NUViT: Vision Transformers for Fourier-Space Data

Anonymous authors
Paper under double-blind review

Abstract

Vision Transformers (ViTs) have demonstrated strong performance in many computer vision tasks but do not naturally extend beyond the natural image domain. One important example is Fourier-space data, which is used to represent a rich and diverse set of modalities, including medical imaging and microscopy. Standard ViT tokenization partitions an image into square patches on a Cartesian grid, which poorly fit the frequency and phase structure of Fourier-space data. We introduce **N**on-**U**niform **ViT**, a simple, general-purpose vision transformer that redesigns image tokenization, data representation, and positional encoding to respect this structure. Experiments on Fourier-domain image classification, 2D crystal lattice identification, MRI annotation, and cryo-EM protein heterogeneity analysis show that NUViT consistently captures the structure of Fourier data better than standard architectures.

1 Introduction

Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have become a widely-used architecture across a broad range of computer vision tasks. Motivated by their strong performance on natural images, there is growing interest in extending ViT-style models to other data modalities. However, Fourier-domain or Fourier-transformed representations are not inherently compatible with standard ViT designs, even though Fourier data is common in applications such as 3D sensing, medical imaging, and microscopy.

General-purpose architectures for Fourier-space data remain underexplored, and existing frequency-space ViTs are largely domain-specific (Buchholz & Jug, 2022; Duan et al., 2022; Alshaabi et al., 2025). Such a general transformer architecture would be valuable for 3D reconstruction methods that exploit the Fourier Slice Theorem (Bracewell, 1956) to avoid costly real-space integration, a standard practice in cryo-electron microscopy (cryo-EM), and in experimental modalities where sensor data is collected natively in frequency space, such as magnetic resonance imaging (MRI). It would also benefit hybrid methods that incorporate Fourier transforms into their processing pipeline, including a growing body of work on frequency-space token mixing as an alternative to attention (Li et al., 2020; Rao et al., 2021; Guibas et al., 2021; Lee-Thorp et al., 2022; Nguyen et al., 2022; Lingsch et al., 2023; Long et al., 2025).

Natural phenomena display a wide range of behavior in Fourier space. For example, speech and audio waveforms, natural images and video (Field, 1987; Van der Schaaf & van Hateren, 1996), 3D sensor measurements, and medical imaging such as MRI (Lustig et al., 2007) all follow a $1/f^{\alpha}$ spectral energy decay with spatial frequency f (Figure 1). This Fourier power law is not the only common example of frequency-dependent behavior: also common are resonance- and dispersion-based phenomena. Resonance produces sharp peaks in a signal's Fourier spectrum, which is observed in atomic spectra and crystal structures. Dispersion describes how a wave's propagation speed varies with frequency, causing phenomena such as a prism splitting white light into its constituent colors.

In this paper, we design a general-purpose transformer architecture for Fourier-space data that better captures these patterns. We observe that the standard ViT tokenization strategy that partitions images into square patches on a uniform Cartesian grid is ill-suited for capturing the wide variety of structures present in Fourier-space data. We argue that the non-uniform frequency and phase coordinates of the Cartesian grid hamper the encoder's ability to understand Fourier space structure, with uniform frequency and phase coordinates providing better inductive bias. Furthermore, in the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

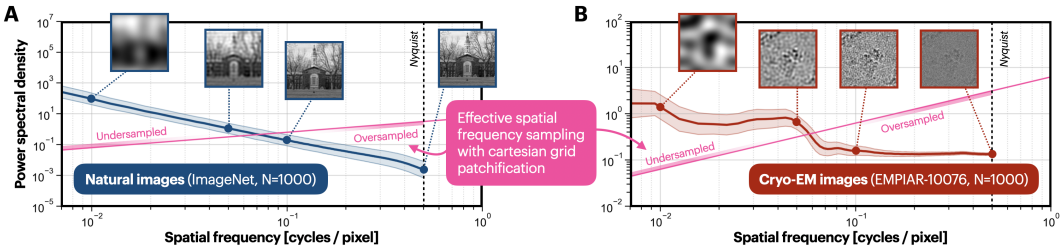

Figure 1: **Power spectral density of natural and cryo-EM images.** Log-log plots of radially-averaged power versus spatial frequency for **(A)** natural images (Deng et al., 2009) and **(B)** cryo-EM images (Davis et al., 2016) demonstrate a characteristic Fourier power-law decay. Shaded regions indicate one standard deviation. Pink lines show the spatial frequency weighting induced by Cartesian grid patching in standard ViTs, which undersamples informative low-frequency components. Insets show low-pass filtered images at different spatial frequencies. In both natural and cryo-EM images, lower spatial frequencies capture broad semantic information. Note that cryo-EM images inherently contain imaging aberrations (CTF), producing characteristic power oscillations.

common case that a signal follows the Fourier power law, traditional ViT tokenization oversamples high spatial frequencies and undersamples the information-dense low frequency components.

Based on this insight, we propose **N**on-**U**niform **Vi**sion **T**ransformers (NUViTs), a general-purpose transformer architecture for Fourier data. NUViTs sample the Fourier transform on a non-uniform grid using the non-uniform Fast Fourier Transform (NUFFT) (Fessler & Sutton, 2003), producing a grid that has uniform frequency and phase coordinates. To better capture phase information, we represent complex Fourier coefficients with amplitudes and phase components. Absolute positional encoding and channel encoding are applied, and the tokens are then processed by a typical transformer network (Vaswani et al., 2017).

We evaluate NUViT's effectiveness on a variety of experiments. In the more common case of data that follows the Fourier power law, we evaluate Fourier-space natural image classification, MRI disease classification, and cryo-EM conformation classification and 3D reconstruction. To evaluate NUViT for the opposite case when data that does not conform to the power law, we test 2D crystal lattice identification, for which spectral energy is concentrated at higher frequencies (Torquato & Stillinger, 2003; Torquato, 2018). Across these settings NUViT consistently outperforms standard ViT and MLP baselines, with clear improvements on tasks requiring fine-grained discrimination under high noise. Our experiments verify the effectiveness of NUViT over different data domains and Fourier structures, showing the importance of non-uniform tokenization for capturing Fourier space information.

2 Methods

2.1 Motivation

Data in Fourier space displays a wide variety of structures which are ill-captured by traditional ViT tokenization. As a motivating and important example, the variance of many natural signals is heavily concentrated in low spatial frequencies (Figure 1). With the standard discrete Fourier transform (DFT) convention that places the origin at the center of the Fourier image, most of the data variation is concentrated around the center. Additionally, the variance of many natural signals scales as a power law in frequency, such that low-frequency variances are orders of magnitude greater than high-frequency variances (Field, 1987; Van der Schaaf & van Hateren, 1996).

Standard ViT tokenization partitions an image into square patches on a uniform Cartesian grid. In Fourier space, this is poorly justified, since each square patch spans a mix of spatial frequencies rather than aligning with concentric frequency bands from the spectrum. Moreover, for many data modalities, this effectively undersamples the information-dense low-frequency components of a Fourier-space signal and oversamples the high-frequency periphery. NUViT addresses both issues by tokenizing on a non-uniform polar grid that allocates more samples to the low-frequency center (Figure 2).

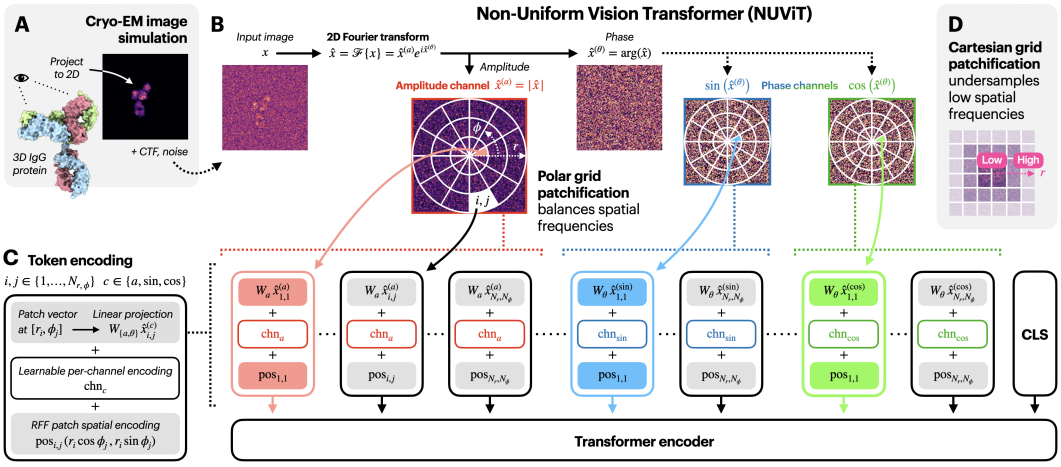

Figure 2: **NUViT architecture. A)** The cryo-EM image formation model is used to generate synthetic benchmark datasets. **B)** NUViT operates on Fourier-transformed 2D images using the NUFFT, where the data is represented in polar coordinates with amplitude a and phase θ (sin and cos) channels. Each channel map is partitioned into radial-angular patches and tokenized. **C)** Patches are encoded with channel and spatial position features. **D)** Traditional Cartesian grid patches undersample the high-variance low-frequency center of the Fourier-transformed data.

2.2 NUViT

2.2.1 Tokenization

To address the sampling mismatch described above, we represent the Fourier data on a polar grid with N_r radial shells and N_ϕ angular samples per shell, yielding coordinates $\{(r_i, \phi_j) : r_i = i/N_r,\ \phi_j = \pi j/N_\phi\}$ for $i = 1, \ldots, N_r$ and $j = 1, \ldots, N_\phi$. We use the NUFFT to evaluate the Fourier transform at these non-uniform coordinates, producing a $N_r \times N_\phi$ grid in polar space. This grid is treated as a regular 2D image and partitioned into square patches of size $P \times P$ using the standard ViT patching procedure. The total number of tokens is thus $N = (N_r/P) \times (N_\phi/P)$.

2.2.2 Data Representation

Fourier-space data $\hat{x}$ lacks a canonical representation for deep learning. Fourier data $\hat{x}$ can be represented naively as complex numbers $\hat{x} \in \mathbb{C}^D$, as two channels of real and imaginary data $\hat{x} \in \mathbb{R}^{2D}$, or as amplitudes $a = |\hat{x}| \in \mathbb{R}^D$ and phases $\theta = \arg(\hat{x}) \in [0, 2\pi)$. Since amplitudes follow a power law, log-amplitudes are also reasonable. If x is real-valued, one can alternatively use the Hartley transform, which produces a real-valued output with no information loss.

In NUViT, we use an amplitude-and-phase representation. For each complex value $\hat{x}_j = a_j e^{i\theta_j}$, we construct three channels: the raw amplitude and a circular encoding of the phase. Concretely, the representation is:

$$\mathbf{r}(\hat{x}) = \big[\, [a_1, \ldots, a_D],\ [\cos(\theta_1), \ldots, \cos(\theta_D)],\ [\sin(\theta_1), \ldots, \sin(\theta_D)] \,\big] \tag{1}$$

On a polar grid, each radial shell corresponds to a band of frequencies at a fixed magnitude. The amplitude-and-phase representation aligns naturally with this radial structure: the amplitude channel captures the power-law decay across shells, while the phase channels encode the within-shell variation.

2.2.3 Positional Encoding

NUViT uses two types of additive positional encoding: a spatial positional encoding that denotes the 2D coordinates of each patch, and a per-channel positional encoding that identifies the amplitude and phase channels.

Spatial positional encoding We encode the 2D center coordinate $\mathbf{x}_{ij} = (r_i, \phi_j)$ of each patch using random Fourier features (RFF) (Tancik et al., 2020):

$$\gamma_{\text{spatial}}(\mathbf{x}) = [\cos(2\pi\mathbf{B}\mathbf{x}),\ \sin(2\pi\mathbf{B}\mathbf{x})]^T, \quad \mathbf{B} \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{d_{\text{pe}} \times 2}$ is a matrix of random frequencies sampled once at initialization and held fixed during training, and σ is a hyperparameter controlling the frequency bandwidth. We use RFF rather than the learned 1D embeddings of a standard ViT (Dosovitskiy et al., 2020) because the polar grid coordinates are inherently 2D and continuous-valued, making them a poor fit for a lookup table of discrete learned embeddings. RFF also allows the encoding to generalize to different grid resolutions at test time.

Channel positional encoding Our data representation produces three channels (amplitude, $\cos\theta$, $\sin\theta$) that have very different value ranges and semantic meanings. To preserve this channel-specific information, we follow Bao et al. (2023) and associate each channel $c \in \{a, \{\cos, \sin\}\}$ with a learnable embedding $\mathbf{e}_c \in \mathbb{R}^D$, which is added to the token embedding after projection.

2.2.4 Forward Pass

Given an input image x, we first compute its 2D Fourier transform and evaluate it on the polar grid using the NUFFT, then construct the three-channel representation $\mathbf{r}(\hat{x})$ (Equation 1). The resulting $N_r \times N_\phi \times 3$ tensor is partitioned into $N = (N_r/P) \times (N_\phi/P)$ patches. Let $\mathbf{p}_k^{(c)}$ denote the k-th patch of channel c, flattened to a vector in $\mathbb{R}^{P^2}$.

Each patch is linearly projected to the transformer hidden dimension D, and both spatial and channel positional encodings are added:

$$\mathbf{z}_k^{(c)} = \mathbf{W}_{\text{proj}}^{(c)}\,\mathbf{p}_k^{(c)} + \mathbf{b} + \gamma_{\text{spatial}}(\mathbf{x}_k) + \mathbf{e}_c \tag{3}$$

where $\mathbf{W}_{\text{proj}}^{(c)} \in \mathbb{R}^{D \times P^2}$ is a per-channel linear projection, $\mathbf{b}$ is a bias term, $\gamma_{\text{spatial}}(\mathbf{x}_k)$ is the RFF encoding of the patch center coordinate (Equation 2), and $\mathbf{e}_c$ is the learnable channel embedding.

A learnable `[CLS]` token $\mathbf{z}_{\text{cls}} \in \mathbb{R}^D$ is prepended to the sequence. The full input sequence is:

$$\mathbf{z}_0 = [\mathbf{z}_{\text{cls}};\ \mathbf{z}_1^{(a)}, \ldots, \mathbf{z}_N^{(a)},\ \mathbf{z}_1^{(\theta)}, \ldots, \mathbf{z}_N^{(\theta)}] \tag{4}$$

This sequence of $2N+1$ tokens is then processed by a standard transformer encoder (Vaswani et al., 2017) consisting of L layers of alternating multi-head self-attention (MSA) and MLP blocks with pre-norm layer normalization (Dosovitskiy et al., 2020):

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1, \ldots, L \tag{5}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell \tag{6}$$

The output representation is taken from the `[CLS]` token: $\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$. For classification, a linear head is applied to $\mathbf{y}$. For cryo-EM reconstruction, $\mathbf{y}$ replaces the MLP encoder in cryoDRGN (Zhong et al., 2020; 2021a) to produce a latent embedding per input image, which models the image class and conditions a 3D neural field for reconstruction.

Inductive bias Compared to a standard ViT, NUViT introduces additional inductive biases. An important one for machine learning is the assumption that the low spatial frequencies of the input data carry information content that is better sampled with a polar grid. This assumption is appropriate whenever the data obeys the Fourier power law, as is the case for natural images, MRI images, cryo-EM images, and many other modalities. The use of a polar grid also introduces a bias towards isotropy, which is common in many types of Fourier data, including Fourier power law data. The polar grid also respects the *spectral autoregression* structure of Fourier data: if we group pixels into Fourier shells $[g_1, g_2, \ldots]$ by distance from the origin, the subsequences $[g_1], [g_1, g_2], \ldots$ encode the image at monotonically increasing resolution. Each ring of patches on the polar grid corresponds naturally to a frequency shell, so the model can attend to coarse-to-fine information in a structured way.

Table 1: **Natural image classification in Fourier space**. Top-1 test accuracy on CIFAR-10/100 and their corrupted counterparts CIFAR-10-C/100-C, evaluated using models trained only on the corresponding clean datasets.

	CIFAR			
Method	**10**	**10-C**	**100**	**100-C**
FIT	23.35	16.55	5.21	2.76
ViT-S (Hartley)	53.76	49.84	23.25	**20.13**
NUViT-S (Ours)	**68.63**	**52.99**	**27.20**	18.69

Table 2: **Synthetic 2D crystal lattice classification.** Top-1 test accuracy. All encoders achieve 100% train accuracy.

Method	**Crystal**
Random chance	25.0
MLP (real)	39.8
MLP (Hartley)	33.3
ViT-S (real)	66.0
ViT-S (Hartley)	58.7
NUViT (Ours)	**97.5**

Table 3: **NUViT ablation study**. Key architectural components of NUViT are evaluated by top-1 classification accuracy on Fourier-transformed natural images.

Ablation	**CIFAR-10**	**CIFAR-100**
Cartesian grid	43.49	14.95
Real + Imaginary	49.73	18.79
Hartley	46.04	18.12
No channel positional encoding	66.04	**27.33**
NUViT (Ours)	**68.63**	27.20

3 RESULTS

We evaluate NUViT in five settings. To validate our design choices against simpler baselines, we test (1) Fourier-domain natural image classification and (2) 2D crystal lattice identification. To demonstrate NUViT's utility for different imaging modalities, we test (3) supervised MRI classification, (4) supervised cryo-EM image classification, and (5) heterogeneous cryo-EM reconstruction, which tests whether NUViT can serve as an effective encoder in a downstream generative model.

3.1 NATURAL IMAGE CLASSIFICATION

Setup We first use Fourier-domain classification of natural images as a simple setting to evaluate NUViT and ablate its design choices. Input RGB images are converted to grayscale and Fourier transformed. We evaluate on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) using top-1 accuracy. We measure robustness by performing test-time inference on CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019). As Fourier-space baselines, we use a FIT (Buchholz & Jug, 2022)-derived model and a ViT-S (Dosovitskiy et al., 2020) with a Hartley transform representation.

Results NUViT achieves the best overall performance on both datasets (Table 1). FIT performs poorly, likely due to its one-token-per-pixel design, which produces extremely long sequences even for modestly sized images, leading to slow convergence. We additionally ablate NUViT's components on the CIFAR task (Table 3). The polar grid is the single most important design choice (-25.14 points on CIFAR-10), followed by the data representation, where the amplitude and phase representation is most effective, though we note it requires more parameters. Using a channel positional encoding has a small impact on performance.

3.2 2D CRYSTAL LATTICE CLASSIFICATION

Setup The domains discussed so far follow the Fourier power law with a negative exponent. To test whether NUViT generalizes to the opposite case, we benchmark on hyperuniform patterns, whose power spectra rise with frequency because they suppress fluctuations at large spatial scales (Torquato & Stillinger, 2003) — as observed in cosmological superhomogeneity (Gabrielli et al., 2002) and crystal lattices. While perfect crystals do not follow a power law, imperfect crystals with random perturbations approximately follow $S(k) \propto k^{\alpha}$ with $\alpha = 2$ (Gabrielli, 2004; Kim & Torquato, 2018; Klatt et al., 2020). We generate a dataset of such imperfect crystals with four types of symmetry, randomly displacing each lattice point (Figure S1). As Fourier-space baselines, we use an

Table 4: **MRI classification.** Test QWK (↑) for different image representations.

Method	Test QWK (↑)
ViT (real)	0.266
ViT (Hartley)	0.232
NUViT (Ours)	**0.348**

MLP (Zhong et al., 2020) and a ViT-S (Dosovitskiy et al., 2020), each with real and Hartley transform representations.

Results In Table 2, we find that NUViT substantially outperforms alternative encoders on 2D crystal lattice classification. Because every encoder reaches 100% train accuracy, this gap reflects generalization rather than capacity. Each lattice class can be uniquely identified by its Bragg peaks, which are a set of peaks in the power spectrum of each crystal. However, identification is complicated by random perturbations, which dampen the peaks by the Debye-Waller factor $e^{-\sigma^2 k^2}$, where σ is the noise and k is the frequency (Ashcroft & Mermin, 1976). Therefore, Bragg peaks at higher frequency are more dampened than those at lower frequency. The higher test accuracy of NUViT compared to baseline methods provides evidence that our polar grid tokenization better matches the geometry of the spectrum than standard ViTs, capturing subtle peak information.

3.3 MRI Classification

Setup To evaluate NUViT on MRI data, we use the OASIS Alzheimer's Detection Multi-Class dataset (Mohanty, 2025), which contains 2D brain MRI images labeled by Alzheimer's disease severity. To increase realism, we add MRI-specific corruptions and noise at a realistic signal-to-noise ratio (SNR), adapted from fastMRI (Zbontar et al., 2018). The task is challenging because it requires fine-grained distinctions between adjacent severity levels. Because these labels are ordinal, each encoder is paired with a CORAL head and trained with a class-weighted CORAL loss (Cao et al., 2020). We evaluate models with quadratic weighted kappa (QWK), which measures agreement with the ground truth relative to chance while penalizing errors in proportion to the squared distance between predicted and true labels.

Results Under realistic MRI conditions, NUViT achieves a meaningful gain in QWK over alternative encoders (Table 4). Distinguishing adjacent severity levels requires detecting subtle anatomical differences in noisy images. NUViT's improvement provides evidence that polar grid tokenization is better at extracting information from the low signal-to-noise ratio (SNR) high frequency bands. Because MRI data is acquired directly in the frequency domain, these results suggest that NUViT could operate on raw MRI data without first reconstructing an image, which we leave to future work.

3.4 Cryo-EM Image Classification

Setup We next evaluate NUViT on cryo-EM heterogeneity classification, where the goal is to classify noisy 2D projection images of dynamic proteins by their underlying 3D state. Cryo-EM images are characterized by extremely low signal-to-noise ratios and are commonly represented in the Fourier domain, making this a natural application for NUViT. We compare against two baselines corresponding to encoders used in cryo-EM reconstruction: the cryoDRGN (Zhong et al., 2020; 2021a) MLP encoder and a ViT-S (Dosovitskiy et al., 2020) encoder, each tested with both real-space and Hartley transform representations. Hartley transform values are the default for many cryo-EM models (Zhong et al., 2020; 2021a;b; Levy et al., 2022; Gu et al., 2026). We do not test FIT (Buchholz & Jug, 2022) due to its computational cost and slow convergence.

We evaluate on one experimental and four synthetic datasets with a 90/10 train-test split. The synthetic datasets are from CryoBench (Jeon et al., 2024): `IgG-1D` and `IgG-RL` contain 100 conformational states of the IgG antibody, while `Ribosembly` and `Tomotwin-100` contain 16 and 100 distinct biomolecular species, respectively. The experimental dataset is `EMPIAR-10076` (Davis et al., 2016), containing 5 major states of the assembling 50S ribosome.

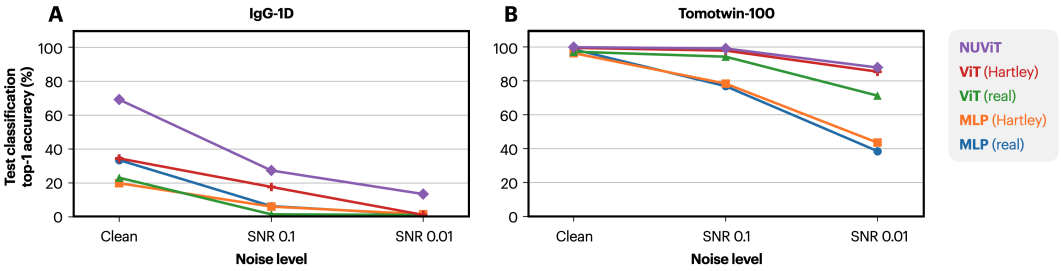

Figure 3: **Cryo-EM image classification noise robustness**. Comparison of test classification top-1 accuracy at different noise levels for simulated datasets **(A)** `IgG-1D` and **(B)** `Tomotwin-100`.

Results NUViT outperforms all Fourier-space baselines on heterogeneity classification (Table 5). The improvements are largest on tasks requiring fine-grained discrimination. On `IgG-1D` and `IgG-RL`, which require distinguishing small differences in the angle of a flexible linker region, all baseline encoders achieve near-random accuracy (1–8% on 100 classes), whereas NUViT reaches 13.30% and 66.79%, respectively. While 13.30% on `IgG-1D` is still low in absolute terms, it represents a qualitative shift: NUViT is the only encoder to extract meaningful conformational signal from these images. On `Ribosembly`, which requires distinguishing more complete from less complete ribosome assemblies, NUViT achieves 80.62% compared to 72.56% for the best baseline. On `Tomotwin-100`, a coarser-grained compositional classification task, NUViT achieves 88.86% compared to 84.32% for ViT-S (Hartley), while all other methods performed worse. On the experimental dataset `EMPIAR-10076`, NUViT achieves a small gain.

We also examine robustness to noise (Figure 3). High noise heavily degrades the performance of existing encoders, with the MLP encoder failing almost entirely at a signal-to-noise ratio (SNR) of 0.01. NUViT is substantially more robust to noise, likely because the polar grid allocates more capacity to low-frequency components, which have higher SNR than high-frequency components.

3.5 Cryo-EM 3D Reconstruction

The goal of heterogeneous cryo-EM reconstruction is to recover multiple 3D structures from a dataset of noisy 2D projection images. Variational autoencoder (Kingma, 2013) (VAE)-based methods such as cryoDRGN (Zhong et al., 2020; 2021a) use an encoder to map each Fourier-space 2D image to a latent code, which a decoder uses to generate the corresponding Fourier-space 3D volume.

Setup We next evaluate whether NUViT can serve as an effective encoder in a downstream generative model for cryo-EM reconstruction. Specifically, we replace the MLP encoder in the cryoDRGN (Zhong et al., 2020; 2021a) model with either a ViT (Hartley) encoder or a NUViT encoder, keeping the neural field decoder and all other components fixed. Beyond standard reconstruction quality, we test whether NUViT can enable amortized inference in cryo-EM. We note that existing encoders in cryo-EM reconstruction do not generalize beyond their training set: they are vulnerable

Table 5: **Cryo-EM image classification.** This supervised classification task assigns labels to heterogeneous protein images. The datasets contain projection images of a compositional mixture of distinct structures (`Tomotwin-100`, `Ribosembly`, `EMPIAR-10076`) or conformational variations from the shape deformation of a single structure (`IgG-1D`, `IgG-RL`) (Jeon et al., 2024; Davis et al., 2016). `EMPIAR-10076` is a published experimental dataset, while the others are simulated. Performance is measured by top-1 accuracy. Results are reported as `train / test accuracy`.

Method	IgG-1D	IgG-RL	Ribosembly	Tomotwin-100	EMPIAR-10076
Random chance	1.00	1.00	6.25	1.00	6.25
MLP (real)	99.33 / 1.25	99.60 / 2.10	99.61 / 42.40	99.50 / 42.12	99.65 / 39.94
MLP (Hartley)	99.10 / 1.43	99.06 / 2.61	99.58 / 44.58	99.18 / 32.71	99.77 / 38.55
ViT-S (real)	100.00 / 1.05	100.00 / 1.05	97.55 / 45.49	99.73 / 75.70	98.98 / 46.47
ViT-S (Hartley)	1.66 / 1.41	57.82 / 8.11	98.01 / 72.56	98.84 / 84.32	98.73 / 50.89
NUViT (Ours)	36.54 / **13.30**	94.77 / **66.79**	98.80 / **80.62**	99.36 / **88.86**	96.84 / **51.52**

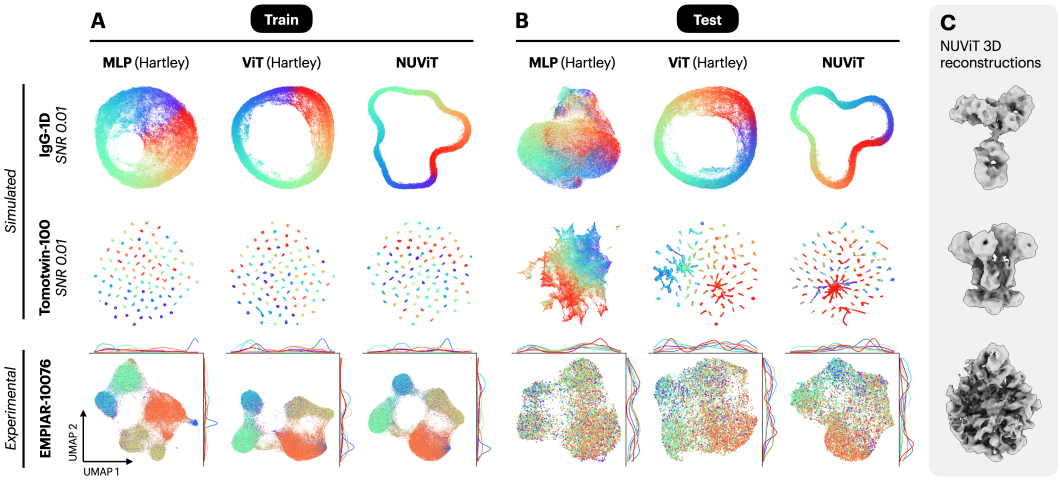

Figure 4: **Latent space visualizations of cryoDRGN with different Fourier image encoders. A,B)** UMAP (McInnes et al., 2018) visualizations of train and test latent spaces. `IgG-1D` latent embeddings are colored by the ground truth Fab angle, `Tomotwin-100` is colored by the ground truth protein, and `EMPIAR-10076` is colored by the published major class labels (Davis et al., 2016). **C)** Example reconstructed density maps from our cryoDRGN+NUViT reconstruction model.

to memorization (Levy et al., 2025) and are not robust to rotations or shifts (Edelberg & Lederman, 2023), with the extreme noise in cryo-EM images being a major contributing factor.

For these experiments, we evaluate on two synthetic datasets, `IgG-1D` and `Tomotwin-100` (Jeon et al., 2024), and the experimental `EMPIAR-10076` dataset of the assembling 50S ribosome (Davis et al., 2016). Reconstruction quality is measured by per-image FSC_{AUC} (Jeon et al., 2024). For amortization experiments, we use a 90/10 train-test split for `EMPIAR-10076`. For the synthetic `IgG-1D` and `Tomotwin-100` datasets, we generate a new test set of 100k images following the cryo-EM image formation model (Jeon et al., 2024).

Reconstruction quality Replacing the encoder with NUViT improves reconstruction quality across all three datasets (Table 6). We compare against additional methods in Table S4. On `IgG-1D`, we find that using cryoDRGN-NUViT achieves state-of-the-art performance among all deep learning methods, trailing only RECOVAR (Gilles & Singer, 2025), a linear subspace method that is able to decode multi-views at inference time via adaptive kernel regression, giving it an advantage over cryoDRGN's decoder.

Amortization We next evaluate each encoder on held-out test images. These images are passed through the encoder to generate latent embeddings (Figure 4), which are then passed through the decoder to reconstruct volumes that are compared with ground truth maps via FSC_{AUC} (Table 7). The MLP encoder shows the largest train-test gap, with FSC_{AUC} dropping from 0.346 to 0.139 on `Tomotwin-100`. ViT encoders improve amortization substantially, and NUViT improves further, achieving the best test performance on all three datasets. The latent space visualizations in Figure 4 are consistent with these findings: NUViT's latent space is better organized on both train and test images, whereas the MLP encoder fails to disentangle different conformations or compositions on test images. The improved amortization is plausibly because NUViT's polar grid tokenization

Table 6: **Cryo-EM reconstruction.** Effect of encoder architecture on cryoDRGN (Zhong et al., 2020; 2021a) reconstruction quality, measured by $\text{FSC}_{\text{AUC}}(\uparrow)$. All methods use the same decoder; only the encoder differs. FSC_{AUC} statistics are calculated over the reconstructed volumes, and do not represent multiple model runs.

Encoder	IgG-1D		Tomotwin-100		EMPIAR-10076	
	Mean (std)	Median	Mean (std)	Med	Mean (std)	Med
MLP (Hartley)	0.350 (0.019)	0.357	**0.346 (0.033)**	**0.349**	0.418 (0.027)	0.422
ViT (Hartley)	0.363 (0.006)	0.365	0.337 (0.027)	0.340	0.423 (0.021)	0.397
NUViT (Ours)	**0.367 (0.003)**	**0.368**	**0.345 (0.039)**	**0.348**	**0.433 (0.020)**	**0.435**

Table 7: **Amortization of cryo-EM reconstruction models.** We show the ability of different Fourier-space encoder architectures with cryoDRGN (Zhong et al., 2020) for reconstructing unseen cryo-EM images of the proteins the reconstruction model was trained on, measured by $\mathrm{FSC}_{\mathrm{AUC}}$.

Encoder Method	IgG-1D		Tomotwin-100		EMPIAR-10076	
	Mean (std)	Median	Mean (std)	Med	Mean (std)	Med
MLP (Hartley) (original)	0.337 (0.018)	0.336	0.139 (0.104)	0.120	0.400 (0.041)	0.413
ViT (Hartley)	0.354 (0.016)	0.359	0.293 (0.101)	0.331	0.399 (0.035)	0.411
NUViT (Ours)	**0.361 (0.011)**	**0.365**	**0.301 (0.105)**	**0.341**	**0.409 (0.037)**	**0.422**

concentrates capacity on the low-frequency components that carry the most structural information, making the encoder less susceptible to overfitting to the high-frequency noise that dominates individual training images.

4 Related Work

Fourier transforms in deep learning Fourier transforms have seen a variety of uses in deep learning, including as a data representation for modalities including audio, medical imaging, and microscopy, a way to do efficient convolution (Gu et al., 2021; Fu et al., 2022) or token mixing (Rao et al., 2021; Guibas et al., 2021; Lee-Thorp et al., 2022; Nguyen et al., 2022), positional encodings (Tancik et al., 2020), a shortcut for efficient 3D reconstruction via the Fourier Slice Theorem (Zhong et al., 2021a), and more. However, architectures designed for Fourier-space data remain underexplored, especially for transformer architectures (Vaswani et al., 2017; Dosovitskiy et al., 2020). Fourier Image Transformers (FIT) (Buchholz & Jug, 2022) considers Fourier-space data arising from tomographic reconstruction. FIT proposes to unroll a Fourier-space image into a 1D sequence of pixels ordered by frequency, with one token per pixel. FIT focuses on tomographic reconstruction and super resolution and does not apply a polar grid or NUFFT for building a general-purpose architecture. (Alshaabi et al., 2025) proposes a two-stage Fourier transformer for aberration correction, but retains the Cartesian ViT (Dosovitskiy et al., 2020) patching scheme. While not specifically designed for Fourier-space data, another line of work focuses on designing architectures for complex-valued data (Trabelsi et al., 2017; Venkatasubramanian et al., 2024), which could also be applied to the frequency domain. However, these methods suffer from challenges in optimization, lack robust activation functions, and are not well-validated for frequency-space data.

Non-uniform grids for Fourier space in deep learning In deep learning, the use of non-uniform grids for Fourier space data has been mainly explored in the context of Fourier Neural Operators (FNOs) (Li et al., 2020), which are resolution-invariant networks designed to solve PDEs. FNOs use repeated blocks of residual global convolutions which consist of a Fourier transform, linear transform, and inverse Fourier transform. FNOs do not use all frequencies, and in practice the high frequency modes are truncated, leaving only the low frequency modes. This approach is feasible because many PDEs follow a Fourier power law. Since FNOs implement the Fourier transform as an FFT, they are limited to Cartesian grids. Lingsch et al. (2023) examines the case of arbitrary grids, and shows that FNOs on irregular grids computed using the NUFFT can outperform FNOs on equispaced Cartesian grids. Our method extends this to general transformer architectures.

5 Conclusion and Limitations

We propose **Non-Uniform Vision Transformers (NUViT)**, a customized transformer architecture for Fourier space data. Motivated by the Fourier power law, NUViT introduces the idea of sampling and tokenizing on a non-uniform polar grid, providing stronger inductive biases such as better sampling of important low-frequency regions of Fourier space that are undersampled by a traditional ViT. Extensive experiments show that NUViT is a more expressive, generalizable, and robust encoder for Fourier data. Despite these encouraging results, NUViT has the limitation that it is more computationally expensive than ViTs. This may be addressed by using more performant NUFFT libraries, such as the one proposed in (Lingsch et al., 2023), as well as reducing the oversampling rate in the NUFFT (Beatty et al., 2005). This work could have long-term societal benefits through improved image reconstruction and analysis techniques for applications like cryo-EM and MRI, contributing to deeper insight into biological and physical systems.

AI use statement In this work, we used generative AI tools to generate the synthetic crystal dataset, to help design or provide feedback on research methodology or experiments, implement and run code, and suggest metrics for data analysis. We have not used generative AI tools to assist with translation or interpret results, and theoretical models and mathematical proofs are not applicable to this work. Additionally, we used generative AI tools to summarize or analyze existing literature, source/search for information, edit the paper, and identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was checked by the authors, and LLM-suggested literature was verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement Details on the method can be found in Figure 2, Section 2, and Appendix A. Information about the datasets and details on how to reproduce the experiments can be found in Appendix A. Code will be released upon acceptance.

REFERENCES

Thayer Alshaabi, Daniel E Milkie, Gaoxiang Liu, Cyna Shirazinejad, Jason L Hong, Kemal Achour, Frederik Görlitz, Ana Milunovic-Jevtic, Cat Simmons, Ibrahim S Abuzahriyeh, et al. Fourier-based 3d multistage transformer for aberration correction in multicellular specimens. *ArXiv*, pp. arXiv–2503, 2025.

Neil W Ashcroft and N David Mermin. Solid state. *Physics (New York: Holt, Rinehart and Winston) Appendix C*, 1, 1976.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023.

Philip J Beatty, Dwight G Nishimura, and John M Pauly. Rapid gridding reconstruction with a minimal oversampling ratio. *IEEE transactions on medical imaging*, 24(6):799–808, 2005.

Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2): 198–217, 1956.

Tim-Oliver Buchholz and Florian Jug. Fourier image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1846–1854, 2022.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Joseph H Davis, Yong Zi Tan, Bridget Carragher, Clinton S Potter, Dmitry Lyumkis, and James R Williamson. Modular assembly of the bacterial large ribosomal subunit. *Cell*, 167(6):1610–1622, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hufei Duan, Yiqing Liu, Hui Yan, Qiming He, Yonghong He, and Tian Guan. Fourier vit: A multi-scale vision transformer with fourier transform for histopathological image classification. In *2022 7th International Conference on Automation, Control and Robotics Engineering (CACRE)*, pp. 189–193. IEEE, 2022.

Daniel G Edelberg and Roy R Lederman. Using vaes to learn latent variables: observations on applications in cryo-em. *ArXiv*, pp. arXiv–2303, 2023.

J. Ryan Feathers, Robert C. Heeter, Geoffrey Woollard, Sonya M. Hanson, Pilar Cossio, Joel Greer, Tom Burnley, and Ellen D. Zhong. The cahra challenge: A community-wide assessment of cryo-em heterogeneous reconstruction algorithms. *bioRxiv*, 2026. doi: 10.64898/2026.09.15.751515.

Jeffrey A Fessler and Bradley P Sutton. Nonuniform fast fourier transforms using min-max interpolation. *IEEE transactions on signal processing*, 51(2):560–574, 2003.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Andrea Gabrielli. Point processes and stochastic displacement fields. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 70(6):066131, 2004.

Andrea Gabrielli, Michael Joyce, and Francesco Sylos Labini. Glass-like universe: Real-space correlation properties of standard cosmological models. *Physical Review D*, 65(8):083523, 2002.

Marc Aurèle Gilles and Amit Singer. Cryo-em heterogeneity analysis using regularized covariance estimation and kernel regression. *Proceedings of the National Academy of Sciences*, 122(9): e2419140122, 2025.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Jeffrey Gu, Minkyu Jeon, Ambri Ma, Serena Yeung-Levy, and Ellen D Zhong. Cryohype: Reconstructing a thousand cryo-em structures with transformer-based hypernetworks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35280–35290, 2026.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M Hanson, Pilar Cossio, and Ellen D Zhong. Cryobench: Diverse and challenging datasets for the heterogeneity problem in cryo-em. *Advances in Neural Information Processing Systems*, 37:89468–89512, 2024.

Jaeuk Kim and Salvatore Torquato. Effect of imperfections on the hyperuniformity of many-body systems. *Physical Review B*, 97(5):054105, 2018.

Dari Kimanius, Kiarash Jamali, and Sjors Scheres. Sparse fourier backpropagation in cryo-em reconstruction. *Advances in Neural Information Processing Systems*, 35:12395–12408, 2022.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Michael A Klatt, Jaeuk Kim, and Salvatore Torquato. Cloaking the underlying long-range order of randomly perturbed lattices. *Physical Review E*, 101(3):032118, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4296–4313, 2022.

Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022.

Axel Levy, Rishwanth Raghu, J Ryan Feathers, Michal Grzadkowski, Frédéric Poitevin, Jake D Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D Zhong. Cryodrgn-ai: neural ab initio reconstruction of challenging cryo-em and cryo-et datasets. *Nature Methods*, pp. 1–9, 2025.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Levi Lingsch, Mike Y Michelis, Emmanuel De Bézenac, Sirani M Perera, Robert K Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains. *arXiv preprint arXiv:2305.19663*, 2023.

Da Long, Shandian Zhe, Samuel Williams, Leonid Oliker, and Zhe Bai. Stft: Spatio-temporal fourier transformer for long-term dynamics prediction. *arXiv preprint arXiv:2503.11899*, 2025.

Zhenwei Luo, Fengyun Ni, Qinghua Wang, and Jianpeng Ma. Opus-dsd: deep structural disentanglement for cryo-em single-particle analysis. *Nature Methods*, 20(11):1729–1738, 2023.

Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi: 10.1002/mrm.21391.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Shreyan Mohanty. Oasis alzheimer's detection multi-class dataset, 2025. URL `https://www.kaggle.com/dsv/14143392`.

M. J. Muckley, R. Stern, T. Murrell, and F. Knoll. TorchKbNufft: A high-level, hardware-agnostic non-uniform fast Fourier transform. In *ISMRM Workshop on Data Sampling & Image Reconstruction*, 2020. Source code available at https://github.com/mmuckley/torchkbnufft.

Tan Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. *Advances in Neural Information Processing Systems*, 35:29319–29335, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Ali Punjani and David J Fleet. 3d variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-em. *Journal of structural biology*, 213(2):107702, 2021.

Ali Punjani and David J Fleet. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nature Methods*, 20(6):860–870, 2023.

Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Salvatore Torquato. Hyperuniform states of matter. *Physics Reports*, 745:1–95, 2018.

Salvatore Torquato and Frank H Stillinger. Local density fluctuations, hyperuniformity, and order metrics. *Physical Review E*, 68(4):041113, 2003.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. *arXiv preprint arXiv:1705.09792*, 2017.

van A Van der Schaaf and JH van van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shyam Venkatasubramanian, Ali Pezeshki, and Vahid Tarokh. Steinmetz neural networks for complex-valued data. *arXiv preprint arXiv:2409.10075*, 2024.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *International conference on artificial neural networks*, pp. 485–491. Springer, 2001.

Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6-7):889–899, 2006.

Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors H W Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, 2014. doi: 10.7554/eLife.03080.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-em images. In *International Conference on Learning Representations*, 2020.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021a.

Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4066–4075, 2021b.

A TRAINING DETAILS

A.1 IMPLEMENTATION DETAILS

Unless otherwise stated, all results are for the NUViT-S (Small) configuration, which is analogous to the ViT-S: $L = 12$ layers, hidden dimension $D = 384$, 12 attention heads, and MLP dimension $4D = 1536$. For the polar grid, we use $H = W = 128$ with patch size $P = 16$, yielding $N = 64$ patches per channel and $2 \times 64 + 1 = 129$ tokens. We set the RFF bandwidth to $\sigma = 1$. We also find it useful to control the max radius of the radial shells, denoted R. Unless otherwise stated, all NUViT models process polar grids of size $N_r = N_\phi = 128$. All our experiments are done in PyTorch (Paszke et al., 2019). We implement NUFFTs using the `TorchKbNUFFT` package (Muckley et al., 2020).

A.2 COMPUTE DETAILS

Models were trained on a shared research cluster. Most models were trained with a single NVIDIA H100 GPU, 32GB of memory, and 32 CPUs. When training on larger datasets such as `Ribosembly`, we use 128GB of memory.

Table S1: **Computational cost.** Wall-clock time per training epoch and peak memory usage.

Encoder	Params	Time / epoch	GPU Mem.	CPU Mem.
MLP	19.9 M	2:19.62	12.3 GB	7.9 GB
ViT	24.9 M	1:40.42	13.6 GB	7.9 GB
NUViT	25.1 M	2:11.84	16.7 GB	7.8 GB

In Table S1, we provide the wall-clock time and memory for a cryo-EM reconstruction model trained on the ADH dataset, using an NVIDIA H100 GPU. NUViT takes 31% more wall-clock time per epoch and 23% more memory than a ViT. Compared to an MLP, NUViT is slightly faster in wall clock time but uses more GPU memory. All encoders have similar CPU memory usage. We believe that there are still untapped potential gains in efficiency for NUViT encoders, including using a more efficient implementation, such as that of (Lingsch et al., 2023). Although the speed of calculating one NUFFT is fast, reducing the oversampling rate during NUViT from 2x to 1.25x could provide further speedups.

A.3 NATURAL IMAGE CLASSIFICATION

For the natural image experiments we process the data as follows:

1. First, we resize all images to 128×128 resolution so we can use the same model for all experiments.
2. We apply `RandomCrop`, `RandomHorizontalFlip`, and `ColorJitter` data augmentations and normalize the data.
3. We transform the images from RGB to grayscale, apply the Fourier transform, and put the Fourier transformed-data into the desired data representation.

All models were trained with batch size 64 with the Adam optimizer (Kingma & Ba, 2014) with learning rate 1e-4 for 100 epochs. For NUViT on CIFAR-10, we use a log amplitude and phase representation and set R to 2. On CIFAR-100, we use an amplitude and phase representation and set R to 4. Both models were trained with a cosine learning rate schedule. NUViT takes approximately 18 hours to train. Hyperparameters were chosen by grid search.

A.4 MRI CLASSIFICATION

The OASIS Alzheimer's Detection Multiclass Dataset (Mohanty, 2025) consists of the brain MRI scans of 416 subjects. The dataset is designed to be data-leakage free, so any single patient's data is contained in a single training split. Alzheimer's severity was assessed based on the Clinical

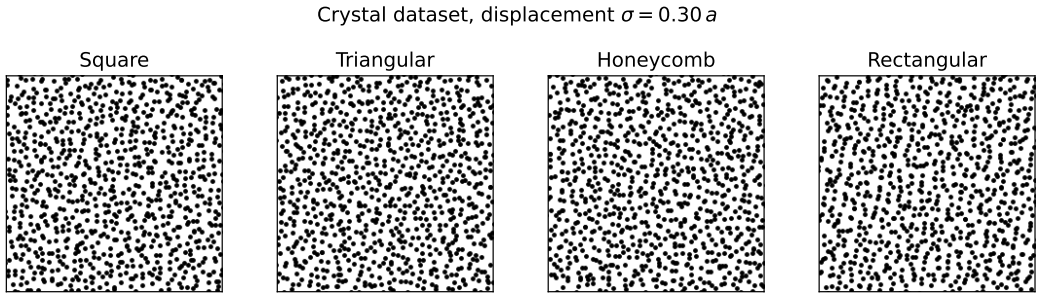

Figure S1: **Simulated 2D crystal lattice examples.**

Dementia Rating (CDR) into four distinct classes: non-demented (CDR 0.0), very mild demented (0.5), mild demented (1.0), and moderate demented (2.0). The OASIS dataset is unbalanced towards non-demented patients, with a class breakdown of 316/70/28/2.

A.5 2D Crystal Lattice Classification

The synthetic 2D crystal lattice dataset consists of four classes of crystal lattices (square, triangular, honeycomb, rectangular) with 1000 train examples and 250 test examples per class, for a total of 4000 training examples and 1000 test examples. Imperfect lattices were generated by randomly perturbing each point by 30% of the lattice spacing (Figure S1). The perturbation value was calibrated on ViT-S (real) to ensure the dataset was neither too easy nor too difficult, with the final value being 30% of the lattice spacing. Each model was trained for 100 epochs without data augmentation. Model-specific hyperparameters were chosen based on the best parameters from natural image classification.

A.6 Cryo-EM Image Classification

For cryo-EM heterogeneity classification, we use a random 90/10 train/test split for each of the datasets. All models were trained with batch size 64 with the Adam optimizer (Kingma & Ba, 2014) with learning rate 1e-4. We use R values of $8, 16, 16, 16, 16, 4, 8$ for `IgG-1D`, `IgG-RL`, `Ribosembly`, `Tomotwin-100`, `EMPIAR-10076`, `RyR`, respectively, and $N_\phi = 128$ for all datasets.

A.7 Cryo-EM 3D Reconstruction

For the synthetic datasets (`IgG-1D` and `Tomotwin-100`), all models were traind on the whole dataset. To assess amortization, we generated an additional 100K images based on CryoBench (Jeon et al., 2024). Models were trained on 20, 50, and 100 epochs for `IgG-1D`, `EMPIAR-10076`, and `Tomotwin-100`, respectively. Otherwise, cryoDRGN-MLP used cryoDRGN's default parameters were used. CryoDRGN-NUViT used the best model parameters from the previous section.

B Metrics

Following CryoBench (Jeon et al., 2024), we use per-image FSC to jointly evaluate reconstruction quality and heterogeneity. In cryo-EM, the Fourier Shell Correlation (FSC) curve is a standard tool for comparing two volumes by measuring their correlation across concentric shells in Fourier space. Since cryo-EM methods typically reconstruct a 3D maps from a single 2D image, per- image FSC quantifies how well each individual reconstruction matches the groundtruth. $\mathrm{FSC}_{\mathrm{AUC}}$ summarizes performance using the area under the FSC curve. To capture the distribution of reconstruction quality across conformations, we sample one image per conformation, as in CryoBench. The maximum value of $\mathrm{FSC}_{\mathrm{AUC}}$ is 0.5.

C ADDITIONAL RESULTS

C.1 CRYO-EM IMAGE CLASSIFICATION

Table S2: **Cryo-EM image classification (top-5 accuracy).** Performance is measured by top-5 accuracy. Results are reported as `train / test accuracy`.

Method	IgG-1D	IgG-RL
Random chance	5.00 / 5.00	5.00 / 5.00
MLP (real)	99.99 / 5.57	100.00 / 9.08
MLP (Hartley)	99.99 / 6.59	99.98 / 11.54
ViT-S (real)	100.00 / 4.77	100.00 / 5.84
ViT-S (Hartley)	7.96 / 7.19	92.30 / 28.54
NUViT (Ours)	86.63 / **59.01**	99.95 / **96.78**

In Table S2, we show top-5 accuracy results for cryo-EM conformation classification on the IgG-1D and IgG-RL datasets, providing context where most results are close to random chance. These datasets are challenging because they require the discrimination of very similar conformations at different viewing angles and in the presence of high noise, with IgG-1D being more difficult since each conformation of IgG-1D has a more similar neighboring conformation than IgG-RL. We find that most baselines for IgG-1D perform near random chance, due to the difficulty of the classification problem. For IgG-RL, most methods perform significantly better than random chance, unlike for top-1 accuracy, indicating some discrimination ability. As with top-1 accuracy, for both datasets our method significantly outperforms the baselines.

C.2 CRYO-EM 3D RECONSTRUCTION

Table S3: **Additional experimental datasets.** $\mathrm{FSC}_{\mathrm{AUC}}(\uparrow)$ by training dataset from EMPIAR-10028 (Wong et al., 2014) and CAHRA (Feathers et al., 2026).

Dataset	MLP	ViT	NUViT (Ours)
EMPIAR-10028	0.471	0.478	**0.480**
CAHRA-ADH	0.390	0.448	**0.481**
CAHRA-Aldolase	0.363	0.437	**0.477**
CAHRA-PKM2	0.402	0.448	**0.462**
CAHRA-ADH/Aldolase/PKM2	0.437	0.448	**0.455**

Additional experimental datasets Table S3 shows the performance of different encoders for cryo-EM reconstruction on the EMPIAR-10028 (Wong et al., 2014) and subsets of the CAHRA Heterogeneity Challenge (Feathers et al., 2026) datasets. The ADH, Aldolase, PKM2 label refers to a dataset composed of a mixture of the three proteins. We find that NUViT encoders outperform ViT encoders, which in turn outperform MLP encoders.

Expanded comparison to prior methods Table S4 shows the performance of our method compared to state-of-the-art deep learning and non-deep learning models. On `IgG-1D`, NUViT achieves state-of-the-art performance among deep learning methods, trailing only RECOVAR (Gilles & Singer, 2025), a PCA-based model which uses a more powerful adaptive kernel regression decoder. On Tomotwin-100, NUViT reaches the performance of the best methods. All baseline numbers, except those for CryoDRGN, were drawn from CryoBench (Jeon et al., 2024) and CryoHype (Gu et al., 2026). Dashes indicate that there is no prior evaluation of that baseline on the given dataset.

Latent space quality To assess the latent spaces of Figure 4 quantitatively, we use two sets of metrics. To evaluate continuous heterogeneity with a known groundtruth coordinate structure, such as IgG-1D, we evaluate trustworthiness (Venna & Kaski, 2001) and continuity (Venna & Kaski, 2006), two neighborhood-preservation metrics designed to assess continuous latent spaces such as IgG-1D's ring latent space. Both metrics take values in $[0, 1]$ with higher values being better. We compute these metrics between a latent space and an idealized ring structure.

Table S4: **Cryo-EM reconstruction on IgG-1D and Tomotwin-100**. We show the ability of NU-ViT to serve as a Fourier image encoder for a VAE-based cryo-EM reconstruction model (cryoDRGN (Zhong et al., 2020; 2021a)). Performance is measured by per-image area under the Fourier shell correlation curves ($\mathrm{FSC}_{\mathrm{AUC}}(\uparrow)$).

Method	**IgG-1D**		**Tomotwin-100**	
	Mean (std)	Median	Mean (std)	Med
3D classification (Punjani et al., 2017)	0.297 (0.019)	0.291	0.046 (0.026)	0.037
3DVA (Punjani & Fleet, 2021)	0.349 (0.004)	0.350	0.088 (0.040)	0.077
RECOVAR (Gilles & Singer, 2025)	0.386 (0.005)	0.388	0.258 (0.109)	0.254
CryoDRGN (Zhong et al., 2020)	0.350 (0.019)	0.357	0.346 (0.033)	0.349
CryoDRGN-AI (Levy et al., 2025)	0.364 (0.002)	0.364	0.202 (0.044)	0.321
OPUS-DSD (Luo et al., 2023)	0.335 (0.026)	0.339	0.237 (0.049)	0.251
3DFlex (Punjani & Fleet, 2023)	0.335 (0.003)	0.335	-	-
SFBP (Kimanius et al., 2022)	-	-	0.036 (0.011)	0.036
CryoHype (Gu et al., 2026)	-	-	0.346 (0.033)	0.353
CryoDRGN + ViT (Hartley)	0.363 (0.006)	0.365	0.337 (0.027)	0.340
CryoDRGN + NUViT (Ours)	**0.367 (0.003)**	**0.368**	0.345 (0.039)	0.348

Table S5: **Embedding quality.** Trustworthiness and continuity (↑) of the learned embedding space.

	Train		**Test**	
Encoder	Trust.	Cont.	Trust.	Cont.
MLP	0.9260	0.8467	0.7534	0.6910
ViT	0.9454	0.9429	0.8791	0.8910
NUViT (Ours)	**0.9634**	**0.9692**	**0.9111**	**0.9323**

In Table S5, we find that NUViT's train and test latent space is much closer to the ideal ring structure of IgG-1D's groundtruth reaction coordinate than the latent spaces corresponding to the other encoders. We believe this is also reflected visually by NUViT's latent space being a nice, thin loop, whereas other latent spaces look thicker and noisier.

Datasets exhibiting significant compositional heterogeneity, such as Tomotwin-100, ideally form many discrete clusters in latent space. To assess these types of datasets, we use Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), which are commonly used to assess these datasets in cryo-EM.

Table S6: **Clustering quality.** Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), ↑ higher is better.

	Train		**Test**	
Encoder	ARI	AMI	ARI	AMI
MLP	**0.9952**	**0.9968**	0.1356	0.3601
ViT	0.9892	0.9944	0.6924	0.8194
NUViT (Ours)	0.9838	0.9942	**0.7067**	**0.8222**

In Table S6, we find that although all encoders perform extremely well on the training data, the latent space of the MLP encoder collapses when performing inference on test data, which matches the non-clustered latent space we observe in Figure 4. NUViT again has the best-performing latent space, the ViT encoder's performance is only slightly less. Since one of the main drivers of reconstruction error is assigning an image the incorrect conformation (i.e., the wrong latent), this partially explains the reconstruction results of Tables 6 and 7.

Table S7: **Ablation studies.** Top-1 accuracy (%) on Tomotwin-100 and IgG-1D.

Method	Tomotwin-100	IgG-1D
Channel positional encoding		
No channel PE	**90.45**	7.62
Channel PE	88.86	**12.84**
Feature representation		
Hartley	67.68	4.91
Real/Imaginary	**90.01**	11.04
Amplitude/Phase	88.86	**13.30**

C.3 Ablations

In Table S7, we repeat the ablation study of Table 3 for cryo-EM conformation classification on the IgG-1D and Tomomtwin-100 datasets. We find that while channel PE and an amplitude-phase representation may not always be the optimal choice, they perform better on difficult datasets such as IgG-1D. As such, our default method may represent a strong starting point while various features may need to be tuned for optimal performance on any single dataset.