TYPE Original Research
PUBLISHED 09 July 2024
DOI 10.3389/fmolb.2024.1393564

Check for updates

OPEN ACCESS

EDITED BY
Edward T. Eng,
New York Structural Biology Center,
United States

REVIEWED BY
Chang Liu,
Biogen Idec, United States
Marc Delarue,
Institut Pasteur, France

*CORRESPONDENCE
David A. Klindt,
✉ klindt@cshl.edu

RECEIVED 29 February 2024
ACCEPTED 22 May 2024
PUBLISHED 09 July 2024

CITATION
Klindt DA, Hyvärinen A, Levy A, Miolane N and Poitevin F (2024), Towards interpretable Cryo-EM: disentangling latent spaces of molecular conformations.
Front. Mol. Biosci. 11:1393564.
doi: 10.3389/fmolb.2024.1393564

Towards interpretable Cryo-EM: disentangling latent spaces of molecular conformations

David A. Klindt[1,2]*, Aapo Hyvärinen[3], Axel Levy[1,4], Nina Miolane[2] and Frédéric Poitevin[1]

[1]LCLS, SLAC National Accelerator Laboratory, Stanford University, Palo Alto, CA, United States, [2]Department of Electrical and Computer Engineering, University of California Santa Barbara (UCSB), Santa Barbara, CA, United States, [3]Department of Computer Science, University of Helsinki, Helsinki, Finland, [4]Department of Electrical Engineering, Stanford University, Palo Alto, CA, United States

Molecules are essential building blocks of life and their different conformations (i.e., shapes) crucially determine the functional role that they play in living organisms. Cryogenic Electron Microscopy (cryo-EM) allows for acquisition of large image datasets of individual molecules. Recent advances in computational cryo-EM have made it possible to learn latent variable models of conformation landscapes. However, interpreting these latent spaces remains a challenge as their individual dimensions are often arbitrary. The key message of our work is that this interpretation challenge can be viewed as an Independent Component Analysis (ICA) problem where we seek models that have the property of identifiability. That means, they have an essentially unique solution, representing a conformational latent space that separates the different degrees of freedom a molecule is equipped with in nature. Thus, we aim to advance the computational field of cryo-EM beyond visualizations as we connect it with the theoretical framework of (nonlinear) ICA and discuss the need for identifiable models, improved metrics, and benchmarks. Moving forward, we propose future directions for enhancing the disentanglement of latent spaces in cryo-EM, refining evaluation metrics and exploring techniques that leverage physics-based decoders of biomolecular systems. Moreover, we discuss how future technological developments in time-resolved single particle imaging may enable the application of nonlinear ICA models that can discover the true conformation changes of molecules in nature. The pursuit of interpretable conformational latent spaces will empower researchers to unravel complex biological processes and facilitate targeted interventions. This has significant implications for drug discovery and structural biology more broadly. More generally, latent variable models are deployed widely across many scientific disciplines. Thus, the argument we present in this work has much broader applications in AI for science if we want to move from impressive nonlinear neural network models to mathematically grounded methods that can help us learn something new about nature.

KEYWORDS

cryo-EM, machine learning, ICA, AI for science, disentanglement, physics-based models

1 Introduction

Molecules such as proteins or nucleic acids make up the building blocks of life. Living organisms contain a plethora of molecules that often comprise thousands of atoms. Biomolecules change their *conformation* (i.e., shape) to fulfill important biological functions such as enzymatic reactions or cellular communication. Understanding the conformational heterogeneity of biomolecules is crucial for deciphering their functional mechanisms and designing targeted interventions. Cryo-Electron Microscopy (cryo-EM) has emerged as a powerful technique for visualizing molecular structures at high resolution. Recent advancements in computational cryo-EM have demonstrated the potential of latent variable models to capture the diverse conformations adopted by biomolecules (reviewed in Donnat et al., 2022). However, interpreting these learned latent spaces and extracting biologically meaningful information from them remains a significant challenge.

In this paper, we propose a fruitful approach to unravel the complexities of conformational latent spaces in cryo-EM by framing this as an Independent Component Analysis (ICA) problem. In its original linear formulation, the high level goal of ICA is to discover linear projections of the data that are as statistically independent as possible (Hyvärinen and Oja, 2000). A common observation in ICA applications is that these linear projections discover underlying factors of variation in the data that give insight into the underlying processes. A few prior works have tested the application of *linear* ICA to molecular imaging (Borek et al., 2018; Gao et al., 2020) finding more meaningful separation of molecular conformation changes. However, the transformation between meaningful factors and the data is inherently nonlinear in cryo-EM. Therefore, we need theory and models that work for the *nonlinear* models used in modern cryo-EM (e.g., Zhong et al., 2021a). Building on recent theoretical work in identifiable nonlinear ICA Hyvärinen et al. (2024), in disentanglement models and their benchmarks in machine learning Locatello et al. (2019), we suggest a path to bridging the gap between theoretical advancements and practical applications in cryo-EM research. We argue that nonlinear ICA methods have the potential to provide a powerful framework to disentangling the latent representations of biomolecular conformations from cryo-EM datasets, overcoming the limitations of traditional volume visualization approaches and ultimately allowing to delve deeper into the structural dynamics of biomolecules. Moreover, we argue that the establishment of

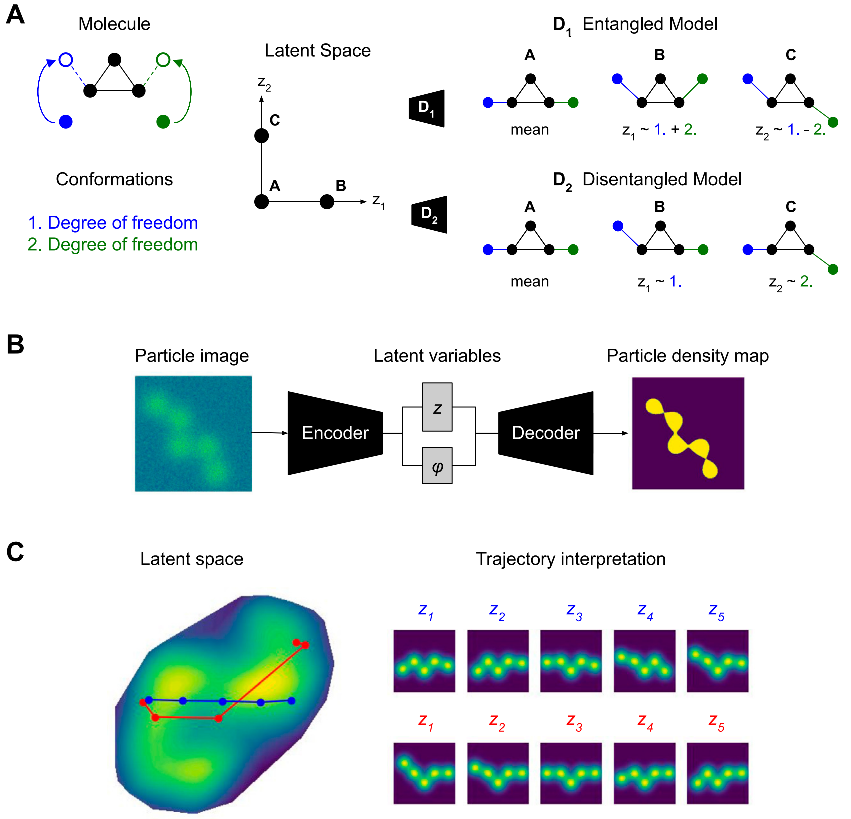

FIGURE 1
Overview. *What does it mean to have a disentangled representation of molecular conformations?* **(A)** The example (left) shows a simple molecule with two degrees of freedom 1. and 2. for changing its conformation. An entangled model (right, top) represents mixtures of both movements on each of its latent dimensions z_1 ~ 1. + 2. and z_2 ~ 1. − 2. A disentangled model (right, bottom) represents pure movements on each of its latent dimensions z_1 ~ 1. and z_2 ~ 2.; actually, $z_2 = -2.$ but the sign flip, incorporated in ~, does not compromise interpretability. Note that disentangling conformations from cryo-EM measurements requires additional information (e.g., time, temperature or physics), as discussed in section 4. **(B)** Training a VAE with separate pose ϕ and conformation z latent spaces on cryo-EM particle images, without any intervention. **(C)** Interpreting the learned latent space of a model. An axis traversal (blue) results in a complex motion of both arms, i.e., fails at disentangling the two degrees of freedom. A simple transformation, moving only the left arm, corresponds to a curved trajectory.

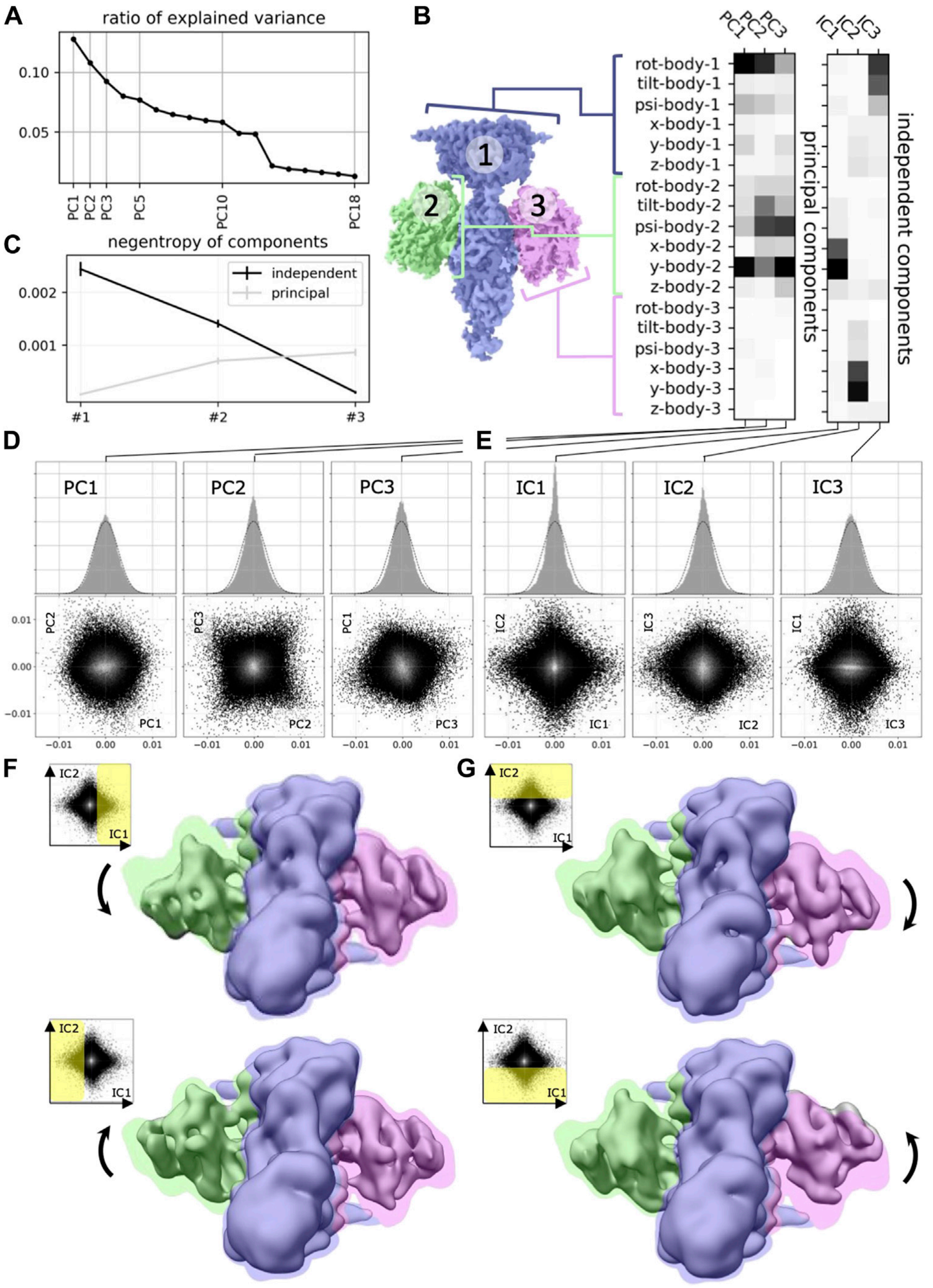

FIGURE 2
Application of ICA to cryo-EM data (reproduced with permission from (Gao et al., 2020). Original caption: **(A)** Principal component analysis of the 18 multi-body parameters refined for each particle image yields 18 principal components (PC) displayed here in decreasing order of explained variance. The first three components explain more than 30% of the variability in the particle images. **(B)** (left) definition of the multi-body segmentation: the central PDE6 stalk in blue corresponds to Body 1, while the 2 GαT· GTP subunits correspond to Bodies two and 3. (right) The motion of each body is parameterized with three translational parameters and three rotational parameters. Each of the 18 principal and three independent components is a linear combination of the resulting 18 rigid-body parameters, and their weights are shown here for the first three principal components (from negligible to larger weight as the shade of grey becomes darker). **(C)** Negentropy (i.e., reverse entropy) of the first three principal and independent components. **(D)** (resp. **(E)**)–(top) histogram of the projection of all image particle parameters on the first three principal (resp. independent) components PC1, PC2 and PC3 (resp. IC1, IC2 and IC3). (bottom) 2D histograms of the projections of all image particle parameters on all pairs of the first three principal (resp. independent) components. **(F)** (resp. **(G)**–Maps illustrating the motions carried by IC1 (resp. IC2). (top) map reconstructed from the particles whose projections belong to the last bin along IC1 (resp. IC2). (bottom) map reconstructed from the particles whose projections belong to the first bin along IC1 (resp. IC2). All maps are shown overlaid on the consensus map, with threshold set at a lower density value, colored according to the scheme in **(B)**.

TABLE 1 *Glossary*. Whenever a distinction is necessary in a given context, we use a $*$ (e.g., g^*) to highlight that we are referring to the *ground truth* model (g^*) or *ground truth* latent variables (z^*, ϕ^*). For instance, we have the *ground truth* generator g^* of the data, in contrast to the *learned* generator g (i.e., decoder) from our model of the data.

Notation	Description
$z \in \mathcal{Z} = \mathbb{R}^K$	Conformation latent variables (vector)
$\phi \in \Phi = SO(3)$	Pose latent variables
$g\colon \mathcal{Z} \times \Phi \rightarrow \mathcal{X}$	Mixing function/generator/decoder
$v \in \mathcal{V}\colon \mathbb{R}^3 \times \mathcal{Z} \rightarrow \{0, 1\}$	Volume function (implicit representation)
$\pi\colon \mathcal{V} \times \Phi \rightarrow \mathbb{R}^N$	Projection/pose function
$x \in \mathcal{X} = \mathbb{R}^N$	Observations (e.g., images)
$f\colon \mathcal{X} \rightarrow \mathcal{Z} \times \Phi$	Learned representation/encoder
ICA	Independent Component Analysis
identifiable	A model with a unique (set) of solutions
disentangle	Separate different factors (e.g., pose and shape)

benchmarks and metrics specific to cryo-EM disentanglement models is of paramount importance. Adapting and extending existing benchmarks from the machine learning field should allow to objectively evaluate the performance of different disentanglement approaches and track progress in the development of interpretable cryo-EM methods. Ultimately, this interdisciplinary approach will enhance our understanding of complex biological processes and open up new avenues for therapeutic interventions and drug discovery.

The paper is structured as follows. We first provide a general background on the cryo-EM computational problem (Section 2) and how it can be framed as an ICA problem (Section 2.1). We then discuss the two fundamental challenges associated with cryo-EM: firstly, separating the conformation and the pose representations (Section 2.2) and, secondly, finding the right (disentangled) representation of conformations (Section 2.3). We then go into more detail on both challenges by providing quantitative metrics to measure progress and modeling suggestions to improve current frameworks. To disentangle poses and shapes, we propose intervention based metrics (Section 3.1) and training schemes (Section 3.2) that can be added to existing models. For the larger problem of disentangling conformation representations (Section 4.1), we discuss existing disentanglement benchmarks and metrics (Locatello et al., 2019). We then discuss three potential approaches for solving this problem (Section 4.2), based on temporal information (Section 4.2.1), temperature control (Section 4.2.2) and atomic models (Section 4.2.3). Finally, we discuss the path forward and the broader implications for this framework to take computational approaches from neural network based curve fitting to actual understanding of the mechanisms in nature.

2 Interpretable heterogeneous reconstruction—a disentanglement problem

Heterogeneous cryo-EM reconstruction methods aim to model the different conformations that a molecule may assume (Donnat et al., 2022). For instance, we can think of a molecule with a fixed central structure and two adjustable "arms" (see Figure 1). Clearly, any conformation that this molecule may assume can be described by providing the position of both arms. Thus, these independently moving parts may be thought of as the fundamental *degrees of freedom* of this molecule's conformations.

We can parameterize them by a two dimensional *latent variable* $z \in \mathcal{Z}$ where $\mathcal{Z} \coloneqq \mathbb{R}^2$ is some degree of movement. The volume that the molecule occupies in three dimensional space can be thought of as a function $v \in \mathcal{V}$, $v\colon \mathbb{R}^3 \times \mathcal{Z} \rightarrow \{0, 1\}$ that is parameterized by z and indicates for any position in space ($\mathbb{R}^3$) whether it is part of the molecular volume or not, known as an *implicit representation* of the volume (Sitzmann et al., 2020; Donnat et al., 2022). That means, for different values of z, $v_z = v(., z)$ would describe a different volume. Crucially, this function is not known and it is a central goal in heterogeneous cryo-EM reconstruction to learn and study it. For instance, one approach would be to train a neural network (v_θ) to approximate the true volume function $v_\theta \approx v^*$ (Zhong et al., 2021a).

Furthermore, in cryo-EM we typically see a projection $\pi\colon \mathcal{V} \times \Phi \rightarrow \mathbb{R}^N$ to a gray-scale pixel image (represented, to keep notation uncluttered, as a vector with N entries). This projection depends on the pose parameters $\phi \in \Phi = SO(3)$, so we will also refer to π as the *pose function* (Table 1). The pose parameters may also need to be inferred (typically, the cryo-EM image formation model would also include camera parameters such as the microscope defocus—we are skipping those for simplicity). That means, for different values of ϕ, $\pi_\phi = \pi(., \phi)$ would describe a different projection. Importantly, the function π does not have to be learned because we know the physics, i.e., optics behind this projection, thus, we know that the projection in our model π_ϕ must be the same one as the *ground truth* projection $\pi^*_\phi = \pi_\phi$.

Putting this together we can write the combined cryo-EM generative model (i.e., the abstract process that yields the data we observe). That is, the observed data x is modeled as being generated by the *ground truth* model as $x = g^*(z^*, \phi^*)$, which is, crucially, a function of the *ground truth* latent quantities (z^*, ϕ^*)

$$g^*\colon \mathcal{Z} \times \Phi \rightarrow \mathcal{X}, \quad g^*(z^*, \phi^*) \coloneqq \pi\left(v^*_{z^*}, \phi^*\right) = \left(\pi_{\phi^*} \circ v^*\right)(z^*) \in \mathbb{R}^N. \tag{1}$$

Usually, we would be measuring very noisy signals $x = g^*(z^*, \phi^*) + \epsilon$ where the noise can be modeled as additive Gaussian $\epsilon \sim \mathcal{N}(0, \sigma^2)$ in image space. The essential problem of heterogeneous reconstruction in computational cryo-EM can now be stated as follows.

Given only noisy observations of $x = g^*(z^*, \phi^*) + \epsilon$, *can we recover* v^**?*

This would be the true volume function v^* that shows us how the independent degrees of freedom change the molecule's conformation. Many cryo-EM models actually learn a probabilistic $p(x|z, \phi)$ representation of the observed data x conditioned on the latent variables. Thus, it becomes necessary to perform inference such as maximum *a posteriori* estimation of the latents, conditioned on some observed data $p(z, \phi|x)$. Alternatively, it is common to approximate the posterior distribution itself with an amortized variational method such as a variational autoencoder (VAE) (Kingma and Welling, 2013). In this work we will be agnostic about the inference procedure (maximum *a posteriori* probability (MAP), or mean of the amortized variational posterior) and just assume that there exists a mapping $f\colon \mathcal{X} \rightarrow \mathcal{Z}$ from data to latent variables.

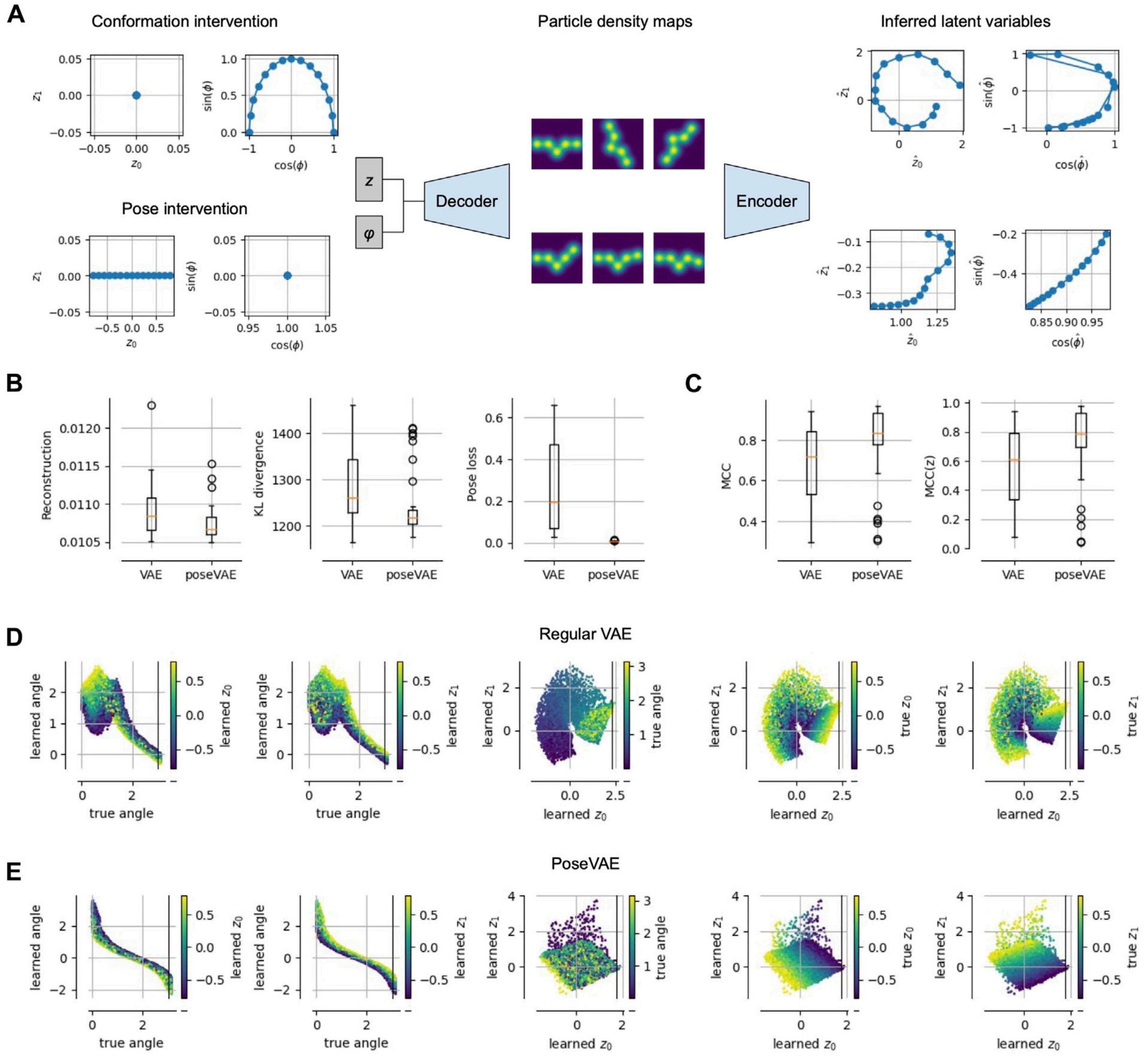

FIGURE 3
Physics-based Pose and Conformation Disentanglement. **(A)** We perform an intervention, i.e., changing only the pose (conformation) of a latent pair; these changed latents are decoded and encoded again to measure the consistency and invariance of our model. **(B)** Compared to a vanilla VAE, a model PoseVAE trained with interventions (i.e., Alg. 1) achieves lower reconstruction error, lower KL divergence to the prior and a lower pose disentanglement loss (Eq. 2). **(C)** Disentanglement, measured as mean correlation coefficient (MCC (Hyvarinen and Morioka, 2017),), increases not only between pose and conformation variables (left), but also among the conformation variables (right). **(D)** Visualizations of the learned latents for the vanilla VAE model, showing that the learned angle is not perfectly representing the true angle (plots one and two from the left); the three plots on the right show the learned conformation latents, representing mixtures of true conformation and pose. **(E)** Same as **(D)** but for PoseVAE, showing a perfect monotonic relationship between learned and true angle; also the conformation latents contain little information about the true angle (third plot) and disentangle the true conformation variables up to a 45° rotation.

2.1 Heterogeneous reconstruction in Cryo-EM is an ICA problem

Let us compare this to a standard independent component analysis (ICA) setting (Comon, 1994). In ICA we assume that there are $K > 1$ independent variables collected in the random vector $z = (z_1, \ldots, z_K)$. As an example to illustrate this, we may think of a public space where K different speakers proclaim their prophecies z_i, completely independently of one another $p(z_i, z_j) = p(z_i)p(z_j)$, $\forall i \neq j$. However, we do not observe those z directly. Instead, we observe K linear combinations of those variables

$$x = Az + \epsilon,$$

with $A \in \mathcal{R}^{K \times K}$ some *unknown* full rank matrix and, again, with additive Gaussian $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In our example, this may correspond to K microphones placed in the space and each recording some linear combination $A_i^T z$ of the speech signals. This scenario is also called *blind source separation*, the term "blind" referring to the idea that we know almost nothing about the "sources" z_i, apart from some general statistical properties. In linear ICA, the function g: $\mathcal{Z} \rightarrow \mathcal{X}$, $g(z) = Az$ that maps from *sources* z to *observations* x is called the *mixing function*. This basic

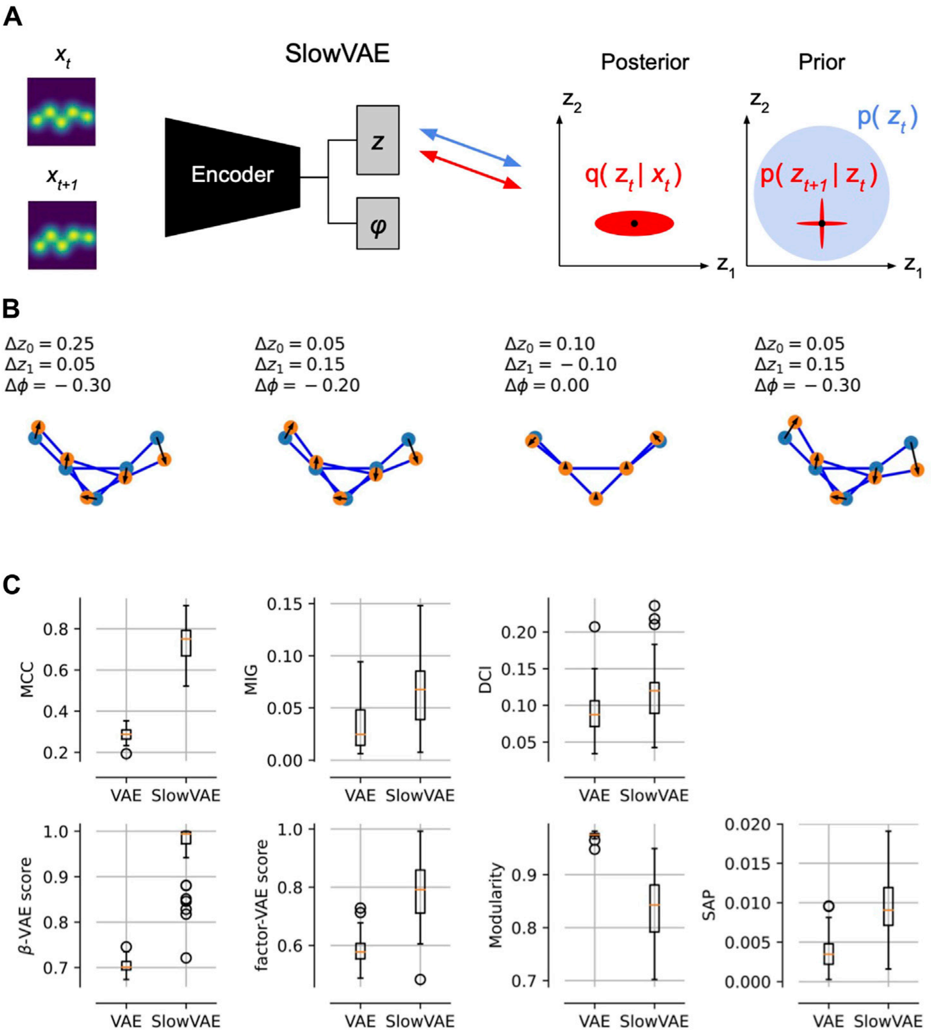

FIGURE 4
Temporal Conformation Disentanglement. **(A)** Schematic of temporal data pairs (x_t, x_{t+1}) and minimal changes to VAE training. To obtain SlowVAE, all we need to do is change the conformation prior for the second time step to be a Laplace distribution centered around the posterior mean of the previous time step (Klindt et al., 2020). **(B)** Example temporal transitions in the two conformation (z_0, z_1) and one pose ϕ latents, drawn from a Laplace distribution (default parameters from Klindt et al. (2020): data rate λ =1, VAE (β = 1, γ = 10), VAE prior rate λ' = 10. **(C)** Most common disentanglement metrics, for details see (Locatello et al., 2019). In seven out of eight metrics we see that SlowVAE learns a more disentangled representation of the conformation latents than a regular VAE without any adaptations.

case of linear ICA has been well-studied in the machine learning and signal processing literature (Hyvärinen and Oja, 2000). Briefly, under the simple assumption that at most one of the sources z_i follows a Gaussian distribution, we can find an *unmixing function* $f: \mathcal{X} \rightarrow \mathcal{Z}$ that approximately inverts the mixing function, in practice up to $(f \circ g)(z) \sim_C z$, i.e., some simple equivalence class $\sim_C$ such as permutations and scalings.[1]

If $g^*(z^*, \phi)$ in Eq. 1 was linear in z and ϕ, then the cryo-EM reconstruction problem would amount to a simple linear ICA problem with the (extended) sources $(z \| \phi)$ where $\|$ denotes concatenation. Unfortunately, the cryo-EM mixing function $g(z, \phi)$ in Eq. 1 is nonlinear. This can be easily appreciated, e.g., by noting that a multiple of some latent αz will not produce the same output as an equally scaled image $g(\alpha z, \phi) \neq \alpha g(z, \phi)$ which would be just the same image but changed in brightness. In the case of a nonlinear mixing function g, (Hyvärinen and Pajunen, 1999) showed that it is possible to construct many functions $f: \mathcal{X} \rightarrow \mathcal{Z}$ that turn the data into independent variables. However, most of these independent variables have no intelligible relationship with the true sources z. This problem is called the lack of *identifiability* of the model, which in general mathematical terms means lack of uniqueness of the solution.

For cryo-EM models this would mean that we can learn *latent spaces* whose individual dimensions have no principled relationship with the true degrees of freedom in molecular conformations. As an example, we may end up with a representation of the simple two dimensional molecule from above where traversing any single dimension in the latent space of our model corresponds to complex

[1] Precisely, writing $f(x) = Wx, \forall z \in \mathcal{Z}: (f \circ g)(z^*) \sim_C z^* \Leftrightarrow WA = DP$ where D is a diagonal matrix and P is a permutation.

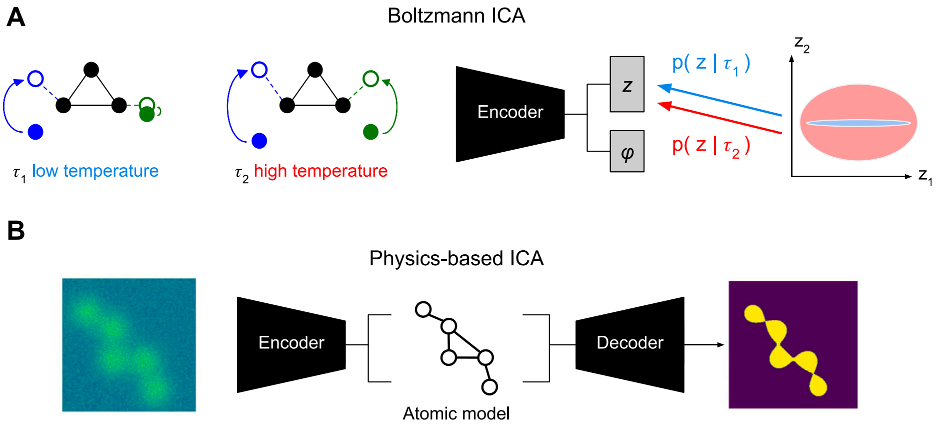

FIGURE 5
Alternative Conformation Disentanglement. **(A)** Proof-of-concept illustration of Boltzmann ICA for the use of temperature as a conditioning variable (Khemakhem et al., 2020). At low temperature τ_1 (left) only the left arm of the molecule shows significant movement. At high temperature τ_2 both arms show significant movement. Conditioning the prior of the conformation latent with this information may allow identifying the distinct sources. **(B)** Proof-of-concept illustration of Physics-based ICA using physics-based decoder with atomic models that assign mechanistic meaning to latent variables, e.g., in terms of atomic coordinates, (potentially, sparse) movement of volume (Punjani and Fleet, 2021), or local normal mode analysis (NMA) deformations (Nashed et al., 2022; Koo et al., 2023).

combinations of the two arm movements (Figure 1). This would, likely, bias our interpretation of how they are articulated together to carry out their function. Thus, without further restrictions on our model, we would fail to discover the simple and elegant structure where the molecule just changes conformation along two independent degrees of freedom, i.e., left and right arm. In the modern machine learning context, finding a latent space that separates the underlying factors of variation is often called *disentanglement* (Bengio et al., 2013), but it has to be noted that the meaning of that term is quite vague.

Fortunately, recent advances propose ways to solve this problem with nonlinear ICA (Hyvärinen et al., 2024). For example, Khemakhem et al. (2020) adds conditioning ("auxiliary") variables u that change the source distributions $p(z_i|u)$. Such a u could represent extra measurements by another modality, or it could be defined by interventions. The model then becomes identifiable if the u modulates the distribution of z strongly enough. This is possible because then the z_i are conditionally independent for any u, which provides much stronger constraints than the mere (unconditional) independence of the z_i as in the basic ICA framework. Khemakhem et al. (2020) further propose to estimate this model using variational methods, leading to an algorithm which is a variant of VAEs. An alternative approach is possible by assuming temporal dependencies of the source time series (Hyvarinen and Morioka, 2017; Klindt et al., 2020; Hälvä et al., 2021); spatial dependencies can also be used (Hälvä et al., 2024). In this case, independence of the components over time lags leads, again, to more constraints, and thus to identifiability under some conditions. A very different approach can be developed by constraining the nonlinear function g, parameterizing it with such a small number of parameters that identifiability is obtained (Hyvärinen et al., 2024; Section 5.4); for example, if we know the physics underlying g (i.e., pose transformations and projections) we may also be able to obtain identifiability. Finally, we point out that the independence assumption can be relaxed (Träuble et al., 2021); even causal relationships between the independent components have been modeled, but this requires further constraints and assumptions (Träuble et al., 2021; Morioka and Hyvärinen, 2023; Yao et al., 2023). Any such learning is easier if interventions on the system are possible (Ahuja et al., 2023) or if it is assumed that the system undergoes sparse, discrete state changes like in robotics experiments (Locatello et al., 2020b), but the theory mentioned above is specifically *unsupervised*, thus not necessarily requiring interventions.

In this work, we will argue that *the heterogeneous reconstruction problem in cryo-EM should be framed as a non-linear ICA problem* to help us build better and more interpretable models that separate the independent degrees of freedom with which molecules change conformation in nature. Few prior works have applied *linear* ICA to molecular imaging (Borek et al., 2018; Gao et al., 2020), finding more meaningful separation of molecular conformation changes (Figure 2). However, to the best of our knowledge, none of the nonlinear ICA approaches mentioned above have so far been applied to the field of computational cryo-EM. Below we will propose three promising candidate approaches for solving the nonlinear ICA problem in cryo-EM.

2.2 Disentangling pose and conformation

The first challenge in computational cryo-EM is that of separating the pose ϕ of a molecule from its conformation z. Again, looking at the cryo-EM mixing function (Eq. 1) $x = g^*(z^*, \phi^*) + \epsilon$, this means we want to find a representation

$$f: \mathcal{X} \to \mathcal{Z} \times \Phi, \quad f(x) = \left(f_z(x) \| f_\phi(x)\right) = (z \| \phi)$$

that separates the estimated conformation z and the pose ϕ.[2] In other words, as a first step, we could find just two latent *subspaces* without specifying the individual components, or the bases, inside those

[2] In the common framework of VAEs, $f_z(x) = \mu_z(x)$ could be defined as the mean of the variational posterior; in an auto-decoding framework this could be the MAP outcome of inference, i.e., $f_z(x) = \text{argmax}_z\ p(x|z)p(z)$.

subspaces. Again, if $g^{\star}$ were linear, we might use the well-developed methods of independent subspace analysis, or subspace ICA, to approach this problem (Hyvärinen and Hoyer, 2000; Theis, 2006). This might also help with the pose variables' topology that is, typically, not Euclidean. For instance, a circular pose variable $\phi_i \in S^1$ that lives on a circle and encodes rotations around one axis could not be represented, by a single dimension, in a typical latent variable model that maps to real valued scalars $f: \mathcal{X} \rightarrow \mathbb{R}^K$. However, some subspace variants of ICA provide exactly such a transformation into spherical coordinates (Hyvärinen et al., 2009, Ch. 10).

Many cryo-EM models use separate latent spaces to represent conformation and pose (Donnat et al., 2022). However, that does not mean that models learn, during nonlinear optimization, to actually use those separate spaces in the intended way. Recent work by Edelberg and Lederman (2023) demonstrated that this is a problem in popular cryo-EM models such as CryoDRGN (Zhong et al., 2021a). In particular, they showed that a 90° rotation of an image causes a different prediction in the space of conformation latent variables, even though those should be *invariant* to pose transformations (see below, 3.1).

2.3 Disentangling independent factors of conformations

A more fundamental challenge is that of separating the independent degrees of freedom of a molecule. Specifically, we want to find a representation f_z of the molecular conformation that inverts, up to some equivalence class $\sim_C$ like permutations and scaling (see above), the *ground truth* generative model $(f_z \circ g^{\star})(z) = z \sim_C z^{\star}$.

A popular approach (see CryoDRGN tutorial) consists in fitting a nonlinear model to cryo-EM data (Figure 1B) followed by manual investigation of the learned latent space that represents conformational heterogeneity (Zhong et al., 2021a) (Figure 1C), thus limiting our ability to quantitatively compare models. Here, we propose a possible remedy in the shape of benchmarks where we simulate data using the generative model (Eq. 1) to assess how close different methods get to the *correct* (i.e., up to $\sim_C$) representation of conformational latent spaces. This taps into a rich, recent literature in nonlinear ICA methods (Hyvärinen et al., 2024) including benchmarks and metrics for model comparisons (Locatello et al., 2020a).

Once we have benchmarks and metrics, we can measure quantitative progress. However, none of the existing heterogeneous reconstruction approaches in computational cryo-EM are identifiable—mirroring the state of the disentangled representation machine learning field in 2019 (Locatello et al., 2019). To actually make progress, in this perspective, we propose three potential approaches to apply nonlinear ICA method for the unsupervised discovery of molecular conformational changes:

1. *Time-resolved single particle imaging.* Observing conformational changes over time, such as a sparse change in a single conformational degree of freedom, provides valuable information; this relies on nonlinear ICA methods that use temporal autocorrelations of the sources (Section 4.2.1).
2. *Boltzmann ICA.* It may be possible to disentangle conformational degrees of freedom by sampling at different temperatures; this relies on nonlinear ICA methods that use additional conditioning variables u, like temperature (Section 4.2.2).
3. *Atomic models.* Building constraint models, with knowledge of the physical mechanism, may exclude faulty solutions; this relies on reducing the hypothesis class provided by the (typically over-parameterized) neural network models (Section 4.2.3).

3 Disentangling pose and conformation

In this section we will first discuss the problem of separating pose and conformation in cryo-EM latent variable models. Recent experiments by Edelberg and Lederman (2023) demonstrated that this desired disentanglement is, unfortunately, violated in the case of CryoDRGN (Zhong et al., 2021a). We start by proposing more systematic evaluations and metrics to measure progress on this task (Section 3.1). Note that those are not metrics in a strict mathematical sense, but rather indices that allow us to measure progress. These metrics inspire simple supervised intervention experiments that can be executed in simulation and added to existing training pipelines to disentangle pose and conformation in cryo-EM latent spaces (Section 3.2).

3.1 Evaluating disentanglement of pose and conformation

We are interested in the cryo-EM *ground truth* generative function $g^{\star}(z^{\star}, \phi^{\star}) = \pi_{\phi^{\star}}(v_{z^{\star}}^{\star})$ (Eq. 1), which consists of a *known* pose function $\pi_{\phi^{\star}} = \pi(., \phi^{\star})$, an *unknown* volume function $v_{z^{\star}}^{\star} = v^{\star}(., z^{\star})$, an *unknown* pose $\phi^{\star}$ and an *unknown* conformation $z^{\star}$. Now, for any specific image, we have full control over the pose function $\pi_{\phi^{\star}}$ but do not know the pose $\phi^{\star}$; however, for the conformation we have neither control over the volume function $v_{z^{\star}}^{\star}$ nor knowledge of the *true* conformation $z^{\star}$ (Shannon et al., 1959). Consequently, in this section and in Section 3.2 we leverage the fact that we have complete knowledge about the pose function, to measure and constrain the flexibility of the conformation z and volume function v_z that we learn in our model.

Put simply, what we want is that, for a molecule with fixed conformation, our model predicts the same conformation even if we change the pose of the molecule. That is, we want the conformation representation f_z to be *invariant* to pose changes. Additionally, we want the pose representation f_ϕ to be invariant to conformation changes. Mathematically, the requirements of invariance can be written as

$$(f_z \circ g)(z, \phi) = z \quad \text{and} \quad \left(f_\phi \circ g\right)(z, \phi) = \phi,$$

for all possible poses $\phi \in \Phi$ and conformations $z \in \mathcal{Z}$. When we train a model, this can go wrong both in our *encoder* $f(x)$ (if it fails to separate pose and conformation), or in our *decoder* $g(z, \phi) = \pi(v_z, \phi)$ (if the volume function v_z learns to represent pose changes). Moreover, it may be necessary to add observation noise to the generated images $g(z, \phi) + \epsilon$ to mitigate for domain shift between the training data and these simulations. To measure progress in this challenge, we can turn this into six different evaluation metrics. We introduce those six in Appendix A. In practice, a single metric (Alg. 1, Eq. 2) seems to suffice as we will discuss in the next section.

3.2 Correcting disentanglement of pose and conformation

```
1. **Encode** a batch $(x_i)_{i=1,\ldots,N}$ of images into conformations and poses $(z_i \parallel \phi_i) = f(x_i)$
2. Detach all $z_i$ and $\phi_i$ from the computation graph.[a]
3. Random shuffle the poses $\phi_i' = \phi_{\sigma(i)}$ [b]
4. Decode and encode the new conformation and pose pairs into $(\hat{z}_i \| \hat{\phi}_i) = (f_{\theta_1} \circ g_{\theta_2})(z_i, \phi_i')$
5. Measure the distances $d(.,\ .)$ to the original latents[c]

   $$\mathcal{L}(f_{\theta_1}, g_{\theta_2}) = \frac{1}{N} \sum_i^N d(z_i, \hat{z}_i) + d\left(\phi_i', \hat{\phi}_i\right) \qquad (2)$$

6. Optimize the encoder and decoder $(f_{\theta_1}, g_{\theta_2})$ along the derivatives $(\frac{\partial \mathcal{L}}{\partial \theta_1}, \frac{\partial \mathcal{L}}{\partial \theta_2})$
7. **Repeat** 1. to 6. until convergence; or add $\mathcal{L}(f_{\theta_1}, g_{\theta_2})$ to total loss function in regular training

[a] We treat these as given latents and do not differentiate with respect to their initial computation
[b] Using a random permutation $\sigma$ instead of a perturbation $\delta_\phi$ ensures that we stay within the posterior distribution $p(\phi|x)$ of poses
[c] In the conformation space, this could just be Euclidean; in pose space we would have to compute, e.g., the geodesic distance in $SO(3)$.
```

Algorithm 1 Interventions for Pose and Conformation Disentanglement.

Based on the ideas proposed in the previous section and the metrics in App. 5.1, we are now going to propose a simple penalty term that can be added to existing cryo-EM models to disentangle pose and conformation. The logic behind these intervention experiments is illustrated in Figure 3. This procedure relies on the physics-based decoder g with an explicit pose representation π_ϕ. Typically, the representation f_{θ_1} and the generator g_{θ_2} are parameterized as neural networks with learnable parameters θ_1 and θ_2 (Kingma and Welling, 2013; Zhong et al., 2021a). Clearly, we can compute the gradients of all metrics with respect to those parameters. In practice, we observed that good results can be achieved simply by following Algorithm 1. Note that this is a straightforward, supervised learning objective that is a relatively standard problem in modern machine learning and should present little difficulty. Thus, we can just add $\mathcal{L}(f_{\theta_1}, g_{\theta_2})$ (Algorithm 1) as an additional penalty term to the loss function of any of the existing models with separate pose and conformation representation to encourage disentanglement.

Importantly, we are only able to write this approach in such a concise and easy form because of the physics-based decoder g. By this we mean the fact that we know the physics, i.e., optics behind the projection π_ϕ in the image formation model (Eq. 1). We could imagine a different cryo-EM generative model where both the conformation and the pose change are modeled by the implicit volume representation v: $\mathbb{R}^3 \times \mathcal{Z}'$ with some extended latent space $\mathcal{Z}'$. Or, in even more general terms, we could just train a standard VAE (Kingma and Welling, 2013) on cryo-EM images to learn a neural network encoder f: $\mathcal{X} \rightarrow \mathcal{Z}''$ and decoder g: $\mathcal{Z}'' \rightarrow \mathcal{X}$ back to image space with some, potentially, even more abstract latent space $\mathcal{Z}''$ Miolane et al. (2020). However, such abstract models would not have the built-in physics of objects in space, their poses ϕ and their projections π_ϕ onto a two dimensional image, which we assume *a priori* in our standard cryo-EM decoder (Eq. 1). In other words, such more abstract models would lack the architectural distinction between z and ϕ which we need in our intervention experiment to disentangle pose and conformation. Thus, we would not be able to manipulate distinct parts of the extended latent spaces ($\mathcal{Z}'$ or $\mathcal{Z}''$), knowing that those represent distinct physical manipulations of the image.

For models using an implicit representation of the volume, the reason we have to use this interventional approach to disentangle pose and conformation in the first place is that the implicit representation $v(z)$ is a highly flexible neural network that can easily model pose changes (Sitzmann et al., 2020). Only by combining this with the constraints physics (i.e., the image formation optics) are we able to disentangle pose and conformation representation. This is akin to disentanglement approaches that use the assumption of sparse manipulations, i.e., pairs of data points where only subsets of the latents are modified (Locatello et al., 2020b). Those models have been demonstrated to solve the nonlinear ICA problem theoretically and practically. Thus, whenever we know something about the physics of the world it makes our representation learning task much simpler if we can run intervention experiments that test the causal dependencies between our latent variables (Ahuja et al., 2023; Squires et al., 2023).

We performed a small *proof-of-concept* experiment to test these predictions and report the results in Figure 3. We train a standard VAE (with separate pose and, implicit, volume representation) on pseudo cryo-EM data and compare it to the same model but with the additional training step in Algorithm 1. We refer to that model as PoseVAE. In Figure 3B, we see that the additional penalty term in PoseVAE does, indeed, succeed at lowering the pose disentanglement metric (Eq. 2). Inspecting the latent representations (Figure 3E, middle), we observe that the pose is now fully confined to the pose variable. Moreover, we observe that the conformation space z is, itself, becoming more disentangled (Figure 3E, right). Intuitively, this makes sense because less can go wrong now in encoding two instead of three variables into it. Quantitatively, this observation is confirmed by standard disentanglement metrics showing that PoseVAE achieves higher mean correlation coefficient (MCC) both across all latents (Figure 3C, left) but also within the conformation latents z alone (Figure 3, right). This is encouraging for the next task of disentangling the conformations.

4 Disentangling conformations

Analogously to the previous section, in a second step, we propose a theoretical framework with metrics and benchmarks concerning the further disentanglement of the individual components inside the conformation vector z (Figure 1). This addresses the essential challenge of interpretable cryo-EM conformational representations for heterogeneous reconstruction (Section 4.1). These benchmarks will help us measure true progress in the budding field of computational cryo-EM (June 2023: Cryo-EM Heterogeneity Challenge). Lastly, we discuss different methods to leverage recent development in nonlinear ICA that have the potential to build the next-generation of cryo-EM models that get closer to the true answer (Section. 4.2). These models may require future technological advancements such as temperature dependent cryo-EM (Bock and Grubmuller, 2022) or time-resolved single particle (X-ray) imaging (Shenoy et al., 2023a; b).

4.1 Evaluating disentanglement of independent factors of conformations

We consider the disentanglement of the individual components or dimensions inside the conformation space z. We propose a metric, in the form of a computational procedure, to evaluate whether the independent components of conformational variations are disentangled in the latent space. Essentially, this proposal relies on simulated data that exactly fulfills the generative model (Eq. 1) which we assume for cryo-EM data. Having full control over the generative model is important, not only to measure progress, but also to simulate extended datasets (e.g., time-resolved imaging), because we know that only those future datasets with additional assumptions will, provably, allow progress in disentangling conformations. Otherwise, this challenge is hopeless (Locatello et al., 2019). Thus, in the ideal case, we have access to a good cryo-EM simulator $g^{\star}$, so we can just use this.

However, if we do not have a good cryo-EM simulator, we can just use the existing state-of-the-art model and see if we can recover its latents. More precisely, we can do the following: Train a regular cryo-EM model with the additional training loss (Algorithm 1) that ensures that pose and conformation are disentangled. Then, check that the model is approximately, invertible. This is a common assumption in ICA theory (Hyvärinen et al., 2024) to make sure that the task of recovering the sources is well-defined. For this, we basically want to make sure that no two distinct points in conformation latent space $z_1 \neq z_2$ would lead to the same volume representation $v(z_1) = v(z_2)$. Once this is, approximately, validated we can use the model as a cryo-EM simulator. Intuitively, we now treat this first model as *ground truth* generator $g^{\star}{:=}g$ and see if we can recover its latents $z^{\star}{:=}z$.

The procedure to evaluate and benchmark heterogeneous cryo-EM latent variable models would then be to assess how well they learn the same (up to equivalences) conformation latent space as the original model. Thus, we would effectively sample a ground truth pose $z^{\star}$ and some random pose $\phi^{\star}$ and feed them into the *ground truth* model to obtain an image $x = g^{\star}(z^{\star}, \phi^{\star}) + \epsilon$. We then process this image with a candidate model $f_z(x) = z$ to obtain the learned conformation representation z. Consequently, we have to compare the two vector representations $z^{\star}$ and z. Depending on the equivalence class ($\sim_C$) that we are interested, there are many different metrics to assess how well $z^{\star}$ is disentangled in z. Intuitively, we want some kind of one-to-one correspondence between the two representations where changing a single entry in $z^{\star}$ corresponds to changing a single entry in z, and *vice versa*. Fortunately, this problem has been studied extensively in the machine learning subfield of disentangled representation learning (Bengio et al., 2013), with many proposed metrics and standardized benchmarks (Locatello et al., 2020a). We can build on those advances to get better quantitative measures on progress in heterogeneous cryo-EM reconstruction than volume based comparisons.

As an example metric measuring disentanglement, we will discuss the Mean Correlation Coefficient (MCC) (Hyvarinen and Morioka, 2017). Intuitively, we want each learned latent variable to be perfectly correlated (or anti-correlated, since sign flips do not compromise interpretability) with a single source variable. To measure this we can just compute the (absolute) correlation coefficient between all *ground truth* latents $z^{\star}$ and all learned latents z. To account for permutations, we have to solve a linear sum assignment with a permutation σ: $\{1, \ldots, K\} \rightarrow \{1, \ldots, K\}$, which basically finds the best matching $z^{\star}_{\sigma(i)}$ for each z_i. The MCC is then, simply, the mean over those matches

$$MCC(z, z^{\star}) = \max_{\sigma} \frac{1}{K} \sum_{i}^{K} |\mathrm{corr}(z_i, z^{\star}_{\sigma(i)})|$$

with corr() denoting correlation. Other metrics focus on decodability, or informational independence (Locatello et al., 2019) and there is no agreed-upon consensus on the optimal disentanglement metric. Thus, we simply report scores across all metrics–these can be further grouped by rank ordering to get overall model comparison scores (see Klindt et al., 2020).

4.2 Correcting disentanglement of independent factors of conformations

Let us now discuss the hardest task, i.e., finding the independent degrees of freedom that determine the conformation of a molecule (Figure 1). This is a hard problem in the sense that, for a nonlinear function g (Eq. 1), without any additional assumption it has been known for the last 2 decades that this is, practically, impossible (Hyvärinen and Pajunen, 1999). Moreover, the field of disentangled representation learning (Bengio et al., 2013) has spent multiple years proposing methods that were, ultimately, unidentifiable (Locatello et al., 2020a). Going forward, computational cryo-EM should learn from those lessons and avoid the same pitfalls. As a very basic example, if our conformation latent space has an isometric Gaussian prior, as in standard VAEs (Kingma and Welling, 2013), we can always perform a random rotation on the learned latents without changing the likelihood of the model (Hauberg, 2018). Thus, any direction in latent space may be representing the actual isolated change in conformation of the molecule. Fortunately, recent years have seen the development of different methods that solve the problem of nonlinear ICA (Hyvärinen et al., 2024). Below, we propose different approaches that are in technological reach (Section 4.2.1), or that make additional statistical assumptions that fit cryo-EM data (Sectioin 4.2.2) or that integrate additional physical knowledge to constrain the problem (Section 4.2.3).

4.2.1 Time-resolved single particle imaging

If we had temporal data of conformation changes for the same molecule over time, we could start applying nonlinear ICA methods that depend on temporal autocorrelations of the sources (Hyvarinen and Morioka, 2017; Hälvä et al., 2021). Specifically, these methods operate under the assumption that we are able to record a time series like

$$(x_t, x_{t+1}) = (g(z_t, \phi_t), g(z_{t+1}, \phi_{t+1}))$$

with temporal dependencies in the sense that

$$p(z_t, z_{t+1}) = p(z_{t+1}|z_t)p(z_t) \neq p(z_t)p(z_{t+1}).$$

For instance, SlowVAE (Klindt et al., 2020) assumes that the transitions

$$\Delta_t^{(i)} := z_{t+1}^{(i)} - z_t^{(i)}$$

follow some sparse distribution, like a Laplace $\Delta_t^{(i)} \sim \mathrm{Lap}(\mu = 0, \lambda > 0)$, and the transitions are independent between components, i.e.,

$$p\left(\Delta_t^{(i)}, \Delta_t^{(j)}\right) = p(\Delta_t^{(i)})p\left(\Delta_t^{(j)}\right) \quad \forall i \neq j.$$

Klindt et al. (2020) showed that those assumptions are often verified on natural video data, which is important since making additional statistical assumptions to obtain identifiable models is only useful if those assumptions are, actually, aligned with the statistics of real world data.

Practically, such a model leads to a minimal modification to standard VAE training, where now the temporal difference of latents is also penalized to follow the specified transition distribution. However, the crucial difference in this learning paradigm is having access to temporal data (x_t, x_{t+1}). While this is not routinely feasible experimentally, efforts to develop time-resolved cryo-EM (Mäeots and Enchev, 2022; Lorenz, 2024) will eventually enable the direct observation of protein dynamics in the microseconds to seconds range, yielding datasets where each particle image will be associated to a timestamp that can be readily deployed in the modified modeling approach above.

We performed these disentanglement experiments in Figure 4. In the first row, we have a demonstration of sparse transitions (drawn from a Laplace distribution) that show changes in some of the latent variables (z_0, z_1, ϕ). Below, we trained $N = 50$ models with and without temporal prior (Klindt et al., 2020) and measure seven typical disentanglement metrics (Locatello et al., 2019). We observe that in six out of seven of those metrics, the model with temporal prior, SlowVAE, does, indeed, achieve higher disentanglement scores. Further improvements could be achieved by including the pose loss from the previous section. However, this is already a promising proof-of-concept for future disentanglement of conformation latents based on temporal data.

4.2.2 Controlling the Boltzmann distribution

The idea above applies to time-resolved experiments studying transient dynamics, triggered by some process such as mixing with a ligand or light excitation. Another class of experiments is concerned with steady-state dynamics where timestamps labels are not helpful. For those experiments, a potentially useful knob that could help solve the non-linear ICA problem is knowledge of the *temperature* associated with each particle in the dataset. The effect of cooling has been reviewed (Bock and Grubmuller, 2022) and different studies have used temperature to change the conformation distribution to obtain insights (Chen et al., 2019; Mehra et al., 2020). Experimentally, this could either be achieved by freezing grids with different cryogens, although a preferable approach would follow the development of thermochromic molecular probes able to report on the local temperature on the cryo-EM grid (Kortekaas and Browne, 2019). This way, precise temperature labeling of each particle in the dataset could be achieved.

Formally, manipulating the temperature τ would provide us with control over the Boltzmann distribution of molecular conformations

$$p(z|\tau) = \frac{1}{Q(\tau)}\exp\left(-\frac{\varepsilon(z)}{k(\tau)}\right),$$

with $Q(\tau)$ the canonical partition function and $\varepsilon(z)$ the energy of being in conformation z. This conditional distribution, where we assume knowledge or experimental control over the temperature, maps onto the theoretical framework of iVAE (Khemakhem et al., 2020) with $u = \tau$. Future theoretical investigations are needed to verify if the additional assumptions for their identifiability results are fulfilled in this setting.

However, to build intuition, we can walk through a thought experiment to see how control of the temperature can suffice to discover the independent degrees of freedom in molecular conformation changes (Figure 5A). Assume, again a molecule with two degrees of freedom $z_1, z_2 \in \mathbb{R}$ that both follow temperature-dependent normal distributions

$$p(z_1|\tau_a) \sim \mathcal{N}(\mu_1, \sigma_1^2(\tau_a)) \quad \text{and} \quad p(z_2|\tau_a) \sim \mathcal{N}(\mu_2, \sigma_2^2(\tau_a)).$$

Now, suppose that at low temperature τ_a, we only see variation in the first component z_1 while the second component is nearly constant, i.e., $\sigma_2^2(\tau_a) \ll \sigma_1^2(\tau_a)$. By contrast, at high temperature $\tau_b > \tau_a$, we see that the second component also starts moving, i.e., $\sigma_2^2(\tau_a) \gg 0$. Thus, using temperature alone, we can successfully isolate the different degrees of freedom. Intuitively, this should make it possible to solve the disentanglement task. We could simply fit a model to the data at temperature τ_b with the additional constraint that the same model also has to be able to encode the data at temperature τ_a, albeit, with only the first latent dimension z_1. Whether those assumptions bare out in real molecules is not clear, yet, recording at different temperatures is within closer technological reach than time-resolved SPI.

4.2.3 Atomic models

While the previous two proposals require different data, we may also hope to make progress with different models. We saw how the implicit volume representation needs additional care to disentangle pose and conformation information (Section 3.1). Imbuing the generative model with physics inspired structure, allowed us to separate pose from conformation (section 3.2). Maybe, even more physics can help us solve the harder problem of finding the conformational degrees of freedom. In particular, if we replace the highly expressive implicit volume representation v with an atomic model (Zhong et al., 2021b; Rosenbaum et al., 2021; Nashed et al., 2022; Koo et al., 2023), then the pose latent variable z will have to encode how an atomic reference structure (maybe the mode of the conformation space) is deformed into an observed conformation (Figure 5B).

One existing work by Punjani and Fleet (2021) proposes to learn a convection field that deforms a reference volume. This comes with the elegant property of *volume preservation* which is not always the case in implicit conformation representations, but obeys our knowledge of the underlying physics. However, the learned convection fields as well as the reference volume model are still over-parameterized compared to an ideal atomistic model with movement vectors for each atom. The problem in building smaller and more constraint models is that modern machine learning methods have, to some extent, proven so powerful because they allow heavily overparameterized hypothesis classes that still generalize well beyond the training data. The question then becomes whether we can combine the best of both worlds, i.e., the non-convex optimization and generalization properties of deep neural networks (e.g., for implicit volume or convection representations) with the physical detail of constraint atomistic models (Zhong et al., 2021b; Rosenbaum et al., 2021; Nashed et al., 2022; Koo et al., 2023).

5 Discussion

In recent years, the integration of computational models, particularly VAEs (Kingma and Welling, 2013), has revolutionized research across various natural sciences, including cryo-EM (Zhong et al., 2021a). This perspective piece underscores the critical importance of understanding conformational latent spaces in cryo-EM by drawing on cutting-edge theoretical advancements in identifiable nonlinear ICA (Hyvärinen et al., 2024). By bridging the gap between theoretical frameworks and practical applications in cryo-EM, we are suggesting a significant advance in the interpretability and utility of latent variable models. Furthermore, our study advocates for the adoption of better quantitative measures to assess progress in heterogeneous cryo-EM reconstruction, transcending traditional volume-based comparisons. This aligns with recent initiatives such as the Cryo-EM Heterogeneity Challenge emphasizing the need for refined evaluation metrics to accurately gauge advancements in this field. Nevertheless, this work is merely an opinion piece and proof-of-concept demonstration. Significant technical (e.g., time resolved SPI) and engineering challenges (e.g., identifiable nonlinear ICA models that work in low SNR regimes) lie ahead on this path towards interpretable cryo-EM conformations spaces.

One of the key limitations to the approach that we are proposing in this paper is the assumption that there exist a limited number of factors of variation that determine the possible conformations and conformation changes out of all the possible rearrangements of constituent atoms. Moreover, we assume that those factors are independent which means that the molecule has parts that move independently of each other. Prior work in nonlinear ICA has considered the effect of dependencies (Träuble et al., 2021) and how to mitigate them, but that might not even be necessary. It is conceivable that, for instance in the little cartoon Figure 1, both arms of the molecule always move up and down together so that they are not independent. However, if both arms always move together, then this is, likely, to fulfill some biological function. In this scenario, our model would presumably learn to describe the combined motion by a single latent variable which would, thus, represent the motion required to perform this biological function. Consequently, such dependencies might unveil (more complex) molecular motions and biological function that we can, thus, extract.

The history of ICA's emergence in the 1990s, and in particular its early adoption in neuroimaging (McKeown et al., 1998), shows its capacity to evolve into a cornerstone of data-based modeling. This trend, moving even further away from hypothesis-driven research (Friston, 1998) toward data-centric approaches (Beckmann et al., 2005), also underscores the importance of incorporating principles like nonlinear ICA to ensure meaningful model outputs. While modern machine learning techniques, including VAEs or nonlinear dimensionality reduction methods such as `t-SNE` (Van der Maaten and Hinton, 2008) or `UMAP` (McInnes et al., 2018), have become ubiquitous in data-based modeling, they often overlook source recovery, i.e., identifiability considerations. To fully harness the potential of latent spaces, it is paramount to ensure their alignment with meaningful representations of the underlying data.

In conclusion, our approach integrates nonlinear ICA principles into the development and analysis of cryo-EM latent variable models, ensuring more interpretable representations that encapsulate the intrinsic structure of the data. Unlocking latent spaces aligned with the underlying fundamental factors governing complex phenomena is pivotal for gaining deep insights into biological processes, expediting drug discovery, and facilitating targeted interventions. This progress extends beyond cryo-EM, resonating with diverse scientific disciplines such as computer vision, natural language processing, and generative modeling, where (VAE) latent spaces play a pivotal role in data representation and the generation of new scientific hypothesis as part of initiatives such as AI4Science. Our interdisciplinary approach, embracing nonlinear ICA and disentanglement models, holds promise in generating meaningful representations that *carve nature at the joints*, thereby propelling transformative discoveries.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DK: Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing–original draft, Writing–review and editing. AH: Conceptualization, Methodology, Validation, Writing–review and editing. AL: Investigation, Methodology, Writing–review and editing. NM: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing–original draft, Writing–review and editing. FP: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the U.S. Department of Energy, under DOE Contract No. DE-AC02-76SF00515. DK was supported by the SLAC LDRD project "AtomicSPI: Learning atomic scale biomolecular dynamics from single-particle imaging data" (PI: FP). AH was supported by a CIFAR Fellowship and the Academy of Finland. NM acknowledges funding from grant NIH 1R01GM144965-01.

Acknowledgments

This work was inspired by discussions at the Flatiron CryoEM Summer Workshop 2023. Specifically, we acknowledge Yifan Cheng's determination to really understand conformational latent spaces. We hope that the present work will help the field of computational cryo-EM to this next level where they can provide truly novel insights into Biology.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. (2023). "Interventional causal representation learning," in *International conference on machine learning* (PMLR), 372–407.

Beckmann, C. F., DeLuca, M., Devlin, J. T., and Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Trans. R. Soc. B Biol. Sci.* 360, 1001–1013. doi:10.1098/rstb.2005.1634

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. pattern analysis Mach. Intell.* 35, 1798–1828. doi:10.1109/TPAMI.2013.50

Bock, L. V., and Grubmuller, H. (2022). Effects of cryo-em cooling on structural ensembles. *Biophysical J.* 121, 148a. doi:10.1016/j.bpj.2021.11.1981

Borek, D., Bromberg, R., Hattne, J., and Otwinowski, Z. (2018). Real-space analysis of radiation-induced specific changes with independent component analysis. *J. Synchrotron Radiat.* 25, 451–467. doi:10.1107/S1600577517018148

Chen, C.-Y., Chang, Y.-C., Lin, B.-L., Huang, C.-H., and Tsai, M.-D. (2019). Temperature-resolved cryo-em uncovers structural bases of temperature-dependent enzyme functions. *J. Am. Chem. Soc.* 141, 19983–19987. doi:10.1021/jacs.9b10687

Comon, P. (1994). Independent component analysis, a new concept? *Signal Process.* 36, 287–314. doi:10.1016/0165-1684(94)90029-9

Donnat, C., Levy, A., Poitevin, F., Zhong, E. D., and Miolane, N. (2022). Deep generative modeling for volume reconstruction in cryo-electron microscopy. *J. Struct. Biol.* 214 (4), 107920. doi:10.1016/j.jsb.2022.107920

Edelberg, D. G., and Lederman, R. R. (2023). *Using vaes to learn latent variables: observations on applications in cryo-em*. arXiv preprint arXiv:2303.07487.

Friston, K. J. (1998). Modes or models: a critique on independent component analysis for fmri. *Trends cognitive Sci.* 2, 373–375. doi:10.1016/s1364-6613(98)01227-3

Gao, Y., Eskici, G., Ramachandran, S., Poitevin, F., Seven, A. B., Panova, O., et al. (2020). Structure of the visual signaling complex between transducin and phosphodiesterase 6. *Mol. Cell* 80, 237–245. doi:10.1016/j.molcel.2020.09.013

Hälvä, H., Corff, S. L., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., et al. (2021). "Disentangling identifiable features from noisy data with structured nonlinear ICA," in *Advances in neural information processing systems (NeurIPS2021) (virtual)*.

Hälvä, H., So, J., Turner, R. E., and Hyvärinen, A. (2024). "Identifiable feature learning for spatial data with nonlinear ICA," in *Proc. Artificial intelligence and statistics (AISTATS2024)* (Valencia, Spain: PMLR).

Hauberg, S. (2018). *Only bayes should learn a manifold (on the estimation of differential geometric structure from data)*. arXiv preprint arXiv:1806.04994.

Hyvärinen, A., and Hoyer, P. O. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* 12, 1705–1720. doi:10.1162/089976600300015312

Hyvärinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural image statistics*. Springer-Verlag.

Hyvärinen, A., Khemakhem, I., and Monti, R. (2024). Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Ann. Inst. Stat. Math.* 76, 1–33. doi:10.1007/s10463-023-00884-4

Hyvarinen, A., and Morioka, H. (2017). "Nonlinear ica of temporally dependent stationary sources," in *Artificial intelligence and statistics* (PMLR), 460–469.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.* 13, 411–430. doi:10.1016/s0893-6080(00)00026-5

Hyvärinen, A., and Pajunen, P. (1999). Nonlinear independent component analysis: existence and uniqueness results. *Neural Netw.* 12, 429–439. doi:10.1016/s0893-6080(98)00140-3

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). "Variational autoencoders and nonlinear ica: a unifying framework," in *International conference on artificial intelligence and statistics* (PMLR), 2207–2217.

Kingma, D. P., and Welling, M. (2013). *Auto-encoding variational bayes*. arXiv preprint arXiv:1312.6114.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., et al. (2020). *Towards nonlinear disentanglement in natural data with temporal sparse coding*. arXiv preprint arXiv:2007.10930.

Koo, B., Martel, J., Peck, A., Levy, A., Poitevin, F., and Miolane, N. (2023). *Reconstructing heterogeneous cryo-em molecular structures by decomposing them into polymer chains*. arXiv preprint arXiv:2306.07274.

Kortekaas, L., and Browne, W. R. (2019). The evolution of spiropyran: fundamentals and progress of an extraordinarily versatile photochrome. *Chem. Soc. Rev.* 48, 3406–3424. doi:10.1039/c9cs00203k

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., et al. (2019). "Challenging common assumptions in the unsupervised learning of disentangled representations," in *In international conference on machine learning* (PMLR), 4114–4124.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., et al. (2020a). A sober look at the unsupervised learning of disentangled representations and their evaluation. *J. Mach. Learn. Res.* 21, 8629–8690.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020b). "Weakly-supervised disentanglement without compromises," in *International conference on machine learning* (PMLR), 6348–6359.

Lorenz, U. J. (2024). *Microsecond time-resolved cryo-electron microscopy*. arXiv preprint arXiv:2402.16519.

Mäeots, M.-E., and Enchev, R. I. (2022). Structural dynamics: review of time-resolved cryo-em. *Acta Crystallogr. Sect. D. Struct. Biol.* 78, 927–935. doi:10.1107/S2059798322006155

McInnes, L., Healy, J., and Melville, J. (2018). *Umap: uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:1802.03426.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., et al. (1998). Analysis of fmri data by blind separation into independent spatial components. *Hum. Brain Mapp.* 6, 160–188. doi:10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1

Mehra, R., Dehury, B., and Kepp, K. P. (2020). Cryo-temperature effects on membrane protein structure and dynamics. *Phys. Chem. Chem. Phys.* 22, 5427–5438. doi:10.1039/c9cp06723j

Miolane, N., Poitevin, F., Li, Y.-T., and Holmes, S. (2020). "Estimation of orientation and camera parameters from cryo-electron microscopy images with variational autoencoders and generative adversarial networks," in *Proceedings of* *the IEEE/CVF conference on computer vision and pattern recognition workshops*, 970–971.

Morioka, H., and Hyvärinen, A. (2023). "Connectivity-contrastive learning: combining causal discovery and representation learning for multimodal data," in *Proc. Artificial intelligence and statistics (AISTATS2023)* (Valencia, Spain: ML Research Press).

Nashed, Y., Peck, A., Martel, J., Levy, A., Koo, B., Wetzstein, G., et al. (2022). *Heterogeneous reconstruction of deformable atomic models in cryo-em*. arXiv preprint arXiv:2209.15121.

Punjani, A., and Fleet, D. J. (2023). 3D flex: determining structure and motion of flexible proteins from cryo-EM. *Nat. Methods*. 20, 860–870. doi:10.1038/s41592-023-01853-8

Rosenbaum, D., Garnelo, M., Zielinski, M., Beattie, C., Clancy, E., Huber, A., et al. (2021). *Inferring a continuous distribution of atom coordinates from cryo-em images using vaes*.

Shannon, C. E., et al. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.* 4, 1.

Shenoy, J., Levy, A., Poitevin, F., and Wetzstein, G. (2023a). "Amortized pose estimation for x-ray single particle imaging," in *Machine learning for structural biology Workshop* (NeurIPS 2023).

Shenoy, J., Levy, A., Poitevin, F., and Wetzstein, G. (2023b). *Scalable 3d reconstruction from single particle x-ray diffraction images based on online machine learning*. arXiv preprint arXiv:2312.14432.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Adv. neural Inf. Process. Syst.* 33, 7462–7473.

Squires, C., Seigal, A., Bhate, S. S., and Uhler, C. (2023). "Linear causal disentanglement via interventions," in *International conference on machine learning* (PMLR), 32540–32560.

Theis, F. (2006). Towards a general independent subspace analysis. *Adv. Neural Inf. Process. Syst.* 19.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., et al. (2021). "On disentangled representations learned from correlated data," in *International conference on machine learning* (PMLR), 10401–10412.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.

Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., et al. (2023). *Multi-view causal representation learning with partial observability*. arXiv preprint arXiv:2311.04056.

Zhong, E. D., Bepler, T., Berger, B., and Davis, J. H. (2021a). Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nat. methods* 18, 176–185. doi:10.1038/s41592-020-01049-4

Zhong, E. D., Lerer, A., Davis, J. H., and Berger, B. (2021b). *Exploring generative atomic models in cryo-em reconstruction*.

Appendix A: Pose and Conformation Disentanglement metrics

We consider disentanglement evaluation metrics for all three learned functions, i.e., the volume $v(z)$, the conformation encoder $f_z(x)$ and the pose encoder f_ϕ. For each function, we will consider two metrics that measure their consistency (i.e., does it model the latent it is supposed to model?) and their invariance (i.e., is it invariant to changes in other latents?). Starting with the *volume metrics*, we measure.

1. *Volume consistency*, i.e., how accurately the learned volume $v(z)$ changes the conformation:

$$\mathcal{L}_{\mathrm{VC}}(v) = \mathop{\mathbb{E}}_{z,\phi}\left[\|z - (f_z^\star \circ g)(z,\phi)\|^2\right] = \mathop{\mathbb{E}}_{z,\phi}\left[\|z - \left(f_z^\star \circ \pi_\phi \circ v\right)(z)\|^2\right]$$

2. *Volume pose invariance*, i.e., how accurately the learned volume $v(z)$ changes *only* the conformation:

$$\mathcal{L}_{\mathrm{VI}}(v) = \mathop{\mathbb{E}}_{z,\phi}\left[\|\phi - \left(f_\phi^\star \circ g\right)(z,\phi)\|^2\right] = \mathop{\mathbb{E}}_{z,\phi}\left[\|\phi - \left(f_\phi^\star \circ \pi\right)(v(z),\phi)\|^2\right]$$

where, ideally, one would use the oracle encoder $f^\star(x) = \mathrm{argmax}_{z,\phi}\, p(x|z, \phi)\, p(z, \phi)$. In practice, we can just use the current encoder $f(x) = (f_z(x)\|f_\phi(x))$ and optimize it as well. Again this can be split into two metrics (consistency and invariance), both for the *conformation encoder* f_z

1. *Conformation-encoder consistency*, i.e., how accurately the conformation encoder f_z recovers any conformation, independent of pose:

$$\mathcal{L}_{\mathrm{CC}}(f_z) = \mathop{\mathbb{E}}_{z,\phi}\left[\|z - (f_z \circ g^\star)(z,\phi)\|^2\right]$$

2. *Conformation-encoder pose invariance*, i.e., how invariant the conformation encoder f_z is to pose perturbations:

$$\mathcal{L}_{\mathrm{CI}}(f_z) = \mathop{\mathbb{E}}_{z,\phi,\delta_\phi}\left[\|(f_z \circ g^\star)(z,\phi) - (f_z \circ g^\star)\left(z, \phi + \delta_\phi\right)\|^2\right]$$

as well as for the *pose encoder* f_ϕ

1. *Pose-encoder consistency*, i.e., how accurately the pose encoder f_ϕ recovers any pose, independent of conformation:

$$\mathcal{L}_{\mathrm{PC}}\left(f_\phi\right) = \mathop{\mathbb{E}}_{z,\phi}\left[\|\phi - \left(f_\phi \circ g^\star\right)(z,\phi)\|^2\right]$$

2. *Pose-encoder conformation invariance*, i.e., how invariant the pose encoder f_ϕ is to conformation perturbations:

$$\mathcal{L}_{\mathrm{PI}}\left(f_\phi\right) = \mathop{\mathbb{E}}_{z,\phi,\delta_z}\left[\|\left(f_\phi \circ g^\star\right)(z,\phi) - \left(f_\phi \circ g^\star\right)(z + \delta_z, \phi)\|^2\right]$$

where, again, ideally we would like to use the *ground truth* generator $g^\star$, but we can also just use the current learned decoder. All of these six metrics can be evaluated, in a supervised way, over a sufficient number of randomly sampled conformations z, poses ϕ and perturbations (δ_z, δ_ϕ).