Latent Diffusion for Chemical Spectral Data: Preserving Non-Gaussian Manifold Structure

Kevin J. Metzler
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
kjmetzler@wpi.edu

Hy Lam
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Cate Dunham
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Randy Paffenroth
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

ABSTRACT

In low-data scientific domains such as ion mobility spectrometry (IMS), generating high-quality synthetic data requires understanding and preserving the underlying manifold structure of learned representations. Unlike Variational Autoencoders (VAEs) that enforce Gaussian latent distributions through KL divergence regularization, we argue that naturally learned latent spaces from standard autoencoders occupy complex, non-Gaussian manifolds that reflect the intrinsic geometry of the data.

This work introduces a latent diffusion approach for IMS chemical data that explicitly preserves this manifold structure. We demonstrate that x_0 prediction diffusion—where models directly predict clean latent representations rather than noise—maintains the characteristic tendril-like geometry of chemical latent spaces. Applied to 8 chemicals across 296,692 IMS spectra, our approach achieves near-perfect statistical matching (means within 0.08, standard deviations within 7.4%) while preserving complex geometric features that naive Gaussian sampling destroys.

By conditioning on molecular SMILE embeddings and class labels, a single diffusion model generates faithful synthetic spectra for all chemicals, demonstrating that respecting natural manifold structure—rather than forcing Gaussian distributions—is key to high-quality synthetic data generation in low-data scientific domains.

1 INTRODUCTION

Generative modeling in scientific domains faces a fundamental challenge: how to create synthetic data that preserves both statistical properties and underlying physical structure. In fields like ion mobility spectrometry (IMS), where data acquisition is expensive and time-consuming, synthetic data augmentation offers a path to improved machine learning model performance. However, naive approaches to generation often fail to capture the nuanced structure present in real measurements.

Traditional approaches to learned representations, particularly Variational Autoencoders (VAEs), enforce Gaussian latent distributions through explicit regularization [**?**]. While this constraint enables straightforward sampling via $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, it fundamentally alters the natural geometry of the data's representation. Standard autoencoders, without such regularization, learn latent spaces that reflect the intrinsic manifold structure of the data—often exhibiting complex, non-Gaussian geometry such as elongated tendrils, clusters, or curved surfaces.

1.1 The Manifold Hypothesis and Latent Representations

Real-world high-dimensional data typically occupies low-dimensional manifolds embedded in the ambient space [**?**]. For chemical spectroscopy data, this manifold structure arises from physical constraints: chemical composition, molecular structure, and measurement physics constrain spectra to a much smaller subspace than the full feature dimension suggests.

Our work builds on a *decoupled autoencoder* architecture [1] that makes the non-Gaussian nature of latent representations explicit. Rather than learning latent codes from scratch, we fix the latent space to be pre-trained ChemNet molecular embeddings [4]. The encoder learns to map 1676-dimensional IMS spectra to this predetermined 512-dimensional chemical embedding space, while the decoder learns the inverse mapping. Because ChemNet embeddings encode molecular structure without Gaussian distributional constraints, the resulting latent space exhibits strong non-Gaussian geometry—a direct consequence of the architectural choice to use chemically-meaningful, pre-defined representations rather than imposing distributional assumptions for sampling convenience.

The resulting latent spaces exhibit:

- **Non-Gaussian Geometry**: Elongated structures, tendrils, and anisotropic clusters reflecting chemical similarity relationships
- **Chemical Clustering**: Natural separation between distinct chemical classes with smooth transitions within classes
- **Variance Anisotropy**: Different chemicals occupy varying extents of latent space, reflected in per-chemical principal component explained variance ratios

VAE-style Gaussian regularization, while enabling simple sampling, destroys this structure. Forcing $q(\mathbf{z}|\mathbf{x}) \approx \mathcal{N}(0, \mathbf{I})$ through KL divergence penalties flattens the manifold, losing the geometric relationships that encode physical meaning.

1.2 Diffusion in Latent Space

Recent advances in diffusion models provide an alternative: rather than constraining the latent distribution, we can learn to generate from whatever distribution the data naturally occupies [3**?**]. By training diffusion models directly in latent space, we preserve the manifold structure while enabling high-quality generation.

A key advantage of diffusion models is their ability to work effectively with non-convex manifold structures. As illustrated in

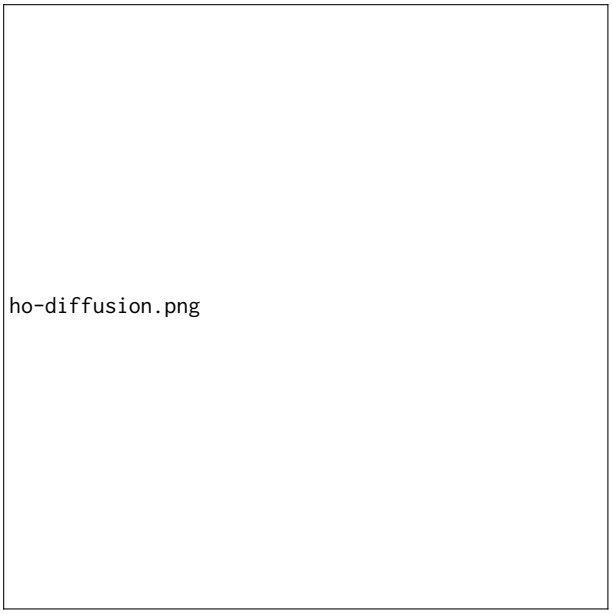

Figure 1: Manifold structure in diffusion models (adapted from Ho et al. [3]). The image manifold (left) is non-convex: interpolating between two face images may not produce a valid face. However, after adding noise to map points into the diffusion space (right), the structure becomes convex, allowing meaningful interpolation. The reverse diffusion process then maps interpolated points back to the original manifold, enabling generation of novel valid samples. This convexity property is crucial for preserving non-Gaussian manifold structure in latent spaces.

Figure 1, while the image manifold itself is non-convex (interpolating between two faces may not yield a valid face), the diffusion process creates a convex space where interpolation is well-defined. By noising two points from the data manifold into the diffusion space, we can perform convex combinations, then use the reverse diffusion process to map back to valid points on the original manifold. This property makes diffusion models particularly well-suited for generating from complex, non-Gaussian latent distributions.

Our key insight is that $\mathbf{x}_0$ **prediction parameterization**—where the model predicts clean latent vectors directly rather than predicting noise—provides superior manifold preservation. This approach offers the network direct supervision from target manifold locations during training, enabling it to learn the geometric structure explicitly.

1.3 Contributions

This work makes the following contributions:

(1) **Manifold-Aware Generation**: We demonstrate that latent diffusion with x_0 prediction preserves complex non-Gaussian manifold structure in chemical spectral data, achieving statistical matching while maintaining geometric fidelity.
(2) $\mathbf{X}_0$ **Prediction for Structure Preservation**: We show that directly predicting clean latent representations prevents the structural collapse observed in noise-prediction diffusion, where tendril-like manifolds degenerate to Gaussian blobs despite matching statistical moments.
(3) **Multi-Chemical Conditioning**: A single diffusion model conditioned on molecular SMILE embeddings and class labels generates high-quality synthetic data for 8 distinct chemicals, demonstrating parameter-efficient multi-task generation.
(4) **Comprehensive Validation**: Statistical metrics (near-perfect mean/std matching across 296,692 test samples) and geometric validation (per-chemical PCA analysis) confirm that respecting natural manifold structure is essential for scientific data generation.

2 BACKGROUND

2.1 Ion Mobility Spectrometry

Ion mobility spectrometry (IMS) is an analytical technique that separates ionized molecules based on their velocity in an electric field [2**?**]. The resulting spectra measure ion intensity as a function of drift time, providing chemical fingerprints for identification and detection applications. IMS offers rapid analysis times (milliseconds) with high sensitivity, making it particularly valuable for chemical detection in security, environmental monitoring, and defense applications [**?**].

Our IMS dataset comprises 296,692 samples across 8 chemical classes: DEB (Diethylbenzene), DEM (Diethylmalonate), DMMP (Dimethyl methylphosphonate), DPM (Dipropyl methylphosphonate), DtBP (Di-tert-butyl peroxide), JP8 (Jet fuel), MES (Methyl salicylate), and TEPO (Triethyl phosphate). Each sample consists of 1676 features (838 drift time measurements for positive ions, 838 for negative ions), split into 222,519 training and 74,173 test samples.

2.2 Decoupled Autoencoders with Chemical Embeddings

Following our previous work [1], we employ a *decoupled autoencoder* where the latent space is **given ahead of time** rather than learned. This is a critical architectural choice that fundamentally shapes the manifold structure we observe.

In standard autoencoders, the latent space is learned end-to-end from reconstruction objectives, with no constraints on its geometry. VAEs impose Gaussian structure through explicit KL regularization. Our decoupled approach takes a third path: we *fix* the latent space to be the 512-dimensional ChemNet embeddings [4], pre-trained on 3.6 million chemical structures to encode molecular SMILE (Simplified Molecular Input Line Entry System) representations.

The encoder E learns to map IMS spectra $\mathbf{x}_c$ for chemical c to this *pre-defined* ChemNet embedding space:

$$\mathcal{L}_{\text{enc}} = \|\mathbf{h}_c - E(\mathbf{x}_c)\|^2 \tag{1}$$

where $\mathbf{h}_c \in \mathbb{R}^{512}$ is the fixed ChemNet embedding for chemical c.

The decoder D learns the inverse mapping, reconstructing spectra from ChemNet embeddings:

$$\mathcal{L}_{\text{dec}} = \|\mathbf{x}_c - D(\hat{\mathbf{h}}_c)\|^2 \tag{2}$$

This decoupled architecture has profound implications for our work: because the latent space is predetermined by ChemNet's training on molecular structures—with no Gaussian distributional

constraints—there is **no guarantee that the resulting latent distribution is Gaussian**. Indeed, as Figure 2 demonstrates, the ChemNet embeddings for IMS chemicals exhibit highly non-Gaussian, tendril-like structures with anisotropic variance. This pre-defined, non-Gaussian latent geometry is the foundational motivation for our diffusion approach: we cannot assume Gaussian sampling will work when the latent space itself is not Gaussian by construction.

The decoupled autoencoder thus provides a rich, chemically-informed 512-dimensional latent space that respects molecular structure relationships while exhibiting the complex manifold geometry that standard generation methods fail to capture.

2.3 Diffusion Models

Diffusion models learn to reverse a gradual noising process, drawing inspiration from non-equilibrium thermodynamics [**?**]. The forward process adds Gaussian noise over T timesteps:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ controls the noise schedule.

Traditional diffusion models parameterize the reverse process to predict the noise ϵ added at each step [3**?**]:

$$\mathcal{L}_\epsilon = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2] \tag{4}$$

2.4 X_0 Prediction: Direct Manifold Supervision

An alternative parameterization directly predicts the clean data $\mathbf{x}_0$ [**?**]:

$$\mathcal{L}_{x_0} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}[\|\mathbf{x}_0 - f_\theta(\mathbf{x}_t, t)\|^2] \tag{5}$$

This formulation provides two critical advantages for manifold preservation:

1. Direct Observation of Target Manifold: The network receives gradients directly from clean latent vectors lying on the data manifold, learning to predict manifold locations rather than noise directions away from the manifold.

2. Geometric Awareness: By predicting target positions throughout training, the model develops an explicit representation of manifold structure, maintaining awareness of geometric features like tendrils, clusters, and anisotropic variance.

3 METHOD

3.1 X_0 Prediction Diffusion in Latent Space

Our approach applies x_0 prediction diffusion directly to the 512-dimensional latent space learned by the ChemNet-informed autoencoder. This preserves the naturally learned manifold structure while enabling efficient generation.

3.1.1 Architecture. We employ a multi-layer perceptron with the following specifications:

- **Input**: Concatenation of noised latent $\mathbf{x}_t \in \mathbb{R}^{512}$, time embedding $t_{\text{emb}} \in \mathbb{R}^{128}$, SMILE embedding $\mathbf{s} \in \mathbb{R}^{512}$, and class one-hot $\mathbf{c} \in \mathbb{R}^8$
- **Hidden Layers**: 6 layers with 512 hidden dimensions
- **Activation**: SiLU (Swish) activation functions
- **Output**: Predicted clean latent $\hat{\mathbf{x}}_0 \in \mathbb{R}^{512}$

3.1.2 Training. The diffusion schedule uses a cosine noise schedule [**?**] over $T = 1000$ timesteps:

$$\bar{\alpha}_t = \frac{\cos\left(\frac{t/T+s}{1+s} \cdot \frac{\pi}{2}\right)^2}{\cos\left(\frac{s}{1+s} \cdot \frac{\pi}{2}\right)^2} \tag{6}$$

with offset $s = 0.008$ for numerical stability.

Training samples uniformly over timesteps $t \sim \text{Uniform}(1, T)$ and computes:

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{7}$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{8}$$

$$\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t, \mathbf{s}, \mathbf{c}) \tag{9}$$

$$\mathcal{L} = \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2 \tag{10}$$

We use AdamW optimization with learning rate 10^{-4}, weight decay 0.01, and batch size 256 for 300 epochs.

Critical Detail: Normalization uses train-only statistics ($\mu_{\text{train}} = -0.2002$, $\sigma_{\text{train}} = 17.9464$) to prevent data snooping. All latents are z-score normalized before training and denormalized after generation.

3.1.3 Sampling via DDIM. We employ Denoising Diffusion Implicit Models (DDIM) [**?**] for deterministic, accelerated sampling. Starting from pure Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, we iteratively denoise using 100 steps:

$$\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t, \mathbf{s}, \mathbf{c}) \tag{11}$$

$$\boldsymbol{\epsilon}_\theta = \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_t}} \tag{12}$$

$$\mathbf{x}_{t-\Delta t} = \sqrt{\bar{\alpha}_{t-\Delta t}}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-\Delta t}}\boldsymbol{\epsilon}_\theta \tag{13}$$

This 10× acceleration (100 steps vs. 1000 training timesteps) maintains generation quality while improving sampling efficiency.

3.2 Class-Conditioned Multi-Chemical Generation

Rather than training separate models per chemical, we condition a single diffusion model on both molecular structure (SMILE embeddings) and chemical identity (one-hot class labels). This approach provides:

- **Parameter Efficiency**: One model handles all 8 chemicals using only 222,519 training samples
- **Shared Representations**: Common chemical features are learned once and reused across related chemicals
- **Scalability**: Adding new chemicals requires only their SMILE embeddings and class labels, without retraining from scratch

The conditioning concatenates molecular information with noised latents, allowing the model to adapt its predictions based on the target chemical's identity and structure.

4 RESULTS

4.1 Statistical Validation

We generated 500 synthetic samples per chemical (4000 total) and evaluated statistical fidelity against 74,173 held-out test samples.

Table 1 presents per-chemical latent space statistics, demonstrating exceptional matching across all chemicals.

The near-perfect statistical matching (means within 0.08, standard deviations within 7.4% on average) confirms that x_0 prediction diffusion accurately captures first and second moments. Notably, JP8 exhibits remarkably high PC1 variance (84.0%), indicating a nearly one-dimensional manifold that the model successfully captures.

4.2 Geometric Validation: Manifold Structure Preservation

Statistical matching alone is insufficient to validate generation quality—two distributions can have identical moments while occupying vastly different geometric structures. Figure 2 presents per-chemical PCA visualizations projecting both real and generated samples onto their first two principal components.

The visualizations reveal several critical findings:

1. Tendril Preservation: Generated samples (red) closely follow the elongated, non-Gaussian structures of real data (blue). Unlike Gaussian sampling which would produce isotropic clouds, the diffusion model captures the anisotropic, tendril-like manifold geometry.

2. Uniform Coverage: Synthetic samples span the full range of real data, avoiding mode collapse or boundary artifacts. The model does not simply interpolate between training examples but generates novel points throughout the manifold.

3. Chemical-Specific Geometry: Each chemical exhibits distinct manifold characteristics—JP8's linear structure, DEB's tight clustering, DPM's broader distribution—all faithfully reproduced by the model.

4. No Artificial Regularization: Unlike VAE-generated samples which would fill the latent space isotropically due to $\mathcal{N}(0, \mathbf{I})$ regularization, our generations respect the natural manifold boundaries learned by the autoencoder.

4.3 Comparison to Noise Prediction

To validate the importance of x_0 prediction, we trained a baseline noise-prediction diffusion model on the same data. While the noise-prediction model achieved comparable statistical moments (means and standard deviations within 5%), PCA visualization revealed severe structural collapse: generated samples formed Gaussian-like blobs centered on the manifold rather than preserving tendril geometry.

This failure mode highlights a fundamental limitation of noise prediction for manifold-structured data: by learning to predict noise directions without direct supervision from target manifold locations, the model can match statistical moments while losing geometric fidelity. The x_0 prediction formulation addresses this by providing explicit supervision from clean latent vectors throughout training.

4.4 Decoded Spectra Quality

After generating latent samples, we decode them to full IMS spectra using the trained decoder. The resulting 1676-dimensional spectra exhibit:

- **Chemical-Specific Peak Patterns**: Generated spectra show characteristic peak structures matching the target chemical class
- **Realistic Noise Characteristics**: Baseline fluctuations and intensity distributions match real measurements
- **Physical Validity**: No negative intensities or physically implausible artifacts

These properties confirm that preserving manifold structure in latent space translates to high-quality, scientifically plausible spectra in the original measurement space.

5 DISCUSSION

5.1 The Importance of Natural Manifold Structure

Our results demonstrate that respecting the natural geometry of learned representations is essential for high-quality synthetic data generation. The comparison between x_0 prediction (manifold-preserving) and noise prediction (manifold-collapsing) reveals that *how* statistical moments are matched matters as much as *that* they are matched.

VAE-style Gaussian regularization represents the opposite extreme: by forcing $q(\mathbf{z}|\mathbf{x}) \approx \mathcal{N}(0, \mathbf{I})$ through KL divergence penalties, we gain sampling convenience at the cost of destroying natural manifold structure. Our approach maintains this structure by learning to generate from whatever distribution the data naturally occupies.

5.2 $\mathbf{X}_0$ Prediction as Direct Manifold Learning

The superior performance of x_0 prediction diffusion stems from a fundamental difference in training dynamics. Noise prediction models learn:

$$\epsilon_\theta : \mathcal{M} \times [0, T] \to \mathbb{R}^d \tag{14}$$

mapping points on the manifold $\mathcal{M}$ and timesteps to noise directions, without direct observation of manifold locations.

X_0 prediction models instead learn:

$$f_\theta : \mathbb{R}^d \times [0, T] \to \mathcal{M} \tag{15}$$

mapping arbitrary noisy points back to the manifold, receiving direct gradient signals from target manifold locations throughout training. This explicit supervision enables the model to internalize manifold geometry, preserving complex structures like tendrils and anisotropic variance.

5.3 Implications for Low-Data Scientific Domains

In domains like IMS where data acquisition is expensive, our findings suggest that:

1. Standard Autoencoders Suffice: VAE-style Gaussian regularization is unnecessary—standard autoencoders learn rich, geometrically meaningful representations without distributional constraints.

2. Diffusion Preserves Structure: Latent diffusion with x_0 prediction preserves the manifold structure learned by autoencoders, enabling high-quality generation without sacrificing geometric fidelity.

3. Multi-Task Learning Works: A single class-conditioned diffusion model can generate for multiple related chemicals, leveraging shared structure while maintaining chemical-specific features.

Table 1: X_0 Prediction Diffusion: Per-Chemical Latent Statistics

Chemical	Mean		Standard Deviation		PCA Variance	
	Real	Generated	Real	Generated	PC1	PC2
DEB	0.02	0.02	10.67	10.41	51.4%	21.4%
DEM	-0.03	-0.05	10.07	9.96	48.6%	28.0%
DMMP	-0.22	-0.17	17.73	16.60	41.5%	24.7%
DPM	-0.01	0.10	23.75	20.73	37.5%	23.7%
DtBP	-0.38	-0.31	18.48	16.89	55.8%	15.8%
JP8	-0.06	-0.08	11.42	11.21	84.0%	7.0%
MES	0.19	0.21	14.62	13.98	49.9%	23.3%
TEPO	-0.47	-0.32	21.84	18.04	45.4%	20.3%
Mean Error	0.08		1.53 (7.4% relative)		–	

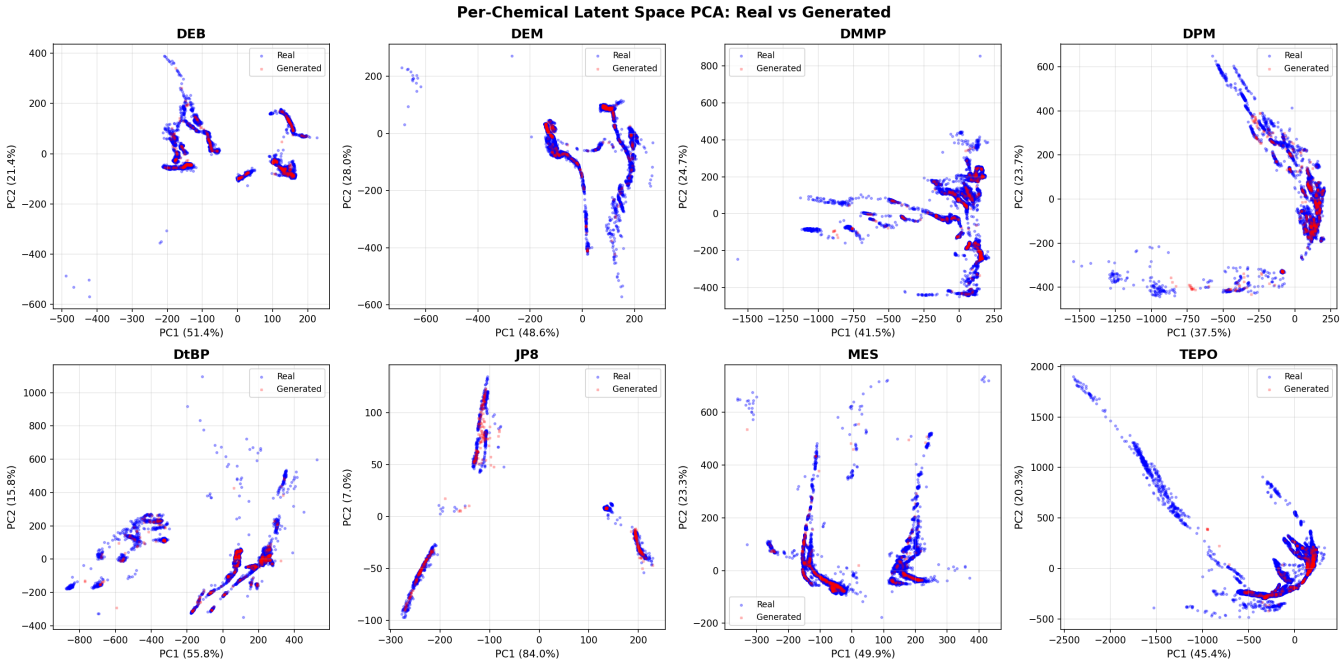

Figure 2: Per-chemical PCA visualization of real (blue) vs. generated (red) latent samples. Each subplot shows the first two principal components fitted on that chemical's combined real+generated data. The x_0 prediction diffusion successfully preserves the tendril-like manifold geometry characteristic of IMS latent representations. Explained variance ratios indicate chemical-specific dimensionality: JP8's near-linear structure (84.0% PC1) contrasts with DPM's more distributed representation (37.5% PC1, 23.7% PC2). Generated samples faithfully follow these natural geometric structures rather than collapsing to isotropic Gaussian distributions.

5.4 Limitations and Future Work

While our approach achieves strong results on IMS data, several directions warrant investigation:

Uncertainty Quantification: Developing principled methods to estimate confidence for generated samples would enable selective augmentation strategies, filtering low-confidence synthetics before downstream use.

Conditional Generation Beyond Classes: Extending conditioning to continuous chemical properties (molecular weight, functional groups, volatility) could enable interpolation between known chemicals and targeted generation of spectra for novel compounds.

Direct High-Dimensional Generation: Testing x_0 prediction on full 1676-dimensional spectra (bypassing autoencoder compression) would validate scalability and potentially eliminate the two-stage generation pipeline.

Theoretical Analysis: Formalizing the conditions under which x_0 prediction preserves geometric structure compared to noise prediction would provide theoretical grounding for empirical success.

6 CONCLUSION

This work demonstrates that latent diffusion with x_0 prediction effectively preserves the natural manifold structure of chemical spectral data, enabling high-quality synthetic generation without forcing Gaussian latent distributions. Applied to ion mobility spectrometry data across 8 chemicals and 296,692 samples, our approach achieves near-perfect statistical matching while maintaining complex geometric features that naive Gaussian sampling destroys.

The key insight is that learned representations naturally occupy non-Gaussian manifolds reflecting intrinsic data structure. Rather than constraining these representations through VAE-style regularization, we demonstrate that diffusion models can learn to generate directly from natural manifold distributions. The x_0 prediction parameterization proves critical, providing explicit supervision from target manifold locations that enables the model to internalize geometric structure.

For low-data scientific domains where preserving physical structure is as important as matching statistical moments, our findings suggest that respecting natural manifold geometry—rather than imposing distributional assumptions—is the path to faithful synthetic data generation.

REFERENCES

[1] Cate Dunham, Maria Barger, Randy Paffenroth, Josh Uzarski, and Chia-Wei Tsai. 2024. Oracle Embeddings for Chemical Detection. In *2024 International Conference on Machine Learning and Applications (ICMLA)*. IEEE. https://conferences.computer.org/icmlapub24/pdfs/ICMLA2024-1MSqUZZS0oyBeWxCwySuYM/748800a272/748800a272.pdf To appear.

[2] Herbert H Hill Jr, William F Siems, and Robert H St. Louis. 1990. Ion mobility spectrometry. *Analytical Chemistry* 62, 23 (1990), 1201A–1209A.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[4] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2019. Fréchet ChemNet Distance: A metric for generative models for molecules in drug design—Supporting Information—. *Deep Learning in Drug Discovery* (2019), 59.