Attention-Informed Diffusion Map (AIDM)

Kelli McCoy*

November 16, 2025

Construction of the global band weights. From the spectral Transformer, each validation sample $\eta_{d,i} \in \mathbb{R}^\nu$ produces an attention vector $a_i \in \mathbb{R}^\nu$ over the $\nu = 285$ bands (post-softmax, $a_{i,j} \geq 0$ and $\sum_j a_{i,j} \approx 1$). The *global* band-importance vector $w \in \mathbb{R}^\nu$ used below is the sample-average

$$w \;=\; \frac{1}{N}\sum_{i=1}^{N} a_i \;=\; \sum_{c=1}^{C} \frac{n_c}{N}\, A_c, \qquad A_c \;=\; \frac{1}{n_c}\sum_{i:\, y_i = c} a_i,$$

where C is the number of mineral classes under consideration (here the 95 Group-1 minerals), A_c is the per-class mean attention (row c of `band_attention_by_class.csv`), and n_c is the number of validation samples in class c. Now w coincides with the *sample-weighted* mixture of per-class means and matches the empirical class frequencies. For numerical robustness, apply a small lower threshold $w_j \leftarrow \max(w_j, \varepsilon_0)$ ($\varepsilon_0 > 0$).

Attention-informed diffusion map (AIDM): global weighting. Let $\{\eta_{d,1}, \ldots, \eta_{d,N}\} \subset \mathbb{R}^\nu$ be the training set in the data space (Ghanem, 2016 notation), with $\nu = 285$ spectral bands. Let $w \in \mathbb{R}^\nu$ be the global band-importance vector obtained from the Transformer (`band_attention_global.csv`), and let $\alpha \in [0.5, 1]$ be a contrast parameter (tempering large weights and lifting small ones). Construct the diagonal scaling

$$[S] \;=\; \operatorname{diag}\left(w^{\alpha/2}\right) \in \mathbb{M}_\nu \quad \Big(\text{elementwise power; apply } w_j \leftarrow \max(w_j, \varepsilon_0) \text{ first}\Big).$$

We can either (i) use the standard isotropic Gaussian kernel on preconditioned data (Option A below), or (ii) equivalently replace Eq. (21) with

$$k_\varepsilon^{\mathrm{att}}(\eta, \eta') \;=\; \exp\Big(-\tfrac{1}{4\varepsilon}\,\big\|[S](\eta - \eta')\big\|^2\Big) \;=\; \exp\Big(-\tfrac{1}{4\varepsilon}\,\big\|(\eta - \eta') \odot \sqrt{w^\alpha}\big\|^2\Big),$$

which preserves symmetry and positive semidefiniteness while inducing band-aware geometry.

End-to-end PLoM workflow with attention (Option A: precondition → PCA/whitening → standard kernel).

1. **Robust scaling in band space.** Given raw spectra $x_i \in \mathbb{R}^\nu$ ($\nu = 285$), apply robust centering/scaling (median/IQR) to obtain $\tilde{x}_i$ and stack $[\tilde{X}] = [\tilde{x}_1 \;\cdots\; \tilde{x}_N]$.

*Notation and workflow follow Ghanem (2016).

2. **Attention preconditioning in band space.** Let $w \in \mathbb{R}^\nu$ and $\alpha \in [0.5, 1]$. Define

$$[S] \;=\; \mathrm{diag}\big((w^\alpha)^{1/2}\big) \in \mathbb{M}_\nu, \qquad \tilde{x}_i^{(\mathrm{att})} \;=\; [S]\tilde{x}_i,$$

and stack $[\tilde{X}^{(\mathrm{att})}] = [\tilde{x}_1^{(\mathrm{att})} \;\cdots\; \tilde{x}_N^{(\mathrm{att})}]$.

3. **PCA/whitening to data space.** Compute the empirical covariance of $[\tilde{X}^{(\mathrm{att})}]$, perform the eigendecomposition $C = \Phi\Lambda\Phi^\top$, keep r components (here $r = \nu$), and define

$$\eta_{d,i} \;=\; \Lambda_r^{-1/2}\,\Phi_r^\top\,\tilde{x}_i^{(\mathrm{att})} \;\in\; \mathbb{R}^r, \qquad [\eta_d] = [\eta_{d,1} \;\cdots\; \eta_{d,N}] \in \mathbb{M}_{r,N}.$$

4. **Kernel and Markov matrix (standard Eq. (21) kernel).** Form $[K] \in \mathbb{M}_N$ with entries

$$[K]_{ij} \;=\; k_\varepsilon(\eta_{d,i}, \eta_{d,j}) \;=\; \exp\Big(-\tfrac{1}{4\varepsilon}\,\|\eta_{d,i} - \eta_{d,j}\|^2\Big).$$

Let $[b] = \mathrm{diag}(\sum_j [K]_{ij})$ and build the symmetric normalization

$$[P_S] \;=\; [b]^{-1/2}[K][b]^{-1/2}.$$

5. **Diffusion-maps basis.** Compute leading eigenpairs $[P_S]\phi_\ell = \lambda_\ell\phi_\ell$; set $\psi_\ell = [b]^{-1/2}\phi_\ell$ with $[\psi]^\top[b][\psi] = [I]$ and define $g_\ell = \lambda_\ell^\kappa\psi_\ell$; collect $[g] = [g_1 \;\cdots\; g_m]$.

6. **Reduced representation and sampling (PLoM).** Project to $[H] \in \mathbb{M}_{m,N}$ as in Eqs. (30)–(32), choose m via the covariance error criterion (Eqs. (33)–(36)), and construct the reduced ISDE in $\mathbb{R}^m$ (Eqs. (37)–(41)). Integrate (e.g. Störmer–Verlet) to generate $[Z(\rho)]$, then decode to data space and invert PCA and preconditioning:

$$[\eta_s] = [Z(\rho)][g]^\top, \quad \tilde{x}_s^{(\mathrm{att})} = \Phi_r\,\Lambda_r^{1/2}\,\eta_s, \quad \tilde{x}_s = [S]^{-1}\tilde{x}_s^{(\mathrm{att})}, \quad x_s = \mathrm{unscale}(\tilde{x}_s).$$

This yields synthetic spectra in the original band units.

End-to-end PLoM workflow with attention (Option B: PCA/whitening → anisotropic kernel in whitened space).

1. **Robust scaling in band space.** Given raw spectra $x_i \in \mathbb{R}^\nu$ ($\nu = 285$), apply robust centering/scaling (median/IQR) to obtain $\tilde{x}_i$ and stack $[\tilde{X}] = [\tilde{x}_1 \;\cdots\; \tilde{x}_N]$.

2. **PCA/whitening to data space.** Compute the empirical covariance of $[\tilde{X}]$, eigendecompose $C = \Phi\Lambda\Phi^\top$, keep r components (here $r = \nu$), and define

$$\eta_{d,i} \;=\; \Lambda_r^{-1/2}\,\Phi_r^\top\,\tilde{x}_i \;\in\; \mathbb{R}^r, \qquad [\eta_d] = [\eta_{d,1} \;\cdots\; \eta_{d,N}] \in \mathbb{M}_{r,N}.$$

3. **Attention-informed anisotropic kernel (replace Eq. (21) with correct metric mapping).** Let $w \in \mathbb{R}^\nu$ be the (lower-thresholded) global band-importance vector and $\alpha \in [0.5, 1]$. Define the band-space metric

$$M_x \;=\; \mathrm{diag}(w^\alpha) \in \mathbb{M}_\nu.$$

Map it to the whitened coordinates via

$$M_\eta \;=\; \Lambda_r^{1/2}\,\Phi_r^\top\,M_x\,\Phi_r\,\Lambda_r^{1/2} \;\in \mathbb{M}_r,$$

so that, for any pair (η, η'), the intended weighted squared distance in band space $\|S_x(\tilde{x}-\tilde{x}')\|^2$ with $S_x = \mathrm{diag}((w^\alpha)^{1/2})$ equals $(\eta-\eta')^\top M_\eta(\eta-\eta')$ in the whitened space. Construct $[K] \in \mathbb{M}_N$ with entries

$$[K]_{ij} \;=\; k_\varepsilon^{\mathrm{att}}(\eta_{d,i}, \eta_{d,j}) \;=\; \exp\Big(-\tfrac{1}{4\varepsilon}\,(\eta_{d,i}-\eta_{d,j})^\top M_\eta\,(\eta_{d,i}-\eta_{d,j})\Big).$$

Let $[b] = \mathrm{diag}(\sum_j [K]_{ij})$ and build the symmetric normalization

$$[P_S] \;=\; [b]^{-1/2}[K][b]^{-1/2}.$$

Practical note: factor $M_\eta = R^\top R$ (e.g., Cholesky) and evaluate distances as $\|R(\eta_{d,i}-\eta_{d,j})\|^2$.

4. **Diffusion-maps basis.** Compute leading eigenpairs $[P_S]\phi_\ell = \lambda_\ell \phi_\ell$; set $\psi_\ell = [b]^{-1/2}\phi_\ell$ with $[\psi]^\top[b][\psi] = [I]$ and define $g_\ell = \lambda_\ell^\kappa \psi_\ell$; collect $[g] = [g_1 \;\cdots\; g_m]$.

5. **Reduced representation and sampling (PLoM).** Project to $[H] \in \mathbb{M}_{m,N}$ as in Eqs. (30)–(32), choose m via the covariance error criterion (Eqs. (33)–(36)), and construct the reduced ISDE in $\mathbb{R}^m$ (Eqs. (37)–(41)). Integrate (e.g. Störmer–Verlet) to generate $[Z(\rho)]$, then decode:

$$[\eta_s] = [Z(\rho)][g]^\top, \qquad \tilde{x}_s = \Phi_r\,\Lambda_r^{1/2}\,\eta_s, \qquad x_s = \mathrm{unscale}(\tilde{x}_s).$$

Note: unlike Option A, there is no $[S]^{-1}$ step at decoding, because M_η was used only in the kernel.

Remark (Option A vs. Option B). Both variants inject attention-driven anisotropy consistently with Ghanem (2016): Option A preconditions the band space before whitening and then uses the standard Eq. (21) kernel; Option B whitens first and uses an anisotropic kernel with the *correctly mapped* metric $M_\eta = \Lambda_r^{1/2}\Phi_r^\top\,\mathrm{diag}(w^\alpha)\Phi_r\Lambda_r^{1/2}$. They are closely related; in practice, Option A keeps the interpretation "per-band scaling" most transparent for decoding, while Option B leaves the forward pipeline unchanged and localizes the change to the kernel construction.

Why AIDM is appropriate. (i) The attention-derived scaling encodes *per-band* relevance while preserving the Gaussian kernel assumptions used in Ghanem (2016).
(ii) Reweighting by w^α increases (decreases) the contribution of highly (weakly) informative bands when constructing the diffusion operator, aligning the manifold with task-relevant spectral structure.
(iii) The exponent α tempers peaky attentions, and the small lower threshold on w prevents numerical collapse.
(iv) All subsequent PLoM steps (normalization, spectral decomposition, reduced ISDE, decoding) remain unchanged—only the preconditioning (Option A) or the kernel metric (Option B) is modified.

Why we do not use the class attention matrix (95×285) in Eq. (21). If we insert the per class matrix $[S]$ from `band_attention_by_class.csv` directly into Eq. (21), the distance would depend on class row. For a pair from different classes, there is no single obvious choice (A's row? B's row? their average?), which breaks the idea of one consistent distance for all pairs. It can also break *symmetry* (distance A→B equals B→A), which diffusion maps rely on. Intuitively, this would make the embedding learn in "*class space*" (it encodes label choices) instead of a single smooth manifold shared by all spectra, which is not what we want for class agnostic generation. It also complicates projecting new, unlabeled points, because you would first have to guess which class weights to use. A global w yields one PSD kernel and one smooth geometry for the whole dataset.

Implementation notes. Choose ε from the reweighted distances (e.g. median kNN squared distance in the preconditioned space for Option A, or in the M_η-induced metric for Option B); sweep $\alpha \in \{0.5, 0.7, 1.0\}$ and select by a downstream criterion (held-out likelihood, reconstruction error, or fidelity of generated spectra). If class balance is desired, w may be taken as a class-balanced mean of per-class attentions; otherwise use the dataset's natural (sample-weighted) mean. A principled alternative would mix metrics with posteriors, e.g. $\sum_c p(c \mid x)\,\mathrm{diag}(A_c^\alpha)$, but this injects estimation error (especially for rare classes) and substantial per-pair computational cost. In contrast, the global weight $w = \frac{1}{N}\sum_i a_i$ yields a single PSD, well-behaved kernel that avoids amplifying noise in small classes while still biasing geometry toward task-relevant bands. Class-conditioned AIDM (per-class PLoM or posterior-weighted mixtures) is viable future work but is out of scope for this class-agnostic pipeline.