IADIS International Journal on Computer Science and Information Systems
Vol. 21, No. 2, pp. 145-160
ISSN: 1646-3692

LARGE-SCALE GENERATION OF ANATOMICALLY REALISTIC *CAENORHABDITIS ELEGANS* PHANTOMS FOR DEEP LEARNING-BASED IMAGE RECONSTRUCTION

Anna-Louise Brun[1,2], Ali Rouwane[1], Guillaume Dévès[1], Hervé Seznec[1], Pascal Desbarats[2] and Claire Michelet[1]
[1]*Univ. Bordeaux, CNRS, LP2I Bordeaux, UMR 5797, France*
[2]*Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, France*

ABSTRACT

Micro-tomography techniques, using proton or synchrotron X-ray beams, enable quantitative three-dimensional mapping of mass density and elemental composition in biological samples at cellular and sub-cellular scales. When acquisition is performed on 2D sample supports, however, grazing angles remain inaccessible, producing a substantial "missing wedge" (~40°) in the sinogram and consequent distortions in the reconstructed density maps. Classical interpolation and iterative reconstruction methods fail to recover this information with the precision required for quantitative biological imaging. We present a pipeline for generating large-scale, biologically realistic synthetic datasets with the objective of training a deep learning-based reconstruction method. Using *Caenorhabditis elegans* as a model organism, we built a topologically consistent, watertight three-dimensional (3D) phantom in Blender comprising 87 organs with associated mass-density and chemical-composition values derived from experimental measurements. Geometric variability was introduced through a custom Free-Form Deformation framework built on an anatomically adapted lattice. For this, an inverse-mapping procedure links the phantom's surface mesh to the lattice's parametric space, enabling deformations that follow experimentally observed *C. elegans* postures while conserving its total length and its topology. Residual mesh defects are corrected via automated repair and remeshing. Boolean subtraction resolves organ nesting into disjoint volumes suitable for projection. Tomographic projections are then simulated with an adapted gVirtualXray (gVXR) framework, with imaging modality-appropriate noise added in post-processing. This pipeline generated roughly 2.4 million anatomically realistic, deformation-consistent phantom instances with paired ground truth. The resulting dataset provides the scale and biological fidelity needed to train inpainting networks for missing-wedge correction, ultimately improving the quantitative reliability of 3D reconstructions in proton micro-tomography of biological structures.

KEYWORDS

Micro-Tomography, Synthetic Data Generation, *Caenorhabditis elegans*, Free-Form Deformation, gVXR

1. INTRODUCTION

1.1 Quantitative Mapping of Mass Density and Chemical Elements

Micro-tomography is a powerful tool for investigating biological samples at cellular and subcellular scales. Various radiation sources are utilized, such as proton beams (~1 MeV) and (synchrotron) X-rays (~20 keV), with spatial resolutions ranging from the nanometer to several tens of micrometers, to study cells, organs and multicellular micro-organisms. Several microscopy techniques can be implemented, by collecting either the transmitted signal (*e.g.,* measuring X-ray phase contrast, or proton residual energy after passing through the sample using STIM: Scanning Transmission Ion Microscopy) or the X-ray emission spectrum (SXRF: Synchrotron X-ray Fluorescence, PIXE: Particle Induced X-ray Emission) (Yang et al., 2019, Michelet, et al., 2017). These physical quantities relate directly to the sample's mass density and chemical composition, respectively. In this way, the three-dimensional (3D) internal structure of the object and the distribution of chemical elements can be revealed (Michelet et al., 2015). More specifically, the physical quantity being reconstructed is the mass density, which provides two types of information: *(i)* the mass density of the total material; *(ii)* the mass density of each individual chemical element, showing its specific localization. At LP2IB, the development of tomography using the proton microbeam delivered at the AIFIRA facility (https://www.lp2ib.in2p3.fr/aifira/) enables quantitative, three-dimensional characterization of microscopic samples (Gordillo, N. et al., 2011) and has three main strengths: *(i)* multi-elemental analysis, *(ii)* quantitative measurement in terms of mass, and *(iii)* high sensitivity, reaching a few $\mu g \cdot g^{-1}$. The distribution of inorganic chemical elements provides crucial information, as they serve as markers for many biological mechanisms, thereby revealing them.

1.2 Addressing the Missing Wedge Limitation

The design of reconstruction codes that produce accurate tomographic images, in terms of mass density, is a key factor in drawing meaningful biological conclusions. However, experimental acquisition is often constrained by limited beam time, especially at accelerator facilities; radiation exposure, particularly for fragile biological specimens; and specific sample support configuration, that can limit the range of accessible measurement angles. The acquired data are stored in a "sinogram" representation, where the rotational motion of the sample manifests as projected sinusoidal patterns. An optimal condition for obtaining high-quality tomographic images is to perform the acquisition over a "complete" rotation, *i.e.,* at least 180° accounting for symmetry. This requires mounting the sample at the end of a rotation axis to cover this full range. However, because most biological samples (cells, microorganisms) are deposited on 2D substrates, optimal experimental conditions must enable 3D imaging directly on the 2D sample support. In practice, acquisitions are performed by rotating the sample in front of the beam over a typical angular range of -70° to +70° relative to the support normal. One difficulty with this approach is the inability to obtain data at grazing angles, *i.e.* where the beam is nearly tangent to the sample support. This "missing wedge" problem is particularly challenging to address because the missing angular interval is substantial, approximately 40° in total. This results in

large portions of data missing in the sinograms and, consequently, errors in the reconstructed images.

Two main strategies can be considered for estimating the missing information to enhance the quality of the reconstructed images. The first recovers the missing data directly in the projection domain, *i.e.*, in the sinograms. The second aims to improve the reconstruction process itself. Working in the projection domain is particularly challenging because the "missing wedge" causes a symmetrical loss of information in the sinogram, so classical interpolation methods are inadequate. Patch-based methods (Newson et al., 2017) attempt to estimate plausible values by identifying similar patterns, but they remain limited due to the significant loss of geometric information. In the reconstruction process, classical iterative reconstruction techniques such as SART or SIRT perform poorly in this context. Even when combined with regularization approaches such as total variation (*e.g.,* SART-TV (Boudjelal et al., 2017)), the resulting precision remains insufficient to meet our requirements.

As artificial intelligence has become a powerful tool for imaging applications, its potential for data completion, particularly for incomplete 3D images, has been widely explored. Initial approaches were developed primarily for 2D data, including context-based autoencoders (Pathak et al., 2016), multi-scale generative adversarial networks (GANs) (Iizuka et al., 2017), and more recent methods such as diffusion models (Lugmayr et al., 2022) and vision transformers (Deng et al., 2022). In the 3D domain, GANs have been notably applied to tasks such as correcting metallic artifacts (Wang et al., 2018) and, more specifically, to the missing-wedge problem using joint-GAN architectures (Ding et al., 2019). However, GAN-based methods often suffer from limited output quality and can generate unrealistic features. In the context of CT imaging, the first prize of the 2022 Helsinki limited-angle tomography challenge highlighted the effectiveness of artificial intelligence methods trained on synthetic datasets. The winning approach relied on a U-Net architecture (Germer et al., 2023) and incorporated prior information into the loss function to improve reconstruction quality. More recently, diffusion models have demonstrated strong potential for limited-angle tomography, as shown in studies trained on large-scale datasets such as airport scans (Liu et al., 2023). The issue is that those models are developed to meet a certain level of precision for imaging techniques that mostly rely on qualitative correctness. In our case, synchrotron X-ray and proton micro-tomography are quantitative imaging techniques, so they require greater precision, as they are intended to reflect biological reality and estimate values in each organ accurately.

1.3 Synthetic Data to Support Deep Learning-Based Inpainting

At LP2IB, key areas of research rely on the biological model nematode *Caenorhabditis elegans* (*C. elegans*) (Le Trequesser et al., 2014). This complex multicellular microorganism, approximately 1 mm in length, exhibits reproducible anatomical and functional characteristics, with consistent parameters for cell number, life cycle, and development (Altun et al., 2002–2024). This organism has thus become an established model for investigating major biological functions (*e.g.,* genetics, aging, and neuroscience). In this context, tomographic imaging yields novel data complementary to conventional imaging modalities (Michelet et al., 2017, McColl et al., 2012). These considerations motivated our interest in using *C. elegans* as a model organism to develop our new sinogram interpolation methods. However, our pipeline is designed to be adaptable to other biological or even non-biological objects with similar

topological constraints. Our objective is to train a neural network for data completion (inpainting) using supervised learning. This approach requires a large training dataset, yet the amount of available experimental data is limited. To overcome this limitation, we propose generating realistic simulated data using 3D modeling of *C. elegans*. This specific objective imposes several constraints; the numerical phantom must meet the following criteria: *i)* anatomically realistic, incorporating both the spatial and quantitative information visible in tomographic imaging, *i.e.* the mass density and chemical content; *ii)* topologically consistent; and *iii)* deformable, to account for geometrical variability.

A detailed numerical phantom was developed in Blender by Caltech for the OpenWorm Project (Szigeti et al., 2014, Sternberg et al., 2024). Based on WormAtlas (McColl et al., 2012), this model captures the complexity of *C. elegans*' internal structure remarkably well. However, it has several limitations for our purposes: *i)* it is not topologically consistent, as some organs intersect; and *ii)* it is not entirely watertight because some structures contain holes *iii)* although it was built in a natural, biologically representative pose, it was not deformable in any way, offering no framework for generating alternative postures. Yet, a correct topology is essential for accurate mesh representation and deformation. Moreover, ensuring that each organ is watertight is critical for assigning mass density and chemical composition to the organ's enclosed volume. This is why, to address these limitations, we developed a custom *C. elegans* model in Blender that prioritizes topological integrity and watertight geometry. Such conditions are essential both for biological realism, since organs do not intersect in vivo, and for imaging, where accurate projection simulation depends on well-defined geometries. Moreover, the processing chain includes a deformation framework that mimics the natural range of motion and shapes observed experimentally in *C. elegans*, ensuring geometric variability consistent with biological realism.

2. MATERIALS AND METHODS

2.1 Overview of the Synthetic Data Generation Pipeline

We constructed a numerical phantom of a young adult hermaphrodite *C. elegans* following several steps (Fig. 1), which are briefly presented here, and described in detail in the following sections. First, the geometry was defined using a Blender model, resulting in a surface mesh that is both topologically consistent, watertight, and free of self-intersections. Moreover, data related to the global mass density and elemental composition of the biological structures were incorporated, based on experimental results obtained at LP2IB (Michelet et al., 2017). This quantitative information is stored in a separate file, which is automatically associated with the different organs during the export process. This export step also includes all additional metadata required to integrate the *C. elegans* model into the simulation pipeline. The next step is to introduce geometric variability into the numerical phantom. For this, a Free Form Deformation (FFD) process is applied on the surface mesh. Deformations were constrained to prevent self-intersections. Specific algorithms were used to verify watertightness, like vertex splitting / edge manifold repair or spatial hashing. Different versions of the deformed *C. elegans* phantoms can be generated depending on the user's needs: a surface or volume mesh for simulation purposes, or a voxelized version for quantitative image analysis. The final step in the

database generation pipeline consists of simulating the acquisition process to generate synthetic projected images. This is performed using the gVirtualXray (gVXR) library (Vidal et al., 2025), producing a sinogram in projection space.

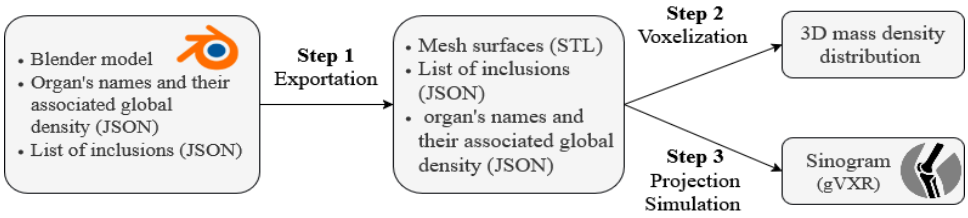


Figure 1. Architecture of the synthetic sinogram generation framework

2.2 Modeling of a Topologically Correct *C. elegans* Phantom

The *C. elegans* model developed in the OpenWorm project (Szigeti et al., 2014, Sternberg et al., 2024) is a highly detailed, comprehensive representation in Blender that includes numerous biological structures of the digestive, nervous, and reproductive systems, hereafter referred to as "organs". Each component is assigned a distinct RGB color to aid visualization. However, as explained in section 1.3, it did not meet our needs because: *i)* it was not topologically correct, *e.g.*, significant intersections between organs were present, and *ii)* it was not deformable. In a first attempt to overcome these limitations, we simplified the geometry (reducing the number of "organs" from 739 to 319) and manually removed all intersections in Blender (Vidal et al., 2025). However, a major limitation remained: the biological medium between the surfaces was treated as empty and could not be represented in this way. A second version of the phantom was therefore developed to address this issue. Each "organ" was then represented by a closed surface enclosing a homogeneous material, *i.e.,* of uniform mass density and composition. A list of inclusions was defined to organize the structures relative to one another. Therefore, the empty spaces between organs could be modeled, using Boolean subtraction operations, to better represent the biological medium and allow the assignment of non-zero mass density and composition between organs. The list of inclusions must also be considered during the voxelization step (step 2 in Fig. 1).

2.3 Anatomically Consistent Deformation using a Custom-Lattice FFD

2.3.1 General Principle

Standard mesh deformation tools in Blender proved inadequate and often introduced additional geometric intersections. Since our objective is to produce smooth, macroscopic deformations while preserving the model's original volume and length, we adopted an FFD method (Sederberg and Parry Scott, 1986). FFD is a flexible technique for smoothly transforming complex 3D shapes by embedding them in a deformation lattice and adjusting its control points. To generate realistic geometric variations, we extend the classical approach by using a lattice adapted to the object's anatomy, enabling more complex and biologically consistent deformations. The process can be summarized in four main steps (Fig. 2). First, the initial and

target shapes are defined by the coordinates of a medial curve of the *C. elegans* phantom (Fig. 3), which will hereafter be referred to as the "spine". From these "spine" coordinates, a smooth lattice representation is constructed. Second, an inverse mapping establishes the correspondence between the phantom's surface mesh and the lattice's parametric space. Third, the displacement between the initial and target lattice control points is computed. Finally, these displacements are applied via FFD to deform the phantom.

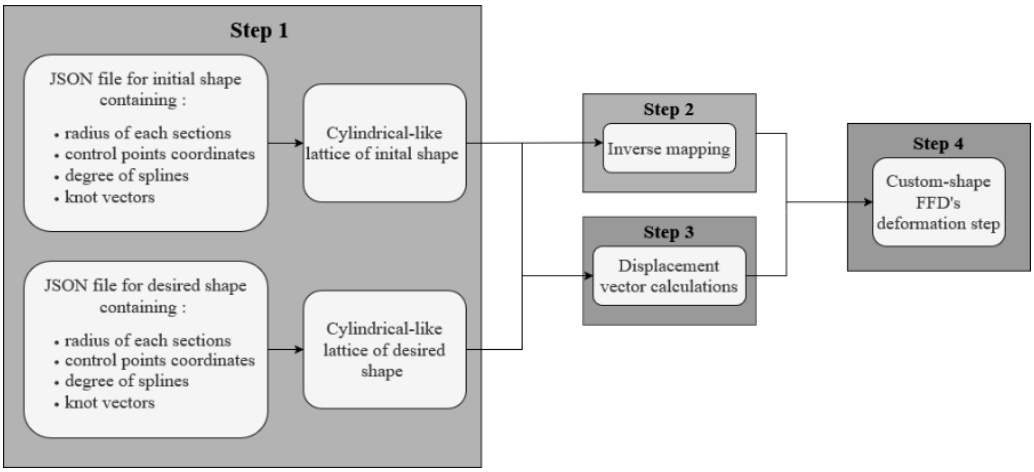


Figure 2. General overview of the custom shape FFD process

2.3.2 Lattice Construction

In the first step, two "spines" are defined as B-spline curves, one in the initial shape of *C. elegans* and the other in the desired target shape (thick central black line in Fig. 3A), both parameterized by the user. The control points (round dots in Fig. 3B) are evenly distributed along circular cross sections whose positions are linked to the "spine" coordinates (round dots in Fig. 3A). The diameters of these circular cross sections are parameterized to follow the shape of the *C. elegans* phantom resulting in a "cylinder-like" mesh. The user has also to specify the spline degree (we used cubic splines) and knot vector. These parameters determine how much influence a specific control point has over a specific coordinate in space, so refining these parameters enables finer control over the deformation's level of detail. However, a trade-off between the regularity of the deformation and its complexity must be considered, as increasing the density of the deformation space (*e.g.,* the number of control points) would enhance precision but make the deformation more challenging.

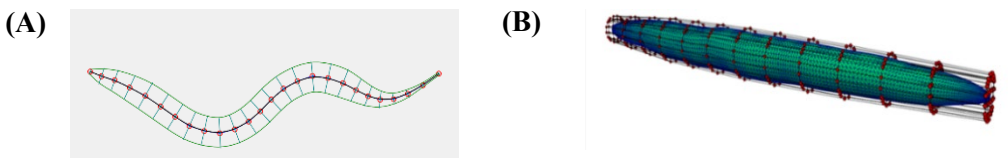


Figure 3. (A) Simplified 2D visualization of the parameters used to build the FFD lattice: medial curve ("spine") coordinates, diameters of circular cross sections supporting the control points;
(B) Example of a cylinder like lattice generated

After setting all parameters, we construct the two lattices (initial shape in Fig. 3B). The lattice is a trivariate tensorial product of B-spline curves. Once the target parametric space is properly defined and the physical positions of the final control points targeted, we can then proceed to the find the correspondence between the physical space and the parametric space in order to apply the deformation. It is important that the meshes be watertight before applying the FFD, as this method will preserve their original properties.

2.3.3 Inverse Mapping for a Custom-Shape FFD Mesh Embedding

Meshes are defined in a physical space of coordinates, whereas FFD is defined and operates in the parametric space. Usually, FFD is set in a simple cubic parametric space, as presented in Fig. 4A. In our case, the parametric space corresponds to the set of all cylindrical configurations with varying cross-sections, constituting our initial lattice. The physical points $\boldsymbol{X} = (x, y, z)$ correspond to the *C. elegans* phantom's surface mesh. To determine the parametric coordinates of $\boldsymbol{u} = (\xi, \eta, \zeta)$ corresponding to the physical points, the inverse mapping has to follow specific steps, since the forward mapping function $\boldsymbol{X} = V(\boldsymbol{u})$ is non-linear (Fig. 4). The inverse mapping will allow us to determine the B-splines basis functions, that will allow a correspondence between the physical space and the parametric space.

For that, the inverse mapping begins with the discretization of the parametric domain into a dense Cartesian grid to precompute a Look-Up Table (LUT) of the forward map. For any target point in physical space, a spatial search identifies the k-nearest neighbors within the LUT to provide an initial estimate $\boldsymbol{u}_0$ via Inverse Distance Weighting (IDW) (Shepard, 1968), which is a deterministic method for multivariate interpolation, where the unknown points are calculated thanks to a weighted average of the values available at the known points. This initial estimate $\boldsymbol{u}_0$ serves as the starting point for a final refinement stage, in which the solution to the equation $\boldsymbol{x} = V(\boldsymbol{u})$ is found using the iterative Least Squares Minimal Residual (LSMR) method. This algorithm allows for sparse least-squares problems that minimizes the residual norm $f(\boldsymbol{u}) = \frac{1}{2}\|V(\boldsymbol{u}) - \boldsymbol{X}\|^2$ at each step, making it particularly efficient for ill-conditioned systems. This step allows us to add a length conservation constraint. This constraint is necessary to keep the realism of the deformation, as the *C. elegans* retains its original length even when it deforms.

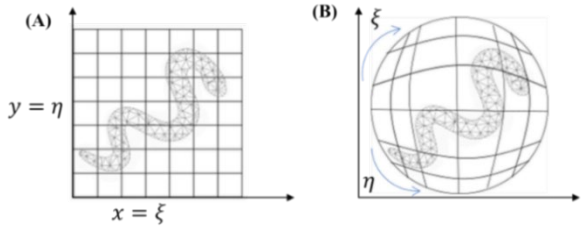


Figure 4. (A) Difference between regular FFD; and (B) custom-shape mesh FFD mapping of coordinates in the two spaces of definition

This iterative process is applied to the surface mesh points of each organ, in order to map them into the parametric space. Then, the B-spline basis functions are evaluated at those parametric coordinates to build the spline operator $\boldsymbol{N}$ specific to this organ, thereby establishing a reliable mapping between physical and parametric spaces. The deformation is then applied to each organ independently and sequentially using its corresponding spline

operator, as follows: the initial lattice shape is defined by the vector of control points $P_{initial}$ and the target deformed lattice shape by P_{final}. The control-points displacement vector is therefore computed as $\Delta \boldsymbol{P} = P_{final} - P_{initial}$ (cf. Fig 4 and 5). For each organ, the displacement field $\boldsymbol{U}$ at its surface points is obtained via the matrix multiplication $\boldsymbol{U} = \boldsymbol{N} \cdot \Delta \boldsymbol{P}$, where the spline operator $\boldsymbol{N}$ encodes the influence of the lattice's control points on the corresponding organ surface points. The deformed surface coordinates are then computed as $\boldsymbol{Q_{new}} = \boldsymbol{Q_{old}} + \mathbf{U}$, where $\boldsymbol{Q_{old}}$ and $\boldsymbol{Q_{new}}$ denote the original and deformed physical coordinates of the surface points, respectively. Repeating this procedure for each organ yields a deformed version of the entire *C. elegans* phantom, ensuring consistency and preserving the smooth and realistic visual aspect of the deformation enabled by the use of B-splines. Intuitively, N acts as a weighted average: each surface vertex's displacement is a combination of nearby control points' displacements, weighted according to their local influence. These weights depend on the degree of the B-splines defining the parametric space at that location and on the distance to the surrounding control points. Because all organs are deformed using the same ΔP, their relative positions are preserved after deformation, ensuring the resulting configurations remain anatomically plausible.

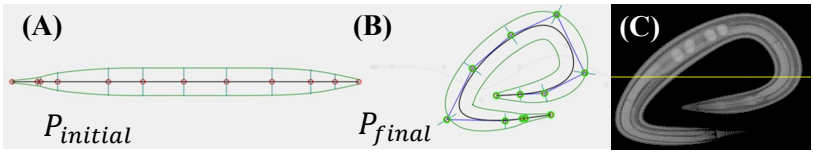


Figure 5. Example of a deformation – here a so-called "omega-turn" – achieved using the custom FFD procedure. (A) initial and (B) target lattice shapes; (C) Sagittal view of the resulting deformed phantom

2.4 Realistic *C. elegans* Deformation from Experimental Observations

With the aim of ensuring that the database captures the full range of plausible poses adopted by *C. elegans*, we reviewed several studies to identify mathematical formulations describing its motion, based on biological observations. There is a broad consensus that *C. elegans* adopts sinuous shapes, which can be approximated either by a single sinusoid or by a combination of sinusoids. Foundational work by Stephens et al. (2008) demonstrated that the space of natural postures of *C. elegans* is in fact low-dimensional: projections onto just four principal components, the so-called "eigenworms", capture approximately 95% of the shape variance of worms crawling off food, and the dynamics along these components characterize classical behaviors such as forward crawling, reversals, and a more distinctive and complex posture commonly referred to as "omega-turn".

Building on the eigenworm framework, one particularly detailed study (Broekmans et al., 2016) was based on the observation the dynamic movement of *C. elegans* in liquids with different viscosity coefficients and also modeled the distinction between crawling and swimming: depending on whether it is attracted to food, the locomotion mode of *C. elegans* changes accordingly. While swimming resembles the undulatory behavior typically associated with the worm, crawling exhibits more distinctive kinematics. Based on this wide range of

observed deformations, the study provides simulations of *C. elegans* using a spine-based representation. They include both sinuous locomotion and omega-turn, the latter derived from kymogram analysis. However, these simulations were limited to 2D motion.

We therefore extended this 2D representation into 3D for our model, in three steps. First, we used the 2D model of Broekmans et al. (2016) to generate several *C. elegans* conformations and extract the corresponding 2D coordinates of their 25 control points. Because these shapes were generated at different positions along the motion trajectory, we then re-centered each configuration by expressing its coordinates relative to the *C. elegans* center of mass, ensuring that only shape-related variations were preserved. The resulting 2D coordinates were then integrated into our framework. The final step consisted in extending the model to 3D. To do so, we assumed that the worm's "spine" remained in a horizontal plane, *i.e.*, that the third coordinate was constant. To introduce additional variability in 3D space, before doing the deformations, we applied rotations of the *C. elegans* phantom from 0 to 360 degrees and combined these rotations with the previously computed coordinate sets. This resulted in a diverse set of shapes with varying orientations and positions.

In our case, particular attention was given to small deformations, in order to remain consistent with the actual experimental conditions. Indeed, the *C. elegans* specimens selected for image acquisition are chosen among those exhibiting limited curvature and avoiding complex postural deformations, in order to facilitate organ visualization and segmentation. As a result, the range of deformations actually observed in the experimental images remains relatively limited, which justifies restricting our simulated deformations to a similarly small amplitude.

Before the FFD process, a mesh cleaning and regularization pre-processing step is performed, which can be also applied as a post-processing step if needed. This step is systematically applied to the organ meshes in order to make sure that they do not contain geometric artifacts, that could potentially be worsened by the FFD, such as non-manifold edges, duplicated or unreferenced vertices, duplicated faces, or small holes in the surface. For this purpose, the PyMeshLab library (Muntoni and Cignoni, 2021) is used, as it provides a wide range of functions dedicated to mesh repair and correction. This cleaning procedure is applied independently to each organ mesh and follows a two-stage pipeline. In a first stage, a series of repair operations is performed to identify potential geometric artifacts and, if any occur, correct them: non-manifold edges are fixed, holes in the surface are closed, and unreferenced vertices as well as duplicate faces and vertices are removed. This initial cleaning ensures that the mesh is topologically valid before any further geometric processing is applied.

In a second stage, an isotropic explicit remeshing is performed on the cleaned surface. This step regularizes the mesh by redistributing its vertices so as to obtain a more uniform triangulation, based on a target edge length expressed as a percentage of the mesh's bounding-box diagonal. The remeshing can be run for a configurable number of iterations, can optionally be made adaptive (allowing the local edge length to vary depending on surface curvature), and preserves sharp features above a specified feature-angle threshold, so that anatomically relevant edges are not smoothed away. Because this remeshing step can itself reintroduce minor topological defects, the same repair operations (non-manifold edge correction and so on) are applied a second time to the resulting mesh, ensuring that the final surface is both geometrically regular and topologically correct.

Importantly, the adaptive remeshing parameters, such as target edge length and feature-angle threshold, can be tuned globally for all organs or specified individually on a per-organ basis, which makes it possible to adapt the level of refinement to the geometric complexity of each anatomical structure. For example, we applied a finer target edge length to structures with

intricate geometry, such as the pharynx, while using coarser settings for simpler organs. All parameters used for a given run are logged automatically for traceability, ensuring that the exact remeshing configuration applied to each organ can be retrieved and reproduced.

2.5 Generation of Synthetic Projection Data

To generate synthetic data for tomographic imaging, we simulated the projection process. As detailed in Section 1.1, the tomographic image intended for reconstruction represents a mass density distribution. This distribution corresponds to either: *i)* the material mass density, derived from transmission tomography, *e.g.,* X-ray attenuation, synchrotron X-ray phase contrast, proton energy loss…; or *ii)* the chemical composition of the object, specifically the mass density of individual chemical elements, derived from X-ray emission tomography, *e.g.,* (S)-XRF or PIXE. The acquired data consist of projections of the local mass density along the incident beam path. Mathematically, the projections are calculated using the Radon transform *R* of the function *f*, which models the object's mass density. For a given tomographic slice, the Radon transform is defined as the integral of the object function along straight lines, called projection rays. In our case, the projection is thus an areal density (mass thickness in $g \cdot cm^{-2}$). Each projection ray is defined by its orientation θ and its perpendicular distance t from the origin, allowing the Radon transform to be expressed as:

$$Rf(t,\theta) = \int_{-\infty}^{+\infty} f(t\cos\theta - s\sin\theta\,, t\sin\theta + s\cos\theta)ds$$

From an experimental perspective, each of these integrals corresponds to a detector measurement, representing the cumulative signal along a single beam path. By collecting these measurements for all projection angles $\theta \in [0°, 180°[$ and all offsets *t*, one obtains a sinogram. The term derives from the fact that any given point within the object traces a sine wave as the projection angle varies. The sinogram (cf. Fig. 6) encodes the complete set of projection data required to reconstruct the original object *f*. In the idealized case of infinite projection angles and rays, the Radon transform remains a continuous function, as described above, and admits a mathematically exact inverse. In practice, however, data acquisition is inherently constrained by a finite number of projections and rays. Consequently, the continuous framework must be discretized for computational implementation and image reconstruction.

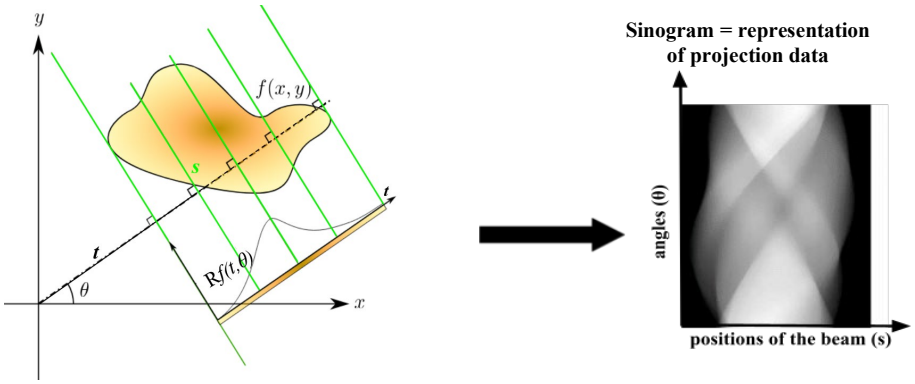


Figure 6. Representation of the Radon transform for parallel beam projection

Simulating accurate projection data rapidly constitutes a significant challenge. On the one hand, Monte Carlo methods provide high physical fidelity by tracking individual particle interactions; however, on the other hand, their computational cost is prohibitive for data generation. A practical alternative is to use deterministic models that prioritize execution speed by simplifying certain physical interactions, such as scattering, while maintaining the high level of accuracy required for quantitative imaging. The gVXR library, developed by F. Vidal *et al.* (Vidal et al., 2025), provides an open-source framework for high-performance X-ray transmission simulation by leveraging Graphics Processing Units (GPUs). The accuracy of gVXR has been rigorously validated against both Monte Carlo simulations and experimental data, making it a valuable tool for digital twinning and data augmentation in conventional X-ray CT. We adapted gVXR for micro-tomography to compute the Radon transform and produce sinograms. Because gVXR operates directly on 3D surface-mesh objects, it was ideally suited for our *C. elegans* phantom. While traditional ray tracing through these meshes is computationally intensive, requiring intersection tests between every ray and triangle, gVXR uses an optimized rasterization technique. This approach significantly reduces computational complexity, enabling the large-scale data generation required for our study.

To serve as an object in gVXR, the digital *C. elegans* phantom is defined as a hierarchy of closed triangular surface meshes, in which each internal organ is fully contained inside a parent structure (*e.g.*, the intestine inside the body cavity, itself inside the cuticle). Because each surface encloses a solid volume, this nesting implies that these volumes must not be counted multiple times during the projection simulation. Indeed, any point inside a child organ is simultaneously inside all of its ancestors. To obtain correct projection results, the *C. elegans* volume must therefore be partitioned into mutually exclusive regions. For this, a Boolean subtraction step is applied to each organ prior to projection. Following the inclusion tree, each organ's mesh is combined with the union of its direct children meshes. A constructive solid geometry (CSG) difference operation (parent minus children) is then performed using the *trimesh* Python library, designed for the processing and manipulation of triangular surface meshes. This CSG operation replaces the parent surface by a new closed surface enclosing only the material located between the parent boundary and its inclusions, effectively carving child-shaped cavities into the parent. After this step, every point of the phantom belongs to exactly one organ. This preprocessing is essential for the simulation of X-ray projection, as the gVXR engine assigns a single linear attenuation coefficient to each loaded mesh and accumulates, for every detector pixel, the attenuation along the ray path through each closed volume independently. With overlapping meshes, a ray crossing a nested organ would traverse both the child volume and the (still solid) parent volume, so the attenuation of the parent would be erroneously added to that of the child over the shared path length. By making the organ volumes disjoint, the CSG subtraction step guarantees the resulting line integrals correspond to the true piecewise-constant attenuation map of the phantom.

3. RESULTS

All reported computation times were obtained on a 2.20 GHz 13th Gen Intel® Core™ i9-13950HX CPU with 24 cores and 52 GB of RAM.

3.1 Blender Model

A topologically consistent young adult hermaphrodite *C. elegans* phantom was constructed following the methodology described in Section 2.2. It consists of 87 "organs" (Fig. 7), grouped inside 6 different collections: "Pharynx", "Digestive"," Hypodermis", "Muscle", "Reproductive" and "Other". For example, the cuticle - the largest organ - features a mesh resolution of 27980 vertices and 55974 faces, whereas smaller organs such as oocytes contain 3840 vertices and 128.

3.2 Deformed *C. elegans* Phantom

We applied the phantom deformation methodology described in Section 2.3 to generate a variety of anatomically realistic shapes. In the custom FFD pipeline, the inverse mapping step is the most computationally expensive, requiring approximately 50 minutes. In contrast, the deformation step completes in just 56 seconds. After deformation, as explained in section 2.4, the meshes may need post-processing to ensure proper topology and watertightness. The cleaning time depends on user-defined parameters, such as the number of iterations.

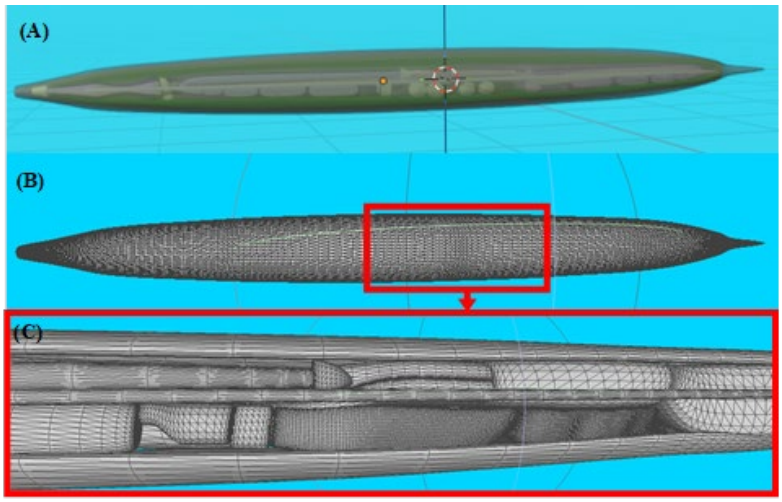


Figure 7. (A) Visualization of the *C.elegans* phantom in Blender. (B) and (C) Visualization of the resulting surface mesh in Meshlab, including a zoom (C) on a specific zone of the intestine and gonads, chosen as an example to visualize the inner content.

The obtained surface mesh was used to generate a voxelized version of the phantom, for subsequent image analysis purposes, the voxel size is a user-defined parameter (Fig. 8 and 9). The voxelization step takes approximately 1 minute for a voxel size of $1\times1\times1$ μm^3.

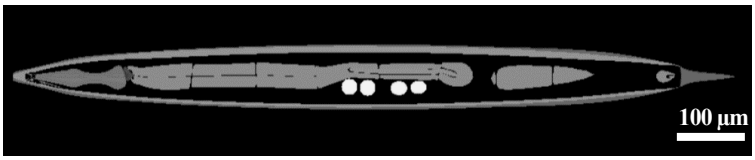


Figure 8. Straight voxelized *C. elegans* phantom in sagittal view, with a voxel size of 1×1×1 µm^3

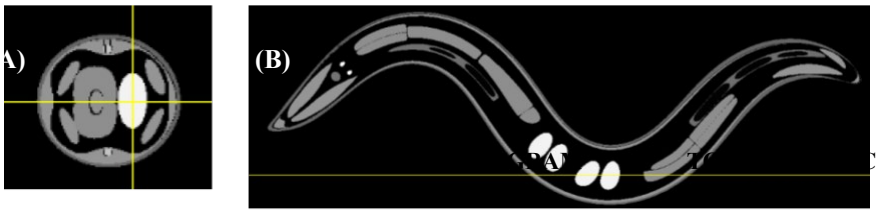


Figure 9. Deformed voxelized phantom in a transversal (A) and sagittal (B) view, with a voxel size of 0.6×0.6×0.6 µm^3

3.3 Generation of Synthetic Projection Data

The synthetic projection data were generated using gVXR, as described in Section 2.5, and represented as sinograms. In the sinogram, the horizontal axis corresponds to the incident beam position (310 steps in Fig. 10), while the vertical axis displays the angle of view (here from 0 to 180° with an angular step of 1.8°. At each position and angle, the projected mass density value is displayed in grayscale, ranging from zero (in black) to the maximum value (here normalized as 1, in white). While a complete sinogram enables the reconstruction of an accurate tomographic slice ("ideal case" Fig. 10), the missing wedge at angles near 0° and 180° (here ± 20°) introduces significant artifacts.

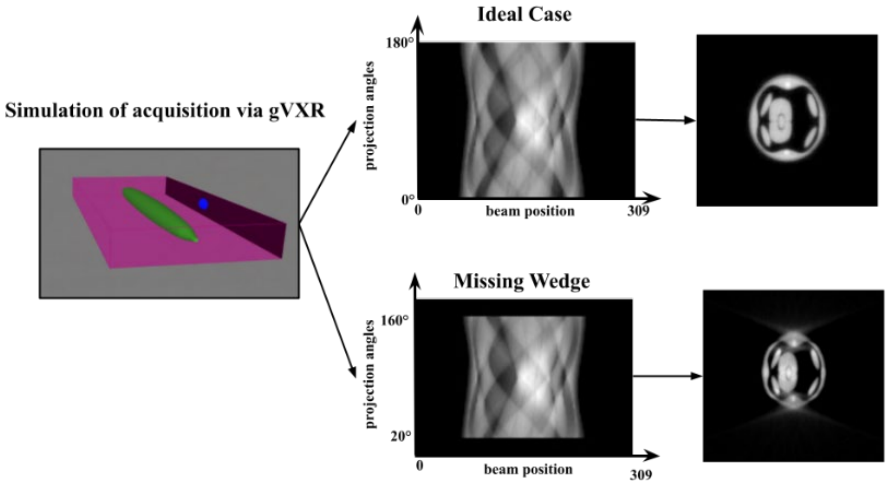


Figure 10. Visualization of the gVXR simulation setup (left) that generates the sinogram (middle), from which the reconstructed tomographic slice is obtained (right), respectively for the ideal case (top) and the missing wedge case (bottom)

The generation of a set of 900 sinograms for the whole *C. elegans* phantom using gVXR takes about 10 seconds. The reconstruction was performed using the Simultaneous Iterative Reconstruction Technique (SIRT) via the SIRT_CUDA function of the ASTRA toolbox (Van Aarle et al., 2016). The reconstruction of 900 tomographic slices with a resolution of 310×310 pixels take 53 seconds for 100 iterations. As described in Section 1.3, the purpose of generating this dataset is to support supervised deep learning tasks, which require a large amount of data. The rapid generation pipeline that we designed meets this requirement by enabling the creation of multiple phantoms. Specifically, we are able to generate $2.4 \cdot 10^6$ biologically realistic deformations, all of which satisfy the necessary constraints. Our ability to generate such a large amount of data in such a short time is enabled by the deterministic approach of gVXR, which supports GPU-based computations for accelerated projection calculations on the *C. elegans* meshed phantoms generated using FFD. By doing so, we significantly reduce computation times compared to traditional Monte Carlo methods, which would have required about 1 hour per tomographic slice (Michelet et al., 2022).

4. DISCUSSION

As our goal is to constitute a database for supervised learning of the inpainting task, it is important that this database contains sufficient variability to enable the neural network to generalize effectively. The aim is that, when applied to experimental data, the network can properly reconstruct the missing information, by following the features learnt from the numerical model. For this reason, we paid particular attention to generating data with significantly different deformations and rotations. Another important aspect for achieving greater realism is the addition of noise to the simulated images. During the experimental acquisition process, transmission imaging is based on the collection of a large number of detected events (*e.g.*, transmitted photons or protons), since the detector is placed in the incident beam direction just after it passes through the object. This generally results in a small amount of Gaussian noise in the obtained images. On the opposite, emission imaging (*e.g.*, X-ray fluorescence) is constrained by the low probability of X-ray emission. As a direct consequence, the amount of exploitable signal is smaller and the noise greater than in transmission imaging. Accounting for this noise, described by a Poisson distribution, is therefore essential to faithfully reproduce the statistical characteristics of experimental emission acquisitions. Therefore, the next step will be to apply the adapted noise for the different cases and use the noised sinograms for a supervised learning process to complete the missing wedge, while respecting the reality of the biological model.

5. CONCLUSION

In this study, we developed a comprehensive pipeline to address the loss of information caused by the missing wedge inherent to micro-tomography experimental setups. Existing models exhibit structural inconsistencies incompatible with such imaging applications. In contrast, the pipeline designed here produces topologically correct 3D meshes of *C. elegans*. The implementation of a custom shape FFD, refined through a high-precision inverse mapping process, enabled the generation of a vast synthetic dataset of about $2.4 \cdot 10^6$ deformations.

These deformations were designed to follow the sinuous shapes characteristic of the observed *C. elegans* postures reported in the literature, thereby preserving the required biological realism of the phantom. Furthermore, by adapting the gVXR framework to mimic proton microtomography acquisition via the Radon transform, we have reduced computation times significantly compared to traditional Monte Carlo methods, without sacrificing the accuracy required for neural network training. The amount of generated data was determined as a compromise between providing sufficient coverage of the documented *C. elegans* posture repertoire and remaining tractable for supervised learning. Because the FFD parameters were constrained to reproduce the sinuosity ranges reported in the literature, dataset completeness was assessed relative to this bounded biological variability rather than to an unconstrained shape space. This is consistent with the nature of the learning task: rather than requiring the network to model the full generative distribution of possible *C. elegans* morphologies, the inpainting task only requires sufficiently dense sampling of the mapping between visible sinogram context and the corresponding missing-wedge content across organ types and postures.

As a result, the integrated approach that we developed provides the necessary data density for supervised deep learning inpainting tasks, ultimately allowing for the robust 3D reconstruction of mass density and trace element distribution in biological samples. The results obtained here for proton micro tomography can be adapted to other types of incident radiation, especially synchrotron X-rays. Furthermore, the framework can be adapted to any mesh phantom. The ultimate goal of generating this synthetic data is to develop a deep learning model capable of transfer learning, thereby improving the inference of missing values in real experimental datasets and effectively mitigating the impact of the missing wedge.

ACKNOWLEDGEMENT

This project has received financial support from the Agence Nationale de la Recherche (ANR CES2010, no. CESA 009 01, TITANIUMS), the Région Aquitaine (TOX-NANO no. 20111201003 and POPRA no. 14006636-034), and the Université de Bordeaux through the “Exploratory Interdisciplinary Research” program (SHAPE 2023-2025 LP2IB, LaBRI, IMS). The authors gratefully acknowledge Philippe Barberet, Nuria Gordillo and Ayyoub Khrifi (LP2IB) for their participation in the tomography experiments and/or data processing at the AIFIRA facility of LP2IB. They also thank Jean-François Giovannelli (IMS, Univ. Bordeaux) and Franck Vidal (Daresbury Laboratory) for valuable discussions and advice.

REFERENCES

Altun, Z. F. et al. (2002–2024). *WormAtlas*.

Boudjelal A. et al. (2017). Improved Simultaneous Algebraic Reconstruction Technique Algorithm for Positron-Emission Tomography Image Reconstruction via Minimizing the Fast Total Variation. *Journal of Medical Imaging and Radiation Sciences, 48*(4), 385–393.

Broekmans, O. et al. (2016). Resolving coiled shapes reveals new reorientation behaviors in C. elegans. *ELife, 5*, e17227.

Deng, Y. et al. (2022). T-former: An efficient transformer for image inpainting. *Proceedings of the 30th ACM international conference on multimedia*, pp. 6559–6568.

Ding, G. et al. (2019). A joint deep learning model to recover information and reduce artifacts in missing-wedge sinograms for electron tomography and beyond. *Sci Rep, 9*, 12803.

Germer, T. et al. (2023). Limited-angle tomography reconstruction via deep end-to-end learning on synthetic data. *Applied Mathematics for Modern Challenges, 1*(2), 126–142.

Gordillo, N. et al. (2011). Technical developments for computed tomography on the CENBG nanobeam line. *Nucl. Instr. Meth. B, 269*(20), 2206–2209.

Iizuka, S. et al. (2017). Globally and Locally Consistent Image Completion. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1670–1678.

Le Trequesser, Q. et al. (2014). In situ titanium dioxide nanoparticles quantitative microscopy in cells and in C. elegans using nuclear microprobe analysis. *Nucl. Instr. Meth. B*, *341*, pp. 58–64.

Liu, J. et al. (2023). DOLCE: A model-based probabilistic diffusion framework for limited-angle CT reconstruction. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10498–10508.

Lugmayr, A. et al. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471.

McColl, G. et al. (2012). Caenorhabditis elegans maintains highly compartmentalized cellular distribution of metals and steep concentration gradients of manganese. *PLoS One, 7*(2), e32685.

Michelet, C. et al. (2017). An implementation of the NiftyRec medical imaging library for PIXE-tomography reconstruction. *Nucl. Instrum. Meth. B, 404*, pp. 131–139.

Michelet, C. et al. (2015). Development and applications of STIM- and PIXE-tomography: A review. *Nucl. Instr. Meth. B, 363*, 55–60.

Michelet, C. et al. (2022). A Geant4 simulation of X-ray emission for three-dimensional proton imaging of microscopic samples. *Physica Medica, 94*, 85–93.

Muntoni, A. and Cignoni, P. (2021). *PyMeshLab*. Zenodo. https://doi.org/10.5281/zenodo.4438750

Newson, A. et al. (2017). Non-Local Patch-Based Image Inpainting. *Image Processing on Line, 7*, 373–385.

Pathak, D. et al. (2016). Context encoders: Feature learning by inpainting. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544.

Sederberg, T. W. and Parry Scott, R. (1986). Free-form deformation of solid geometric models. *SIGGRAPH Comput. Graph, 20*, 151–160.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 ACM National Conference*, pp. 517–524.

Sternberg, P. W. et al. (2024). WormBase: status and transitioning to Alliance infrastructure. *Genetics*, *227*(1), iyae050. The Virtual Worm Project. http://caltech.wormbase.org/virtualworm/

Szigeti, B. et al. (2014). OpenWorm: an open-science approach to modeling Caenorhabditis elegans. *Frontiers in Computational Neuroscience, 8*, 137.

Van Aarle, W. et al. (2016). Fast and Flexible X-ray Tomography Using the ASTRA Toolbox. *Optics Express, 24*(22), 25129–25147.

Vidal, F. P. et al. (2025). X-ray simulations with gVXR in education, digital twining, experiment planning, and data analysis. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *568*, 165804.

Vidal, F. P. et al. (2025). X-ray simulations with gVirtualXray in medicine and life sciences. *3rd Dirk Bartz Prize for Visual Computing in Medicine and Life Sciences, EuroVis*, Luxembourg. https://hal.univ-lorraine.fr/hal-05129468

Wang, J. et al. (2018). Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear. *Med Image Comput Comput Assist Interv.*, *11070*, 3–11.

Yang, Y. et al. (2019). Three-Dimensional Correlative Imaging of a Malaria-Infected Cell with a Hard X-ray Nanoprobe. *Anal. Chem., 91*(10), 6549–6554.