Ghosh, A., Barman, D., and Sarkar, P. 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.

Moskowitz, M. *A course in complex analysis in one variable*. World Scientific Publishing Company, 2002.

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

Nakata, M. and Shimazaki, T. 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.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Roth, K., Akata, Z., Damen, D., Balazevic, I., and Hénaff, O. J. Context-aware multimodal pretraining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4267–4279, 2025.

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., and Friede, T. On weakly informative prior distributions for the heterogeneity parameter in bayesian random-effects meta-analysis. *Research Synthesis Methods*, 12(4):448–474, 2021.

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S., and Klambauer, G. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023.

Shankar, R. *Principles of quantum mechanics*. Springer Science & Business Media, 2012.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

Shubin, M. A. Almost periodic functions and partial differential operators. *Russian Mathematical Surveys*, 33(2):1, 1978.

Vosoughi, A., Shahnazari, A., Xi, Y., Zhang, Z., Hess, G., Xu, C., and Abdolrahim, N. Openxrd: A comprehensive benchmark and enhancement framework for llm/mllm xrd question answering. *arXiv preprint arXiv:2507.09155*, 2025.

Wu, Q., Yao, G., Feng, Z., and Shuyuan, Y. Peri-midformer: Periodic pyramid transformer for time series analysis. *Advances in Neural Information Processing Systems*, 37: 13035–13073, 2024.

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

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., 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.

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

Yang, C., Chen, X., Sun, L., Yang, H., and Wu, Y. Enhancing representation learning for periodic time series with floss: A frequency domain regularization approach. *arXiv preprint arXiv:2308.01011*, 2023.

Yuan, M., Zou, Z., Luo, Y., Jiang, J., and Hu, W. Qme14s: A comprehensive and efficient spectral data set for small organic molecules. *The Journal of Physical Chemistry Letters*, 16(16):3972–3979, 2025.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. 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.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Zhuo, Y., Mansouri Tehrani, A., and Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673, 2018.

A. Sampling from Closed-Form Marginal Complex Normal Distributions

Normal-PCSP defines the distribution of the perturbed structure embedding $\mathbf{s}_t$ as:

$$\mathbf{s}_t \sim \mathcal{CN}(\alpha_t \mathbf{s}_{t-1}, \gamma^2(1-\alpha_t^2)\mathbf{I}), \tag{34}$$

Based on the reparameterization trick, we can rewrite $\mathbf{s}_t$ as a closed-form formula given by:

$$\mathbf{s}_t = \alpha_t \mathbf{s}_{t-1} + \gamma\sqrt{1-\alpha_t^2}\boldsymbol{\epsilon}_{t-1} \tag{35}$$

Recursively, $\mathbf{s}_t$ is described by the initial data $\mathbf{s}_0$ as:

$$\begin{aligned}
\mathbf{s}_t &= \alpha_t \mathbf{s}_{t-1} + \gamma\sqrt{1-\alpha_t^2}\boldsymbol{\epsilon}_{t-1} \\
&= \alpha_t\alpha_{t-1}\mathbf{s}_{t-2} + \gamma\sqrt{1-\alpha_t^2\alpha_{t-1}^2}\boldsymbol{\epsilon}_{t-2} \\
&= \cdots \\
&= \bar{\alpha}_t \mathbf{s}_0 + \gamma\sqrt{1-\bar{\alpha}_t^2}\boldsymbol{\epsilon}_0,
\end{aligned} \tag{36}$$

where $\boldsymbol{\epsilon}_0, ..., \boldsymbol{\epsilon}_{t-1} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \mathbf{I})$. Therefore, we can sample $\mathbf{s}_t$ from a closed-form marginal distribution $p(\mathbf{s}_t|\mathbf{s}_0)$ without time-consuming autoregressive chain as follows.

$$\mathbf{s}_t \sim \mathcal{CN}\left(\bar{\alpha}_t \mathbf{s}_0, \gamma^2(1-\bar{\alpha}_t)\mathbf{I}\right), \tag{37}$$

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

For a given periodicity criterion β and error level ζ, Eqs. (12) and (13) indicate that CDFs of D_R and D_I should greater than β for all $t \in \{0, 1, ..., L-1\}$ as follows.

$$\text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\cos(2\delta t))}}\right) \geq \beta, \tag{38}$$

$$\text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\sin(2\delta t))}}\right) \geq \beta, \tag{39}$$

where $\zeta_h = \zeta/2$. For the real part, Eq. (38) can be rewritten with respect to the minimum value over t as:

$$\min_{t\in\{0,1,...,L-1\}} \left\{\text{erf}\left(\frac{\zeta_h}{\gamma\sqrt{2(1-\cos(2\delta t))}}\right)\right\} = \beta \tag{40}$$

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. (40) into the following maximization problem as:

$$\text{erf}\left(\frac{\zeta_h}{\gamma \max_t\left\{\sqrt{2(1-\cos(2\delta t))}\right\}}\right) = \beta \tag{41}$$

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

$$\gamma = \frac{\zeta_h}{2\text{erf}^{-1}(\beta)\max_t\left\{\sqrt{2(1-\cos(2\delta t))}\right\}} \tag{42}$$

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

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

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

$$\gamma = \begin{cases} \frac{\zeta_h}{2\text{erf}^{-1}(\beta)\max_t\left\{\sqrt{2(1-\cos(2\delta t))}\right\}}, & \text{for real} \\ \frac{\zeta_h}{2\text{erf}^{-1}(\beta)\max_t\left\{\sqrt{2(1-\sin(2\delta t))}\right\}}, & \text{for imaginary.} \end{cases} \tag{43}$$

C. Sample Variances and Generalized Chi-Squared Distributions

By the definition in Eq. (8), the i-th real part of the generated sample is given by:

$$Var(\Re(\mathbf{s}_t)_i) = \frac{\gamma^2}{2}(1 - \cos(2\delta t)). \tag{44}$$

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:

$$\begin{aligned} Var(\Re(\mathbf{z})_i) &= Var(\mathbf{w}_{0,i}\Re(\mathbf{s}_0)_i + \cdots + \mathbf{w}_{L-1,i}\Re(\mathbf{s}_{L-1})_i) \\ &= \frac{\gamma^2}{2}\sum_{t=0}^{L-1}\mathbf{w}_{t,i}^2(1 - \cos(2\delta t)) \end{aligned} \tag{45}$$

However, we assume $\mathbf{w}_0, \mathbf{w}_2, ..., \mathbf{w}_{L-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$ 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:

$$\frac{\gamma^2}{2}(1 - \cos(2\delta t))\mathbf{w}_{t,i}^2 \sim \mathcal{X}_1'^2 \tag{46}$$

Therefore, $Var(\Re(\mathbf{z})_i)$ eventually follows a generalized chi-squared distribution (Davies, 1980), and its expected value is calculated by:

$$E[Var[\Re(\mathbf{z})_i]] = \sum_{t=0}^{L-1}\frac{\gamma^2}{2}(1 - \cos(2\delta t)) \tag{47}$$

$$= \frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1}\cos(2\delta t)\right). \tag{48}$$

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. Hence, the period (= process length) L for the real part should satisfy the following inequality.

$$\frac{\gamma^2}{2}\left(L - \sum_{t=0}^{L-1} cos(2\delta t)\right) \geq \frac{1}{2d}. \tag{49}$$

By substituting γ_{dcp} into γ, we can derive an inequality to calculate an analytical solution to estimate L as follows.

$$L \geq \frac{1}{\gamma^2 d} \tag{50}$$

D. 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** (Yuan et al., 2025): 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** (Yuan et al., 2025): 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** (Na & Rho, 2025): 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 (Linstrom & Mallard, 2001). While the NIST dataset provides a smaller number of molecules than the and QMe14S-RM datasets, it includes a broader and more diverse set of molecules and their experimental IR spectra.

- **OpenXRD** (Vosoughi et al., 2025): 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 syntheized materials.

- **2DNMRGym** (Li et al., 2025): 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** (Cheng et al., 2023): 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.

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

- **ESOL** (Delaney, 2004): 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** (Hu et al., 2021): The PCQM4Mv2 dataset is a large benchmark dataset originally curated under the PubChemQC project (Nakata & Shimazaki, 2017). 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** (Jain et al., 2013): 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.

E. 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 (Yuksekgonul et al., 2023; Roth et al., 2025; Kanakala et al., 2024; Lu et al., 2025b). The structure encoders in the retrieval methods were implemented by AttentiveFP (Xiong et al., 2019) and CGCNN (Xie & Grossman, 2018) for molecules and crystalline materials, respectively. All source codes and experiment scripts of this work are publicly available at `GithubURL`.

F. Ablation Study

We conducted an ablation study to assess the effectiveness of each module in CVCR in cross-modal retrieval tasks. We generated four 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. (26) and setting γ to 0.01 manually. (3) w/o L_{dcp}: removing the L estimator in Eq. (28) and setting L to 30 manually.

Table 8. 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)	0.949 (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)	0.443 (0.051)	0.605 (0.041)	0.647 (0.142)	0.812 (0.035)	0.833 (0.036)

Table 8 presents recall@1 and recall@9 of the ablation methods and CVCR 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, the retrieval accuracy was significantly degraded after removing PCSP, and this result strongly supports the necessity of the perturbed structure embeddings in cross-modal retrieval with analytical data. (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. (26) and (28) 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.

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 9 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 9. 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 9, 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. (26) and (28). 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 10. 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 (0.018)
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 10 shows recall@1 of CVCR implemented with $\beta = 0.90, 0.99$, 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

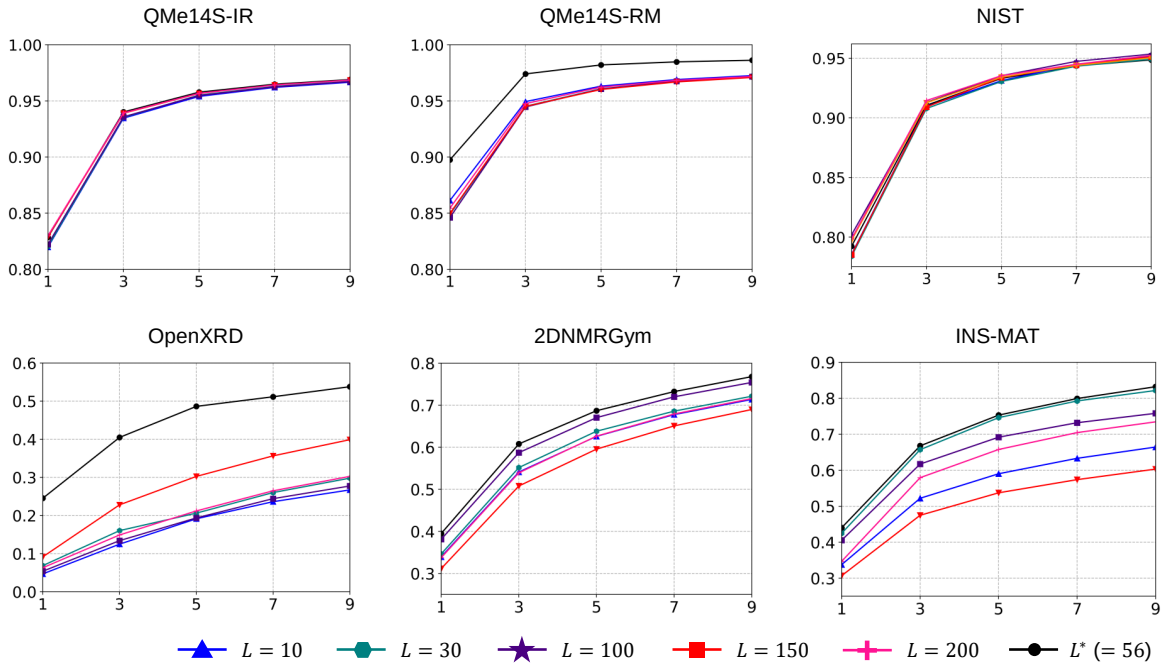

Figure 4. Recall@k of CVCR with Normal-PCSPs for different values of L.

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.

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. 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 (Gilmer et al., 2017; Xie & Grossman, 2018). 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

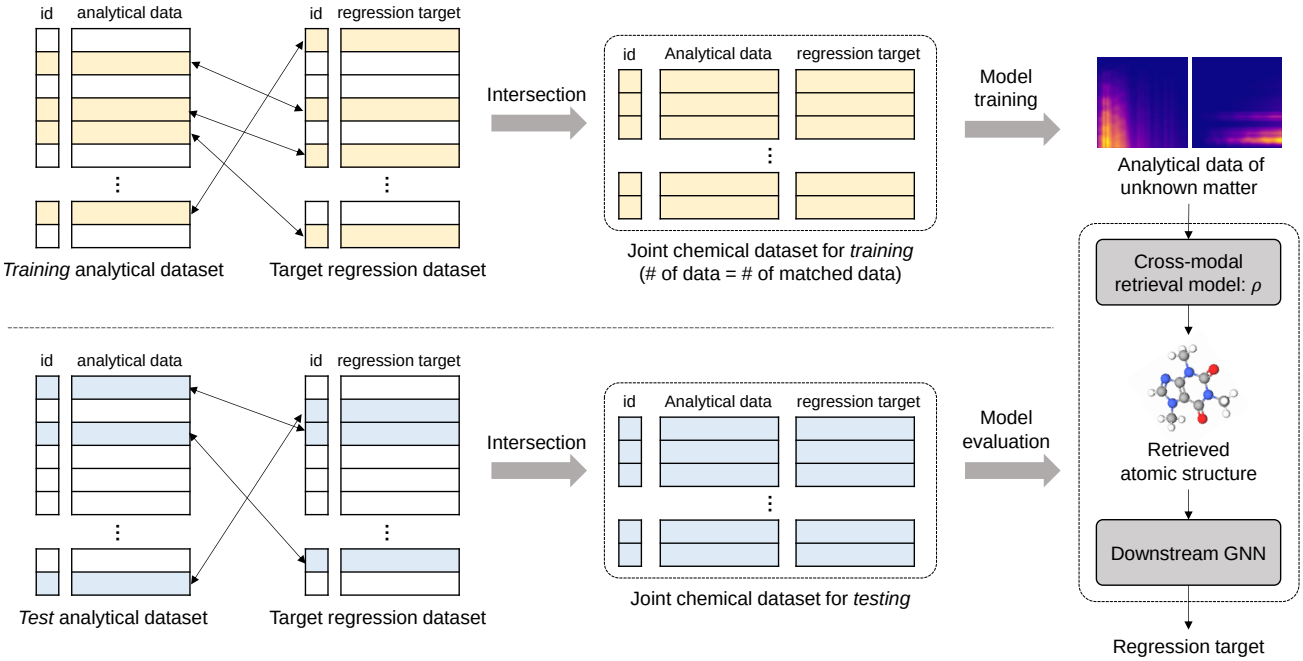

Figure 5. The overall processes to generate the joint chemical datasets and molecular regression in the complete real-world settings. The yellow rows indicate the matched data in both the training analytical dataset and the target regression dataset, and a set of the matched data compose the joint chemical 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.

target regression datasets for molecular and materials property predictions are presented in Appendix D. Table 7 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.