

# Galactic Alchemy: Deep Learning Map-to-Map Translation in Hydrodynamical Simulations

Philipp Denzel<sup>1,2,3</sup>\*, Yann Billeter<sup>1,2</sup>, Frank-Peter Schilling<sup>1,3</sup>, and Elena Gavagnin<sup>1,3</sup>

<sup>1</sup>Centre for Artificial Intelligence, Zurich University of Applied Sciences ZHAW, Technikumstrasse 71, Winterthur 8400, Switzerland

<sup>2</sup>Institute of Science Technology and Policy, ETH Zurich, Universitätstrasse 41, Zurich 8092, Switzerland

<sup>3</sup>Institute of Business Information Technology, Zurich University of Applied Sciences ZHAW, Theaterstrasse 17, Winterthur 8400, Switzerland

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

We present the first systematic study of multi-domain map-to-map translation in galaxy formation simulations, leveraging deep generative models to predict diverse galactic properties. Using high-resolution magneto-hydrodynamical simulation data, we compare conditional generative adversarial networks and diffusion models under unified preprocessing and evaluation, optimizing architectures and attention mechanisms for physical fidelity on galactic scales. Our approach jointly addresses seven astrophysical domains – including dark matter, gas, neutral hydrogen, stellar mass, temperature, and magnetic field strength – while introducing physics-aware evaluation metrics that quantify structural realism beyond standard computer vision measures. We demonstrate that translation difficulty correlates with physical coupling, achieving near-perfect fidelity for mappings from gas to dark matter and mappings involving astro-chemical components such as total gas to H I content, while identifying fundamental challenges in weakly constrained tasks such as gas to stellar mass mappings. Our results establish GAN-based models as competitive counterparts to state-of-the-art diffusion approaches at a fraction of the computational cost (in training and inference), paving the way for scalable, physics-aware generative frameworks for forward modelling and observational reconstruction in the SKA era.

**Key words:** hydrodynamics – galaxies: structure – dark matter – galaxies: stellar content – software: machine learning

## 1 INTRODUCTION

The spatial matter distribution of galaxies is the result of a complex and chaotic interaction between its individual components, such as dark matter (DM), stellar populations, (predominantly hydrogen) gas, super-massive black holes (SMBHs), and their surrounding environment. Interactions between these components are governed by their mutual gravitational and electromagnetic forces, and hydrodynamical processes which collectively shape the structure and properties of galaxies over cosmic time (Binney & Tremaine 2011; Tinsley 2022; Conselice 2014; D’Onofrio et al. 2016). These interactions imprint subtle signatures in the phase-space distribution of a galaxy, retaining the various feedback mechanisms that have influenced its formation and evolution (Binney & Vasiliev 2023; Bassini et al. 2024). Thus, the physical components encode distinct aspects of galaxy evolution:

- **DM haloes** of galaxies dominate the gravitational potential into which baryonic matter flows and forms visible substructure (White & Frenk 1991; Moore et al. 1999; Frenk & White 2012).
- **Stellar mass** reflects the cumulative outcome of star formation and feedback, but its distribution is temporally highly non-local and entropic (McKee & Ostriker 2007; Kennicutt & Evans 2012; Hwang et al. 2019), while star formation is spatially localized

in H<sub>2</sub> clouds within the interstellar medium (ISM; Colman et al. 2024; Schinnerer & Leroy 2024).

- **Gas** traces the baryonic backbone of galaxies, regulating cooling, heating, and star formation through radiative feedback cycles (Gavagnin et al. 2017; Luisi et al. 2021) and turbulence induced by active galactic nuclei (AGNs; Biernacki & Teyssier 2018; Valentini et al. 2019), supernovae (Fielding et al. 2017; Ibrahim & Kobayashi 2023), stellar winds (Krumholz et al. 2014; Bally 2016), galaxy-galaxy mergers (Hopkins et al. 2006; Cibinel et al. 2019), or interaction with the intergalactic medium (IGM; Muratov et al. 2017; Poggianti et al. 2019).
- **Neutral hydrogen** and **21cm brightness** are key observational tracers of the ISM for low-redshift galaxies, critical for radio surveys with MeerKAT, ASKAP, or the upcoming SKA-Mid (such as WALLABY, MIGHTEE-H I, MHONGOOSE, or the MeerKAT Fornax Survey; Maccagni & Blok 2024; O’Beirne et al. 2025; Maddox et al. 2021; de Blok et al. 2024; Maccagni & Serra 2025).
- **Temperature** captures thermodynamic states shaped by shocks, cooling, and AGN-driven outflows (Zubovas et al. 2024; Ward et al. 2024).
- **Magnetic fields** emerge from turbulent amplification and influence gas dynamics (e.g., Beck 2015; Rieder & Teyssier 2017), yet remain poorly constrained observationally.

Recovering these domains from limited information is challenging; observationally due to technical limitations: for most instruments, signals beyond the local Universe – especially from HI – become

\* E-mail: philipp.denzel@zhaw.ch

too faint due to intrinsic dimming (Messias et al. 2024), foreground contamination (Collaboration et al. 2025), and low-frequency RFI (Harper & Dickinson 2018; Engelbrecht et al. 2024).

Conversely, theoretical inference of these domains has computational challenges: the understanding of the distribution of matter in the Universe remains largely driven by numerical simulations. Among these, (magneto-)hydrodynamical simulations present the most principled approach to model and capture the non-linear co-evolution of dark and baryonic matter fields across cosmological and astrophysical scales (for recent reviews, see Crain & van de Voort 2023). However, this quality comes at steep computational costs or forces detrimental trade-offs between resolution and volume.

To mitigate these challenges, simpler alternatives such as dark-matter-only (DMO) simulations (e.g., Potter et al. 2017; Ishiyama et al. 2021; Cheng et al. 2020) reproduce large-scale structure and halo statistics at reduced cost but omit baryonic physics. Semi-analytical models (SAMs) attempt to compensate by applying post de facto prescriptions to approximate baryonic effects on top of DMO outputs (e.g., Berlind et al. 2003; Somerville et al. 2008; Schneider et al. 2019; Obuljen et al. 2023). While these methods enable exploration of cosmological parameter space, they lack the fidelity needed to capture the full complexity of galaxy-scale feedback and morphology. In particular, their intrinsic post-hoc nature often ignores the gravitational back-reaction caused by the redistribution of baryons on the DM field.

With the proliferation of deep generative models, a complementary line of research has emerged that seeks to emulate aspects of these simulations rather than compute them from first principles. Recent efforts have explored enhancing simulations and augmenting galaxy models through scalable deep learning techniques in various ways. For instance, Perraudeau et al. (2019) use scalable GANs (for details see Section 2.3.1) to produce entire N-body 3D cubes of the cosmic DM distribution in a multi-scale approach. Still, techniques aiming for the full 3D reconstruction of cosmological simulations often face challenges in scaling to resolutions where individual galaxies can be resolved. Alternatively, Bernardini et al. (2021) employ Wasserstein-GANs to paint baryons onto thin slices of simulation boxes from the FIRE simulation suite. Li et al. (2021); Schanz et al. (2024) use StyleGAN and denoising diffusion models, respectively, to super-resolve cosmic large-scale structure predictions. Thiele et al. (2020) applied a U-Net architecture (for details see Section 2.4) to infer observable thermal and kinematic Sunyaev-Zel'dovich maps of haloes from DMO simulations, explicitly linking theory to observations. Similarly, Chadayammuri et al. (2023) use a U-Net for image-to-image translation of ILLUSTRIS-TNG galaxy cluster haloes to the corresponding baryonic fields.

Most studies focus on a single aspect of a simulation's galaxy formation or feedback model and do not fully reproduce (or harness) all physical modes of simulated galaxies (for details see Section 2, Equation 2).

In this paper, we introduce a novel application of deep generative models for map-to-map translation across multiple astrophysical domains in cosmological simulations on the galaxy-scale level. In contrast to other works, we propose a more comprehensive representation of a formation scenario by fitting various permutations of galaxy properties, without explicit heuristics or phenomenological tuning. Using high-resolution magneto-hydrodynamical simulation data from the ILLUSTRIS-TNG suite (TNG50-1; Springel et al. 2017; Nelson et al. 2017; Pillepich et al. 2017; Marinacci et al. 2018; Naiman et al. 2018), we systematically compare conditional generative adversarial networks (GANs) and diffusion models under unified preprocessing and evaluation. Our approach goes beyond prior

work by jointly addressing multiple domains and introducing physics-aware metrics – such as asymmetry, clumpiness, concentration, and power spectra – that assess structural realism and astrophysical fidelity beyond standard computer vision measures. We show that GAN-based models can achieve performance comparable to diffusion models at a fraction of the computational cost (in training and inference), in particular for map-to-map translations involving astrochemical components. Moreover, a set of deep generative models including all domain translations provides a comprehensive representation of a galaxy's formation scenario (for details see Section 2 and Equation 2). Finally, the generative models establish a bridge between theory and observation by incorporating domains that are directly observable, such as 21-cm brightness, into the translation process. This is particularly relevant for upcoming large-scale radio surveys with the Square Kilometre Array (SKA; Braun et al. 2015; Staveley-Smith & Oosterloo 2015) telescopes, which will probe the cosmic distribution of H<sub>1</sub> through 21cm emission. By enabling the reconstruction of astrophysical quantities from observational proxies and forward modelling of instrument-specific effects, our approach provides a scalable pathway to interpret SKA data within the context of galaxy formation scenarios.

The remainder of this paper is structured as follows: Section 2 details our methodology, models, evaluation metrics, and data, Section 3 presents the results, and Section 4 discusses implications and future directions.

## 2 DATA & METHODOLOGY

Our work aims to address the limitations identified above by leveraging high-resolution simulation data as the foundation for a generative modelling approach. To this end, we require a dataset that captures the full complexity of baryonic and DM interactions (i.e. magneto-hydrodynamics) at galaxy scales. In the following Section 2.1, we detail the selection criteria and preprocessing steps applied to construct our dataset of galaxy maps.

### 2.1 Dataset

The ILLUSTRIS-TNG project is a series of publicly released, cosmological magneto-hydrodynamical simulations of galaxy formation, run with the AREPO (Weinberger et al. 2020) moving-mesh code (Springel et al. 2017; Nelson et al. 2017; Pillepich et al. 2017; Marinacci et al. 2018; Naiman et al. 2018). Each simulation self-consistently solves the coupled evolution of DM, cosmic gas, luminous stars, and SMBHs. The TNG50-1 simulation was run with a total of  $2 \times 2160^3$  resolution elements, a DM mass resolution of  $3.1 \times 10^5 M_{\odot}/h$ , and a baryon mass resolution of  $5.7 \times 10^4 M_{\odot}/h$ , providing a rare combination of large volume and fine resolution in a simulation released to the public. Galaxies were selected from snapshots between  $z = 1$  and  $z = 0$  with a required minimum number of resolution elements of  $10^4$ , to ensure sufficient resolution even for larger satellite and dwarf galaxies.

Projections onto images for each selected galaxy were performed in multiple domains (galaxy properties  $\zeta$ ):

- DM mass (DM; column density)
- stellar mass (STARS; column density)
- total gas mass (GAS; column density)
- H<sub>1</sub> gas mass (HI; column density)
- (mock) 21-cm brightness temperature (21CM)
- gas temperature (TEMP)
- magnetic field strength (BFIELD)

**Table 1.** The preprocessing transformation parameters:  $c$  is the normalization constant (in the respective units of the corresponding maps) and  $\gamma$  the power scaling. The Boolean  $b$  deciding whether the transformation maps to a symmetric or non-negative interval was always 1 for diffusion models and 0 for GANs.

	DM	STARS	GAS	HI	21CM	TEMP	BFIELD
$c$	$2 \times 10^{10}$	$8 \times 10^{11}$	$10^{10}$	$10^8$	165	$10^8$	$10^{-1}$
$\gamma$	8	16	8	8	8	8	8

The projections extend to two half-mass radii of a galaxy’s total gas mass, ensuring each domain image has the same spatial resolution for a given galaxy. All but the 21-cm brightness temperature maps are directly simulated quantities; the former were generated following Villaescusa-Navarro et al. (2018). The map projections were performed using an adapted version Pylians3 code (Villaescusa-Navarro 2018). The resulting dataset of multiple domains counts 504,000  $512 \times 512$  images in total (72,000 images per domain), produced from roughly 3000 galaxies per snapshot (6 in total), each galaxy randomly rotated (on all axes) four different ways before projection for data augmentation. Note that the first iteration of the dataset contained fewer samples with a slightly higher average total halo mass; this dataset was used where explicitly stated in Section 2.7 and 3.

In summary, the dataset contains a set of galaxy projections in multiple domains (different physical modes of a galaxy) which jointly approximate the fiducial TNG model, i.e. ILLUSTRIS-TNG’s formation scenario.

Deep learning networks often work best on non-peaked data distributions, numerically standardized in intervals between  $[0, 1]$  (for uniform priors) or  $[-1, 1]$  (for Gaussian priors). Inspired by common transformation used in the high-energy physics domain (see e.g. Finke et al. 2021), we use the following scaling for all maps

$$\tilde{x} = (b + 1) \cdot \left(\frac{x}{c}\right)^{\frac{1}{\gamma}} - b \quad (1)$$

where  $c \neq 0$  is the normalization constant (around the maximum of the data distribution),  $\gamma \sim \mathcal{U}\{0, O(10)\}$  the power scaling, and the Boolean parameter  $b \in \{0, 1\}$  depending on whether the interval should map to  $[0, 1]$  or  $[-1, 1]$ . The exact values for the  $\gamma$  parameters were found via grid search (such that the median of the dataset distribution is  $\geq 0.3$  and  $\leq 0.6$ ) per domain as listed in Table 1;  $b$  was 0 for all models with a uniform prior (GANs) and 1 for all models with a Gaussian prior (diffusion models). This transformation normalizes the data ranges and stabilizes the variances in the data, making them more Gaussian-like.

## 2.2 Galaxy formation scenario

Capturing the complex interplay between baryonic components and DM distributions at the galaxy scale is computationally the most expensive task in any numerical simulation. Often, a trade-off between simulation size and resolution is required to make a hydrodynamical treatment even feasible. Additionally, surrogate techniques, so-called sub-grid models, are employed to capture effects of baryonic components below the resolution limit. The ill-constrained parameters of such sub-grid models are calibrated to match observed properties at the simulated scales, leading to degeneracy and difficulties in the interpretation of outcomes (cf. Crain & van de Voort 2023).

For this reason, different simulation suites produce similarly realistic galaxies with a wide variety of formation “recipes”. Notable, publicly available (and thus for this work relevant) examples of such

suites are the EAGLE (Schaye et al. 2014; Crain et al. 2015; McAlpine et al. 2016), HORIZON-AGN (Dubois et al. 2014), ILLUSTRIS-TNG (Springel et al. 2017; Nelson et al. 2017; Pillepich et al. 2017; Marinacci et al. 2018; Naiman et al. 2018), and SIMBA (Davé et al. 2019) suites.

The summary of all these physical effects characterizing a simulated population of galaxies, we will abstractly describe as a galaxy *formation scenario*  $\Phi$ . In Bayesian terms, a simulation describes galaxy samples from a *population*  $\Gamma_i$  by the marginalization

$$P(\Gamma|\Phi) = \sum_{\zeta \in \Omega} P(\Gamma|\zeta)P(\zeta|\Phi) \quad (2)$$

where  $\zeta \in \Omega$  represents a *galaxy property* from the set of galaxy characteristics  $\Omega$ . There will also be *nuisance parameters*  $v$  which lead to the expression of a galaxy distribution but are not related to any physical galaxy property, such as orientation

$$P(\Gamma|\zeta) = \sum_v P(\Gamma|\zeta, v)P(v). \quad (3)$$

A major inconvenience of simulations is the impracticality of drawing new samples from the galaxy population  $g \sim P(\Gamma|\Phi)$ , as this would require re-running an entirely new simulation at repeated computational expense. We pose that one or a set of deep generative models can properly encapsulate a simulation’s formation scenario  $\phi$  by learning individual galaxy properties  $\zeta$ , enabling in-painting a learnt formation scenario onto DMO simulations.

Recent advancements in deep learning techniques have demonstrated their efficacy in performing generative tasks that involve complex functional mappings between images. Given that simulated galaxies are typically reduced to 2D for comparison with observational data, this study will focus on image-based deep learning techniques.

## 2.3 Deep generative modelling

As general function approximators, deep learning neural networks have proven extremely useful for data processing across various scientific disciplines (Hornik et al. 1989; Goodfellow et al. 2016). Their ability to beat the curse of dimensionality allows for extraction of subliminal signals from complex data, finding hidden patterns or concepts that are difficult to (manually) formalize. Especially deep learning generative models have demonstrated unparalleled results, creating high-quality synthetic data, modelling complex systems and processes (Whang 2023; Bengesi et al. 2024). The goal of deep generative models is to learn an implicit (true) data distribution from which a finite number of samples is available for training (cf. Bond-Taylor et al. 2022); this usually means fitting an over-parametrized model  $p_\theta(x) \approx p(x)$  such that new samples  $\hat{x} \sim p_\theta(x)$  can be drawn and/or the likelihood  $p_\theta(x)$  be evaluated. *Conditional* generative models additionally include control variables  $c$  which guide the generative process such that  $p_\theta(x|c) \approx p(x|c)$ . *Image-to-image translation* is a particular application of conditional generation, where an input image of one domain is transformed into a corresponding output image in a different domain (Pang et al. 2022); in this study, the domains are defined by individual galaxy properties  $c \equiv \zeta$  where  $\zeta \in \Omega$  (see Section 2.1). Examples of image-to-image tasks include style transfer, image colourization, denoising, super-resolution, or semantic segmentation. Within the sciences, such tasks have been adapted in and across many disciplines, showing impressive performance in modelling the distribution of atomistic systems, proteins,

and biomolecules (e.g., [Ingraham et al. 2023](#); [Rives et al. 2021](#); [Schneuing et al. 2024](#); [Rønne et al. 2024](#)), particle jets (e.g., [Leigh et al. 2024](#); [Golling et al. 2024](#)), or for medical imaging enhancements (e.g., [Amirian et al. 2024](#); [Bullock et al. 2019](#)).

The conditional probability distributions approximated by these models directly correspond with the terms in Equation (2); thus such methods are particularly well-suited for this investigation. We examined GAN and diffusion-based approaches, as detailed in Section 2.3.1 and 2.3.2. Both approaches are known to produce high-quality samples. While *diffusion models* are considered state-of-the-art in scientific applications of image generation, they are intrinsically inefficient in their inference process, even when applied in latent space, even more so in pixel space (cf. [Dhariwal & Nichol 2021](#)). On the other hand, *GANs* can efficiently generate samples with a single forward pass, but generally have poorer training stability and distribution coverage. Therefore, we investigated both approaches for this work’s use case and compared their results, advantages, and challenges. As a secondary objective, we assess whether GAN-based models can achieve performance comparable to diffusion models, as this would substantially reduce computational costs and enable scalable deployment in large-scale simulation pipelines. Demonstrating such parity would not only accelerate inference but also substantially reduce the time required to iterate over all image translation directions, enabling more comprehensive exploration of domain mappings within practical computational budgets.

### 2.3.1 Generative Adversarial Networks

*GANs* are two-component models where a generative network, the *generator*  $G$ , and a discriminative network, the *discriminator*  $D$ , compete in an adversarial game; first introduced by [Goodfellow et al. \(2014\)](#).  $G$  aims to map<sup>1</sup> out of an implicit distribution  $p_G(z)$  to samples indistinguishable from the true data distribution  $p(x)$ , while at the same time  $D$  is optimized to distinguish between generated samples from  $p_G$  and real samples from the true data distribution. The adversarial game simultaneously invokes the minimization of the objective  $\mathcal{L}_{\text{adversarial}}(G, D)$  by  $G$  and the maximization of the same by  $D$ . These seemingly diametrical goals give rise to an efficient mechanism which optimizes  $G \rightarrow G^*$  leading to plausible, high-quality samples

$$G^* = \arg \min_G \max_D \mathcal{L}_{\text{adversarial}}(G, D). \quad (4)$$

This effectively eliminates the need to formulate an explicit loss function, as the discriminator will take that role; in other words, the loss function is learnt.

[Isola et al. \(2016\)](#) furthermore demonstrated a conditional version (cGAN) of this adversarial game as a general-purpose solution to image-to-image translation dubbed *pix2pix*. Although very similar to the classical GAN formulation, both cGAN networks are additionally conditioned on an input image  $x$ . The generator learns a mapping from input to output image domain space  $G : (x, z) \mapsto y$ . The discriminator is also additionally shown the input image with the corresponding generator output  $G(x, z)$ . This subtly changes the interpretation of its task from originally judging the realness of generated images to a judgment on the plausibility of the domain mapping. Accordingly, the GAN optimization objective is constructed as follows

$$\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{x,y} [\log D(x, y)] + \mathbb{E}_{x,z} [\log (1 - D(x, G(x, z)))] \quad (5)$$

<sup>1</sup> in a classical GAN the input  $z$  is typically a noise variable  $z \sim \mathcal{N}(0, \mathbb{I})$

where the first term is the average prediction strength of the discriminator when the images are sampled from the data distribution. The second term establishes the actual adversarial game, describing the average discriminator’s prediction strength when the images are sampled from the generator.

Moreover, [Isola et al. \(2016\)](#) proposed to mix the GAN objective with a traditional  $L_p$  loss term

$$\mathcal{L}_{L_p} = \mathbb{E}_{x,y,z} [||y - G(x, z)||_p] \quad (6)$$

where  $p = 1$  was found to be optimal by the authors whereas  $p = 2$  lead to blurriness in the predicted images.

The final adversarial objective is then given by

$$\mathcal{L}_{\text{adversarial}}(G, D) = \mathcal{L}_{L1}(G) + \lambda \cdot \mathcal{L}_{\text{cGAN}}(G, D) \quad (7)$$

where the objective weighting factor  $\lambda$  can be treated as a fixed hyper-parameter or adaptively tuned similar to [Esser et al. \(2020\)](#).

Finally, note that the noise variable  $z$  is necessary to learn a stochastic mapping, matching a distribution other than a delta function. However, [Isola et al. \(2016\)](#) have found noise input ineffective as cGAN models tend to simply ignore the noise and suggested to use dropout at test time instead to capture the full entropy of the modelled conditional distributions.

In practice, GANs are notoriously difficult to train despite their proven ability to generate high-quality samples. Two major challenges are *vanishing gradients* and *mode collapse*, which can be mitigated through architectural and objective modifications. Architectural strategies include residual skip connections to improve gradient flow ([He et al. 2015](#)), experimenting with normalization layers (batch [Ioffe & Szegedy \(2015\)](#), group [Wu & He \(2018\)](#), layer [Ba et al. \(2016\)](#), or none), and refining deconvolution operations near the generator output ([Odena et al. 2016](#)). Objective-based approaches involve alternative loss formulations for  $\mathcal{L}_{\text{cGAN}}(G, D)$ , such as DCGAN ([Radford et al. 2015](#)), LSGAN ([Mao et al. 2016](#)), or Wasserstein-GAN variants (WGAN, WGAN-GP; [Arjovsky et al. 2017](#); [Gulrajani et al. 2017](#)). Due to the minimax nature of GANs, losses often oscillate rather than converge, making diagnosis difficult. Overall, balancing generator and discriminator remains inherently unstable (cf. [Arjovsky & Bottou 2017](#)), requiring alternating gradient updates or separately scheduled learning rates.

In this study, we closely followed the implementation of the Pix2Pix model by [Isola et al. \(2016\)](#), including the aforementioned techniques and best practices. The generator is implemented as a standard *U-Net* ([Ronneberger et al. \(2015\)](#); architecture modifications are detailed in Section 2.4), paired with a *PatchGAN* discriminator which evaluates the plausibility of an image in sub-regions rather than a classical full-image discrimination. The discriminator can be restricted to enforce the correctness in local patches because the L1 loss in Equation (7) motivates the model to correctly predict low-frequency features in images, ultimately leading to more details in generated samples.

### 2.3.2 Diffusion-based models

Diffusion models have emerged as the de facto state of the art in computer vision (CV), surpassing in stability, distribution coverage, and arguably in sample quality models like GANs, normalizing flows or variational auto-encoders. They, colloquially speaking, learn to iteratively denoise a corrupted version of the data. More precisely speaking, diffusion models include a *forward* (noising) process which is designed to push samples off the data manifold and a *backward* (denoising) process for which a model is trained to produce trajectories back to that data manifold, generating plausible samples. There

are various framings for diffusion models, leading to slightly different expressions for these forward and backward processes. Here, we give a high-level overview of the formalisms relevant to this study.

Following Ho et al. (2020)’s description of Denoising Diffusion Probabilistic Models (DDPMs), the forward and backward processes take the form of Markov chains. The forward process starts from the input  $x_0$  and step-wise transitions to latent variables  $\{x_1, \dots, x_T\}$  (and vice versa for the backward process). Each forward transition at a particular time step only depends on the previous step and its probability is parametrized as a diagonal Gaussian

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbb{I}) \quad (8)$$

where the variance is  $\beta_t \in (0, 1)$  and typically scheduled as  $\beta_{t-1} < \beta_t$ . In the limit of infinitesimal step sizes, the true reverse process has the same functional form as the forward process, a well-known fact from Brownian diffusion in physics (see Equations 76 and 77 in Feller 1949). Thus, learning to approximate the backward process for small (enough) step sizes becomes feasible and can be analogously parametrized as

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (9)$$

Like the forward process, the backward process is a Markov chain for which its joint probability is given by the product of individual step conditionals

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) \quad (10)$$

where the marginal probability is a pure Gaussian  $p(x_T) = q(x_T) = \mathcal{N}(x_T; 0, \mathbb{I})$ .

The actual objective of the diffusion process, the sample probability  $p_\theta(x_0)$ , is generally intractable, as it would require marginalization over all possible trajectories. However, akin to latent space models such as VAEs (Kingma & Welling 2013), an *evidence lower variational bound* (ELBO) can be estimated (see Kingma et al. 2021; Sohl-Dickstein et al. 2015, and references therein)

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[ \log p(x_T) + \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] \quad (11)$$

$$\geq \mathbb{E}_q [\log p_\theta(x_0|x_1)] - \mathbb{E}_q [D_{\text{KL}}(q(x_T|x_0)||p(x_T))] - \sum_{t>1} \mathbb{E}_q [D_{\text{KL}}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))] \quad (12)$$

where  $D_{\text{KL}}$  denotes the Kullback-Leibler divergence. The lower bound (11) can have high variance and hence limited training efficiency and stability, compared to (12). Note that the first term in (12) ensures sample reconstruction quality while the second term matches the priors (assumed Gaussian), analogous to a classical VAE. The third summation term is known as the diffusion loss  $\mathcal{L}_{\text{diffusion}}$  and can be calculated in closed form given  $x_0$  is known (as it is during training).

The reverse step  $p_\theta(x_{t-1}|x_t)$ , that is the neural network, has various implementations. Ho et al. (2020) observed more stable training when the network only predicted  $\mu_\theta$  and assumed the variances to be time-dependent constants  $\Sigma_\theta(t) = \beta_t\mathbb{I}$ . Through the *reparametrization trick*, it is also possible to predict the added noise  $\epsilon$  through a neural network  $\epsilon_\theta$  rather than the mean of the Gaussian. Other forms of diffusion models directly predict the original data point  $x_0$ , or some combination of both (Salimans & Ho 2022).

In any case, the diffusion loss in Equation (12) can be shown to

generally reduce to

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbb{I}), t \sim \mathcal{U}_{[0, T]}} [\gamma'_\eta(t) \|\epsilon - \epsilon_\theta(x_t, t)\|^2] \quad (13)$$

where  $\gamma'_\eta(t)$  is an optional weighting pre-factor (with learnable bounds) evaluated via automatic differentiation. The noise schedule  $\gamma_\eta(t)$  also has various forms with the simplest schedule linearly increasing between two extremal bounds  $\eta = \{\gamma_{\min}, \gamma_{\max}\}$ .

For conditional generative tasks, conditioning variables  $c$  (here images of the original domain) are fed as additional inputs to the network during training  $\epsilon_\theta(x_t, t, c)$ . The conditioning can be further enforced by guiding the diffusion process, pushing the backward process in the direction of the gradient of the target condition probability (Ho & Salimans 2022). *Classifier-free diffusion guidance* achieves this through a modified training procedure by linearly combining null-labelled  $\emptyset$  diffusion and conditioned diffusion  $\tilde{\epsilon}_\theta(x_t, t, c) = \epsilon_\theta(x_t, t, \emptyset) + s(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset))$  given a guidance strength  $s$ . At inference time, samples can be artificially pushed towards the conditional direction by increasing the guidance strength  $s \geq 1$ .

In this study, various noise schedules and objective variations have been investigated, see Section 2.7 for details.

## 2.4 Neural Network Architectures

All network implementations can be found in our `chuchichaestli` package<sup>2</sup> published on PyPI and publicly available on GitHub. Here, we give an overview of their architecture, but for details we refer to the correspondingly listed sources.

**U-Net.** GAN as well as diffusion models implement their generative networks using the U-Net convolutional architecture, first introduced by Ronneberger et al. (2015). It was initially designed for segmentation of biomedical images, but has since been adapted to generative tasks for many other scientific fields (e.g., Bianco et al. 2025; Andersson et al. 2019; Yao et al. 2018). Its basic structure consists of a contracting (encoder) and an expansive (decoder) path, resulting in characteristically U-shaped graphs. While the architectural blocks in a U-Net have seen various updates since its inception, the basic encoder level follows the typical convolutional network structure with repeated 3x3 convolutional layers each followed by activations (LeakyReLU or ReLU) and a downsampling layer (a convolutional layer with stride 2); more recent versions additionally include residual block connections to improve gradient flow (He et al. 2015). With multiple levels, this leads to image compression, feature extraction, and ultimately representational learning. For image-to-image domain translation tasks, the structure of the decoder blocks is typically mirrored using deconvolutional layers to recover the input image resolution. Due to the repeated application of downsampling convolutional operations, spatial information is lost in deeper levels of the encoder. To this end, U-Nets additionally include skip connections between the corresponding levels which directly pass the encoder output information, concatenated to the output from lower decoder levels, and effectively integrate spatial information in the expansive path of the U-Net.

The basic U-Net structure in this work resembles the implementation by Isola et al. (2016) with a few notable updates:

- we opted for Swish activation functions (Ramachandran et al. 2017) instead of ReLU and LeakyReLU

<sup>2</sup> release version v0.2.13

- each block optionally includes a self-attention (Vaswani et al. 2017) or convolutional self-attention layer (Yang et al. 2019)
- dropout regularization in hidden layers (with a probability of 0.2)

Self-attention enables the handling of global interactions between pixels regardless of their relative position in the image and nicely complements the inherently local convolutional pixel treatment. Originally applied to language tasks, it quickly became an essential ingredient of any state-of-the-art neural network for image processing. However, since attention increases the computational complexity quadratically with sequence length, transformer networks become quickly infeasible, especially for high-dimensional data like images. Parmar et al. (2018); Weissenborn et al. (2019); Ho et al. (2019) proposed various solutions to this problem which usually entail reducing the receptive field and long-range interactions as compromise. We tested such *convolutional self-attention layers*<sup>3</sup> in our U-Nets, but have not noticed any significant improvements in performance or efficiency over classical self-attention (see Section 3).

Moreover, for the use in diffusion models the U-Net additionally contains a sinusoidal time embedding (aka positional embedding) to keep track of the time step in the diffusion process. The time embedding is injected in all residual blocks via linear projection layers whose outputs are added to the blocks' first convolutions.

**PatchGAN.** As introduced in Section 2.3.1 a PatchGAN is a patch-based discriminator which models an image as a Markovian field where each probability depends on neighbouring patches within a patch diameter. This concept was initially explored by Li & Wand (2016) in the context of texture synthesis and later implemented for general image-to-image translation by Isola et al. (2016). Our discriminator networks for adversarial training were adapted from Isola et al. (2016) with a 70x70 pixel receptive field. The implementation follows a simple convolutional block pattern consisting of batch normalization, activation (LeakyReLU), and two-dimensional downsampling convolutional layers.

## 2.5 Image-based evaluation metrics

To evaluate the similarity and quality of generated galaxy maps during and after training these neural networks, we first employ a set of widely used metrics from the CV domain. These metrics provide a baseline for assessing pixel-level accuracy (distortion), perceptual fidelity, and statistical realism in image synthesis tasks. While they are not tailored to astrophysical data, they offer valuable insights into the generative performance of deep learning models and can be used for initial hyper-parameter tuning.

**Mean Squared Error (MSE)** quantifies the average squared difference between corresponding pixels in two images

$$\text{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (14)$$

where  $x_i$  and  $\hat{x}_i$  are pixel values in the reference and generated images, respectively, and  $N$  is the total number of pixels.

It is sensitive to small pixel-level deviations and is often used to measure reconstruction accuracy. However, it does not account for perceptual or structural similarity.

<sup>3</sup> where key, query, and value representations are mapped using two-dimensional convolutions instead of fully connected linear layers

**Peak Signal-to-Noise Ratio (PSNR)** expresses the ratio between the maximum possible pixel value and the power of the error signal

$$\text{PSNR}(x, \hat{x}) = 10 \cdot \log_{10} \left( \frac{c^2}{\text{MSE}(x, \hat{x})} \right) \quad (15)$$

where  $c$  is the maximum pixel value range (typically 1 for normalized images, or 2 if the data range from -1 to 1).

Higher PSNR values indicate better fidelity. It is commonly used in image compression and denoising tasks.

**Structural Similarity Index (SSIM)** evaluates perceptual similarity by comparing luminance, contrast, and structural information between two images (Wang et al. 2004):

$$\text{SSIM}(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + k_1)(2\sigma_{x\hat{x}} + k_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + k_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + k_2)} \quad (16)$$

where  $\mu$ ,  $\sigma$ , and  $\sigma_{x\hat{x}}$  are the means, variances, and covariances of the images, and  $k_1$ ,  $k_2$  are stabilizing constants.

SSIM ranges from 0 to 1, with higher values indicating greater structural similarity. It is more aligned with human visual perception than MSE or PSNR.

**Fréchet Inception Distance (FID)** measures the distance between the distributions of real and generated images in a feature space extracted by a pre-trained neural network (such as Szegedy et al. 2015):

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr} \left( \Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right) \quad (17)$$

where  $(\mu_r, \Sigma_r)$  and  $(\mu_g, \Sigma_g)$  are the mean and covariance of real and generated image set features.

Lower FID scores indicate that the generated images are statistically similar to real ones in terms of feature distribution. FID is widely used to evaluate generative models such as GANs and diffusion models. However, since it is evaluated with model backbones typically pre-trained on ImageNet (Deng et al. 2009, an extensive dataset consisting of 3-channel, natural images), its application on scientific maps may be problematic. In our use case, each map is replicated on 3-channels before its features are extracted and thus does not exhibit the same colour variation as for natural images. Moreover, critics argue that FID's reliance on ImageNet-trained embeddings and its assumption of Gaussianity in high-level feature space render it ill-suited for domains with drastically different image statistics, such as scientific or medical imaging (Kynkänniemi et al. 2022; Jayasumana et al. 2023). In contrast, some studies have shown that using ImageNet-trained features can still correlate better with human perception than domain-specific feature extractors, even in, e.g., medical image synthesis tasks (Woodland et al. 2024). In any case, these findings caution against uncritical use of FID in non-natural image domains and suggest careful validation supported also by alternative metrics.

The metrics mentioned above serve as a foundational measures for evaluating image-to-image translation tasks. While they offer general-purpose assessments of distortion and perceptual image quality, they do not capture the domain-specific physical properties of galaxies. To address this, we complement them with a set of astrophysical metrics tailored to the structural and morphological characteristics of galaxy maps.

## 2.6 Astrophysical evaluation metrics

To assess the physical plausibility of the generated galaxy maps beyond pixel-wise similarity, we introduce a set of astrophysically

motivated metrics. These metrics are designed to quantify structural, morphological, and distributional properties of galaxies, enabling a more rigorous comparison between generated and ground truth samples. Each metric captures a distinct aspect of galaxy morphology and mass distribution, reflecting the underlying formation scenario.

**Asymmetry Error (AE).** Asymmetry evaluates the rotational symmetry of a galaxy map by comparing it to its  $180^\circ$  rotated counterpart (centred on the galaxy). This is a standard morphological indicator in observational astronomy (first introduced by Schade et al. 1995), and often used to identify signs of mergers, tidal interactions, or structural disturbances. It is typically defined as part of the concentration-asymmetry-smoothness parameter system (CAS; Conselice et al. 2000; Conselice 2003). Here, we define the AE in a slightly simplified adaptation as the difference between the normalized asymmetry of a ground truth  $I_r$  and generated map  $I_g$

$$\text{AE}(I_r, I_g) = \frac{\sum_{ij} |I_{r,ij} - I_{r,ij}^{180^\circ}|}{\sum_{ij} |I_{r,ij}|} - \frac{\sum_{ij} |I_{g,ij} - I_{g,ij}^{180^\circ}|}{\sum_{ij} |I_{g,ij}|} \quad (18)$$

where  $I^{180^\circ}$  are the  $180^\circ$ -rotated map correspondents. Higher asymmetry errors with respect to generated maps indicate discrepancies in structural symmetry which may indicate unrealistic morphology or artefacts.

**Smoothness/Clumpiness Error (SCE).** The so-called clumpiness quantifies the presence of small-scale structures such as star-forming regions or dense gas clumps (cf. CAS parameter system; Conselice et al. 2000; Conselice 2003). We calculate a proxy by subtracting a smoothed version of the map from the original and measuring the positive residuals, to avoid biasing the result through smoothing artefacts and removal of diffuse regions.

$$\text{SCE}(I_r, I_g) = \frac{\sum_{ij} \max(|I_{r,ij} - S_{r,ij}|, 0)}{\sum_{ij} |I_{r,ij}|} - \frac{\sum_{ij} \max(|I_{g,ij} - S_{g,ij}|, 0)}{\sum_{ij} |I_{g,ij}|} \quad (19)$$

where  $S$  is a smoothed (Gaussian blurred) version of the corresponding map  $I$ . A high SCE value may indicate excessive noise or unrealistic fragmentation, while a low error suggests smooth, well-resolved distributions. This metric is particularly relevant for evaluating the realism of baryonic, frictional components like gas and stars and substructure in DM haloes.

**Centre-Of-Mass Distance (COMD)** measures the Euclidean distance between the centre of mass of the generated map and that of the ground truth. The center of mass of a galaxy reflects the spatial alignment of its component distribution.

$$\text{COMD}(I_r, I_g) = \left\| \frac{\sum_{ij} x_{ij} I_{r,ij}}{\sum_{ij} I_{r,ij}} - \frac{\sum_{ij} x_{ij} I_{g,ij}}{\sum_{ij} I_{g,ij}} \right\|_2 \quad (20)$$

where  $x_{ij}$  are the spatial coordinates of distribution elements (pixels). Misalignment may indicate translation artefacts, structural inconsistencies, or failure to preserve spatial coherence. This metric is particularly important for tasks involving domain translation where positional accuracy is critical.

**(Cumulative) Radial Curve Errors (CRCE/RCE).** This metric compares the radial intensity or mass profile of the generated map to that of the ground truth. The radial profile is computed by averaging pixel values in concentric radial bins centred on the galaxy's centre of mass. The *Radial Curve Errors* capture deviations in the spatial

distribution of matter, such as incorrect central concentration or scale mismatch. It is essential for validating the structural integrity of generated galaxies. As a complement, the analogous comparison of cumulative radial distributions of generated and ground truth maps measures cumulative content deviations at a given radius. Errors from cumulative profiles are sensitive to global mass conservation and spatial allocation.

**Power Spectrum Errors (PSE)** compares the radially averaged 2D power spectra (i.e. squared magnitude of the Fourier coefficients at each frequency) of two maps. For each map, we compute the two-dimensional discrete Fourier transform and derive the power spectrum as the squared modulus of the Fourier coefficients. The resulting two-dimensional power spectrum is then radially averaged in Fourier space to obtain a one-dimensional power spectrum curve  $P(k)$  which characterizes the distribution of power as a function of spatial frequency. Finally, the normalized power spectrum curve residuals can be reduced by means of summation or averaging. This approach assesses similarity in spatial structure, texture, and particularly characteristic, second-order (filamentary) scales, independent to normalization, translation, or rotation. The PSE is especially useful when validating a model's accurate reproduction of multi-scale spatial features.

The aforementioned metrics collectively provide a robust framework for evaluating the astrophysical realism of generated galaxy maps. They complement traditional image similarity metrics from Section 2.5 by incorporating domain-specific knowledge and physical constraints, thereby enabling a more meaningful assessment of generated samples.

Finally, beyond pixel-level and morphological assessments, we also perform an inter-model consistency analysis based on integrated physical quantities (such as e.g. average magnetic field strength or total mass content). By substituting individual components with model predictions while keeping others at ground truth, we quantify biases and scatter, as well as cross-source disagreement for a fixed target domain. These metrics reveal whether different translation models preserve masses and energy globally and maintain physically plausible component fractions, independent of local image fidelity. Furthermore, they test whether models can be chained in cycles, potentially avoiding the need to train all model translation permutations if the goal is to complete a physical model from an arbitrary galaxy property. This approach provides a complementary, physically grounded perspective on model performance, ensuring that generated maps respect fundamental conservation principles and astrophysical scaling relations.

## 2.7 Experiments

Given that both generative methodologies described in Sections 2.3.1 and 2.3.2 employ U-Net architectures as their backbone, it is essential to optimize the architectural hyper-parameters for the specific characteristics of the dataset. However, exhaustive hyper-parameter searches across all possible configurations are computationally prohibitive, especially for generative models. We therefore constrain our ablation studies to architectural components that have demonstrated the most significant impact on conditional image generation performance. For these ablations, a coarse grid search across diverse optimizers, learning rates, and loss term weights have been carried out beforehand to find good values/choices. The hyper-parameters for the final network architectures used in Section 2.8 have been optimized using the *Optuna* and *Ray Tune* frameworks.

**Training.** All models were implemented in PyTorch. The experiments were conducted on Nvidia V100/A100/H100/H200 GPUs depending on the specific VRAM requirements. Adam optimizers (separate ones in the case of GAN-based models) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ , and weight decay of  $10^{-5}$  were used. Unless otherwise stated, the maximum learning rate was set to  $5 \times 10^{-5}$  for generators and  $1 \times 10^{-5}$  for discriminators (where used), with a one-cycle policy schedule (following Smith & Topin 2017). It provides a smoother warm-up phase at lower learning rates, a ramp up to the maximum learning rate, and a cosine-annealing phase to  $10^{-4}$  of the maximum value. When attention layers are included, this schedule was found to lead to less instabilities during GAN training. All model experiments were trained for 30 epochs with a batch size of 8. The datasets were split into 85% training, 10% validation, and 5% test sets, ensuring that galaxies from the same halo did not appear in multiple splits. All reported metrics for experiments in Section 2.7 were evaluated on the validation set whereas astrophysical validation was performed on the test set, and reported after the final epoch.

**U-Net size experiments.** Previous work has identified model capacity, defined by depth (number of U-Net levels) and width (number of hidden feature channels) as a key factors of generative fidelity (Ronneberger et al. 2015; Ho et al. 2020; Isola et al. 2016). Additionally, the choice of normalization layers (Dhariwal & Nichol 2021), and the inclusion of residual and attention mechanisms (Zhang et al. 2018a; Dhariwal & Nichol 2021), have consistently shown to enhance both training stability and output quality. In contrast, other design choices, such as the specific up-sampling scheme or minor variations in skip connections, tend to yield marginal improvements and diminishing returns. To systematically assess these factors, we first examine the impact of model capacity on generative performance, limiting experiments to high-impact components to keep computational costs tolerable.

Table 2 summarizes the U-Net configurations tested in this initial set of targeted experiments. Each experiment varies only the architectural parameters under investigation, while all other training settings are held constant. For bench-marking, we selected the GAS $\rightarrow$ DM translation task, which exhibited intermediate difficulty across all domain pairs in preliminary tests.

All models were trained with adversarial loss for 30 epochs using the standard discriminator configuration (as described in Section 2.4), with a warm restart technique for stochastic gradient descent and a cosine annealing learning rate schedule (Loshchilov & Hutter 2016). Initial learning rates were set to  $10^{-4}$  for the generator and  $5 \times 10^{-5}$  for the discriminator.

The results of the U-Net size ablation study are summarized in Table 3. Among the tested configurations, MEDIUMU (64 channels, 4 levels) consistently achieved the best overall performance across most evaluation metrics. Notably, the SSIM metric seemed to saturate in all tests quickly, indicating most U-Net configurations yield structurally similar outputs to the ground truth, but may lack sensitivity with smooth, high-resolution distributions like those from simulations and may not capture subtle differences fine-grained textures and localized features.

Deeper and larger U-Net variants started exhibiting artefacts and over-fitting that degraded perceptual quality of generated samples evident by higher PSNR values, but at the cost of increased FID. Moreover, larger models exhibited signs of mode collapse, with unreliable metric results. Conversely, the shallower and smaller performed comparably or worse in PSNR and SSIM but suffered from a substantially worse FID, suggesting insufficient capacity to model the full complexity of the domain mapping.

**Table 2.** Various U-Net configurations with varying sizes in depth and width that were tested. “Width” refers to the base number of feature channel in the first U-Net layer, whereas “Depth” is the number of levels between down- and up-sampling layers. The “encoder” consists of base “DownBlocks” including and “decoder”. “# Params” is the total number of trainable parameters in the U-Net.

Designation	Width	Depth	Levels	# Params
TINYU	16	4	4	4,010,369
SMALLU	32	4	4	16,024,833
MEDIUMU	64	4	4	64,066,049
MEDIUMU_L3	64	3	3	15,814,657
MEDIUMU_L5	64	5	5	257,037,825
LARGEU	128	4	4	256,197,633

**Table 3.** Evaluation results of various U-Net size configurations from Table 2 after training for 30 epochs. All experiments are based on the translation GAS $\rightarrow$ DM, adversarially trained with the same discriminator configuration. Model results in bold are optimal values, and those marked with  $^\dagger$  exhibit mode collapse and are not reliable.

Designation	PSNR $\uparrow$	SSIM $\uparrow$	MSE $\downarrow$	FID $\downarrow$
TINYU	35.31	0.9954	$5.5 \times 10^{-4}$	12.02
SMALLU	39.12	0.9966	$6.6 \times 10^{-4}$	12.75
MEDIUMU	39.76	<b>0.9977</b>	$4.2 \times 10^{-4}$	<b>9.71</b>
MEDIUMU_L3	39.57	0.9967	$6.8 \times 10^{-4}$	30.44
MEDIUMU_L5	<b>48.01</b>	0.9972	<b><math>2.9 \times 10^{-4}</math></b>	18.31
LARGEU	$^\dagger$ 65.28	$^\dagger$ 0.9978	$^\dagger$ $3.1 \times 10^{-3}$	$^\dagger$ 270.0

Based on these findings, we identified the MEDIUMU configuration as the optimal U-Net configuration for this task. It offers a favourable trade-off between performance and computational cost, avoids overfitting, and maintains stable training dynamics. This configuration is therefore used as the default architecture in all subsequent experiments unless stated otherwise.

Note that for diffusion models, spot tests resulted in comparable distortion metrics, however, to offset the substantial increase in computational cost, the U-Net width was reduced to 32 channels, prioritizing the inclusion of additional attention layers.

**Attention layer placement experiments.** Having established an optimal baseline configuration, we investigated the impact of attention layer placement within the U-Net architecture in a subsequent series of experiments. While prior work suggests that attention mechanisms can enhance global context modelling (Vaswani et al. 2017), their effectiveness and efficiency may depend on the resolution level at which they are applied. To this end, we varied the position of self-attention blocks across encoder and decoder stages, including configurations with attention in early layers (high-resolution features), late layers (low-resolution, high-semantic features), and hybrid placements spanning multiple levels. While attention layers are expected to yield superior results no matter the placement, the main purpose of these experiments was to assess the relative performance differences of less computationally demanding placement in late layers to those in high-resolution features. All other architectural and training settings were kept identical to those in the optimal configuration from the U-Net size experiments. Only the learning rate update schedule was changed to a one-cycle policy due to instabilities in the generator-discriminator dynamics and to keep learning rate comparably high. The evaluation focused on image-based metrics to determine whether attention placement influences fine-grained structural fidelity. These

**Table 4.** U-Net configurations with various attention blocks positioning (encoder levels numbered top to bottom, continuing in the decoder bottom up). The second column “# Encoder” indicates how many attention layers are included in the encoder, the third “# Decoder”, how many in the decoder. “# Params” is the total number of trainable parameters in the U-Net. The right-most column is the average time for a forward pass with a single batch.

Designation	# Encoder	# Decoder	# Params	Forward pass
ATTN_U1	1	0	64,082,817	15.3765 s
ATTN_U3	1	0	64,329,729	2.0411 s
ATTN_U4	1	0	65,117,697	1.9542 s
ATTN_U_MID	1	0	68,266,497	1.9796 s
ATTN_U5	0	1	68,266,497	1.9904 s
ATTN_U6	0	1	65,117,697	2.2751 s
ATTN_U8	0	1	64,132,353	27.8642 s
ATTN_3xU3	2	1	69,581,825	5.7335 s
ATTN_ALL	4	4	71,046,529	48.2022 s

**Table 5.** Evaluation results of U-Net attention layer placement experiments from Table 4 after training for 30 epochs. All experiments are based on the translation GAS→DM, adversarially trained with the same discriminator configuration. Model results in bold are optimal values, and those marked with  $\dagger$  exhibit mode collapse and are not reliable.

Designation	PSNR $\uparrow$	SSIM $\uparrow$	MSE $\downarrow$	FID $\downarrow$
ATTN_U1	37.23	0.9968	$3.4 \times 10^{-4}$	6.93
ATTN_U3	37.10	0.9972	$3.5 \times 10^{-4}$	6.64
ATTN_U4	36.70	0.9968	$3.6 \times 10^{-4}$	6.59
ATTN_U_MID	35.27	0.9983	$3.7 \times 10^{-4}$	7.47
ATTN_U5	35.76	<b>0.9983</b>	$3.6 \times 10^{-4}$	6.60
ATTN_U6	36.26	0.9974	$3.6 \times 10^{-4}$	4.57
ATTN_U8	$\dagger 22.55$	$\dagger 0.9$	$\dagger 1.2 \times 10^{-2}$	$\dagger 253.2$
ATTN_3xU3	37.71	0.9970	$3.4 \times 10^{-4}$	<b>4.04</b>
ATTN_ALL	<b>42.55</b>	0.9934	<b><math>3.2 \times 10^{-4}</math></b>	8.70

experiments aim to identify the most effective strategy for leveraging attention without incurring unnecessary computational overhead.

The results of these experiments (Table 5) reveal that the benefits of self-attention layers in U-Net blocks are indeed dependent on both the number and placement within the network. While adding attention universally across all levels (ATTN\_ALL) improved pixel-wise metrics such as PSNR, it comparatively degraded distributional consistency as measured by FID, indicating over-parametrization. Conversely, a moderate number of self-attention layers in the deepest levels seems to generally improve distributional, perceptual fidelity compared to the previous experiments, in trade for distortion (cf. Blau & Michaeli 2018). In particular, adding attention layers near the bottleneck (ATTN\_3xU3) yielded the best FID scores while maintaining competitive PSNR and the other distortion metrics. These findings suggest that for this dataset global interactions are most effectively modelled when attention is applied to low-resolution, high-semantic feature maps, whereas attention in high-resolution layers may lead to equally or better performance but introduces unnecessary inefficiency and instability. Moreover, all experiments have been repeated using the convolutional attention variant, with nearly identical results in each run, and minimally shorter forward pass timings. Based on these results, we adopt a configuration with three deep convolutional attention layers for all subsequent experiments, as it offers the best balance of generative fidelity, stability, and efficiency.

**Table 6.** PatchGAN configurations of various sizes. “Width” refers to the base number of feature channel in the first hidden layer, whereas “Depth” is the number of hidden layers in the network. “# Params” is the total number of trainable parameters in the PatchGAN network.

Designation	Width	Depth	# Params
PGAN_SMALL	64	3	2,765,505
PGAN_MEDIUM	64	4	11,165,377
PGAN_LARGE	64	5	44,742,337
PGAN_WIDE	128	3	178,875,777
PGAN_NARROW	32	3	11,197,281

**Table 7.** Evaluation results of PatchGAN size experiments from Table 6 after training for 30 epochs. All experiments are based on the translation GAS→DM, adversarially trained with the same generator configuration. Model results in bold are optimal values.

Designation	PSNR $\uparrow$	SSIM $\uparrow$	MSE $\downarrow$	FID $\downarrow$
PGAN_SMALL	32.46	0.9894	$8.3 \times 10^{-4}$	18.04
PGAN_MEDIUM	<b>37.16</b>	0.9901	<b><math>4.3 \times 10^{-4}</math></b>	<b>4.62</b>
PGAN_LARGE	36.56	<b>0.9960</b>	$5.7 \times 10^{-4}$	8.47
PGAN_WIDE	34.77	0.9952	$7.1 \times 10^{-4}$	8.65
PGAN_NARROW	35.16	0.9909	$6.0 \times 10^{-4}$	9.75

**Model specifics** With the attention configuration fixed, we proceed to model-specific refinements. In this stage, we tuned discriminator architectures for GAN-based models and evaluate noise scheduling strategies for diffusion models.

For GAN-based models, the PatchGAN discriminator’s width, depth, and number of hidden layer configurations were spot tested (see Table 6 for configurations). Based on the results in Table 7, the PGAN\_MEDIUM configuration was used for the final model training. While there were no clear differences between the various configurations, smaller networks had the tendency to impose checker-board artefacts in the generator outputs and larger, wider ones lead to instabilities during training due to mismatched sizes between discriminator and generator.

For diffusion models, the U-Net includes a sinusoidal time-embedding with 32 channels in each block (as described in Section 2.4). Moreover, linear, quadratic, and cosine noise schedules have been tested, and cosine clearly improved image quality (with a consistent 2-3 dB improvement in PSNR, 0.05-0.1 difference in SSIM), convergence, and provided smoother denoising transitions.

## 2.8 Domain translations

With all model components conservatively optimized, the final stage of experiments extended the map-to-map translation task to encompass all available domains. Given the combinatorial nature of the dataset, exhaustively exploring all 5040 possible domain translations is infeasible. However, it is reasonable to expect that the complexity of translations tasks varies between the astrophysical interactions between the components. For instance, domain translations such as GAS→HI or 21cm→GAS are likely to be less complex, as they represent information completion or reduction. On the other hand, mappings like STARS→DM are inherently more challenging due to the “weak” coupling of these components in the simulation.

To capture this diversity, we selected a representative subset of domain pairs that span a broad range of translation difficulties; the translations are centred around GAS due to its close relation to ob-

servable quantities and, thus, astronomical relevance (as mentioned in Section 1). These included the following mappings:

- within baryonic components
  - $\text{GAS} \rightarrow \text{HI}$ ,
  - $\text{GAS} \rightarrow 21\text{CM}$ ,
  - $21\text{CM} \rightarrow \text{GAS}$ ,
  - $\text{GAS} \rightarrow \text{STARS}$
- baryonic-to-DM translations
  - $\text{GAS} \rightarrow \text{DM}$ ,
  - $\text{DM} \rightarrow \text{GAS}$
- thermodynamic transformations
  - $\text{GAS} \rightarrow \text{TEMP}$
- magnetic field strength reconstructions
  - $\text{GAS} \rightarrow \text{BFIELD}$ .

For each selected pair, models were trained using the optimized U-Net configuration identified in previous experiments, with attention layers placed near the bottleneck.

Both GAN-based and diffusion-based models were evaluated, and their outputs compared using the full suite of CV and astrophysical metrics. This strategy allowed us to assess not only the fidelity of individual translations but also the consistency of physical quantities across domains. In particular, we investigated whether certain domain pairs exhibit systematic biases or structural artefacts, and whether translation difficulty correlates with the intrinsic entropy or sparsity of the source domain. The results of these experiments are summarized in the following Section 3.

## 3 RESULTS

### 3.1 Qualitative assessment of samples

Figure 1 shows representative samples of map-to-map translations across the (unseen) test set of domain pairs. Each triplet shows the input map (left), the ground truth target (middle), and the model prediction (right). For strongly coupled domains such as  $\text{GAS} \rightarrow \text{DM}$ , both GAN and DDPM reproduce global morphology and substructures with high fidelity across various scales and mass ranges. In some cases, smaller satellite haloes are either missing or were generated without any counterpart in the ground truth maps. When present, they are typically plausible domain translations of the input map.

Also, the translations  $\text{GAS} \rightarrow \text{HI}$ ,  $\text{GAS} \rightarrow 21\text{CM}$ , and  $21\text{CM} \rightarrow \text{GAS}$  are consistently in excellent agreement for both models, with only mild over- or underestimation in some systems.

For thermodynamic and field-like targets ( $\text{GAS} \rightarrow \text{TEMP}$ ,  $\text{GAS} \rightarrow \text{BFIELD}$ ), DDPM predictions better preserve global gradients, whereas GANs sometimes sharpen local contrast and slightly overemphasize smaller map features.

The arguably most challenging inverse mappings (e.g.,  $\text{DM} \rightarrow \text{GAS}$ ) reveal residual artefacts and misaligned substructures for both models, underscoring the difficulty of inferring baryonic components from DM alone.

Similarly, both models struggle to faithfully reproduce translations involving the weakly correlated components  $\text{GAS} \rightarrow \text{STARS}$ . Samples from this task exhibit noticeable deviations: predicted stellar maps fail to capture the clumpy, centrally concentrated structures, reflecting the intrinsic (temporal) non-locality and higher entropy of the stellar distribution.

Overall, these examples illustrate that translation quality correlates strongly with the physical coupling between source and target domain, and that GAN models and DDPM reproduce very similar

samples and differ mostly in the details and high-frequency features: GANs often excel in structural sharpness for tightly coupled mappings, whereas DDPMs better maintain global coherence in more weakly constrained tasks.

### 3.2 Overall performance across translation tasks

The measured model performance varies systematically with the physical coupling between source and target domains (Tables 8 & 9). As in previous experiments, the SSIM metric saturates quickly during training and is less discriminative than the other image-based metrics.

Table 8 lists image-based (traditional CV) metric evaluations for all domain translation tasks, grouped in pairs of GAN and DDPM. The best mean value of each metric across all tasks and models is listed in bold. Similarly, Table 9 shows the set astrophysical metric evaluations in the same order and grouping.

Among all tested translations,  $\text{GAS} \rightarrow \text{DM}$  attains the highest overall fidelity: GAN and DDPM models reach best FID scores of  $1.56 \pm 0.36$  and  $2.03 \pm 0.08$ , respectively, with PSNR values above 35 dB and  $\text{SSIM} \gtrsim 0.997$  (see Table 8). The astrophysical metric evaluations listed in Table 9 confirm this trend for  $\text{GAS} \rightarrow \text{DM}$ : asymmetry and clumpiness errors are among the smallest, COM offsets are negligible, and cumulative total mass deviations (evaluated at  $R_{50}$ ) and power-spectrum errors remain modest.

Translations within the baryonic sector also perform strongly when the target is closely tied to the gas morphology.  $\text{GAS} \rightarrow \text{HI}$  and  $\text{GAS} \rightarrow 21\text{CM}$  achieve low FID values of 4–6 and competitive PSNR/MSE. These models show low morphological errors (AE and SCE), minimal COM drift, excellent recovery of the radial profiles, and reproduce the expected near-monotonic relations between the total gas mass, neutral hydrogen mass, and 21-cm brightness temperature.

Moreover, the inverse mapping  $21\text{CM} \rightarrow \text{GAS}$  remains tractable with similar FID values up to 7.6, competitive ranges for the other pixel-wise metrics. The aligned performance with its counterpart across all astrophysical metrics suggests that reconstructing gas maps from observational 21-cm inputs is feasible.

In contrast, mapping  $\text{DM} \rightarrow \text{GAS}$  is substantially harder, only scoring within an FID range between 22 and 45, and slightly but consistently worse results across all astrophysical metrics.

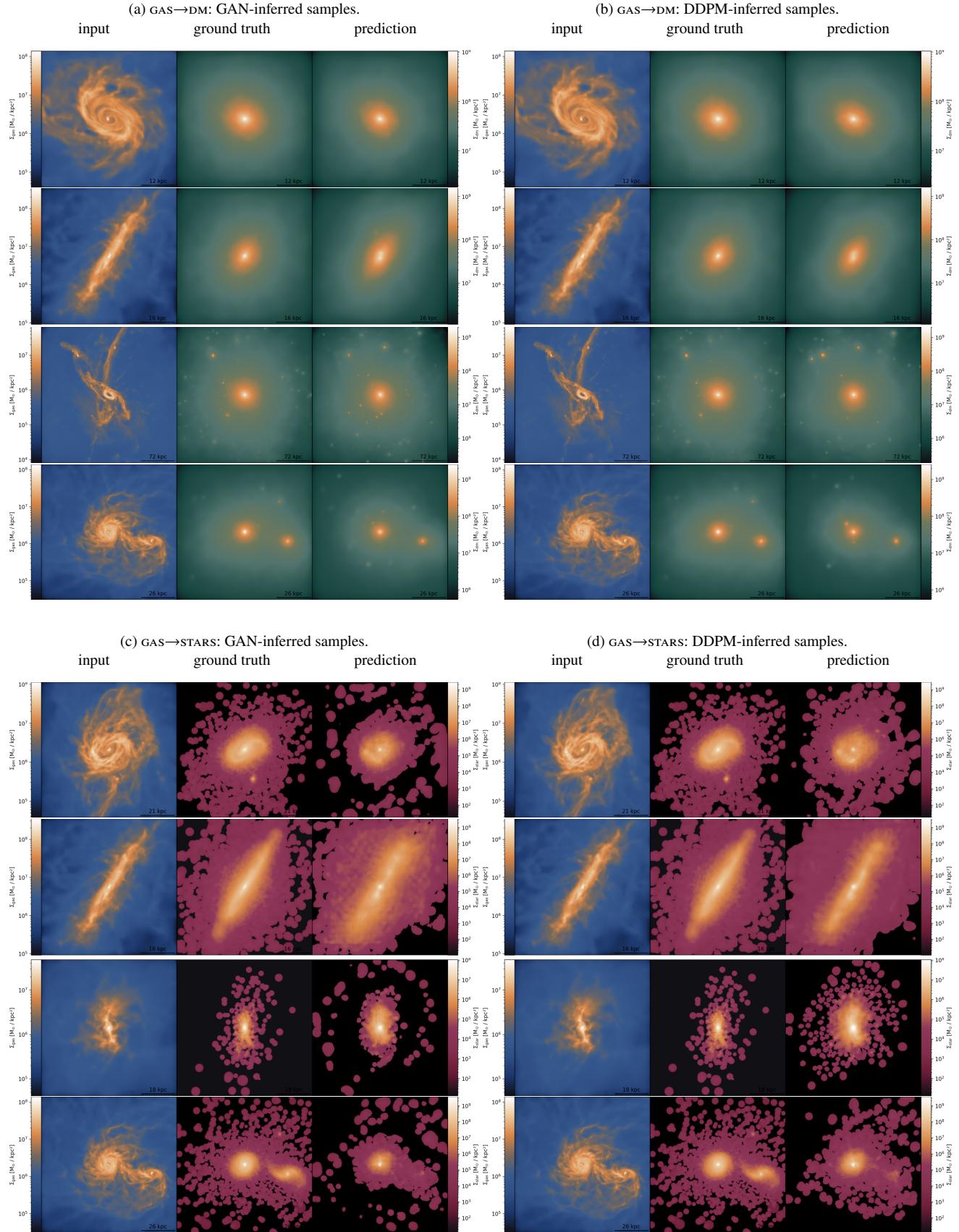
However, the most challenging mapping is clearly  $\text{GAS} \rightarrow \text{STARS}$ , which yields FID scores above well above 50, PSNR below 20 dB, and SSIM values well below the normal saturation levels. The large morphological errors, especially in asymmetry, reflect the models' inability to capture the alignment and ellipticity of the mass distributions, and the clumpiness errors indicate the models' difficulty to cope with the high non-locality of the stellar components.

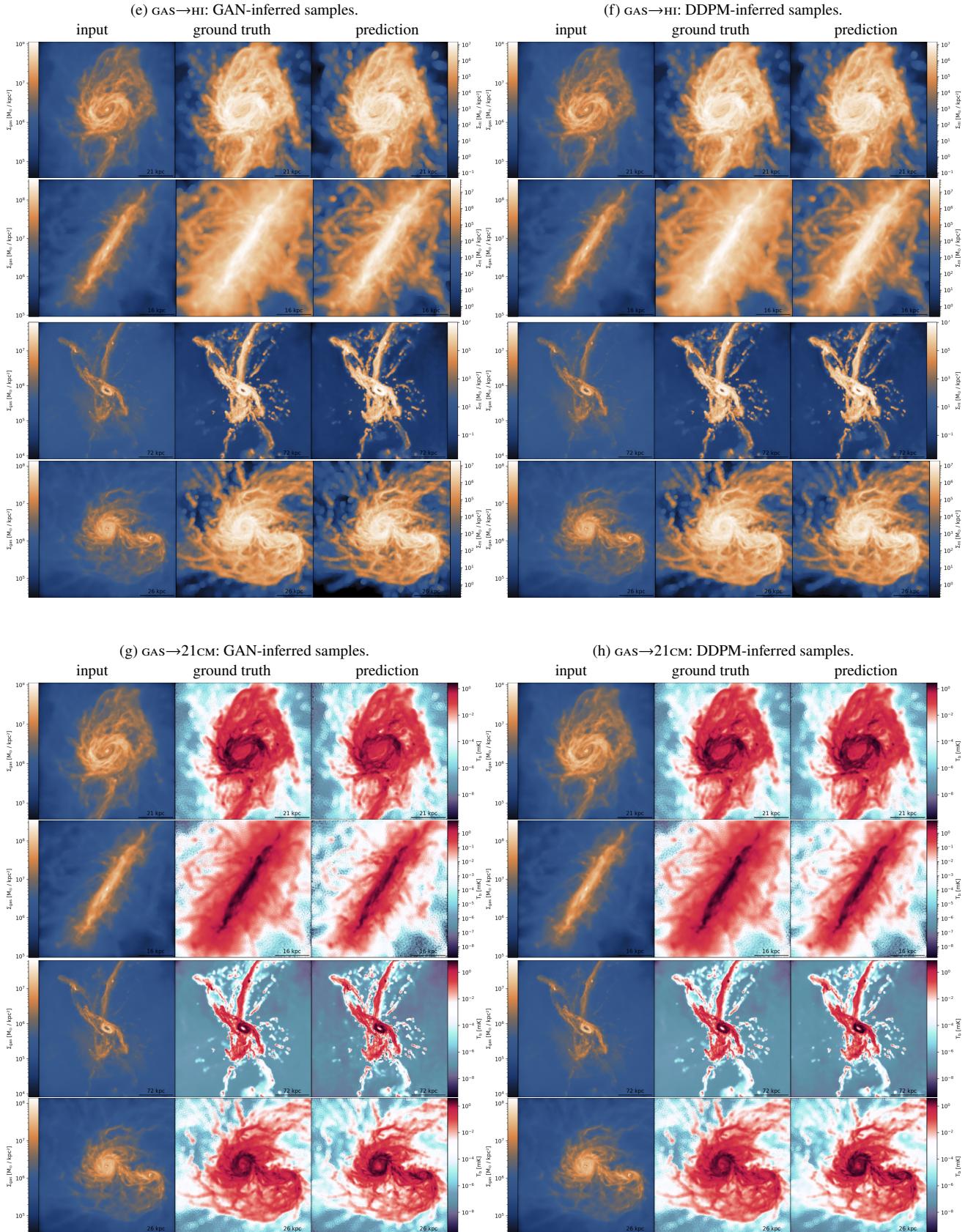
We note that hyper-parameters (U-Net size/attention and PatchGAN settings) were primarily optimized on the  $\text{GAS} \rightarrow \text{DM}$  task (Section 2.7); this may confer a slight advantage to  $\text{GAS} \rightarrow \text{DM}$  in cross-task comparisons. Thus, we repeated a small hyper-parameter sweep for a balanced set of tasks ( $\text{GAS} \rightarrow \text{STARS}$ ,  $\text{GAS} \rightarrow \text{HI}$ , and  $\text{DM} \rightarrow \text{GAS}$ ) to assess possible task-selection bias and performed a regret analysis based on the average FID score. The resulting task ranking was unchanged and the regret of the optimal configuration (as in Section 2.7) remained small across tasks, indicating that the results reflect intrinsic task difficulty rather than tuning alone.

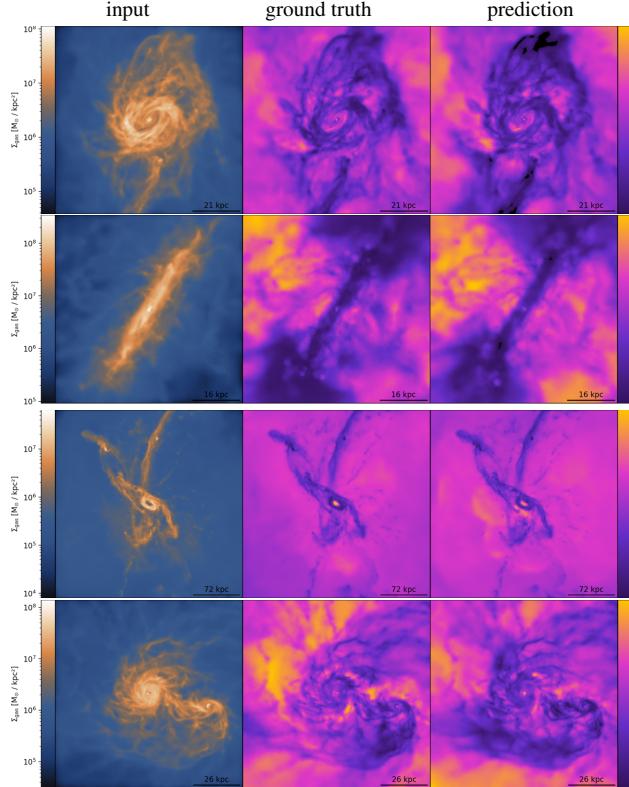
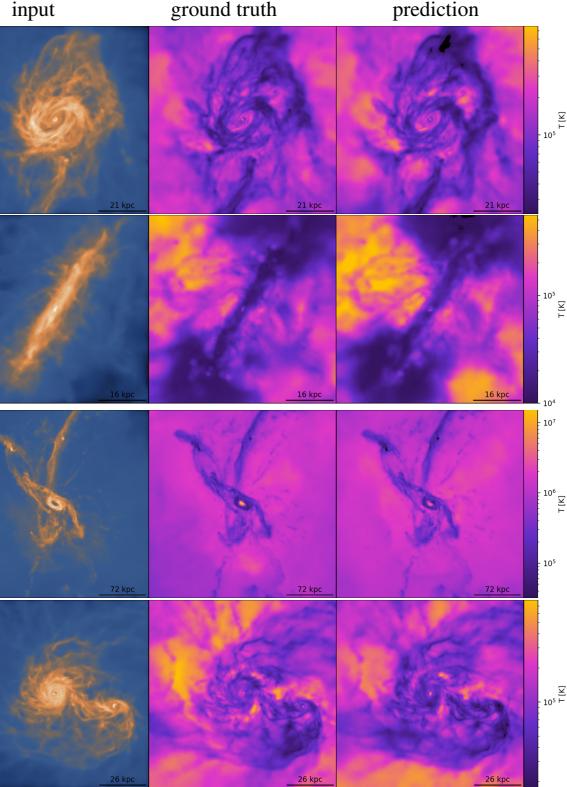
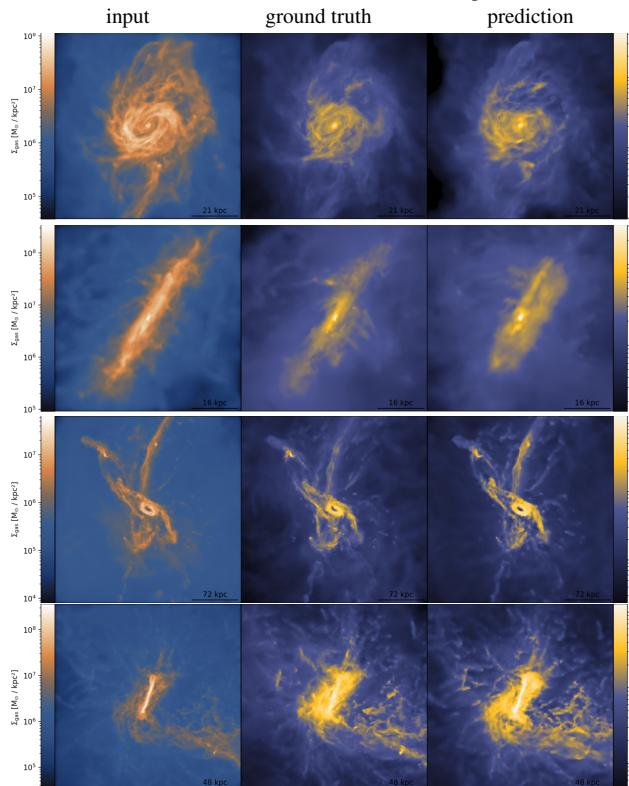
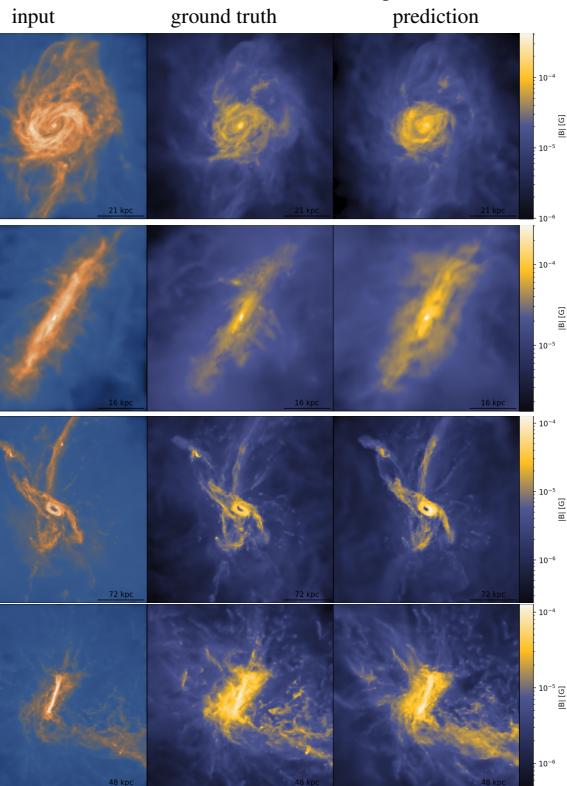
### 3.3 Metric interpretation

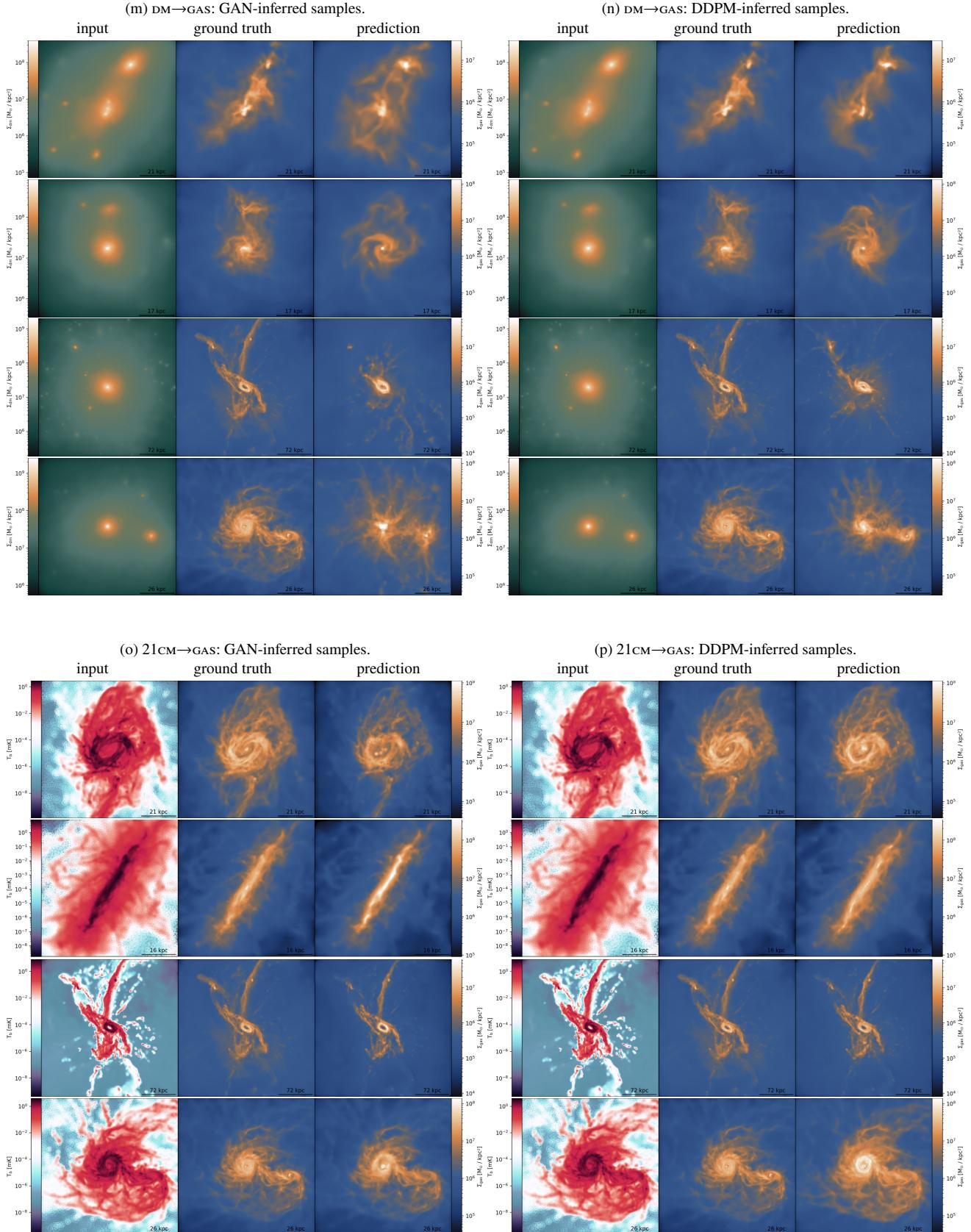
Direct comparison of PSNR and MSE across models require care because of the different preprocessing ranges (see Equation 1): GAN

**Figure 1.** Samples from various models and tasks. Each panel shows a model input map on the left, the corresponding ground truth in the middle, and prediction on the right. Qualitative comparison confirms the alignment of astrophysical plausibility and human perception with astrophysical metrics and FID (see Tables 8 and 9).





(i) GAS $\rightarrow$ TEMP: GAN-inferred samples.(j) GAS $\rightarrow$ TEMP: DDPM-inferred samples.(k) GAS $\rightarrow$ BFIELD: GAN-inferred samples.(l) GAS $\rightarrow$ BFIELD: DDPM-inferred samples.



**Table 8.** Extensive results for the entire suite of map-to-map translation models with image-based metrics (see Section 2.5). The values listed are mean  $\pm$  standard deviation of the respective metrics from the last 5 epochs (duration chosen as patience parameter when testing for convergence), as metric values for GANs fluctuate more. Note that the data ranges differ for GAN and DDPM models, which inherently biases the metric towards DDPMs by  $\sim 6.02$  dB for the same MSE value. Thus, the PSNR values for DDPM models were implicitly unbiased in the discrimination analysis.

Translation	Model	PSNR $\uparrow$	SSIM $\uparrow$	MSE ( $\times 10^{-4}$ ) $\downarrow$	FID $\downarrow$
GAS $\rightarrow$ DM	GAN	35.31 $\pm$ 0.11	<b>0.9974 <math>\pm</math> 0.0002</b>	3.82 $\pm$ 0.00	<b>1.56 <math>\pm</math> 0.36</b>
GAS $\rightarrow$ DM	DDPM	41.17 $\pm$ 0.07	0.9970 $\pm$ 0.0001	4.24 $\pm$ 0.10	2.03 $\pm$ 0.08
GAS $\rightarrow$ STARS	GAN	18.55 $\pm$ 0.36	0.5738 $\pm$ 0.0154	324.12 $\pm$ 3.44	60.56 $\pm$ 15.98
GAS $\rightarrow$ STARS	DDPM	23.34 $\pm$ 0.05	0.5577 $\pm$ 0.0046	324.60 $\pm$ 10.58	56.17 $\pm$ 14.33
GAS $\rightarrow$ HI	GAN	33.27 $\pm$ 0.16	0.9739 $\pm$ 0.0011	15.31 $\pm$ 0.80	4.57 $\pm$ 0.99
GAS $\rightarrow$ HI	DDPM	39.99 $\pm$ 0.14	0.9749 $\pm$ 0.0009	17.11 $\pm$ 0.77	5.86 $\pm$ 0.05
GAS $\rightarrow$ 21cm	GAN	31.60 $\pm$ 0.48	0.7958 $\pm$ 0.0115	17.90 $\pm$ 0.78	3.57 $\pm$ 1.03
GAS $\rightarrow$ 21cm	DDPM	38.55 $\pm$ 0.05	0.8133 $\pm$ 0.0013	17.95 $\pm$ 0.12	5.78 $\pm$ 0.07
GAS $\rightarrow$ TEMP	GAN	37.05 $\pm$ 0.27	0.9973 $\pm$ 0.0002	5.04 $\pm$ 0.19	9.91 $\pm$ 3.18
GAS $\rightarrow$ TEMP	DDPM	41.56 $\pm$ 0.06	0.9967 $\pm$ 0.0001	3.99 $\pm$ 0.32	7.86 $\pm$ 0.13
GAS $\rightarrow$ BFIELD	GAN	<b>38.76 <math>\pm</math> 0.62</b>	0.9964 $\pm$ 0.0003	<b>2.75 <math>\pm</math> 0.51</b>	9.80 $\pm$ 1.67
GAS $\rightarrow$ BFIELD	DDPM	43.39 $\pm$ 0.26	0.9955 $\pm$ 0.0007	3.60 $\pm$ 0.28	8.38 $\pm$ 0.33
DM $\rightarrow$ GAS	GAN	31.28 $\pm$ 0.12	0.9853 $\pm$ 0.0003	12.18 $\pm$ 0.79	36.36 $\pm$ 9.58
DM $\rightarrow$ GAS	DDPM	36.96 $\pm$ 0.03	0.9845 $\pm$ 0.0008	10.62 $\pm$ 0.40	22.87 $\pm$ 0.73
21cm $\rightarrow$ GAS	GAN	35.95 $\pm$ 0.56	0.9904 $\pm$ 0.0013	4.46 $\pm$ 0.55	7.60 $\pm$ 2.24
21cm $\rightarrow$ GAS	DDPM	42.08 $\pm$ 0.07	0.9900 $\pm$ 0.0003	3.75 $\pm$ 0.10	5.63 $\pm$ 0.90

**Table 9.** Extensive results for the entire suite of map-to-map translation models with astrophysical metrics (see Section 2.6). The values listed are mean  $\pm$  standard deviation of the respective metrics from the last 5 epochs.

Translation	Model	AE $\downarrow$	SCE $\downarrow$	COMD $\downarrow$	CRCE (at $R_{50}$ ) $\downarrow$	PSE $\downarrow$
GAS $\rightarrow$ DM	GAN	0.0655 $\pm$ 0.0005	0.0027 $\pm$ 0.0000	0.0211 $\pm$ 0.0173	0.2132 $\pm$ 0.1982	0.0788 $\pm$ 0.0041
GAS $\rightarrow$ DM	DDPM	0.0746 $\pm$ 0.0005	0.0032 $\pm$ 0.0000	0.0215 $\pm$ 0.0168	0.2196 $\pm$ 0.1998	0.0856 $\pm$ 0.0042
GAS $\rightarrow$ STARS	GAN	0.7460 $\pm$ 0.0529	0.0975 $\pm$ 0.0265	0.0657 $\pm$ 0.0446	1.3772 $\pm$ 5.7906	0.0690 $\pm$ 0.0046
GAS $\rightarrow$ STARS	DDPM	0.4466 $\pm$ 0.0494	0.0812 $\pm$ 0.0235	0.0297 $\pm$ 0.0393	1.2875 $\pm$ 3.4577	0.0596 $\pm$ 0.0042
GAS $\rightarrow$ HI	GAN	0.0839 $\pm$ 0.0028	0.0207 $\pm$ 0.0013	0.0128 $\pm$ 0.0178	0.2684 $\pm$ 0.3197	<b>0.0307 <math>\pm</math> 0.0024</b>
GAS $\rightarrow$ HI	DDPM	0.0885 $\pm$ 0.0031	0.0219 $\pm$ 0.0014	0.0136 $\pm$ 0.0191	0.2948 $\pm$ 0.3595	0.0363 $\pm$ 0.0028
GAS $\rightarrow$ 21cm	GAN	0.0713 $\pm$ 0.0027	0.0186 $\pm$ 0.0012	<b>0.0109 <math>\pm</math> 0.0155</b>	0.2192 $\pm$ 0.3051	0.0452 $\pm$ 0.0270
GAS $\rightarrow$ 21cm	DDPM	0.0813 $\pm$ 0.0029	0.0210 $\pm$ 0.0013	0.0120 $\pm$ 0.0167	0.2765 $\pm$ 0.2648	0.0524 $\pm$ 0.0308
GAS $\rightarrow$ TEMP	GAN	0.0901 $\pm$ 0.0001	0.0024 $\pm$ 0.0000	0.0561 $\pm$ 0.0367	0.1754 $\pm$ 0.1909	0.0568 $\pm$ 0.0047
GAS $\rightarrow$ TEMP	DDPM	0.0793 $\pm$ 0.0001	<b>0.0019 <math>\pm</math> 0.0000</b>	0.0484 $\pm$ 0.0332	<b>0.1605 <math>\pm</math> 0.1725</b>	0.0597 $\pm$ 0.0041
GAS $\rightarrow$ BFIELD	GAN	0.0822 $\pm$ 0.0012	0.0093 $\pm$ 0.0002	0.0371 $\pm$ 0.0213	0.2209 $\pm$ 0.1843	0.1000 $\pm$ 0.0554
GAS $\rightarrow$ BFIELD	DDPM	<b>0.0647 <math>\pm</math> 0.0010</b>	0.0072 $\pm$ 0.0002	0.0294 $\pm$ 0.0192	0.1928 $\pm$ 0.1739	0.0875 $\pm$ 0.0497
DM $\rightarrow$ GAS	GAN	0.1093 $\pm$ 0.0022	0.0224 $\pm$ 0.0006	0.0367 $\pm$ 0.0242	0.3143 $\pm$ 0.4294	0.0328 $\pm$ 0.0030
DM $\rightarrow$ GAS	DDPM	0.1085 $\pm$ 0.0021	0.0184 $\pm$ 0.0006	0.0357 $\pm$ 0.0246	0.2946 $\pm$ 0.3300	0.0333 $\pm$ 0.0030
21cm $\rightarrow$ GAS	GAN	0.0891 $\pm$ 0.0019	0.0161 $\pm$ 0.0004	0.0148 $\pm$ 0.0141	0.3483 $\pm$ 0.3448	0.0641 $\pm$ 0.0038
21cm $\rightarrow$ GAS	DDPM	0.0705 $\pm$ 0.0018	0.0131 $\pm$ 0.0004	0.0124 $\pm$ 0.0115	0.3231 $\pm$ 0.3867	0.0621 $\pm$ 0.0037

inputs/outputs are mapped  $[0, 1]$ , while DDPMs use  $[-1, 1]$ . The DDPM pixel value range is a factor of 2 larger, which biases PSNR by roughly 6.02 dB for the same MSE. Thus, throughout the cross-model PSNR comparisons in this Section 3 we implicitly remove this bias.

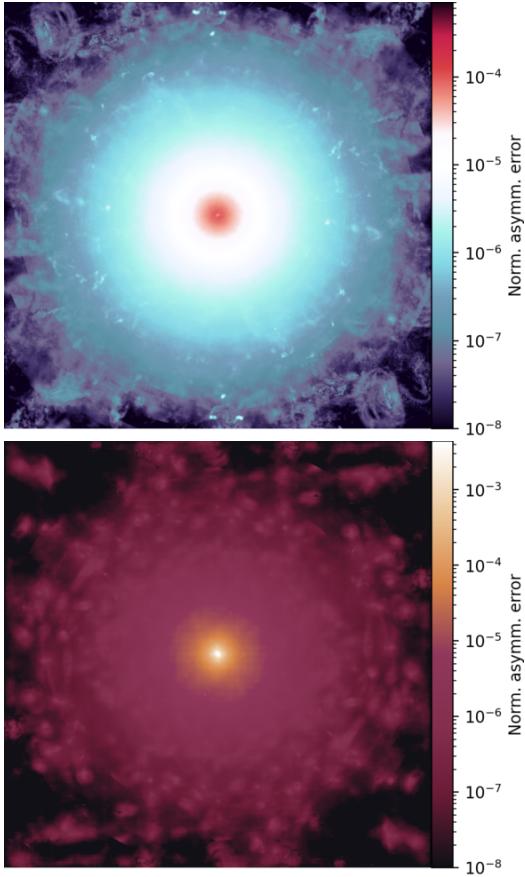
On these data domains, pixel distortion metrics PSNR, MSE, and especially SSIM are near-ceiling for several tasks (e.g. GAS  $\rightarrow$  DM) and can under-discriminate subtle morphological differences in smooth, high-resolution simulation maps. Conversely, the suite of astrophysically motivated metrics (AE, SCE, COMD, CRCE, and PSE) remains sensitive to structural realism and are model-agnostic.

Figures 2 and 3 illustrate how global errors manifest spatially. Both metrics measure important features (e.g., structural symmetry and fine-structure resolution) indicative of morphological realism and overall plausibility of generated samples. Figure 2 shows the average asymmetry error map for GAS  $\rightarrow$  21cm versus GAS  $\rightarrow$  STARS. Errors of the latter task are roughly an order of magnitude larger with a slight chequerboard pattern, indicating unresolved fine-structure and adversarial artefacts. For GAS  $\rightarrow$  DM and DM  $\rightarrow$  GAS (Figure 3) the harder inverse mapping (latter) exhibits higher small-scale residuals consistent with unrealistic fragmentation.

Centre-of-mass drift errors can also be decomposed in more detail (Figure 4). While the COMD only measures the scalar global drift, higher values may have different causes: the upper panel shows a near-uniform distribution of COM drifts (good positional agreement), whereas the lower panel exhibits a noticeable angular bias, signalling a systematic vectorial drift of the inferred mass centroid.

### 3.4 Model types: performance and trade-offs

There is no universal winner between GANs and DDPMs across all tasks. GANs tend to achieve lower FIDs when the target is tightly tied to the gas morphology (e.g., GAS  $\rightarrow$  DM, GAS  $\rightarrow$  HI, and GAS  $\rightarrow$  21cm), while DDPMs often deliver more favourable astrophysical fidelity (lower AE, SCE, and COMD) for less strongly related quantities such as GAS  $\rightarrow$  TEMP, or GAS  $\rightarrow$  BFIELD. Moreover, GANs inherently exhibit more quality fluctuations even long into training due to the adversarial nature of their objective; this is evidenced by the typically higher standard deviations of the metric results from the last five epochs. These complementary behaviours suggest that adversarial training sharpens structural realism in strongly coupled mappings, whereas diffusion-based modelling better preserves global morphology for

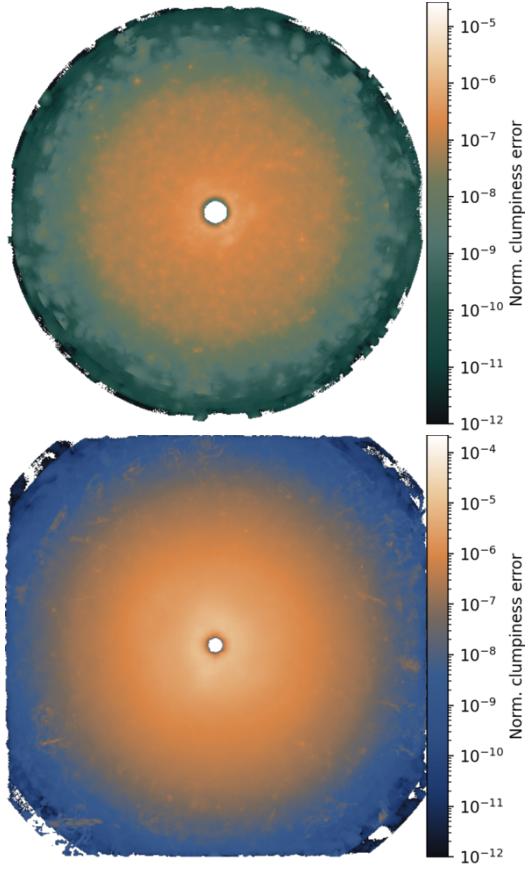


**Figure 2.** Examples of normalized asymmetry error maps for the mappings  $\text{GAS} \rightarrow 21\text{CM}$  (top) and  $\text{GAS} \rightarrow \text{STARS}$  (bottom) in the test set, inferred by GANs. The overall mean error is around an order of magnitude larger for  $\text{GAS} \rightarrow \text{STARS}$  and relatively uniform but exhibits a slight checkerboard pattern, indicating the difficulty to model the fine-grained structure of the stellar mass distribution.  $\text{GAS} \rightarrow 21\text{CM}$  exhibits smaller irregular errors which are noticeable due to overall lower average error.

thermodynamic and field-like targets. From a resource perspective, the GAN models in this work required  $\sim 140$  kWh training energy versus  $\sim 520$  kWh for DDPMs in our setup (summarized read-outs from the GPU monitoring system for all runs, not including ablation tests). Both approaches are orders of magnitude more energy-efficient than re-running comparable hydrodynamical simulations  $\mathcal{O}(\text{GWh})$  (cf. Table 1 in Nelson et al. 2019), but the  $\sim 4\times$  advantage of GANs can be decisive when many map-to-map translation models need to be trained.

### 3.5 Global consistency of inferred quantities

Figure 5 compares integrated inferred properties against ground truth for the unseen test population. For strongly coupled mappings such as  $\text{GAS} \rightarrow \text{HI}$  and  $\text{GAS} \rightarrow 21\text{CM}$ , both GAN and DDPM models recover total masses with minimal bias and scatter, indicating robust conservation of global properties. Notably, 99.9% of errors for all listed mappings, including  $\text{GAS} \rightarrow \text{DM}$ ,  $21\text{CM} \rightarrow \text{GAS}$ ,  $\text{GAS} \rightarrow \text{TEMP}$ ,  $\text{GAS} \rightarrow \text{BFIELD}$ , are within a factor of 10 (see also Table 9). In contrast, the mapping  $\text{GAS} \rightarrow \text{STARS}$  is exceptionally challenging for both models and presents large scatter and bias patterns (Figure 5b): DDPMs over-predict at low masses and under-predict at the high-mass end, while GANs exhibit smaller mean bias but extreme scatter reaching be-

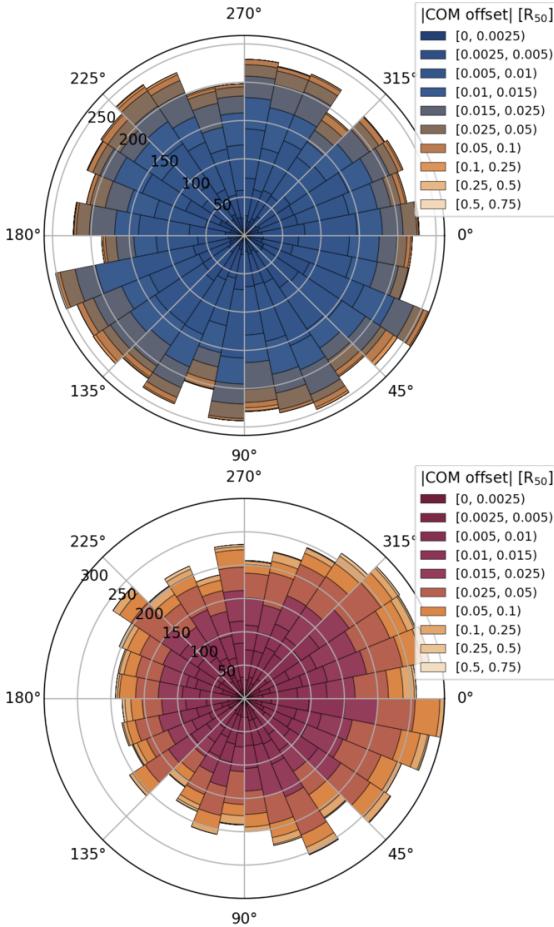


**Figure 3.** Examples of normalized clumpiness error maps for the mappings  $\text{GAS} \rightarrow \text{DM}$  (top) and  $\text{DM} \rightarrow \text{GAS}$  (bottom) in the test set, inferred by GANs. To keep numerical stability, the inner regions of the error maps have been masked to 5% of the map's respective half-mass radius. The overall mean error is around an order of magnitude larger for  $\text{DM} \rightarrow \text{GAS}$ , indicating the increased difficulty of predicting baryonic components from DM compared to the inverse mapping. Moreover, due to the collisionless nature of DM, its distributions tend to be smoother, which also contributes to the lower mean error. For  $\text{GAS} \rightarrow \text{DM}$ , errors mainly arise due to the wrong estimate of DM substructure in the haloes, whereas errors for  $\text{DM} \rightarrow \text{GAS}$  indicate unrealistic fragmentation in small-scale structures.

yond two orders of magnitude. These outcomes mirror the expected entropy and non-local differences among target domains and underline the task difficulty ordering observed in the other astrophysical metrics.

## 4 CONCLUSION

We presented the first systematic study of multi-domain map-to-map translations for galaxy formation simulations, introducing deep generative models as scalable data-driven alternatives that map between seven physical domains (DM, stellar mass, gas mass, neural hydrogen mass, 21-cm mock brightness, temperature, and magnetic field strength), comparing adversarial (GAN) and diffusion (DDPM) deep learning approaches under unified preprocessing and evaluation. Both approaches are able to learn physically plausible solutions to these domain translations, demonstrated on a dataset of galaxy maps extracted from the ILLUSTRISTNG suite (TNG50-1). Across extensive ablations and metrics – distortion (MSE, PSNR, SSIM), perceptual (FID) and astrophysical metrics (asymmetry, clumpiness,



**Figure 4.** Examples of the angular distribution of COM drifts for the mappings  $\text{GAS} \rightarrow \text{HI}$  (top) and  $\text{GAS} \rightarrow \text{STARS}$  (bottom) in the test set, inferred by DDPMs. While the upper wind rose diagram shows a uniform distribution for  $\text{GAS} \rightarrow \text{HI}$  COM drifts,  $\text{GAS} \rightarrow \text{STARS}$  exhibits an angular bias towards 0. The concentration of these errors in lower offset bins (in units of  $R_{50}$ ), as shown for  $\text{GAS} \rightarrow \text{HI}$ , indicates low overall drift and typically good agreement with the ground truth.

centre-of-mass drift, radial/cumulative curves, power spectra) – we find that translation difficulty strongly correlates with the physical coupling of source and target:  $\text{GAS} \rightarrow \text{DM}$  achieves the best fidelity measured by image-based metrics ( $\text{FID} \approx 2.0$ ),  $\text{GAS} \rightarrow \text{HI}$ ,  $\text{GAS} \rightarrow 21\text{cm}$ , and  $21\text{cm} \rightarrow \text{GAS}$  are likewise strong and conserve integrated quantities, while  $\text{DM} \rightarrow \text{GAS}$  is substantially harder but still produces plausible results.  $\text{GAS} \rightarrow \text{STARS}$  remains the most challenging across all measures. GANs tend to excel for tightly coupled targets with sharper structure and lower FID, whereas DDPMs better preserve global morphology and thermodynamic or field-like structure; this complementarity comes with a  $\sim 4\times$  difference in training energy in our setup ( $\sim 140$  kWh vs  $\sim 520$  kWh). These results demonstrate the feasibility of learnt representations that encapsulate aspects of a simulation’s formation scenario  $\Phi$  from different observationally motivated inputs, while underscoring the need for domain-aware metrics and physics-informed inductive biases to tackle weakly constrained mappings. Notably, despite the controversy around the use of FID in scientific domains (cf. 2.5), it correlated surprisingly strongly with the astrophysical metrics which capture structural realism, suggesting it is an appropriate discriminator for our use case.

**Physical couplings.** The empirical task ordering we observe follows the expected information coupling among galaxy components. Gas traces the gravitational potential well and interacts collisionally, so  $\text{GAS} \rightarrow \text{DM}$  is comparatively well-posed: large-scale morphology and substructure are strongly correlated, enabling excellent astrophysical veracity (low FID and morphological errors; see Tables 8 and 9 and Figure 3). In contrast, the inverse mapping  $\text{DM} \rightarrow \text{GAS}$  is under-constrained: while the DM halo delineates the potential, baryon distributions are additionally set by feedback, heating, and cooling; our models thus exhibit more clumpiness residuals and centre-of-mass drift (Figure 3), consistent with fragmentation artefacts. The most difficult case,  $\text{GAS} \rightarrow \text{STARS}$ , reflects the intrinsically non-local nature and higher entropy of stellar mass assembly: star formation depends on history and feedback cycles only weakly encoded in a single gas snapshot, leading to large asymmetry and clumpiness errors, poor FID, and strong biases in integrated stellar mass (see Tables 8 and 9 and Figures 2 and 4). Altogether, the results corroborate the conceptual view in Equations 2 and 3: learning conditional terms is easier when nuisance parameters are few and the conditional entropy of the target given the source is low.

Integrated quantities provide an orthogonal check of global physical plausibility. We find minimal bias and scatter for most mappings. The outlier is  $\text{GAS} \rightarrow \text{STARS}$ , which shows systematic bias and large scatter for both model types (Figure 5b). These findings imply that for a subset of domains with strong coupling, chaining of models (e.g.  $21\text{cm} \rightarrow \text{GAS} \rightarrow \text{DM}$ ) may be feasible without much loss of information (without explicitly training for cycle-consistency).

**Model choice guidance.** No single model type dominates across all translations. Adversarial training yields good high-frequency results (at times mildly exaggerated), especially for targets strongly coupled to the gas morphology, but exhibits larger epoch-to-epoch variability – a hallmark of the minimax optimization game (cf. Tables 8, 9, and Section 2.3.1, Equation 4). Diffusion models tend to preserve global gradients and in more complex couplings and often improve astrophysical plausibility at the cost of slower sampling and higher training time and energy. From a practitioner’s standpoint:

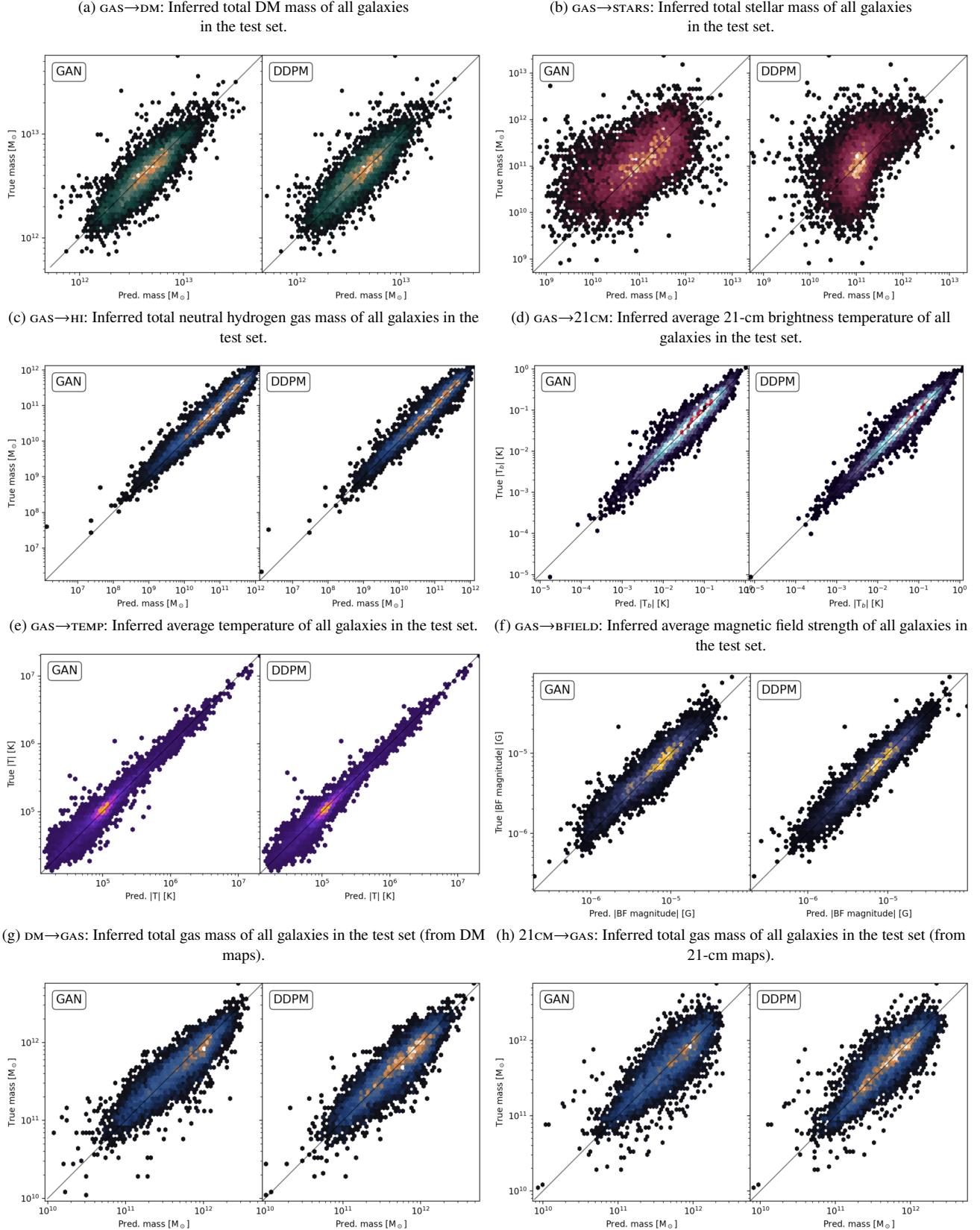
- choose GANs for tight, morphology-driven mappings where fast inference is worth the small trade-off in accuracy;
- choose DDPMs when the target encodes smoother or more complex fields, when robustness in astrophysical accuracy has highest priority.

Importantly, through targeted architectural and training optimizations, including U-Net depth/width tuning, attention placement near the bottleneck, and discriminator sizing (Section 2.7), we demonstrate that GAN-based models can achieve performance on par with state-of-the-art DDPMs for most mappings. This parity, combined with GANs’ lower training energy and single-pass inference, positions them as a competitive and computationally sustainable alternative for large-scale deployment. Moreover, a hybrid approach which draws from each methods advantages while mitigating their disadvantages, could be an promising avenue for future work.

**Implications for observations.** This work offers a direct path to observational validation by incorporating domains that are measurable in practice, such as 21cm brightness and neutral hydrogen, into the translation process. This capability is particularly critical for the SKA, which will probe the distribution of H I in nearby galaxies to unprecedented precision. Here, two practical applications of our models emerge:

- *Forward modelling:* predicting 21-cm brightness from simulated

**Figure 5.** Global statistics of inferred vs true integrated quantities. The colour scheme qualitatively indicates histogram density and matches the task assignment analogous to Figure 1. In general, GANs and DDPMs show no biases and minimal scatter of integrated quantities (except for GAS→STARS).



gas maps and pass through an instrument response pipeline (such as Karabo; [Sharma et al. 2025](#)) for high-realism mock observations including SKA-like systematics.

- *Reconstruction:* inferring gas distributions and related galactic properties from observed 21-cm maps of nearby galaxies to support feedback and morphological studies.

By embedding observational proxies and incorporating, e.g., beam smoothing, thermal noise, and foreground residuals into the generative framework (during training or via data augmentation), domain-shift robustness is increased; our astrophysical metrics are naturally suited to quantify degradation after instrumental effects. This provides a scalable pathway to interpret SKA data within the context of galaxy formation scenarios.

**Limitations.** Our models learn by design conditional slices of a simulation’s formation scenario  $\Phi$ . Because  $\Phi$  depends on sub-grid physics and calibration, generalization across suites (e.g. ILLUS-TNG, SIMBA, FIRE, or EAGLE) and redshift evolution must be demonstrated rather than assumed.

Furthermore, perceptual metrics such as FID carry domain-mismatch assumptions; fine-tuning feature extractors on domain-specific (astrophysical) data could provide an even better measure for astrophysical veracity. More flexible alternatives to FID such as LPIPS ([Zhang et al. 2018b](#)) could improve evaluation fidelity even further.

Translation with weak couplings could be improved with additional constraints. Models for the GAS→STARS mapping lack sufficient mutual information between input and target domains, making the task particularly challenging.

#### Outlook.

Future work will focus on addressing these limitations.

Weakly constrained mappings could be improved by further extending the dataset domains with intermediates. For instance, since H<sub>2</sub> is more closely tied to star formation, it should provide better constraints for the stellar mass prediction via GAS→H<sub>2</sub>→STARS.

Alternatively, various inductive biases could also provide stronger constraints during training:

- *Regularization of the objective function:* directly physics-informed networks through, e.g., constraining mass within aperture, or penalties on radial-profile mismatch.
- *Structure-aware discriminators:* adversarial heads operating on radial profiles, power spectra, or multi-scale losses.
- *Equivariant architecture:* SO(2)-aware U-Nets can reduce sample complexity, and inherently enforce symmetries, thereby explicitly handling nuisance parameters.
- *Multi-domain training:* predicting several targets at once in multiple channels would increase cross-domain robustness but increase processing time.
- *Cross-suite transfer learning:* cross-suite transfer learning and domain adaptation avoids re-training models on other simulation suites from scratch, requiring only a small amount of fine-tuning on the target simulation.
- *Redshift conditioning:* redshift introduces temporal information to models and helps capture the true galaxy evolution through cosmic time.

Our findings demonstrate that learnt generative surrogates can transform galaxy formation research by bridging simulations and observations, reducing reliance on costly, repeated hydrodynamical runs. By coupling our blueprint for domain-aware assessment of physical realism with computational scalability, this work marks a significant step towards efficient, next-generation modelling

pipelines, automated survey interpretation, and managing the ensuing data deluge in the SKA era.

## ACKNOWLEDGEMENTS

PD, FS, and EG acknowledge support from SERI as part of the SKACH consortium. We would also like to thank the ZHAW Centre for Artificial Intelligence’s science cluster admin M. Stadelmann for his support and management of the high resources this project demanded.

## DATA AVAILABILITY

The original source of the dataset is publicly released by the [IL-  
LUSTRIS-TNG project](#). The extracted dataset and model weights can be shared upon reasonable request. Our PyTorch-based code used for the training of the presented deep learning models is publicly released on GitHub under a GPLv3 license ([chuchichaestli](#)) and ([skais-mapper](#)), including scripts and *hydra* configurations ([Yadan 2019](#)) to re-create the results in this work.

## REFERENCES

- Amirian M., Barco D., Herzig I., Schilling F.-P., 2024, *IEEE Access*, 12, 10281  
 Andersson J., Ahlström H., Kullberg J., 2019, *Magnetic Resonance in Medicine*, 82, 1177  
 Arjovsky M., Bottou L., 2017, preprint ([arXiv:1701.04862](#))  
 Arjovsky M., Chintala S., Bottou L., 2017, preprint ([arXiv:1701.07875](#))  
 Ba J. L., Kiros J. R., Hinton G. E., 2016, preprint ([arXiv:1607.06450](#))  
 Bally J., 2016, *ARA&A*, 54, 491  
 Bassini L., Feldmann R., Gensior J., Faucher-Giguère C.-A., Cenci E., Moreno J., Bernardini M., Liang L., 2024, *MNRAS: Letters*, 532, L14  
 Beck R., 2015, *A&ARv*, 24, 4  
 Bengesi S., El-Sayed H., Sarker M. K., Houkpati Y., Irungu J., Oladunni T., 2024, *IEEE Access*, 12, 69812  
 Berlind A. A., et al., 2003, *ApJ*, 593, 1  
 Bernardini M., Feldmann R., Anglés-Alcázar D., Boylan-Kolchin M., Bullock J., Mayer L., Stadel J., 2021, *MNRAS*, 509, 1323  
 Bianco M., et al., 2025, *MNRAS*, 541, 234  
 Biernacki P., Teyssier R., 2018, *MNRAS*, 475, 5688  
 Binney J., Tremaine S., 2011, *Galactic Dynamics*. Princeton University Press, pp 1–54, doi:[10.2307/j.ctvc778ff](#)  
 Binney J., Vasiliev E., 2023, *MNRAS*, 520, 1832  
 Blau Y., Michaeli T., 2018, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 6228–6237, doi:[10.1109/cvpr.2018.00652](#), <http://dx.doi.org/10.1109/CVPR.2018.00652>  
 Bond-Taylor S., Leach A., Long Y., Willcocks C. G., 2022, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 7327  
 Braun R., Bourke T. L., Green J. A., Keane E., Wagg J., 2015, in Proceedings of Advancing Astrophysics with the Square Kilometre Array - PoS(AASKA14). p. 174, doi:[10.22323/1.215.0174](#), <http://dx.doi.org/10.22323/1.215.0174>  
 Bullock J., Cuesta-Lazaro C., Quera-Bofarull A., 2019, in Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging. p. 69, doi:[10.1117/12.2512451](#)  
 Chadayammuri U., Ntampaka M., ZuHone J., Bogdán Á., Kraft R. P., 2023, *MNRAS*, 526, 2812  
 Cheng S., Yu H.-R., Inman D., Liao Q., Wu Q., Lin J., 2020, in 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). pp 685–690, doi:[10.1109/ccgrid49817.2020.00-22](#)

- Cibinel A., et al., 2019, *MNRAS*, 485, 5631  
 Collaboration M., et al., 2025, *MNRAS*, 537, 3632  
 Colman T., et al., 2024, *A&A*, 686, A155  
 Conselice C. J., 2003, *ApJS*, 147, 1  
 Conselice C. J., 2014, *ARA&A*, 52, 291  
 Conselice C. J., Bershadsky M. A., Jangren A., 2000, *ApJ*, 529, 886  
 Crain R. A., van de Voort F., 2023, *ARA&A*, 61, 473  
 Crain R. A., et al., 2015, *MNRAS*, 450, 1937  
 D’Onofrio M., et al., 2016, *The Physics of Galaxy Formation and Evolution*. Springer International Publishing, pp 585–695, doi:10.1007/978-3-319-31006-0\_8  
 Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, *MNRAS*, 486, 2827  
 Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255, doi:10.1109/cvpr.2009.5206848, <http://dx.doi.org/10.1109/CVPR.2009.5206848>  
 Dhariwal P., Nichol A., 2021, preprint ([arXiv:2105.05233](https://arxiv.org/abs/2105.05233))  
 Dubois Y., et al., 2014, *MNRAS*, 444, 1453  
 Engelbrecht B. N., et al., 2024, *MNRAS*, 536, 1035  
 Esser P., Rombach R., Ommer B., 2020, preprint ([arXiv:2012.09841](https://arxiv.org/abs/2012.09841))  
 Feller W., 1949, in Neyman J., ed., *First Berkeley Symposium on Mathematical Statistics and Probability*, pp 403–432  
 Fielding D., Quataert E., Martizzi D., Faucher-Giguère C.-A., 2017, *MNRAS: Letters*, 470, L39  
 Finke T., Krämer M., Morandini A., Mück A., Oleksiyuk I., 2021, *Journal of High Energy Physics*, 2021, 161  
 Frenk C., White S., 2012, *Annalen der Physik*, 524, 507  
 Gavagnin E., Bleuler A., Rosdahl J., Teyssier R., 2017, *MNRAS*, 472, 4155  
 Golling T., Heinrich L., Kagan M., Klein S., Leigh M., Osadchy M., Raine J. A., 2024, preprint ([arXiv:2401.13537](https://arxiv.org/abs/2401.13537))  
 Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, preprint ([arXiv:1406.2661](https://arxiv.org/abs/1406.2661))  
 Goodfellow I., Bengio Y., Courville A., 2016, *Deep Learning*. MIT Press  
 Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A., 2017, preprint ([arXiv:1704.00028](https://arxiv.org/abs/1704.00028))  
 Harper S. E., Dickinson C., 2018, *MNRAS*, 479, 2024  
 He K., Zhang X., Ren S., Sun J., 2015, preprint ([arXiv:1512.03385](https://arxiv.org/abs/1512.03385))  
 Ho J., Salimans T., 2022, preprint ([arXiv:2207.12598](https://arxiv.org/abs/2207.12598))  
 Ho J., Kalchbrenner N., Weissenborn D., Salimans T., 2019, preprint ([arXiv:1912.12180](https://arxiv.org/abs/1912.12180))  
 Ho J., Jain A., Abbeel P., 2020, preprint ([arXiv:2006.11239](https://arxiv.org/abs/2006.11239))  
 Hopkins P. F., Hernquist L., Cox T. J., Matteo T. D., Robertson B., Springel V., 2006, *ApJS*, 163, 1  
 Hornik K., Stinchcombe M., White H., 1989, *Neural Networks*, 2, 359  
 Hwang H. S., Shin J., Song H., 2019, *MNRAS*, 489, 339  
 Ibrahim D., Kobayashi C., 2023, *MNRAS*, 527, 3276  
 Ingraham J. B., et al., 2023, *Nature*, 623, 1070  
 Ioffe S., Szegedy C., 2015, preprint ([arXiv:1502.03167](https://arxiv.org/abs/1502.03167))  
 Ishiyama T., et al., 2021, *MNRAS*, 506, 4210  
 Isola P., Zhu J.-Y., Zhou T., Efros A. A., 2016, preprint ([arXiv:1611.07004](https://arxiv.org/abs/1611.07004))  
 Jayasumana S., Ramalingam S., Veit A., Glasner D., Chakrabarti A., Kumar S., 2023, preprint ([arXiv:2401.09603](https://arxiv.org/abs/2401.09603))  
 Kennicutt R. C., Evans N. J., 2012, *ARA&A*, 50, 531  
 Kingma D. P., Welling M., 2013, preprint ([arXiv:1312.6114](https://arxiv.org/abs/1312.6114))  
 Kingma D. P., Salimans T., Poole B., Ho J., 2021, preprint ([arXiv:2107.00630](https://arxiv.org/abs/2107.00630))  
 Krumholz M. R., et al., 2014, *Star Cluster Formation and Feedback*. University of Arizona Press, doi:10.2458/azu\_uapress\_9780816531240-ch011  
 Kynkänniemi T., Karras T., Aittala M., Aila T., Lehtinen J., 2022, preprint ([arXiv:2203.06026](https://arxiv.org/abs/2203.06026))  
 Leigh M., Sengupta D., Quétant G., Raine J. A., Zoch K., Golling T., 2024, *SciPost Physics*, 16, 018  
 Li C., Wand M., 2016, preprint ([arXiv:1604.04382](https://arxiv.org/abs/1604.04382))  
 Li Y., Ni Y., Croft R. A. C., Matteo T. D., Bird S., Feng Y., 2021, *Proceedings of the National Academy of Sciences*, 118  
 Loshchilov I., Hutter F., 2016, preprint ([arXiv:1608.03983](https://arxiv.org/abs/1608.03983))  
 Luisi M., et al., 2021, *Science Advances*, 7  
 Maccagni F. M., Blok W. D., 2024, in *Proceedings of the 4th URSI Atlantic RadioScience Conference - AT-RASC 2024*, , doi:10.46620/ursiatrasc24/uxqp4342, <http://dx.doi.org/10.46620/URSIATRASC24/UXQP4342>  
 Maccagni F. M., Serra P., 2025, preprint ([arXiv:2507.18109](https://arxiv.org/abs/2507.18109))  
 Maddox N., et al., 2021, *A&A*, 646, A35  
 Mao X., Li Q., Xie H., Lau R. Y. K., Wang Z., Smolley S. P., 2016, preprint ([arXiv:1611.04076](https://arxiv.org/abs/1611.04076))  
 Marinacci F., et al., 2018, *MNRAS*  
 McAlpine S., et al., 2016, *Astronomy and Computing*, 15, 72  
 McKee C. F., Ostriker E. C., 2007, *ARA&A*, 45, 565  
 Messias H., Guerrero A., Nagar N., Regueiro J., Impellizzeri V., Orellana G., Vioque M., 2024, *MNRAS*, 533, 3937  
 Moore B., Quinn T., Governato F., Stadel J., Lake G., 1999, *MNRAS*, 310, 1147  
 Muratov A. L., et al., 2017, *MNRAS*, 468, 4170  
 Naiman J. P., et al., 2018, *MNRAS*, 477, 1206  
 Nelson D., et al., 2017, *MNRAS*, 475, 624  
 Nelson D., et al., 2019, *MNRAS*, 490, 3234  
 O’Beirne T., et al., 2025, *Publ. Astron. Soc. Australia*, 42, e087  
 Obuljen A., Simonović M., Schneider A., Feldmann R., 2023, *Phys. Rev. D*, 108, 083528  
 Odena A., Dumoulin V., Olah C., 2016, *Distill*  
 Pang Y., Lin J., Qin T., Chen Z., 2022, *IEEE Transactions on Multimedia*, 24, 3859  
 Parmar N., Vaswani A., Uszkoreit J., Kaiser L., Shazeer N., Ku A., Tran D., 2018, preprint ([arXiv:1802.05751](https://arxiv.org/abs/1802.05751))  
 Perraudin N., Srivastava A., Lucchi A., Kacprzak T., Hofmann T., Réfrégier A., 2019, *Computational Astrophysics and Cosmology*, 6, 5  
 Pilipich A., et al., 2017, *MNRAS*, 475, 648  
 Poggianti B. M., et al., 2019, *ApJ*, 887, 155  
 Potter D., Stadel J., Teyssier R., 2017, *Computational Astrophysics and Cosmology*, 4, 2  
 Radford A., Metz L., Chintala S., 2015, preprint ([arXiv:1511.06434](https://arxiv.org/abs/1511.06434))  
 Ramachandran P., Zoph B., Le Q. V., 2017, preprint ([arXiv:1710.05941](https://arxiv.org/abs/1710.05941))  
 Rieder M., Teyssier R., 2017, *MNRAS*, 472, 4368  
 Rives A., et al., 2021, *Proceedings of the National Academy of Sciences*, 118  
 Rønne N., Aspuru-Guzik A., Hammer B., 2024, *Phys. Rev. B*, 110, 235427  
 Ronneberger O., Fischer P., Brox T., 2015, preprint ([arXiv:1505.04597](https://arxiv.org/abs/1505.04597))  
 Salimans T., Ho J., 2022, preprint ([arXiv:2202.00512](https://arxiv.org/abs/2202.00512))  
 Schade D., Lilly S. J., Crampton D., Hammer F., Fèvre O. L., Tresse L., 1995, *ApJS*, 451, L1  
 Schanz A., List F., Hahn O., 2024, *The Open Journal of Astrophysics*, 7  
 Schaye J., et al., 2014, *MNRAS*, 446, 521  
 Schinnerer E., Leroy A., 2024, *ARA&A*, 62, 369  
 Schneider A., Teyssier R., Stadel J., Chisari N. E., Brun A. M. L., Amara A., Refregier A., 2019, *J. Cosmology Astropart. Phys.*, 2019, 020  
 Schneuing A., et al., 2024, *Nature Computational Science*, 4, 899  
 Sharma R., et al., 2025, preprint ([arXiv:2504.00303](https://arxiv.org/abs/2504.00303))  
 Smith L. N., Topin N., 2017, preprint ([arXiv:1708.07120](https://arxiv.org/abs/1708.07120))  
 Sohl-Dickstein J., Weiss E. A., Maheswaranathan N., Ganguli S., 2015, preprint ([arXiv:1503.03585](https://arxiv.org/abs/1503.03585))  
 Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, *MNRAS*, 391, 481  
 Springel V., et al., 2017, *MNRAS*, 475, 676  
 Staveley-Smith L., Oosterloo T., 2015, in *Proceedings of Advancing Astrophysics with the Square Kilometre Array - Pos(AASKA14)*. p. 167, doi:10.22323/1.215.0167, <http://dx.doi.org/10.22323/1.215.0167>  
 Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2015, preprint ([arXiv:1512.00567](https://arxiv.org/abs/1512.00567))  
 Thiele L., Villaescusa-Navarro F., Spergel D. N., Nelson D., Pilipich A., 2020, *ApJ*, 902, 129  
 Tinsley B. M., 2022, preprint ([arXiv:2203.02041](https://arxiv.org/abs/2203.02041))  
 Valentini M., et al., 2019, *MNRAS*  
 Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2017, preprint ([arXiv:1706.03762](https://arxiv.org/abs/1706.03762))  
 Villaescusa-Navarro F., 2018, Pylians: Python libraries for the analysis

- of numerical simulations, Astrophysics Source Code Library, record ascl:1811.008 (ascl:1811.008)
- Villaescusa-Navarro F., et al., 2018, *ApJ*, 866, 135
- Wang Z., Bovik A., Sheikh H., Simoncelli E., 2004, *IEEE Transactions on Image Processing*, 13, 600
- Ward S. R., Costa T., Harrison C. M., Mainieri V., 2024, *MNRAS*, 533, 1733
- Weinberger R., Springel V., Pakmor R., 2020, *ApJS*, 248, 32
- Weissenborn D., Täckström O., Uszkoreit J., 2019, preprint ([arXiv:1906.02634](https://arxiv.org/abs/1906.02634))
- Whang J., 2023, PhD thesis, Computer Science
- White S. D. M., Frenk C. S., 1991, *ApJ*, 379, 52
- Woodland M., et al., 2024, Feature Extraction for Generative Medical Imaging Evaluation: New Evidence Against an Evolving Trend. Springer Nature Switzerland, pp 87–97, doi:10.1007/978-3-031-72390-2\_9, [http://dx.doi.org/10.1007/978-3-031-72390-2\\_9](http://dx.doi.org/10.1007/978-3-031-72390-2_9)
- Wu Y., He K., 2018, preprint ([arXiv:1803.08494](https://arxiv.org/abs/1803.08494))
- Yadan O., 2019, Hydra - A framework for elegantly configuring complex applications, Github, <https://github.com/facebookresearch/hydra>
- Yang B., Wang L., Wong D., Chao L. S., Tu Z., 2019, preprint ([arXiv:1904.03107](https://arxiv.org/abs/1904.03107))
- Yao W., Zeng Z., Lian C., Tang H., 2018, *Neurocomputing*, 312, 364
- Zhang H., Goodfellow I., Metaxas D., Odena A., 2018a, preprint ([arXiv:1805.08318](https://arxiv.org/abs/1805.08318))
- Zhang R., Isola P., Efros A. A., Shechtman E., Wang O., 2018b, preprint ([arXiv:1801.03924](https://arxiv.org/abs/1801.03924))
- Zubovas K., Tarténas M., Bourne M. A., 2024, *A&A*, 691, A151
- de Blok W. J. G., et al., 2024, *A&A*, 688, A109

This paper has been typeset from a  $\text{\TeX}$ / $\text{\LaTeX}$  file prepared by the author.