LyAl-Net：利用神经网络进行高效的莱曼-α森林模拟

Chotipan Boonkongkird[1] ⋆、Guilhem Lavaux[1], 、Sebastien Peirani[1,2], 、Yohan Dubois[1], 、Natalia Porqueres[3], 和Eleni Tsaprazi[4]

[1] 索邦大学、法国国家科学研究中心、UMR 7095、巴黎天体物理研究所，法国巴黎阿拉戈大道98 bis号，邮编75014[2] 蔚蓝海岸大学、蔚蓝海岸天文台、法国国家科学研究中心、拉格朗日实验室，法国尼斯天文台大道，CS 34229，邮编06304[3] 牛津大学物理系，丹尼斯·威尔金森楼，基布尔路，英国牛津 OX1 3RH[4] 奥斯卡·克莱因中心，斯德哥尔摩大学物理系，阿尔巴诺瓦大学中心，瑞典斯德哥尔摩 SE 106 91

接收日期：2023年3月30日

arXiv:2303.17939v1 [astro-ph.CO] 31 Mar 2023

摘要

背景。 宇宙学量的推断需要准确且大体积的宇宙学流体动力学模拟。不幸的是，即使在宇宙学尺度上仅覆盖适中的范围，其计算时间也可能达到数百万CPU小时（$\approx (100h^{-1}$ 兆秒差距$)^3$）。生成大量Lyman-α 观测模拟数据的可能性，为更好地控制宇宙学参数推断中的协方差矩阵估计，以及重子效应导致的系统误差影响，提供了机会。

目标。 我们提出了一种机器学习方法，用于模拟星系际介质物理过程的流体动力学模拟，以生成Lyman-α 森林，该方法称为*LyAl-Net*。这项工作的主要目标是提供高效且低成本的模拟，同时保留对气体场层面信息的解释能力，并将其作为开展其他宇宙学探索的工具。

方法。 我们使用一种基于U-net架构的神经网络来预测中性氢的物理性质、密度和温度；该架构是卷积神经网络的一种变体。我们使用Horizon-noAGN模拟训练*LyAl-Net* 模型，但仅使用了其中9%的体积。我们还通过使用包含不同重子反馈的宇宙学模拟测试迁移学习框架，探究该模型的稳健性。我们通过分析不同情景下模拟场的一点和两点统计量及其随机性质，对结果进行了检验。

结果。 模拟得到的莱曼-α 森林吸收的系综平均随红移的变化，与完整流体动力学模拟得出的结果相差不超过 2.5%。根据暗物质密度计算单个场的结果，与宇宙场的常规物理状态吻合良好。在 IllustrisTNG100 上测试的结果表明，无需任意重新缩放，莱曼-α 森林通量也有了显著改善。

结论。 这类模拟器对于充分利用 WEAVE-QSO 等即将开展的巡天可能至关重要。迁移学习技术有望降低对所学习物理模型的敏感性。这项技术可能对基于当前实验得出的结果产生决定性影响，例如基于 SDSS3 的类星体巡天（其分辨率 R = 1500），以及 SDSS4（R = 2000）数据；这些正是我们的模拟器所适用的分辨率范围。

关键词。 方法：数值方法；类星体：吸收线；星系际介质；宇宙大尺度结构

1. 引言

T莱曼-α 森林是由一系列尖锐的吸收特征组成的，
g通常从遥远的类星体（QSO）
s光谱中观测到，最早由 Lynds（1971）观测到。这一现象源于
f背景源发射的光子，其频率
q高于或等于莱曼-α 频率，并经历红移-
s由于哈勃膨胀而发生红移，并被介质吸收——
i 星系际介质。由于这些吸收发生在视线方向上距观测者不同的距离处，因此也会发生红移。这一机制在观测到的类星体宽带 Lyman-α 发射光中形成了类似森林的特征集合。尽管这一过程仅发生于中性氢（HI）原子，而中性氢只占星系际介质（IGM）的一小部分，但人们可以利用 Lyman-α 森林的性质来推断宇宙中重子物质的空间结构。对于 concordance ΛCDM 模型而言，物质的绝大多数以暗物质（DM）的形式存在（Jarosik 等人 2011年；PlanckCollaboration VI 2020年）。

宇宙中的物质以暗物质（DM）的形式存在（Jarosik 等人 2011年；PlanckCollaboration VI 2020年）。

因此，重子的空间分布主要由暗物质密度所刻画的引力势决定。因此，在这一宇宙学模型中，莱曼-α森林观测数据是总物质的良好示踪物（Petitjean 等人 1995年；Hui与Gnedin 1997年；Croft 等人 1998年）。此外，这对宇宙学至关重要，因为在莱曼-α所探测的相关红移处，我们没有像在低红移（$z \lesssim 1$）时那样多的易于观测的星系。因而，莱曼-α森林成为对$z \simeq 2-3$处物质聚集的互补探针，该时期也被称为宇宙正午时代。

莱曼-α 森林能够为宇宙学模型提供大量信息。该领域以往的研究包括，例如，使用贝叶斯反演方法恢复星系际介质的性质（Pichon 等人 2001年）、约束中微子质量（Palanque-Delabrouille 等人 2015年；Yèche 等人 2017年），以及从一维莱曼-α 吸收中推断三维物质分布，采用贝叶斯

⋆ chotipan.boonkongkird@iap.fr

前向建模（Porqueres等人 2019年；Horowitz等人 2019年；Porqueres 等（2020）；Kraljic 等人 2022年）、约束星系际介质的热历史（Villasenor 2022年），以及对重子声学振荡（BAO）的测量；重子声学振荡能够对宇宙的膨胀历史提供严格约束（Font-Ribera 2014年）。如今，多个大型宇宙学巡天都包括类星体巡天及其光谱测量（BOSS、eBOSS、DESI；Lee 等人 2013年；du Mas des Bourboux 等人 2020年；Levi 等人 2019年）。

为了分析莱曼-α 森林观测数据，我们需要对星系际介质进行流体动力学模拟，并同时运行暗物质N体模拟。尽管在大尺度上，星系际介质与暗物质超密度产生的引力势相耦合，但在小尺度上，它主要受重子激波、冷却和反馈支配，而这些过程没有解析解。因此，必须通过数值方法求解星系际介质的状态。随着 WEAVE-QSO（Pieri 等，2016）等宇宙学巡天即将开展，所需的分辨率和体积都将提高和增大，进而大幅增加所需的计算时间。因此，使用全流体动力学模拟来覆盖相同体积已变得难以实现。另一种方法是采用纯暗物质的N体模拟，其速度远快于流体动力学模拟。这样就可以避免求解流体动力学过程，转而从现象学角度校正小尺度上的影响（$\sim 100h^{-1}$kpc，即发生莱曼-α 现象的金斯尺度）。

涨落冈恩-彼得森近似（FGPA，Gunn 和 Peterson 1965；Weinberg 等人 1997年）是根据暗物质超密度估算模拟莱曼-α 的一种主要框架。然而，这种近似确实存在一些局限性。例如，它无法捕捉小尺度上的重子反馈过程；在密度较高且气体受到强烈加热的区域，幂律会失效（Luki´c 等人 2014；Kooistra 等人 2022）；并且当$z \gtrsim 3$时，幂律开始脱耦。在大多数情况下，暗物质超密度在用于仿真前都会经过平滑。例如，通过将物质密度与高斯核卷积来仿真重子压力（Hui与Gnedin 1997）。其他研究（Sorini 等人 2016）提出了一种名为迭代匹配统计（IMS）的数值模拟方法，旨在提高精度，使其优于高斯平滑。最后，LyMAS-2 技术（Peirani等人2014；Peirani 等（2022））利用条件概率，将通量映射自平滑后的暗物质场，并引入维纳滤波器，以更好地表示沿视线方向的相干性。

近年来，天体物理学界越来越多地采用机器学习模型，这是因为它们能够提供通用的拟合函数，并且可以通过图形处理器（GPU）等加速器快速进行预测，从而在存在训练数据的前提下，同时实现高精度和高速度的模型。受 Peirani 等（2014）工作的启发，我们提出用神经网络替代条件概率密度函数。在本工作中，我们将测试一种*U-Net* 架构，重点模拟描述 Lyman-α所需的最关键星系际介质场。我们介绍了这样一种模拟器，将其命名为*LyAl-Net*。该模拟器经过训练，可将N体模拟中的暗物质密度场映射为源自配套流体动力学模拟的星系际介质场（中性氢密度和温度）。在本工作中，这两项模拟分别为 Horizon-DM 和 Horizon-noAGN，红移为$z = 2.43$（来自 Horizon-AGN 模拟套件；Dubois 等，2014；Peirani等人（2017））。随后，我们从模拟场中推导出 Lyman-α 森林的吸收特征。此前的工作采用了类似的初始策略，以推导星系际介质的发射和吸收特征（Villaescusa-Navarro 等，2013；Harrington 等，2021）。在本工作中，我们通过开展更多测试来评估模型的准确性，并提高模型使用上的灵活性，使其能够应用于观测研究中。

我们还探索了重新校准状态方程的方法，以改进不同气体物理条件下模拟的莱曼-α 吸收，使现有的*LyAl-Net* 能够应用于截然不同的条件，而只需付出少量额外训练集的代价。这样，我们就能直接影响该物理量，而不只是调整平均透射通量。这是快速模拟莱曼-α 森林的强大工具，并有望将机器学习模型推广到训练集之外的不同气体物理过程。本文旨在为生成更精确的气体流体动力学模拟器，尤其是莱曼-α 森林模拟器，提供一个框架和概念验证。

本文的组织结构如下。我们将在第2节讨论所使用的莱曼-α 吸收系统的物理过程和模型。在第3节中，我们阐述并描述用于训练的神经网络架构，包括数据变换和训练超参数。在第4节中，我们使用 Horizon-noAGN 和 Horizon-DM 展示预测结果及准确性基准测试。在第5节中，我们将 IllustrisTNG（Nelson et al. 2019）作为测试集，讨论相关结果以及用于微调状态方程的迁移学习方法。最后，我们将在第6节讨论结果并作出总结。

2. 莱曼-α 吸收的物理模型

本节简要回顾产生观测到的莱曼-α森林的物理机制基础。Meiksin（2009）的早期研究详细阐述了这一推导所依据的原子物理学。这里，我们着重介绍框架试图建模的相关物理量。

观测到的莱曼-α 吸收特征背后的机制可分为几个组成部分：莱曼-α 跃迁共振的本征线宽、原子小尺度随机运动导致的热展宽，以及氢云整体运动造成的多普勒频移。

来自一个类星体的宽带莱曼-α 辐射的吸收强度，使我们能够推断视线上不同距离处中性氢的密度。吸收率取决于相互作用的次数。我们可以将其称为不透明度，它是观测频率 $\nu_{\rm obs}$ 的函数，表示为

$$\tau(D_{\rm QSO}, \hat{u}, \nu_{\rm obs}) = \int_0^{s_{\rm QSO}(D_{\rm QSO})} n_{\rm HI}(s, \hat{u}) \sigma_{\rm HI}(\nu_{\rm obs}, s, \hat{u})\, {\rm d}s\,, \qquad (1)$$

其中，$n_{\rm HI}$ 是中性氢原子的数密度，$\sigma_{\rm HI}$是光子与中性氢原子之间的相互作用截面。我们注意到，s 是本征距离。因此，$s_{\rm QSO}(D_{\rm QSO})$ 是与所考虑视线方向上类星体的共动距离$D_{\rm QSO}$ 相对应的本征距离。

为了推导所有这些量的表达式，我们首先考虑空间中某个原子氢静止参考系下的情形。原子截面可以用洛伦兹线型描述，其表达式如下

$$\sigma_{\rm HI}(\nu) = \frac{\pi e^2}{m_e c} f_{lu} L(\nu) = \frac{\pi e^2}{m_e c} f_{lu} \frac{\Gamma_{ul}/(4\pi^2)}{(\nu - \nu_{lu})^2 + (\Gamma_{ul}/(4\pi)^2)}\,, \qquad (2)$$

其中，$L(\nu)$ 为洛伦兹线型，ν_{lu} 为从较低能级（l）跃迁到较高能级（u）的频率，Γ_{ul} 为较高

能级阻尼宽度，f_{lu} 为振子强度，m_e 和 e 分别为电子质量及其电荷。我们注意到，Lyman-α 系统的跃迁能级为 $l = 1$ 和 $u = 2$，由此得到 $f_{12} = 0.4162$。

除了基本跃迁的概率外，我们还必须考虑另外两个组成部分。中性氢原子具有有限的正温度，这会引起它们的热运动；对于非相对论气体而言，这种热运动可以用麦克斯韦分布很好地描述。这种效应称为多普勒展宽，我们用 $G(\nu, T)$ 表示该分布。这种展宽取决于气体的温度，而温度本身又随位置变化。由于中性氢云团相对于类星体发射的光子处于运动状态，我们还必须施加额外的多普勒增益，将入射光子变换到云团的静止参考系中。

将光子发射器的红移考虑在内后，可得两个参考系中的频率关系为

$$\nu = \nu_{\rm obs}(1+z)\left(1+\frac{v_z}{c}\right) \tag{3}$$

其中 ν 是中心频率，$\nu_{\rm obs}$ 是观测频率，而v_z 是视线速度。因此，在这种情况下，截面是洛伦兹轮廓与多普勒轮廓的卷积，

$$\sigma_{\rm HI}(\nu, T) = \frac{\pi e^2}{m_e c} f_{12}(L * G)(\nu, T) = \frac{\pi e^2}{m_e c} f_{12} V(\nu, T). \tag{4}$$

这两个剖面的卷积称为沃伊特剖面，我们将其写作 $V(\nu, T)$：

$$V(\nu, T) = \int_{\nu'=-\infty}^{\infty} {\rm d}\nu' \frac{\Gamma_{21}/(4\pi^2)}{(\nu'-\nu_{12})^2 + (\Gamma_{21}/(4\pi)^2)} \frac{1}{\sqrt{\pi}\Delta\nu_D(T)} \exp\left[-\left(\frac{\nu-\nu'}{\Delta\nu_D(T)}\right)^2\right], \tag{5}$$

其中

$$\Delta\nu_D = \frac{\nu_{12}}{c}\sqrt{\frac{2k_B T}{m_H}} \tag{6}$$

是多普勒宽度，即由多普勒效应引起的莱曼-α 频率展宽，而 ν_{12} 是莱曼-α 频率。

通过引入 $x = (\nu - \nu')/\Delta\nu_D(T)$，我们可以重新排列各项以简化表达式，其中该量表示以 $\Delta\nu_D$ 为单位的频率偏移。我们令 $a = \Gamma_{21}/(4\pi\Delta\nu_D)$ 表示莱曼-α 谱线宽度与多普勒频率宽度 $\Delta\nu_D$ 之比。现在可以将方程(5)改写为：

$$V(\nu, T) = \int_{x'} {\rm d}x' \left(\frac{1}{\pi\Delta\nu_D}\right)\left(\frac{a}{(x(\nu)-x')^2 + a^2}\right)\frac{1}{\sqrt{\pi}\Delta\nu_D}\exp\left(-x'^2\right). \tag{7}$$

因此，沃伊特剖面可写为

$$V(a, x) = \frac{a}{\pi^{3/2}\Delta\nu_D}\int_{-\infty}^{\infty}\frac{e^{-y^2}}{(x-y)^2+a^2}dy\,. \tag{8}$$

沃伊特剖面的归一化值为 1。我们进一步对表达式进行变换，以沃伊特函数 $V(a,x)$ 表示沃伊特剖面 H，即

$$V(a, x) = \frac{1}{\sqrt{\pi}\Delta\nu_D}H(a, x)\,. \tag{9}$$

下一节中，我们会再谈使用此函数提高计算效率的原因。按照这一约定，我们可以使用吸收的振幅，从而得到

$$H(a, x) = \frac{a}{\pi}\int_{-\infty}^{\infty}\frac{e^{-y^2}dy}{a^2+(x-y)^2}\,. \tag{10}$$

因此，在观测者参考系中，莱曼-α 吸收的截面为：

$$\sigma_{\rm HI}(\nu_{\rm obs}) = \frac{\pi e^2}{m_e c} f_{12}\frac{H(a, x)}{\sqrt{\pi}\Delta\nu}\,. \tag{11}$$

利用公式（11）中的截面，我们随后计算归一化的Lyman-α吸收率，即Lyman-α 森林，表示为

$$F(\lambda) = \exp\left[-\tau\left(\nu_{\rm obs}\right)\right]. \tag{12}$$

因此，我们得到了一个用于计算由$F(\lambda)$编码的吸收特征的明确算法。我们在第3.3节详细介绍数值实现。需要指出的是，我们尚未讨论紫外背景的影响；紫外背景会直接影响中性氢的含量（Becker & Bolton 2013）、自旋温度与气体温度之间的关系（Liszt 2001），以及可能存在的相对论效应（Iršič et al. 2016）。此外，通过考虑函数$F(\lambda)$，我们假设类星体的Lyman-α谱线轮廓是完全已知的，因为这并非本文的研究对象。我们只关注如何以最快的速度模拟流体动力学模拟可能产生的最广泛影响，从而获得函数$F(\lambda)$。

2.1. 基准模型：*modified-FGPA*

为评估我们的模型生成莱曼-α 吸收的性能，我们采用与涨落 Gunn-Peterson 近似（FGPA，Gunn 和 Peterson 1965；Weinberg 等，1997）类似的方法。该解析框架基于暗物质引力势会吸引重子物质这一观点，从暗物质超密度估算模拟的莱曼-α （Bi 和 Davidsen，1997）。该近似表示为

$$\tau(z, \mathbf{r}) \propto (1+\delta(\mathbf{r}))^{\beta}, \tag{13}$$

其中，τ 是红移z 处、共动位置$\mathbf{r}$上的星际介质光学深度，δ 是暗物质超密度，而 β 由温度-密度幂律关系的斜率推导得出，该关系取决于中性氢碰撞复合与光电离之间的平衡态。

我们对每个独立场采用类似的近似，由此得到以下中性氢密度关系：

$$n_{\rm HI}{}^{\rm mock} = \bar{n}_{\rm HI}\left(\frac{\rho_{\rm DM}}{\bar{\rho}_{\rm DM}}\right)^{\alpha_n}, \tag{14}$$

其中，$\bar{\rho}_{\rm DM}$ 是暗物质平均密度，$n^{\rm mock}{}_{\text{中性氢}}$ 是估算的中性氢数密度，而 $\bar{n}_{\rm HI}$ 是幂律系数。

并且，根据温度-密度幂律关系，我们可以通过以下方式估算模拟氢温度：

$$T_{\rm HI}^{\rm mock} = \bar{T}_{\rm HI}\left(\frac{\rho_{\rm DM}}{\bar{\rho}_{\rm DM}}\right)^{\alpha_T} \tag{15}$$

其中，$T^{\rm mock}$ 是估算的氢温度，$\bar{T}_{\rm HI}$ 是幂律的系数。暗物质的参数

表 1. 方程（14）和（15）中暗物质与星系际介质参数幂律关系的拟合参数。我们使用 Horizon-DM 的暗物质超密度，以及 Horizon-noAGN 的气体密度和温度。

气体参数（X）	$\bar{\mathbf{X}}$	$\alpha_{\mathbf{X}}$
温度	3.98	0.58
密度	-10.43	1.27

通过 Horizon-DM 拟合得到的参数，以及 Horizon-noAGN 的中性氢密度和温度，列于表 1。从这里开始，我们将基准模型称为 modified-FGPA（mod-FGPA）。此外，为了将 modified-FGPA 与 *LyAl-Net* 进行比较，我们将估算的密度和温度代入方程（1），该方程还包含沃伊特剖面，并假设暗物质和气体的视线速度近似相同。

我们还要指出，严格来说，该模型并不是 FGPA，但它采用了相同的幂律近似来描述气体状态场与暗物质之间的关系。我们预计，这项测试的结果将优于经典 FGPA，但不如使用 *LyAl-Net* 得到的结果，因为它的灵活性较低，并且忽略了气体的随机成分。

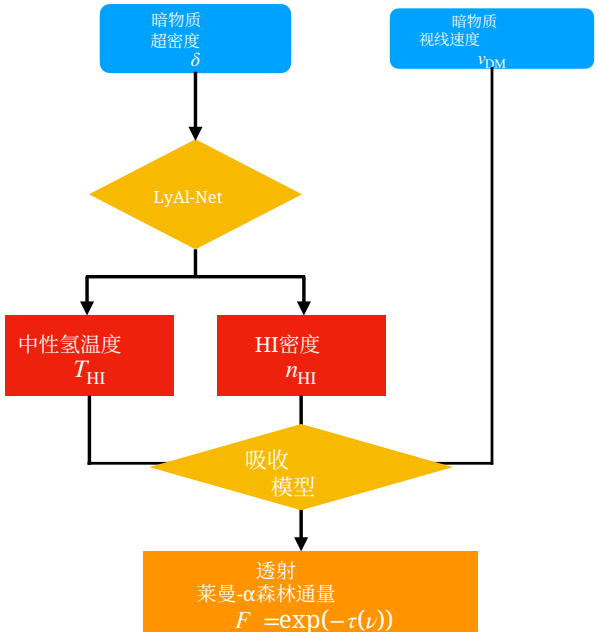

图 1. 该流程图展示了用于莱曼 α 森林模拟的*LyAl-Net* 流程示意图。通量吸收模型（方程（12））需要三种气体物理量，即温度、密度和视线速度。LyNet 仅仿真中性原子氢的气体温度和气体密度。我们使用同一*N*体模拟中的暗物质视线速度，作为近似的气体视线速度。

3. 方法

我们使用的模型需要三个要素来计算莱曼-α 宽发射线在不同频率处的吸收，这些要素用于表征中性氢的物理状态：温度、气体密度和视线方向上的气体速度。因此，本研究的主要目标是构建一个神经网络，用于生成计算莱曼-α森林模型所需的这些参数。我们针对每种气体参数分别训练 *LyAl-Net* ，以便进行优化并解读预测结果。本节将详细介绍作为训练集的宇宙学模拟、生成莱曼-α 吸收模拟数据的数值方法，以及数据预处理流程。

3.1. 模拟数据集

我们使用的模拟是 Horizon-AGN模拟。这是一种基于自适应网格细化代码 Ramses（Teyssier 2002）的宇宙学流体动力学模拟。该模拟包含 1024^3 暗物质粒子，分辨率为$M_{\rm DM,\ res} = 8 \times 10^7\ {\rm M}_\odot$，模拟体积为边长$100h^{-1}$兆秒差距的立方体。该模拟耗费了400万CPU小时才完成，红移低至$z = 1.2$。Horizon-AGN采用标准的 ΛCDM宇宙学进行计算，并采用与 WMAP任务观测数据相符的宇宙学参数（Komatsu 等（2011））：$\Omega_m = 0.272$、$\Omega_\Lambda = 0.728$、$\sigma_8 = 0.81$、$\Omega_b = 0.045$、$H_0 = 70.4\ {\rm kms^{-1}}$兆秒差距$^{-1}$，以及$n_s = 0.967$。此外，模拟将气体建模为绝热指数为 $\gamma = 5/3$的理想单原子气体。气体通过一种依赖金属丰度的氢和氦光子发射模型冷却（Sutherland 与 Dopita（1993））。这使气体温度能够降至 10^4 K。同时，气体还会通过 Haardt 与 Madau（1996）的模型，在红移$z_{\rm reion} = 10$之后受到均匀紫外背景的加热。在本研究中，我们使用 Horizon-noAGN 作为训练集。它是 Horizon-AGN 的姊妹模拟，两者的一个主要区别是：由于缺乏黑洞增长，noAGN 变体不包含 AGN 反馈。使用 noAGN 的原因是，模拟的计算时间因关闭 AGN 反馈、并通过微调紫外背景辐射来模拟其作用以获得相同的气体性质而缩短了（Peirani等人（2022））。

使用 Horizon-noAGN 的另一个优点是，我们能够研究 Lyman-α 森林模型的影响和敏感性。我们还使用 TNG100 模拟作为验证集，以便在第 5.4 节对结果进行适当的交叉核验。

3.2. 用于训练和验证的数据集生成 *ion*

本节将介绍创建训练和验证机器学习模型所需的不同数据集的步骤。

在我们所考虑的宇宙学模拟中，暗物质的相空间分布由粒子采样。为了实现我们的仿真策略，需要将密度映射到规则网格上，以便进行计算。我们将粒子分配到一个 1024^3 网格上，每个体素的分辨率为 ~ $98h^{-1}$ 千秒差距。现有多种质量分配方案，其中最常用的是云团网格（Cloud-In-Cell）分配法（Hockney 与 Eastwood（1988））。遗憾的是，这种方法可能导致某些网格单元完全没有质量，在宇宙空洞中的采样效果也很差，而莱曼-α 森林的信号恰恰在这些区域最为显著（Porqueres 等（2020））。我们采用 Colombi 等人（2007）提出的粒子分配方法来避免这些问题。该方法基于光滑粒子流体动力学（SPH）核导出的自适应滤波器（Monaghan（1992））。

这里简要提醒读者相关步骤。粒子系统的完整经典描述由

相空间分布 $f(\mathbf{r}, \mathbf{v}_L)$ 给出。进行计算时，我们通常需要该分布的矩，对于速度而言，即：

$$\mathbf{v}(\mathbf{r}) = \frac{1}{\rho} \int \mathrm{d}^3\mathbf{v}_L \; \mathbf{v}_L f(\mathbf{r}, \mathbf{v}_L), \tag{16}$$

其中，$f(\mathbf{r}, \mathbf{v})$ 是暗物质粒子的相空间密度。质量密度 $\rho(\mathbf{r})$ 同样通过以下积分获得

$$\rho(\mathbf{r}) = \int \mathrm{d}^3\mathbf{v}_L f(\mathbf{r}, \mathbf{v}_L). \tag{17}$$

实际上，我们将方程 (16) 和 (17) 计算到由小型立方网格块组成的网格中 Δr^3, , 并对所得场进行分段近似。为降低散粒噪声，我们采用类似于光滑粒子流体动力学的方法进行这些计算（Monaghan（1992））。我们可以将每个粒子表示为具有有限尺寸的平滑云团，其尺寸基于局部密度，而局部密度取决于邻近粒子之间的距离。通常，我们将云团尺寸定义为 $2R_{\mathrm{SPH}}$，邻近粒子的数量定义为 N_{SPH}。通过恰当地选择网格节点权重，我们使用的算法能够保持质量和动量守恒，具体如下所述。

为了将粒子插值到网格位置，我们使用以下方程

$$\tilde{A}(i,j,k) = \frac{1}{[R_{\mathrm{SPH}}(i,j,k)]^3} \left[\sum_{l=1}^{N_{\mathrm{X}}-1} A_l W_l \mathcal{S}\left(\frac{d_l}{R_{\mathrm{SPH}}(i,j,k)} \right) \right]. \tag{18}$$

该方程给出了粒子 l 携带的量 A_l 的加权和。每个相邻粒子都乘以 $\mathcal{S}(x)$，即一个 SPH 核函数，其中 x 是相对距离，通过计算粒子 l 与网格节点 (i,j,k) 之间的距离 d_l 与 $R_{\mathrm{SPH}}(i,j,k)$ 的比值得出。它们通过 W_l 因子进行加权，使得每个网格位置上的总和等于 1。此约束得出每个粒子权重的以下恒等式：

$$W_l = 1/S_l = \left[\sum_{i,j,k} \frac{1}{[R_{\mathrm{SPH}}(i,j,k)]^3} \mathcal{S}\left(\frac{d_l}{R_{\mathrm{SPH}}(i,j,k)} \right) \right]^{-1}. \tag{19}$$

对于将质量赋予网格这一特定情况，我们得到网格每个网格节点上的质量。将其除以该网格节点对应的体积，即可得到质量密度。因此：

$$\tilde{\rho}(i,j,k) = \tilde{m}(i,j,k)/\Delta r^3 \; . \tag{20}$$

类似地，插值速度根据方程 (16) 得出，该方程为网格给出了以下恒等式：

$$\tilde{v}(i,j,k) = \tilde{p}(i,j,k)/\tilde{m}(i,j,k) \; , \tag{21}$$

其中 $\tilde{}p$ 是单元的插值动量。

与暗物质不同，Ramses 的气体模拟采用自适应网格细化结构。每个原始气体网格单元具有不同的共动尺寸。最大的网格单元范围为 $0.0977h^{-1}$ 兆秒差距，而最小网格单元的长度为该尺寸的 1/32。我们根据气体网格单元的坐标，将其分配到最近的网格点，从而统一分辨率，这与最近网格点方案类似。某些网格单元可能包含多个子网格单元，例如两个 $L^{\mathrm{largest}}/4$ 网格单元和一个 $L^{\mathrm{largest}}/2$ 网格单元。我们在对这些量取平均时，无论长度如何都使用相同的权重；鉴于我们所研究的尺度，这是一个足够准确的近似，因为我们正在研究 Lyman-α 系统。该算法可以很容易地表示为

$$X_{cell}(i,j,k) = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{22}$$

其中 X_{cell} 是坐标 (i,j,k) 处的目标参数，而 x_i 是目标参数的子网格单元值。

3.3. 模拟莱曼-α 吸收的生成

我们已在第 2 节讨论过莱曼-α 森林的物理过程。现在来介绍如何根据中性氢参数，通过数值方法计算归一化通量。方程 (1) 中定义的给定视线上的气体不透明度是观测频率的连续函数。然而，由于数据集采用网格结构，我们以黎曼和取代方程 (1) 来计算积分。此时不透明度可写为

$$\tau_{\boldsymbol{a},k} = \sum_{j=0}^{j=N} n_{\mathrm{HI},\boldsymbol{a},l_j} \sigma_{\mathrm{HI}}(\nu_{\mathrm{obs},j}, \boldsymbol{a}, j) \times \delta l \; , \tag{23}$$

其中，j 是视线方向上的位置索引，$l_j = j \times \delta l$ 是从观测者到源的本征距离，$\boldsymbol{a}$是垂直于源平面的单位向量，δl 是每个网格单元宽度对应的物理长度，在我们的计算中为 $\delta l = 0.098h^{-1}$ 兆秒差距。沃伊特函数的数值计算开销很大。我们选择使用 Faddeeva 函数，该函数已作为 SciPy 的一部分预先计算（Jones 等，2001年）。关于该函数与我们实际计算之间的关系，请参阅附录A。随后，我们依据方程 (12) 计算归一化吸收通量 $F(\nu)$。我们不考虑模拟类星体自身宽带发射莱曼-α 谱线的确切形状；在观测场景中，这通常需要估算。

将模拟的空间分辨率转换为光谱仪分辨率后，得到的分辨率为 $R = \frac{\langle \Delta\lambda \rangle}{\langle \lambda \rangle} \approx 30000$。这一分辨率远高于当前宇宙学巡天所能达到的水平。例如，WEAVE-QSO 在低分辨率和高分辨率模式下的分辨率分别为$R = 5000$ 和 $R = 20000$（Dalton 等，2014）；斯隆数字巡天第四阶段（BOSS 和 eBOSS）的分辨率为 $R = 2000$。这使我们能够估算模拟相对于实际巡天的误差预算。当然，高分辨率并不意味着模拟中的物理过程准确，也不意味着模拟在该分辨率下包含了所有重子物理。我们将在第 5 节通过与其他模拟所得结果进行比较，考察模型的可靠性。

3.4. 训练前对场进行预变换

暗物质密度场通常用密度对比度来描述，其定义如下：

$$\delta_{\mathrm{DM}} = \frac{\rho_{\mathrm{DM}}}{\bar{\rho}_{\mathrm{DM}}} - 1 \; , \tag{24}$$

其中，$\bar{\rho}_{\mathrm{DM}}$ 是暗物质平均密度。按照构造，δ_{DM} 的一维分布呈严重偏斜，取值范围为 -1 到无穷大。这对模拟器来说未必是最方便的处理方式。此外，只要灵活性足够，偏斜的数据集不会影响神经网络。不过，这种变换有几个好处：它压缩了动态范围，使其成为尺度不变过程，并且有助于

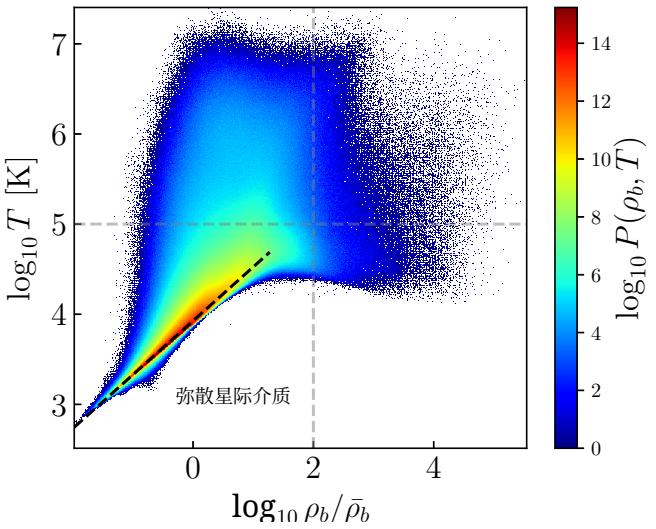

图 2。 Horizon-noAGN模拟中氢密度与温度的密度-温度（$n - T$）图，$z \approx 2.4$。黑色虚线表示弥散星际介质区域内拟合得到的幂律关系，该区域是Lyman-α森林的主要贡献来源，占总体积的99.5%，占总氢质量的76.7%。浅灰色虚线划分了不同的重子区域。我们在该相图中提供了密度图，使用右侧的对数刻度为每个点着色。

提高训练速度，并减少达到给定精度所需的样本量。

因此，我们将提供由密度对比度导出的变换后密度 A_ρ，表示为

$$A_\rho(\rho_{\rm DM}) = \log_{10}\left(\frac{\rho_{\rm DM}}{\bar{\rho}_{\rm DM}}\right) = \log_{10}(1+\delta_{\rm DM})\,. \tag{25}$$

同样，我们对中性氢的密度和温度进行了十进对数变换。

$$A_{n,\rm HI} = \log_{10}\frac{n_{\rm HI}}{1\,{\rm g\ cm^{-3}}}\,, \tag{26}$$

and

$$A_{T,\rm HI} = \log_{10}\frac{T_{\rm HI}}{1\ {\rm K}}\,. \tag{27}$$

选择对密度和温度进行这种变换，不仅是为了使输入与预测所用的尺度保持一致，也有助于我们在预测分析中解读暗物质分布的差异。最重要的是，依据扩散星系际介质（IGM）环境中密度与温度呈幂律关系这一观点（该环境通常位于空洞区域；Hui与Gnedin 1997年；Schaye等1999年；Meiksin，2009），温度—密度关系可表示为

$$T = T_0\left(\frac{\rho_b}{\bar{\rho}_b}\right)^{\gamma-1}\,, \tag{28}$$

其中，γ 为绝热系数，T_0 为平均密度下的温度，ρ_b 为重子物质的密度，$\bar{\rho}_b$ 为重子物质的平均密度。由于我们关注的是中性氢，为简单起见，假设 $\frac{\rho_b}{\bar{\rho}_b} \approx n^{\rm HI}_{\rm HI}$。因此，对这一关系取对数后，可得到如下线性关系：

$$A_{T,\rm HI} = A_{\bar{T},\rm HI} + (\gamma-1)\,(A_{n,\rm HI} - A_{\bar{n},\rm HI})\,. \tag{29}$$

我们在图 2 中展示了密度—温度图，并以温度阈值 $T > 10^5$ K 和密度阈值 $n_{\rm H} < 10^{-4}(1+z)$ cm^{-3} 划定弥散星系际介质区域（Martizzi等，2019）。这一气相包含了大部分重子物质密度，占总体积的 99.5%，并包含 76.7% 的氢总质量。原子氢密度也分布在这一范围内，因此承载着莱曼-α 森林最重要的信息。对于暗物质和气体速度，我们在神经网络训练中不进行重新缩放，因为它们的量级相近。我们希望神经网络能够构建一个唯象模型，将修正量与这一平均关系联系起来。

*3.5. LyAl-Net*架构

本节将介绍本项目采用的神经网络架构及其优势，随后讨论训练过程，包括训练样本的规模、损失函数的选择，以及*LyAl-Net*架构的细节。本研究选择了 U-net 架构作为神经网络，该架构最初由 Ronneberger 等人于 2015年提出，用于生物医学图像分割。U-net 基于卷积神经网络（CNN；O'Shea & Nash 2015年），能够提取并创建特征映射，其工作方式类似于人类视觉皮层。它以模式识别能力、空间不变性和尺度不变性而著称。与传统人工神经网络（人工神经网络；O'Shea & Nash 2015年）相比，例如多层感知器，这种架构速度快得多。它已成为解决各种问题的宝贵技术，尤其适用于图像相关问题。CNN 与人工神经网络之间的显著区别在于卷积层。这些层使机器能够使用卷积核执行卷积运算，并利用三个关键特性：稀疏交互、参数共享和等变表示（Goodfellow 等人，2016年）。U-net架构借助网络跳跃连接，能够提取并汇总小尺度信息，同时保留大尺度信息。鉴于氢气是暗物质的示踪物，反之亦然，与典型的流体动力学模拟相比，U-net 是一种能够在短时间内生成气体场的实用方法；Wadekar等人（2020年）和 Bernardini 等人（2021年）也采用了这种方法。

LyAl-Net 的示意图如图 3 所示。它包括三个主要步骤：卷积侧（左侧）、瓶颈层（底部）和上采样侧（右侧）。我们为每个卷积步骤选择了 $3\times3\times3$ 卷积核，同时使用"SAME"填充来保持维度，并采用丢弃率为 0.2 的丢弃层。[1] 此外，最大池化也用于在每个收缩步骤中降低维度。[2] 在瓶颈阶段之后，我们使用了$3\times3\times3$ 上采样层，随后采用相同的卷积配置。在上采样阶段的每个卷积层中，我们使用了 16 个滤波器，将其应用于输入拼接和来自收缩侧的旁路连接。

输入的整体尺寸为 $81\times81\times81$ 像素，输出为 $27\times27\times27$ 像素，并进行了裁剪，以避免卷积可能造成的边缘效应。因此，我们需要将输出平铺到更大的立方体中，以根据输出获得完整的模拟，这将在第 3.5.3 节中进一步说明。

[1] 采用"SAME"填充的卷积核会在输入周围添加零值，使输出与输入具有相同的维度。[2] 最大池化通过提取$3\times3\times3$ 卷积核内的最大值，在张量中进行扫描。

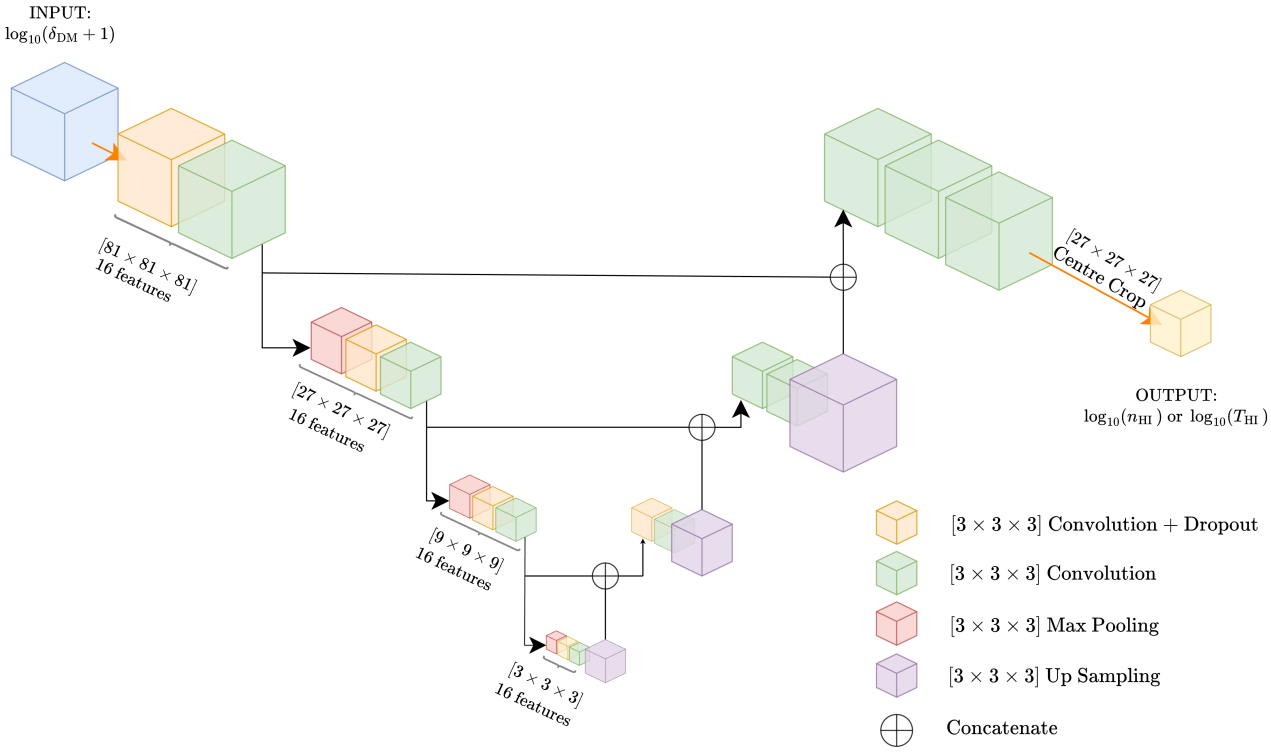

图 3. 这里展示了*LyAl-Net* 架构的示意图。网络的输入是经过变换的暗物质场 $\log_{10}(\delta_{DM}+1)$，大小为 $81\times81\times81$ 体素。输出是目标气体参数（温度或气体密度），大小为 $27\times27\times27$ 像素。我们从最后一个卷积层的中心裁剪出体积。箭头表示数据的方向，每种颜色代表一种使用 $3\times3\times3$ 核大小的不同操作。需要注意的是，所有卷积层都会提取 16 个特征图。为简化优化过程并便于解释，该模型分别预测每个气体参数。

我们注意到，这种架构并非最优；通过更少的训练参数，其他更好的架构也能取得相同或更好的结果。

3.5.1. 损失函数

我们选择均方误差（MSE）作为损失函数。使用该损失函数的主要优势在于，与 GAN 损失（Nguyen 等，2021）相比，它更加稳定。它还能够更好地理解神经网络所学习的内容，从而提高可解释性。其形式为

$$\mathrm{MSE}\,(Y_i,\hat{Y}_i)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2\,, \tag{30}$$

其中，n 为批量大小，Y_i 为预测值，而 $\hat{}Y_i$ 为真实值。

我们注意到，该损失应用于变换后的变量（A_X），并且仅计算中心部分 $27\times27\times27$ 像素的损失。对于速度场，这一关系保持线性。对于其他量，由于$T_{\rm HI}$ 和$n_{\rm HI}$采用了对数变换，有效损失函数为均方对数误差（MSLE），其表达式为

$$\mathrm{MSLE}(Y_i,\hat{Y}_i)=\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{Y_i}{\hat{Y}_i}\right)^2\,. \tag{31}$$

该损失函数对高值和低值一视同仁，有助于简化训练过程。均方对数误差损失对低估的惩罚大于对高估的惩罚。这是一种双向选择，二者都兼具风险与优势。事实上，它可能导致平均响应出现偏差，后文将对此进行探讨。这样做还可以避免对高值偏离过度敏感，因为这与 Lyman-α 森林无关。只要预测值处于合理的误差范围内，并且符合重子观测的分辨率，该损失函数就不应成为主要关注点。我们将在第4节稍后讨论预测偏差和误差。

3.5.2. 训练过程

我们训练*LyAl-Net* 来分别预测每个气体参数，因为这样可以调整训练超参数，并独立于其他场解释预测结果。我们使用Horizon-noAGN模拟作为训练集，其中包含暗物质和气体成分。其体积为$(100h^{-1}$兆秒差距$)^3$，，每条边的分辨率为1024像素。该快照的红移范围为$z=2.44$至$z=2.33$。由于我们受限于单个完整模拟，因此随机选取了5120个子盒，每个子盒的大小为81^3体素，体积为 $\sim(7.9\,h^{-1}$兆秒差距$)^3$，相当于总模拟体积的~ 9% 。这种数据采样方式还允许我们使用训练集中的重复样本和重叠区域，从而促使网络学习平移不变性。子盒的大小旨在适应GPU内存，并促使神经网络学习小尺度上的信息。我们注意到，物理尺寸更大的核大小可能通过更好地捕捉从大尺度到小尺度的功率传递来改进结果。我们将在第6节中进一步讨论这一点。

表2。 该表总结了使用 Horizon-noAGN 进行 *LyAl-Net* 训练时所采用的超参数。

目标氢中性参数	学习率	迭代轮数
密度	0.002	1000
温度	0.002	1000
速度	0.003	100

我们还在全部三个轴上进行旋转增强 [3]，但视线速度仅绕z轴旋转。这种旋转对于训练后的神经网络获得旋转不变性是必要的（He 等，2019年；Kaushal 等人，2022年；Ramanah 等人，2019年），反映了预测中涉及的物理过程。我们预计这一过程将使模拟器更加稳健。该特性将有助于针对不同暗物质模拟进行迁移学习。

表2 汇总了我们所用训练循环参数的详细信息，包括迭代轮数和学习率。我们使用批量大小为64，并采用 Adam 优化器（Kingma 和 Ba 2014）。值得注意的是，*LyAl-Net* 的超参数是通过数值实验确定的，以确保收敛过程可重复且迅速。

3.5.3. 平铺算法

如前所述，模拟的大小被离散化为 1024^3 个体素。由于*LyAl-Net* 对于 81^3 个体素的输入只能生成包含 27^3 个体素的模拟场，因此在评估预测模型的准确性和质量之前，我们必须平铺模拟的基本立方体，以达到所需的体积。我们通过应用一个 27 像素的滑动窗口实现了平铺算法：从 XY 平面开始沿水平方向扫描，每行、每列和每层（Z 方向）扫描 38 次。算法的伪代码见算法1。平铺后的总体积由 $38 \times 38 \times 38$个基本立方体组成（相当于 1026^3 个体素），随后裁剪得到 1024^3 个体素。在 Nvidia Tesla V100 和 Intel(R) Xeon(R) Gold 6230 （ 40 核）上，构建完整体积耗时 ~ 8 分钟。通过将部分操作直接重新安排在图形处理器上，可以减少 GPU 到 CPU 的数据传输，从而进一步降低这一耗时。我们将这一优化留待后续处理；它可能有助于构建完全可微的 Lyman-α森林模型。

4. Horizon-noAGN 的结果

在下文中，我们将在第4.1节介绍用于评估模拟场质量的指标。随后，我们将在以下各节依次讨论各个物理场：第4.2节中的原子中性氢数密度、第4.3节中的温度，以及第4.4节中的气体整体速度。需要强调的是，我们使用 Horizon-noAGN 训练 *LyAl-Net*，并将 Horizon-DM 用作验证集，以考察模型对不同重子反馈的响应。我们还交替使用“预测场”和“模拟场”这两个术语。

4.1. 预测质量评估

我们考虑多个指标，以评估我们对 *LyAl-Net* 预测物理场的信任程度。我们首先考察单点统计，即给定空间点处真实值与预测值之间的关系。我们将考察的第二个统计量是两点相关函数，它将有助于了解两个相应物理场的平均空间结构的再现情况。

4.1.1. 单点统计

我们采用若干指标来评估我们对 *LyAl-Net* 预测物理场的可信程度。首先考察单点统计，即空间中给定点处真实场与预测场之间的关系。接下来考察两点相关函数，以深入了解两个相应场的平均空间结构再现情况。

4.1.2. 单点统计

我们通过考察模拟中同一X 的真实值$X_{\rm true}$ 给定条件下，场 X 的预测值$X_{\rm Pred}$ 的条件概率分布函数，来量化模拟器的性能。根据贝叶斯定理，可以直接得到该条件概率，其表达式为

$$P(X_{\rm Pred}|X_{\rm True}) = \frac{P(X_{\rm Pred}, X_{\rm True})}{P(X_{\rm True})}, \tag{32}$$

其中，$P(X_{\rm Pred}|X_{\rm True})$ 是给定$X = X_{\rm Pred}$真实值时X 预测值的概率，$P(X_{\rm Pred},X_{\rm True})$ 是$X = X_{\rm Pred}$ 与$X = X_{\rm True}$的联合概率，而$P(X_{\rm True})$ 是$X_{\rm True}$的概率。

我们使用核密度估计（KDE，Silverman 1986；Scott 2015）估计了预测值与真实值的联合概率分布函数 $P(X_{\rm Pred},X_{\rm True})$ 以及$P(X_{\rm True})$。这是一种通过使用加权核函数平滑分布，从有限样本推断连续概率密度函数（PDF）的技术。由于该估计器的计算成本较高，我们在数量减少的体素上构建了模拟。因此，我们使用总体积的约 50% 来计算核密度估计器。完整细节见附录 B。

4.1.3. 两点统计量

两点相关函数 $\xi(r)$ 定义为

$$\xi(|\mathbf{r}|) = \langle \delta_A(\mathbf{r}')\delta_B(\mathbf{r}'+\mathbf{r})\rangle \,. \tag{33}$$

与均匀分布相比，该函数度量了相距 $|\mathbf{r}|$ 的两个对象或场涨落出现的超额概率。这些对象或场例如可以是星系或气体物理量。功率谱是相关函数在空间频率域中的图像。其定义为

$$P(|\mathbf{k}|) = \int \mathrm{d}^3\mathbf{r}\, \xi(r) e^{-i\mathbf{k}\cdot\mathbf{r}}. \tag{34}$$

我们用于评估模拟场的幅度和空间信息的指标是传递函数 $T_X(k)$ 和互相关函数 $r_X(k)$。

传递函数 $T_X(k)$（其中 k 为空间共动波数）是一种便捷的表示方式，用于描述预测场相对于真实场 X 在每个 k 尺度上的原始相关结构的偏离程度。该函数通常被文献用作评估模拟器或

[3] 为增加现有数据集的多样性

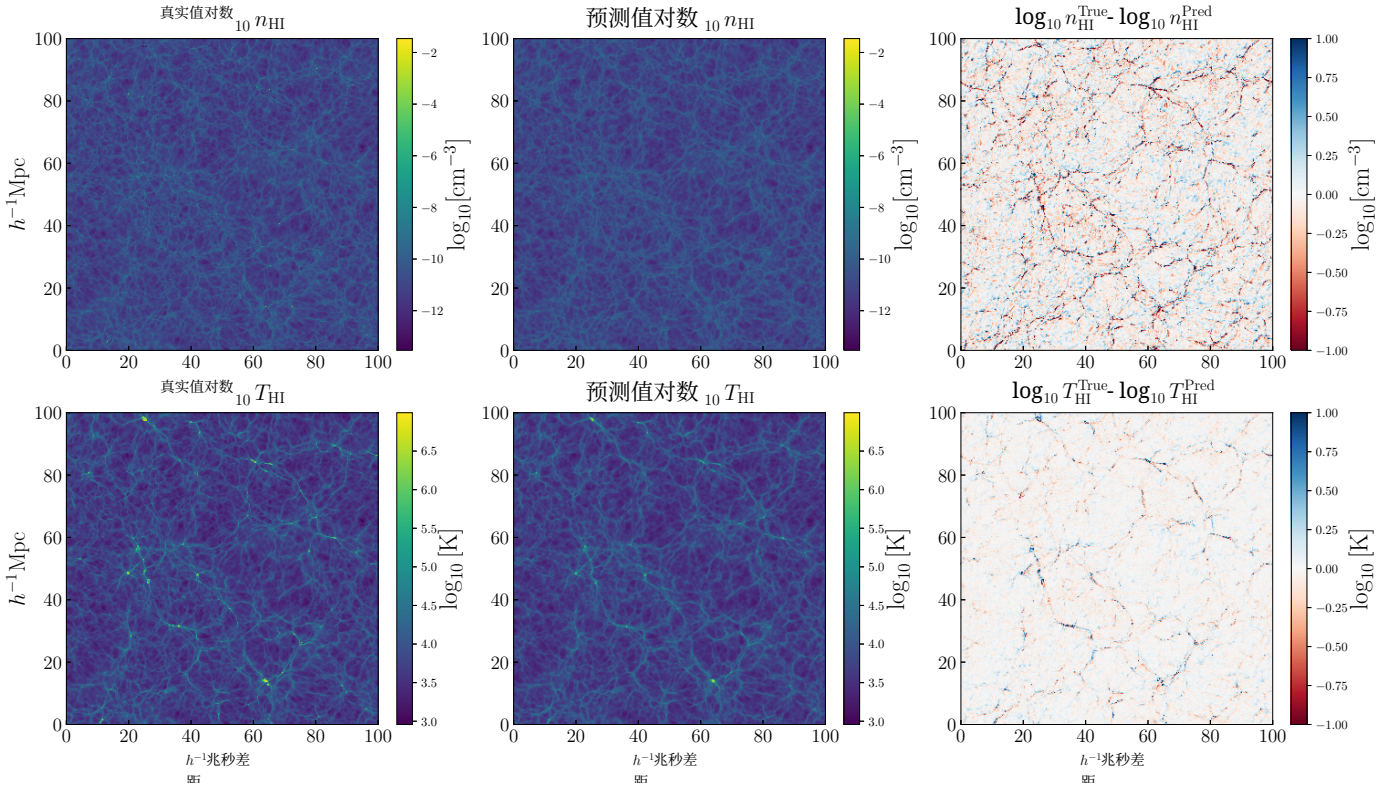

图 4。 不同流体动力学量样本切片在十进对数尺度下的比较。左列：直接从 Horizon-noAGN 中提取的场。中列：将 *LyAl-Net* 应用于暗物质场后获得的场。右列：最左侧面板与中间面板之间的差异。原子氢数密度（温度）分别显示在顶行（底行）。

近似方法性能的基准方法（例如 Bardeen 等，1986年；Leclercq 等，2013年；Vlah 等，2016年；He 等，2019年；Dai & Seljak，2021年）。传递函数定义为

$$T_X(k) = \sqrt{\frac{P_{\rm pred,X}(k)}{P_{\rm true,X}(k)}}\,, \tag{35}$$

其中，$P^{\rm pred}(k)$ 是模拟场的功率谱，$P_{\rm true}(k)$ 是参考场或真实场。

最后，我们考察相关率 $r_X(k)$，以检验线性对应关系

$$r_X(k) = \frac{P_{\rm true,X\times pred,X}(k)}{\sqrt{P_{\rm true,X}(k)P_{\rm pred,X}(k)}}\,. \tag{36}$$

为了计算本文所研究的不同标量场的功率谱，我们使用 nbodykit（Hand 等，2018）。在对 Horizon-noAGN 进行预测质量评估时，我们决定在功率谱计算中纳入数据集的训练部分，因为仅有 ≤ 9%[4] 的模拟体积用于训练*LyAl-Net*。这样做的主要原因是避免用于功率谱计算的体积不规则，同时该体积相较于其余验证体积应足够小。

4.2. 中性氢密度（n_{HI}）

在计算莱曼-α 森林中的吸收率时，需要的不同物理量中，原子氢数密度是最需要准确获取的，因为中性氢云是莱曼-α 森林过程中的吸收体。图4顶行展示了来自基于暗物质超密度的 Horizon-noAGN 预测的原子氢数密度 $n_{\rm HI}$ 切片，以及真实值与预测值之间的（对数）差异。直观来看，两幅切片中的结构吻合良好。我们没有发现由 *LyAl-Net* 造成的明显缺陷，也没有发现由用于覆盖整个体积的平铺过程引入的边缘效应；但在相对误差较高的区域存在一些斑点，这些区域通常与高密度区域相关，并且其中的莱曼-α 吸收已达到饱和。

现在，我们按照第 4.1 节中的模型评估流程，利用 Horizon-noAGN 的真实数密度和模拟数密度，计算真实数密度的联合概率分布和条件概率分布。估计得到的概率分布函数 $P(\log_{10} n^{\rm Pred}_{\rm HI} | \log_{10} n^{\rm True}_{\rm HI})$ 如图 5 所示，其中的等高线表示不同的概率水平。绿色虚线表示条件概率中的平均预测中性氢密度，记作 $\mu\left(\log_{10} n_{\rm HI}{}^{\rm Pred} | n^{\rm True}_{\rm HI}\right)$，绿色带表示 68% 置信区间。这一表示基于如下事实：条件概率非常接近高斯分布。红色虚线作为参考线，其给出为$n^{\rm Pred}_{\rm HI} = n^{\rm True}_{\rm HI}$。此外，我们加入了一个红色带，表示以中点为基准上下 10% 误差的边界，用作定性参考，以检测模拟器的平均偏差，我们将其称为预测偏差。等高线水平和均值表明，在数据集的主要区域内，预测表现良好，因为第 0 和第 99 百分位数分别为 4.5×10^{-9} 和 $3.8\times10^{-13}{\rm cm}^{-3}$。只有当$n_{\rm HI} < 10^{-12}{\rm cm}^{-3}$和$n_{\rm HI} > 3\times10^{-9}{\rm cm}^{-3}$ 时，才开始出现偏差，这可能是由较长的

[4] 训练集的数据采样算法允许重复采样。

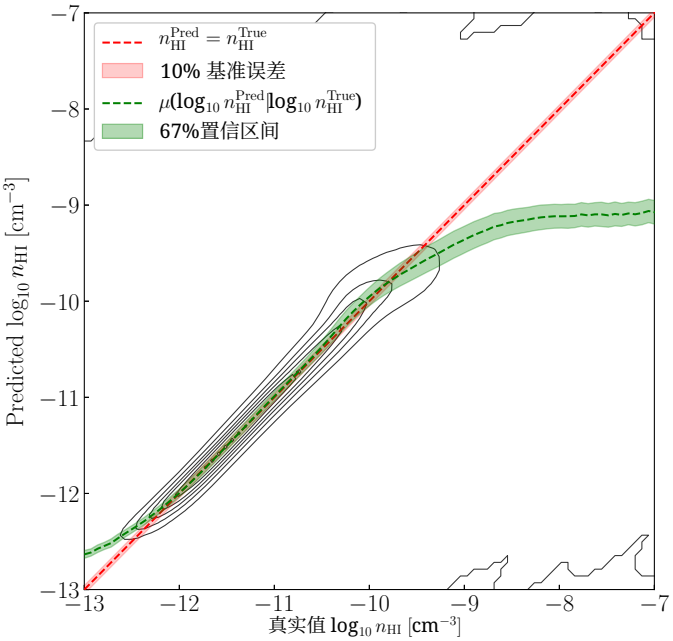

图 5. Horizon-noAGN 中真实密度对应的预测密度十进对数$n_{\rm HI}$的条件概率及其相应的边缘分布，其中黑色实线等高线表示不同的概率水平（见公式（32））。红色虚线对角线表示预测密度与真实密度之间的无偏关系，以及 10% 基准误差预算。绿色虚线表示条件概率分布的均值，并标示出 67%置信区间。我们注意到，在预测气体密度从 1×10^{-12} cm^{-3} 到 6×10^{-10} cm^{-3}的多个数量级范围内，两者吻合良好。对于极低密度$n_{\rm HI} \leq 10^{-12}$ cm^{-3}，会出现明显的饱和现象。

分布尾部和样本方差影响 KDE 本身所导致的。为了更清楚地了解预测偏差，我们将偏差定义为真实值的函数，表示为

$$\text{Bias} = \left[10^{\mu(\log_{10} X_{\rm HI}^{\rm Pred}|\log_{10} X_{\rm HI}^{\rm True})} - X_{\rm HI}^{\rm True}\right] \frac{1}{X_{\rm HI}^{\rm True}}, \quad (37)$$

其中$X_{\rm HI}$ 是一个气体参数。我们在图 6 中展示了模拟得到的$n_{\rm HI}$ 函数，该图显示了当 10^{-12}cm^{-3} $< n_{\rm HI} < 10^{-10}$ cm^{-3} 时，预测偏差保持在 10%基准偏差以内。将*LyAl-Net* 与基准方法进行比较可见，modified-FGPA 的预测仅在密度$n_{\rm HI} \approx 10^{-11}$cm^{-3}附近无偏，且无法捕捉更广泛的密度范围。

虽然单点统计表明*LyAl-Net* 表现良好，但我们将分析扩展到模拟场的两点统计量。在计算功率谱之前，我们将模拟值和真实$n_{\rm HI}$ 中的极端值掩蔽为零。这样做是因为数据点较少，分布尾部开始出现偏差。此外，必须进行掩蔽，因为功率谱对极端值非常敏感（见图 C.1）。因此，我们实施了三种不同的掩膜范围，并将其称为保真度范围：低、中和高。表 3 汇总了这一过程中使用的阈值。这些掩蔽值是任意设定的，因此我们可以确定*LyAl-Net* 表现最佳的区域。我们强调，被掩蔽体素的最大数量不到 1%。因此，这不应人为提高相关率$r(k)$。

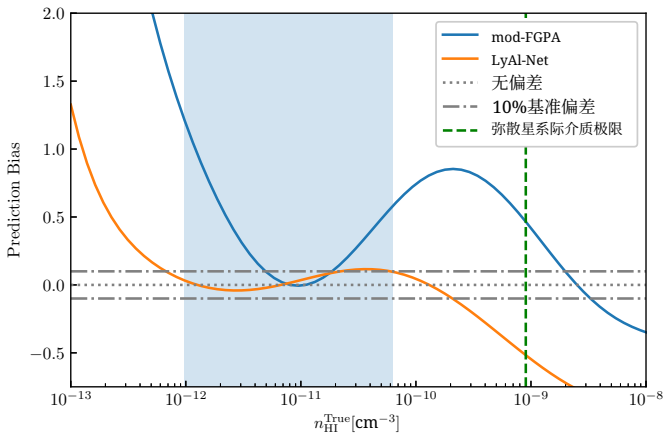

图 6. Horizon-noAGN 在$z = 2.44$ 时，*LyAl-Net*（橙色实线）和 modified-FGPA（蓝色实线）模拟得到的$n_{\rm HI}$ 相对于真实$n_{\rm HI}$ 的偏差。modified-FGPA 仅在$n_{\rm HI} \approx 10^{-11}$ cm^{-3}处无偏。总体而言，*LyAl-Net*的偏差表现出更好的预测效果。其偏差大多保持稳定，在第 1 和第 90 百分位数之间的密度范围内处于 10%基准偏差水平以内；该范围以浅蓝色条带标出，分别对应 1×10^{-12} 和 6×10^{-11} cm^{-3} 。偏差在蓝色条带外出现剧烈变化是意料之中的，因为密度分布两端的极值很少出现。

表 3. 掩蔽过程中使用的不同密度范围，用于评估*LyAl-Net* 的性能。低、中、高保真范围的上限密度边界分别基于第 99.95、99.9 和 99.5 百分位数。中性氢密度的下限为 4.36×10^{-15}cm^{-3}，温度的下限为 600 K。被掩蔽体素的最大数量不到 0.5%，因此，这不应人为提高互相关$r(k)$。

保真度范围	掩蔽值上限	
	密度 [cm^{-3}]	温度 [K]
LOW	5.16×10^{-7}	3.08×10^{6}
中等	1.79×10^{-8}	5.92×10^{5}
HIGH	1.10×10^{-9}	8.03×10^{4}

我们比较了图7所示不同保真度范围下的传递函数和互相关。对于所有保真度范围，传递函数都大致保持稳定，而高保真度的$T_{\rm HI}(k)$表现最佳，在$k \approx 1 h$ 兆秒差距$^{-1}$以内保持在0.9左右。互相关函数$r_{\rm nHI}$ 显示，所有保真度范围均趋于收敛，且振幅在$k \approx 3h$ 兆秒差距$^{-1}$以内保持在0.9以上。基于两点统计量，这项评估表明，*LyAl-Net*的最佳性能可达$n_{\rm HI} \leq 1.10 \times 10^{-9}$ 厘米^3, ，这与估计的弥散星系际介质极限大致相当。 [5]

*4.3. HI*气体温度（T_{HI}）

我们对中性氢气体温度进行了与密度相同的分析。在图4的底部面板中，我们展示了Horizon-noAGN模拟中温度场的一个切片、根据暗物质密度模拟得到的温度，以及这两个场之间的对数差异。模拟值和真实值的温

[5] 我们通过缩放 Martizzi etal.（2019）的 $n_{\rm H}$ 极限，估算了中性氢密度的弥散星系际介质极限；$n_{\rm HI} = \alpha n_{\rm H} = 10^{-4}(1+z)$ cm^{-3}，其中 $\alpha = \text{mean}(n_{\rm HI}/n_{\rm H})$ 是直接从 Horizon-noAGN 模拟中获得的，模拟红移为 $z = 2.43$。

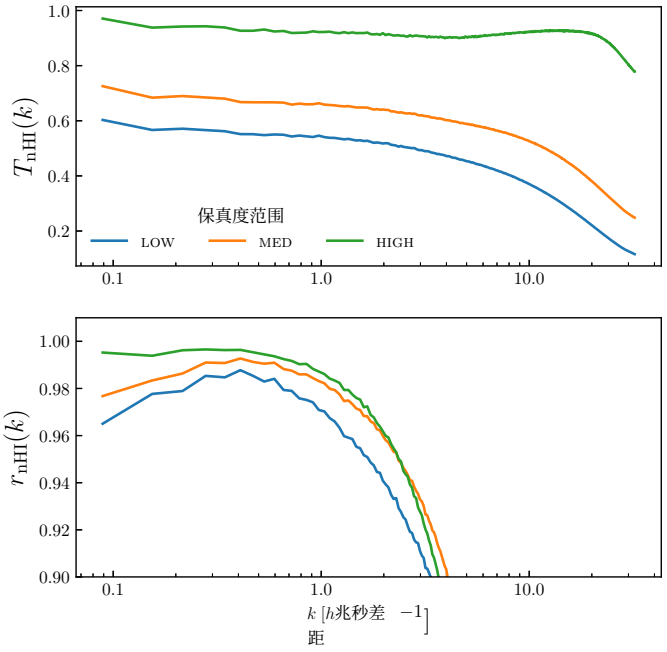

图 7. 采用两点相关性评估 Horizon-noAGN 在不同保真度范围下对n_{HI}的预测质量。**上图：**展示不同低（蓝色实线）、中（橙色实线）和高（绿色实线）档位的传递函数$T_{n_{\mathrm{HI}}}(k)$。具体说明见表 3。**下图：**展示与上图相同保真度范围对应的相关率$r_{n_{\mathrm{HI}}}(k)$，颜色标示方式也相同。

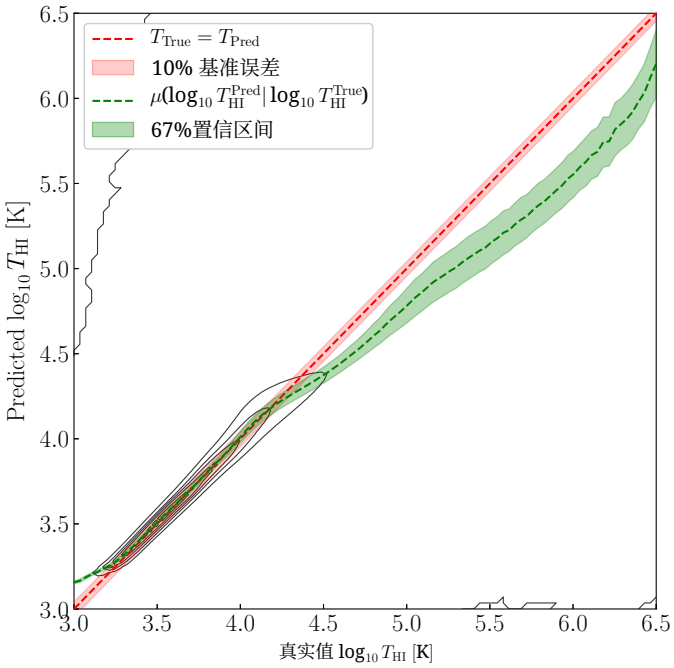

图 8. Horizon-noAGN 中预测温度十进对数相对于真实温度的条件概率，等高线表示不同的概率。红色对角虚线表示这两种温度之间的无偏关系，旁边红色阴影区域表示基准 10% 误差范围。绿色虚线表示条件分布的均值，旁边绿色阴影区域表示 67%置信区间。

度也一致，未发现由*LyAl-Net*造成的明显缺陷。温度的对数差异图显示，总体而言，该图优于密度图，在高密度区域的相对误差更低。

与密度部分一样，我们也计算了模拟温度在给定真实温度条件下的联合概率分布和条件概率分布$P(\log_{10} T^{\mathrm{Pred}}{}_{\text{中性氢}} | \log_{10} T^{\mathrm{True}}{}_{\text{中性氢}})$，如图 8 所示，并沿用$n_{\mathrm{HI}}$中的相同配色方案。等高线表示不同的概率水平，绿色虚线表示预测温度的均值，绿色带状区域表示从其最大值处取得的 68% 置信区间。红色虚线作为参照，表示$T^{\mathrm{Pred}}{}_{\text{中性氢}} = T^{\mathrm{True}}{}_{\text{中性氢}}$；红色带状区域则表示以中点为基准的 10% 基准误差，可作为定性参照。从图中可以看出，模拟温度的均值在整个温度范围内似乎存在较小偏差，这种偏差在该温度区间会更加明显。模拟温度偏差随真实值变化的具体情况以橙色实线绘于图 9。因此，我们认为模拟器在 2×10^3 K $< T_{\mathrm{HI}} < 10^4$ K 范围内近乎无偏且稳定。只有在温度极高、超过 10^4 K时，才会出现过预测。此外，*LyAl-Net* 的表现明显优于改进版 FGPA；后者在所有温度值范围内都存在预测偏差。

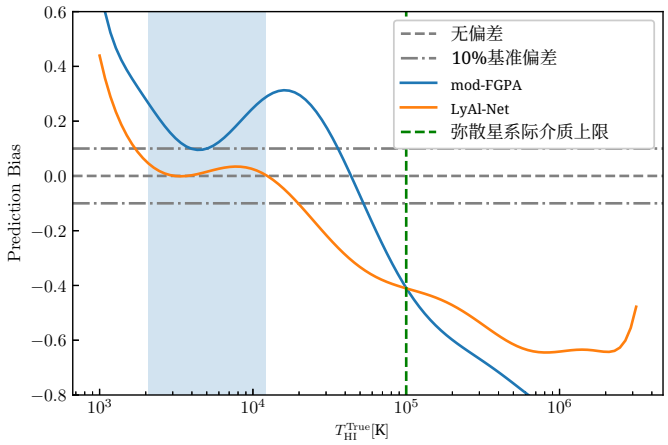

图 9. 在$z = 2.44$处，Horizon-noAGN 的模拟T_{HI}偏差随真实T_{HI}的变化情况：LyAl-Net （橙色实线）和 modified-FGPA（蓝色实线）。modified-FGPA 的偏差在整个温度范围内都超出基准偏差的 10% 水平。而*LyAl-Net* 的偏差总体较为稳定，在以浅蓝色条带标示的第 1 和第 90 百分位数之间的温度范围内，始终处于基准偏差的 10% 水平以内；该范围对应 2×10^3 K 至 2×10^4 K。

为估算*LyAl-Net*的保真度范围，我们还选取了三个掩膜范围，汇总于表 3。由图 10 可见，对于高保真范围，传递函数表明振幅在$k \approx 10\ h$ 兆秒差距$^{-1}$ 以内始终远大于 0.90，但在$k < 0.3h$ 兆秒差距$^{-1}$处略有下降。互相关函数测得的空间信息表明，对于低保真度和中等保真度范围，$r > 0.9$ 可保持至$k = 2h$ 兆秒差距$^{-1}$；对于高保真范围，则可保持至$k = 3h$ 兆秒差距$^{-1}$。这意味着*LyAl-Net*的性能最高

位于$14h^{-1}$兆秒差距的尺度范围内，至

温度范围为 $1.92 \times 10^3 < T$

$_{\mathrm{HI}} <$

$8\,03 \times 10^4$

. K 时，尺度为 $63h^{-1}$兆秒差距。值得指出的是，当n_{HI}和T_{HI}的取值处于第 1 至第 90 百分位数之间时，平均预测偏差保持稳定。这一范围的实证结果表明，*LyAl-Net* 在两个场的主要分布中表现最佳，而较大的偏差只出现在分布尾部罕见且极端的环境中。因此，这表明神经网络在分布尾部对温度和密度的加权存在内在局限，这种局限源于

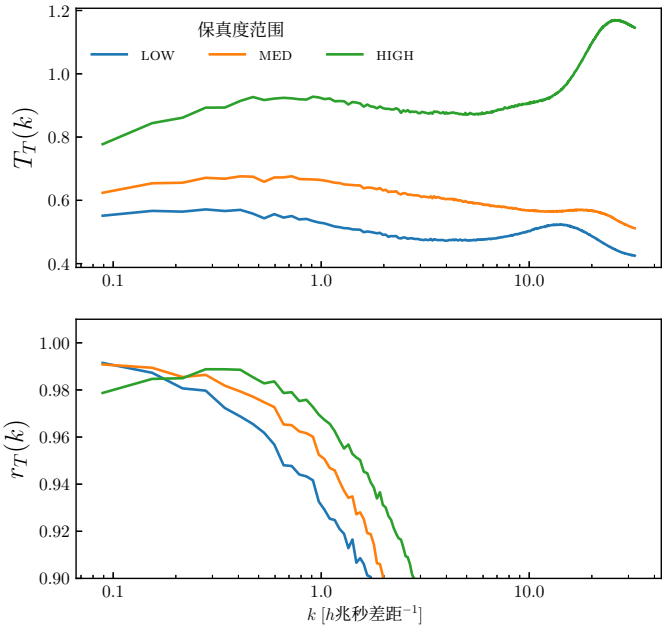

图 10。 使用两点相关性评估 Horizon-noAGN 对模拟的$T_{\rm HI}$在不同保真度范围选择下的预测质量。**上图：** 低、中、高保真度范围的传递函数对比分别以蓝色、橙色和绿色表示。温度范围见表 3。**下图：** 采用相同配色方案对比相同保真度范围的相关率。

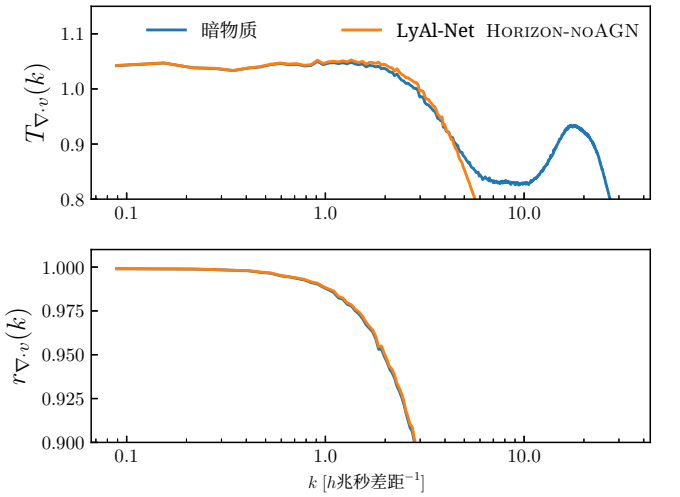

图 11。 使用两点相关性评估 Horizon-noAGN 的预测质量：比较 Horizon-noAGN 暗物质速度的速度散度与模拟的 Horizon-noAGN 中性氢速度的速度散度，分别以蓝色的**暗物质**和橙色的***LyAl-Net*** 标示。**上图：** 传递函数对比显示二者的振幅相近。然而，当$k \gtrsim 6h$ 兆秒差距$^{-1}$时，Horizon-noAGN 暗物质速度散度的表现更好。**下图：** 互相关函数对比显示二者完全相同，这意味着*LyAl-Net* 并未改善空间分布。

数据点不足。幸运的是，在实际应用中，Lyman-α 吸收对高密度区域并不敏感，因为饱和发生在密度偏差出现位置以下（见公式 12）。

4.4. 中性氢气体的视线速度（v_z，HI）

气体的视线速度 $v_{z,\rm HI}$ 在莱曼-α 吸收中起着重要作用，因为多普勒效应会使入射光子的频率发生偏移（在非相对论情况下）。速度场的值不准确时，模拟通量会导致吸收率改变，并使莱曼-α 吸收在某些频率上发生偏移。为估算 $v_{z,\rm HI}$，我们可以直接使用暗物质的视线速度 $v_{z,\rm DM}$，因为二者的速度场几乎相同，这是由于暗物质的引力会影响星系际介质。这种估算在大尺度上有效，但由于各种激波和反馈，星系际介质的速度不可避免地会在小尺度上发生解耦（Pando 等人 2004年）。因此，直接使用 $v_{z,\rm DM}$ 可能会影响小尺度吸收的准确性，这意味着模拟 $V_{\rm HI}$ 仍有改进空间。为通过实证展示解耦效应，我们利用传递函数和互相关，对比 $V_{\rm DM}$ 与 $V_{\rm HI}$ 的速度散度，如图 11 中的蓝色实线所示。暗物质的传递函数 $T_{\nabla\cdot v}(k)$ 显示出稳定的振幅 ≈ 1.1，其中 $k < 2h$ 兆秒差距$^{-1}$，这意味着暗物质速度涨落高于气体。底部面板中的互相关率 $r_{\nabla\cdot v}(k)$ 显示，当 $k > 2h$ 兆秒差距$^{-1}$ 时，其值降至 0.95 以下，这意味着在低于 3 h^{-1}兆秒差距的尺度上，星系际介质的速度开始发生解耦。基于 LyAl-Net 应能够校正小尺度效应这一设想，我们训练 LyAl-Net 使用暗物质视线速度作为输入，预测中性氢的视线速度。我们检验 LyAl-Net *LyAl-Net* 是否能改善结果，而不是仅仅给暗物质速度场增加更多系统效应，并在图 11 中以橙色实线展示。两种方法在速度散度互相关的空间分布方面表现相似。然而，暗物质的传递函数在高k 区域显示出更好的可靠性，而*LyAl-Net*在相同范围内导致振幅损失。尽管如此，我们仍使用$v_{z,\rm DM}$ 作为$v_{z,\rm HI}$ 的替代指标，并将速度预测的改进留待未来工作。

4.5. 缺少重子反馈的影响

由于方程更加复杂且时间步数量大幅增加，宇宙学流体动力学模拟的计算成本通常至少比N体模拟高一个数量级（Peirani等人（2017））。虽然生成纯暗物质场的成本更低，但忽略重子反馈会降低较小尺度上模拟气体场的准确性。由于主要目标是将*LyAl-Net* 应用于纯暗物质N体模拟以提高时间效率，因此在本节中，我们探究纯暗物质N体模拟会如何影响模拟场的质量。随后，我们将在第4.6节考察这种近似对莱曼-α 森林吸收的影响。对于这项测试，我们利用Horizon-DM模拟模拟中性氢密度和温度。该模拟是Horizon-AGN的姊妹模拟，仅考虑纯暗物质动力学。我们使用与*LyAl-Net*训练所用相同的快照，即相同的红移范围。由于神经网络模型是在Horizon-noAGN上训练的，并采用了公式（25）所述的暗物质密度变换，因此 *LyAl-Net* 能够不受纯N体模拟与完整流体动力学模拟之间暗物质粒子质量差异的影响。我们可以将这些模拟场作为直接来自Horizon-DM的模拟场这一情形的基准方法。

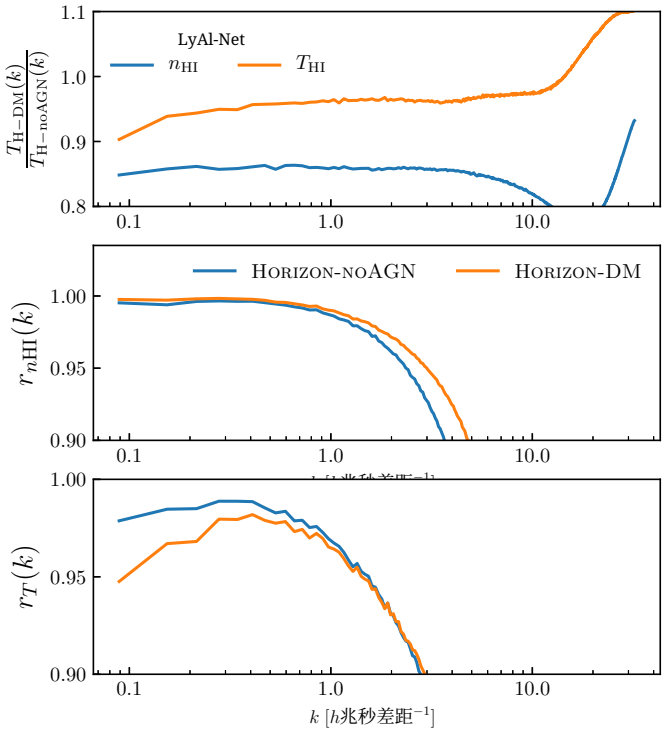

图12。 重子物理对暗物质场的改变对仿真质量的影响。**顶部面板**展示了使用 Horizon-noAGN 和 Horizon-DM 的模拟场传递函数之比，表明温度不受缺失重子反馈的影响。然而，中性氢的密度损失了功率，这意味着密度更加平滑。**中部面板**和**底部面板**展示了来自 Horizon-noAGN（此前在图7和图10中展示）的模拟密度和温度的高保真范围互相关函数的比较，以及与 Horizon-DM 的比较。

下面介绍一个新的指标，即传递函数之比，

$$\frac{T_{x,\rm H-DM}(k)}{T_{x,\rm H-noAGN}(k)} = \sqrt{\frac{P_{x,\rm H-DM}(k)}{P_{x,\rm H-noAGN}(k)}}, \tag{38}$$

其中，带有下标x、H – noAGN 和x、H – DM 的 a 分别表示使用 Horizon-noAGN 和 Horizon-DM 模拟的气体场 x 。我们还对两个场使用了表 3 中引入的高保真掩蔽范围。

图 12 上图展示了模拟密度和温度的传递函数之比。在所考察的中性氢密度尺度范围内，该比值相对稳定。然而，这一稳定值维持在 ≈ 0.85，意味着缺少重子反馈会影响$n_{\rm HI}$的涨落。$n_{\rm HI}$ 的互相关函数（中图）显示，在$k \approx 5h$ 兆秒差距$^{-1}$处，其精度略有提高。这略优于 Horizon-noAGN，也与预期相反：原本预期 Horizon-DM 的模拟场会因缺少反馈过程而表现更差。另一方面，Horizon-DM 模拟得到的温度与使用 Horizon-noAGN 时的结果更为接近，因为比值相对稳定在 0.95。其互相关函数表明，使用 Horizon-DM 会轻微影响模拟场的大尺度空间分布。值得指出的是，当$k > 10h$ 兆秒差距$^{-1}$时，两种场的传递函数比都开始偏离，这意味着缺少重子反馈主要影响大于 $10h$ 兆秒差距$^{-1}$的尺度，即 $0.62h^{-1}$兆秒差距。使用重子效应设置不同的暗物质模拟会降低模拟气体场的精度，这是可以预期的，尤其是在缺少激波和反馈过程时。我们将在第 5.4 节中进一步讨论这一点，以及使用 IllustrisTNG 可能采取的处理方法和改进方案；在 IllustrisTNG 中，模拟参数的差异要复杂得多。

4.6. 莱曼α森林吸收（F）

本节将讨论使用 Horizon-noAGN 的模拟 HI 密度和温度进行莱曼α森林预测的质量，尤其是使用 Horizon-DM 时，在缺少重子反馈的情况下会发生什么。

我们按照方程（23）计算归一化的莱曼-α吸收。图13展示了模拟器根据一个样本视线得出的莱曼-α 吸收，以及根据 Horizon-noAGN 的气体参数计算出的通量（作为真实值）。底部面板显示残差，定义如下：

$$\text{Residual} \equiv \exp(-\tau_{\rm True}) - \exp(-\tau_{\rm Pred}) = F_{\rm True} - F_{\rm Pred}. \tag{39}$$

我们还将 FGPA 估算的通量残差作为基准模型纳入比较[6]。从视觉上看，模拟光谱与整体残差相符，且残差保持在 ±0.1基准误差范围内。在强吸收区域存在一些显著误差，*LyAl-Net* 在极高密度环境中的表现有限。然而，总体而言，其预测性能优于 FGPA 模型。尽管如此，单条视线的残差不足以评估模拟场的质量，原因有几点。吸收过程需要三个不同的参数，它们会影响模拟通量的不同表现：

1. 密度$n_{\rm HI}$ 直接影响吸收量；2. 温度$T_{\rm HI}$ 会因热展宽而影响吸收宽度；3. 视线速度会影响吸收频率，从而在光谱中表现为偏移

因此，为了抑制小尺度涨落并检测仿真中的系统误差，我们沿 XY 平面对透射通量进行集合平均，得到随红变变化的平均透射通量（$\langle F(z)\rangle$），并在图14中展示。图的上面板显示，LyAl-Net 通量与真实通量（称为“真实值”）的涨落在所有红移范围内均一致。唯一主要差异在于模拟场的吸收量。下面板显示，LyAl-Net 使用 Horizon-noAGN 和 Horizon-DM 时的相对误差分别约为 2.5% 和 3.5%。这是直接受到 LyAl-Net 在高物质密度区域低估柱密度的影响。相对误差还显示，在红移$z \approx 235$、2.40 和 2.43 附近存在明显涨落。随机成分可以解释这种涨落，这可能源于 LyAl-Net 忽略的各种反馈过程。要减轻这些影响，需要设计一种新的损失函数，以降低训练期间较大偏差的影响。modified-FGPA 的相对误差约为 1%。然而，Lyman-α 光学深度 τ 的一维分布显示，LyAl-Net 优于基准模型，尤其是在 modified-FGPA 无法预测的低光学深度区域。

对于图15所示的两点汇总统计量，使用 Horizon-noAGN（蓝色）的传递函数显示，

[6] 我们纳入了第2.1节所述的沃伊特剖面

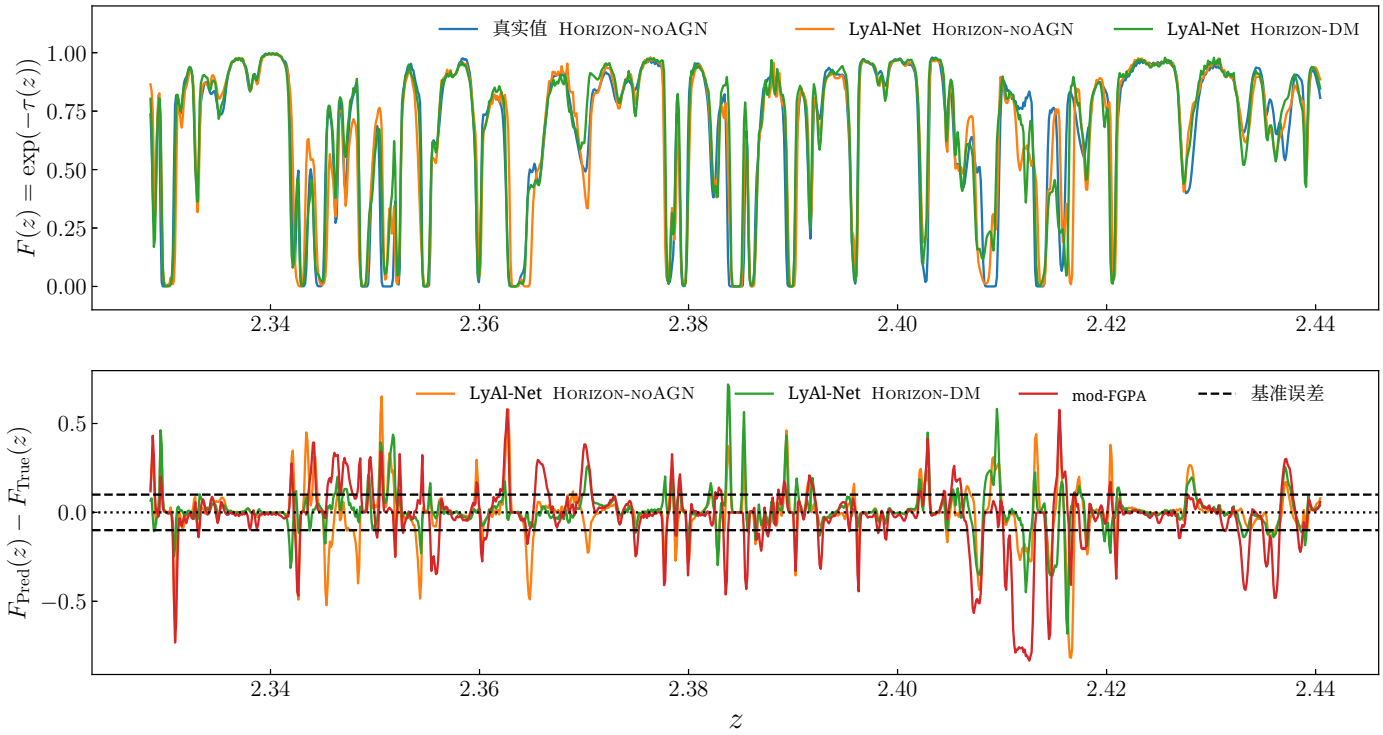

图 13. 上图：来自同一样本视线的真实值和模拟通量的莱曼-α 森林归一化通量比较，分别标记为真实值和 *LyAl-Net*。通量根据 Horizon-noAGN 计算得出，而模拟的中性氢参数则由 *LyAl-Net* 使用 Horizon-noAGN 和 Horizon-DM 暗物质生成。**下图：**模拟通量与真实通量的残差，以及在 −0.1 和 0.1 处以虚线绘制的基准误差。*LyAl-Net* 和 modified-FGPA 均表明，这些数值大多保持在基准区域内；在强吸收区域内，残差和峰值以*LyAl-Net* 表现得更加稳定。

振幅一直保持在 0.90 以上，直到$k \approx 10h$ 兆秒差距$^{-1}$。我们注意到，$T_F(k)$ 在大尺度（$k \approx 0.1h$ 兆秒差距$^{-1}$）上先下降、随后缓慢上升，这意味着预测通量比真实值更加平滑。关闭重子反馈后，Horizon-DM 模拟得到的通量与 Horizon-noAGN 的表现相同，但振幅更小。这一振幅下降直接体现了第4.5节讨论的$n_{\rm HI}$ 低估。同时，modified-FGPA 模型在更大尺度上表现出更高的振幅，这也反映在图14中，其系综平均更接近 Horizon-noAGN。然而，当$k > 3$时，传递函数出现了明显偏离，这意味着*LyAl-Net* 在较小尺度上成功超越了基准模型。底部面板中的互相关函数显示，在使用 Horizon-noAGN 时，模拟通量与真实值在$k \approx 6h$ 兆秒差距$^{-1}$ 以内具有显著的一致性；对于 Horizon-DM，这一范围为$k \approx 9$ h 兆秒差距$^{-1}$。modified-FGPA 也表现出与*LyAl-Net*相近的一致性水平。然而，使用 Horizon-DM 时，空间分布更好。对此我们可以考虑两个可能的原因：Horizon-DM 的互相关略好（见图12），或者其速度场恰好更接近气体的速度场。我们将在未来的工作中对此进行研究。

现在，我们考察模型的不准确性是否会对观测数据的建模造成问题。作为一个实际例子，Bautista 等（2015）对模拟类星体光谱进行了建模，具体为

$$f(\lambda) = \left\{[F(\lambda) \cdot C(\lambda)] * \tilde{W}\left(\lambda, R_p, R_w\right) + N(\lambda)\right\} \cdot M(\lambda) + \delta f_{\rm sky}(\lambda). \quad (40)$$

其中，$f(\lambda)$为类星体通量，$F(\lambda)$为莱曼-α 吸收，$C(\lambda)$为类星体连续谱，$\tilde{W}\left(\lambda, R_p, R_w\right)$为BOSS像素化内核，$N(\lambda)$为Bautista 等（2015）提供的噪声，$M(\lambda)$为校正线性函数，而 $\delta f_{\rm sky}(\lambda)$为添加的天空扣除残差。

我们将全流体动力学模拟得到的莱曼-α 吸收记为 $F(z)$。*LyAl-Net* 的输出可表示为

$$F(z)_{LyAl\text{-}Net} = F(z)B(z) \quad (41)$$

其中，$B(z)$是*LyAl-Net* 对N体模拟输入作出的任意内禀偏差。因此，必须将预测模型产生的系统偏差纳入考虑。因此，方程（40）中的$F(\lambda)$现在多了$B(\lambda)$这一额外项，需要对其进行处理。

如果我们假设最低要求是类星体光谱的信噪比为 SNR = 10，并进一步假设偏差与 2.5% 的偏差呈线性关系，那么信噪比要求将改为 10.3。

5. 使用 IllustrisTNG 进行迁移学习

在前文中，我们测试了将 *LyAl-Net* 直接应用于模拟的性能。然而，真实宇宙不太可能遵循其中任何一种情况。因此，检查该模型的可移植性至关重要。为确保这一点，我们需要一个能够推广到其他物理反馈以及气体物理其他更新的框架。在本节中，我们探讨将在 Horizon-noAGN 上训练的*LyAl-Net*移植到不同气体物理过程中的效率。我们使用 IllustrisTNG 作为基准方法和验证集，以检验预测密度场和温度场的准确性。我们首先通过直接应用 *LyAl-Net*（开箱即用）进行测试，随后我们

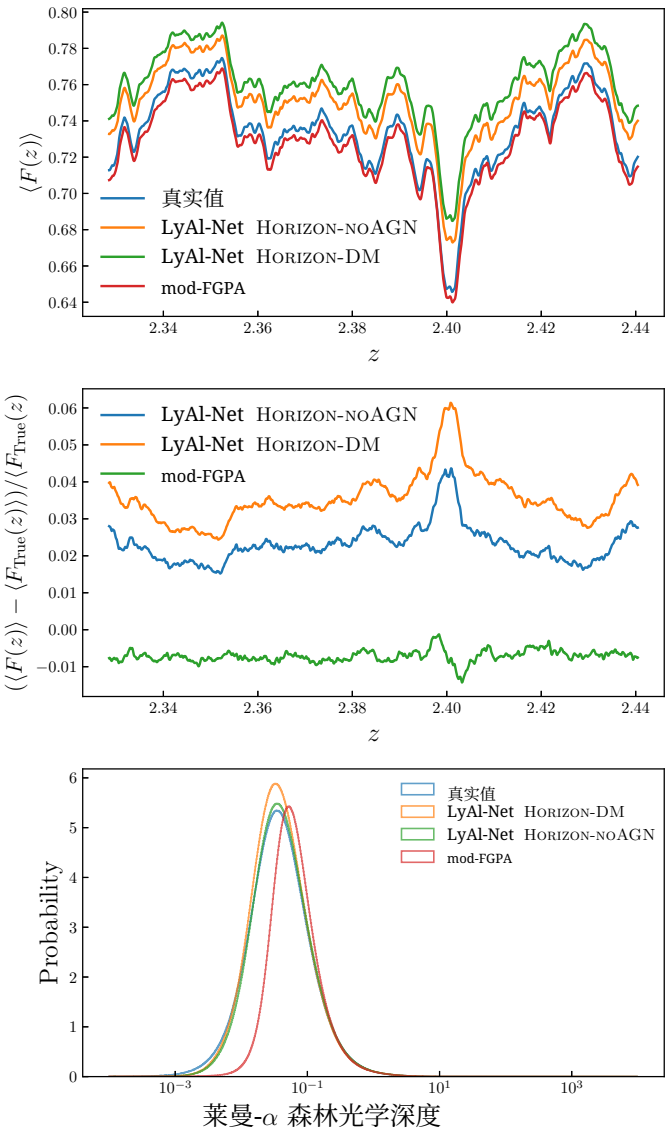

图 14。Horizon-noAGN（真实值）与使用*LyAl-Net* 及基准模型（mod-FGPA）模拟的莱曼-α 吸收之间的单点统计比较。**上图**：平均透射通量随红移的变化，$\langle F(z)\rangle$，显示*LyAl- Net* 与真实值吻合良好。**中图**：比较模拟通量的平均透射通量相对误差随红移的变化。LyAl-Net 的相对误差形状相同，而 Horizon-noAGN 的均值更接近真实值。基准模型的平均透射通量更接近真实值。**下图**：光学深度一维分布的比较显示，使用 Horizon-noAGN 和 Horizon-DM 的*LyAl-Net* 通量表现优于基准模型。

将讨论使用受迁移学习启发的技术改进预测结果的可能性。

5.1. TNG100 规格

IllustrisTNG是一组采用不同设置和分辨率的模拟。我们选择TNG100，因为它的体积与Horizon-noAGN模拟最为接近。TNG100是一项流体力学宇宙学模拟，其体积为边长75h^{-1}兆秒差距的立方体。该模拟采用AREPO，在移动的非结构化网格上求解动力学方程，该网格由一组离散点的Voronoi镶嵌定义（Springel 2010）。

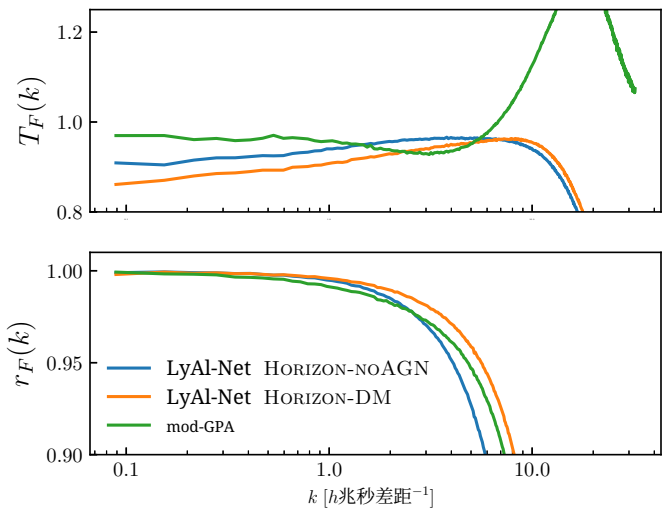

图 15. 使用 *LyAl-Net* 对 Horizon-noAGN、Horizon-DM 和 modified-FGPA 上的模拟结果进行比较：传递函数和互相关函数（上图），以及互相关函数（下图）。当 $k > 3$ 时，modified-FGPA 的传递函数明显偏离，这表明 *LyAl-Net* 在较小尺度上表现得更好，也更稳定。

我们选取了快照28，其对应的红移为$z = 2.58$。这是与使用Horizon-noAGN$z = 2.5$训练的*LyAl-Net* 最接近的快照。

5.2. 预计算

如上所述，TNG100 的数据结构与 Horizon-noAGN 略有不同，因为它基于 Voronoi镶嵌。为了生成用于测试的流体动力学量数据集，我们采用相同的方法和自适应滤波器，将暗物质和气体场高效地映射到网格结构中，其分辨率与 Horizon-noAGN 训练集相同（~ 9.78h^{-1} 兆秒差距），从而得到 768×768×768 个体素。

我们利用关系式 [7]，根据 TNG100 模拟的内能u推导出气体的温度，

$$T = \mu(\gamma - 1)\frac{u}{k_B}\,, \tag{42}$$

其中，T 是每个单元中气体的温度，$\gamma = 5/3$ 是理想单原子气体的绝热指数，k_B 是玻尔兹曼常数，μ 是平均分子量。平均分子量根据电子丰度推导得出；电子丰度是自由电子数密度与总氢数密度之比（Kim 等（2022））：该关系表示为，

$$\mu = \frac{4m_p}{1 + 3X_H + 4X_H x_e}\,, \tag{43}$$

其中，m_p 是质子质量，$x_e = n_e/n_H$ 是电子丰度，即自由电子数相对于总氢数密度的比例；X_H 是氢质量分数，我们假设其为恒定且均匀分布。

IllustrisTNG 模拟数据库并未公开提供中性氢密度。因此，我们选择

[7] 我们遵循了 IllustrisTNG 项目官方网站提供的指南（https://www.tng-project.org/data/docs/faq/#gen5）。

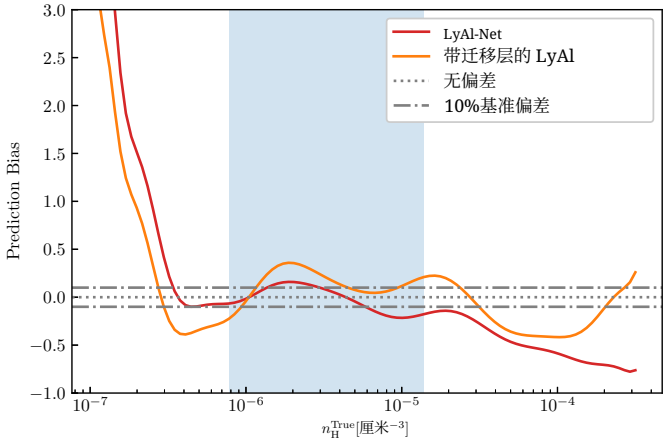

图 16。 在 TNG100 快照 28（第 5.4 节）上应用迁移学习前后预测偏差的比较。浅蓝色带表示处于第 1 和第 90 百分位数之间的密度。与开箱即用的 *LyAl-Net* 相比，在超过第 0 百分位数后，高密度区域的整体表现有所改善，同时在整个密度范围内保持了相近的偏差。

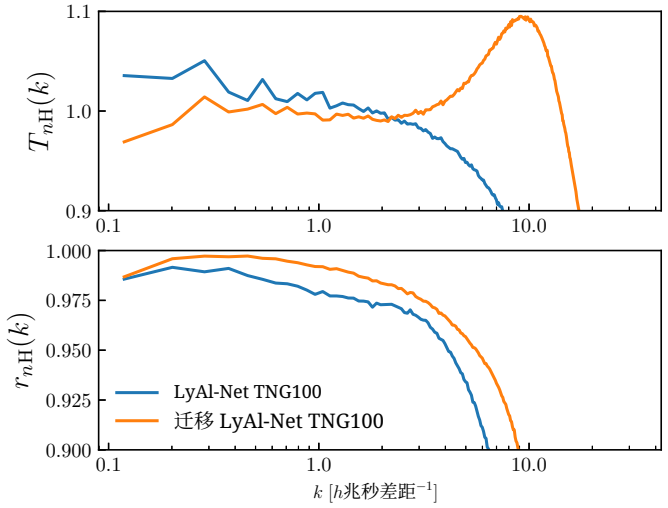

图 17。上图：传递函数；下图：使用 *LyAl-Net*模拟的中性氢温度的交叉相关函数。蓝线表示*LyAl-Net* 开箱即用时的性能，橙线表示*LyAl-Net* 使用迁移层时的性能。我们还使用了（高保真度）第 99.5 百分位数范围掩蔽，对应于 1
.13 × 10^{-7} 至 1.26 × 10^{-4} cm^{-3}
.

根据可获取的总氢密度对其进行估算。为此，我们根据 Horizon-noAGN 模拟中的平均相对丰度，按比例估算中性氢密度n_{HImock}。这可以表示为

$$n_{\mathrm{HI}_{\mathrm{mock}}}^{\mathrm{TNG100}} = \left\langle \frac{n_{\mathrm{HI}}^{\mathrm{HnoAGN}}}{n_{\mathrm{H}}^{\mathrm{HnoAGN}}} \right\rangle n_{\mathrm{H}}^{\mathrm{TNG100}} . \tag{44}$$

上述方程是对氢电离状态复杂物理过程的重要简化。在未来的工作中，我们计划通过充分计算这一平衡态并将其与 Horizon-noAGN 进行匹配，从而更好地为训练集建模这一效应。为了比较性能，我们使用相同的 Horizon-noAGN 训练*LyAl-Net* 预测总氢密度，即n_H^{HnoAGN}，从而可以使用方程（44）缩放密度，以获得 TNG100 模拟中性氢密度。

5.3. LyAl-Net 能力评估

现在，我们讨论使用 *LyAl-Net* 对 TNG100 中暗物质过密度进行模拟的中性氢密度和温度的性能。

5.3.1. 氢密度

我们对 400 个切片（768 × 768 × 400 体素）的一点统计执行相同的核密度估计分析。图16显示，*LyAl-Net* 的预测偏差保持相对稳定，在第一和第九十百分位数处均维持在 10%基准偏差附近，如蓝色部分所示。与 Horizon-noAGN 的对应结果相比，整体预测更加显著，这种情况是预期的。图17展示了传递函数（上图）和相关率（下图），二者分别在最高达到 $k \approx 7$ h 兆秒差距和 $6h$ 兆秒差距之前均保持在 90% 以上。此外，与 Horizon-noAGN 相比，精度从$k \approx 10$ h 兆秒差距处开始下降。

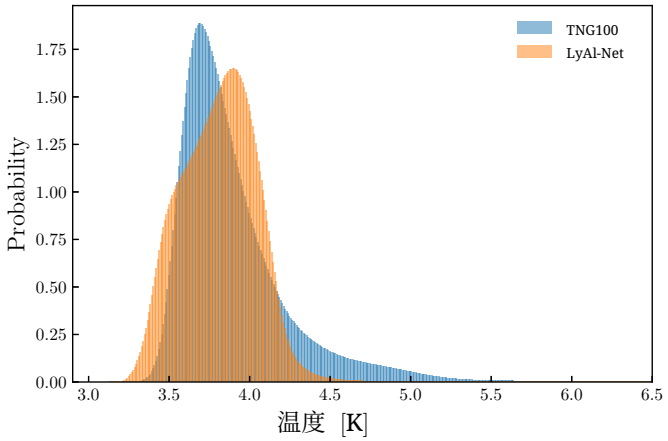

图 18。 LyAl-Net 预测值与 TNG100 真实值之间的温度十进对数一维分布比较

5.3.2. 温度

图18比较了真实温度与预测温度之间的一维分布。与TNG100相比，在高温区域（$T > 10^4$ K）中，预测温度的中位数似乎更高，但分布尾部更短。这种整体低估也体现在图19的偏差中。这主要应是由于缺少AGN 反馈所致。图20中的两点统计量显示，*LyAl-Net* 的传递函数保持在50%以下。功率不足表明，*LyAl-Net* 预测温度的涨落低于真实值，这一点也可从偏差图中看出。另一方面，底部面板中的互相关函数显示，*LyAl-Net* 在达到$k \approx 1h$ 兆秒差距$^{-1}$之前表现最佳，而使用Horizon-noAGN的*LyAl-Net* 在达到$k \approx 3h$兆秒差距$^{-1}$之前表现最佳。

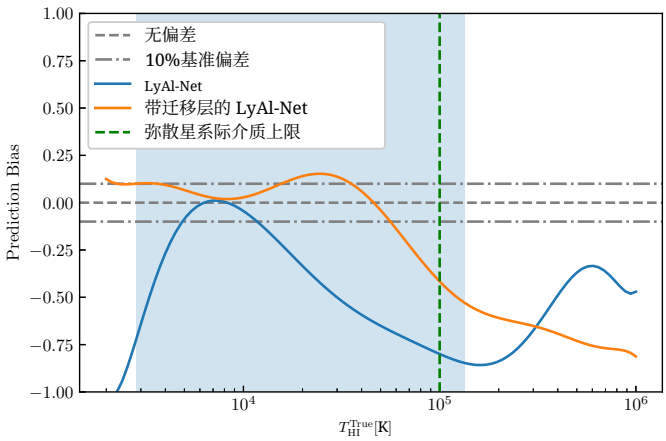

图19。 比较了对 TNG100 的第28号快照（红移 $z = 2.58$）应用迁移学习前后的预测偏差。我们采用与图16相同的颜色和线型约定。结果显示，高温区域整体上有所改善。值得注意的是，通过这一修改，我们几乎达到了弥散星系际介质极限的要求。

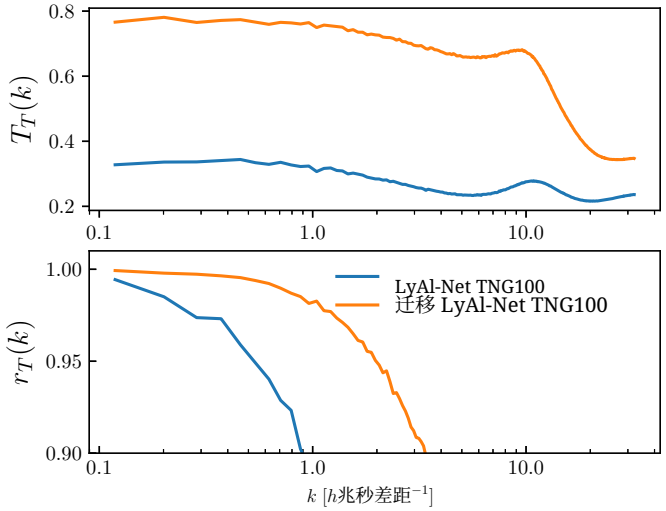

图20。 比较了使用*LyAl-Net*模拟得到的中性氢温度的传递函数和互相关函数（上图）以及互相关函数（下图），同时也采用了高保真范围掩蔽。蓝色实线表示使用原生*LyAl-Net*得到的结果，橙色实线则表示针对四个额外参数进行优化后的迁移结果。值得注意的是，网络对不同物理过程的表现有了显著提升。

5.4. 迁移学习

LyAl-Net 主要将暗物质超密度转换为中性氢密度和温度。正如前几节所示，它依赖于物质密度与气体参数之间的普适空间相关性。尽管暗物质通常能够追踪气体云，但 *LyAl-Net* 能够进行预测的空间范围存在限制。前几节展示的结果表明，将*LyAl-Net* 应用于其他暗物质模拟（本例中为IllustrisTNG）可能会导致模拟气体场不准确。这是预期之中的，因为至少重子反馈与Horizon-noAGN并不相同。应用*LyAl-Net* 的主要目标是建立一个能够泛化并模仿针对不同状态方程的气体物理过程，例如冷却率和紫外指数，并有可能推广到不同的宇宙学模型。实现 *LyAl-Net* 的这种可移植性，最终将使其能够用于宇宙学分析以及计算星系际介质物理中的其他气体物理量。尤其是，宇宙学推断需要大量前向模拟，才能恰当地边缘化气体物理中的系统效应和不确定性。对于未来的巡天项目（例如莱曼-α-森林巡天），一种通用且低成本的星系际介质地图生成方法至关重要。本节将探讨稍微扩展 *LyAl-Net* 的可能性，以获得包含理想宇宙学模型的流体力学模拟。*LyAl-Net* 可被视为一种复杂的非线性变换，其中输出气体场（y）是对输入暗物质（x）的响应，可表示为

$$y = f(x)\,. \tag{45}$$

模拟器 *LyAl-Net*只能在有限的尺度范围内、针对给定的物理过程，实现从暗物质到气体性质的充分变换。在 Horizon-noAGN 上训练的 *LyAl-Net* 具有良好的气体状态方程仿真能力。我们首先应尝试保留的，正是这一物理性质，它相当于对气体物理过程任意解析解进行 n 阶展开。通过调整 *LyAl-Net*的输入和输出，而无需重新训练整个网络，我们或许能提高仿真气体参数的高阶精度。根据方程（45），可得

$$\tilde{y} = g\{f[h(x)]\}\,, \tag{46}$$

其中，$h(x)$ 和 $g(y)$ 分别是输入暗物质超密度和输出气体场的任意变换函数。为了检验这一想法，我们引入一个自定义变换函数，其形式为

$$L(x) = Ax + B, \tag{47}$$

其中，A 和 B是两个自由参数。该函数对*LyAl-Net*在对数空间中进行线性缩放和平移。我们假设物理反馈过程可以归纳并编码为少数几个参数，从而减小偏差，尤其是气体弥散区域的偏差，并最终改善莱曼-α通量。正如引言中所述，预计这在一阶近似下能够改善对气体状态方程和密度的仿真。我们注意到，可能存在更好的变换方式，这将留待未来研究。我们将该函数作为一个不含激活函数的层集成到 *LyAl-Net* 中，并称之为迁移层。迁移层分别置于暗物质输入之前和*LyAl-Net*预测之后，从而将其包裹起来。以这种方式实现迁移层后，函数的线性特性将不再保留

$$y' = L(y; A, B) = L(LyAl[L(x); C, D]; A, B)\,. \tag{48}$$

图21展示了迁移层实现方式的示意图。*LyAl-Net* 的权重被冻结。与网络的原始训练一样，我们使用均方误差来优化这些参数。这使我们能够利用图形处理器和 TensorFlow，根据真实值与预测值之间的 χ^2 损失函数，优化迁移层的系数，即 A、B、C 和 D，其中真实值与预测值来自 IllustrisTNG-100。我们仅抽取了512个子盒，27^3 个体素，相当于总体积的 ~ 1%。表4汇总了通过上述优化得到的参数。随后，我们重新计算了模拟得到的总氢密度$n_{\rm HI}$ 和温度$T_{\rm HI}$。

表4。在红移$z = 2.58$时，IllustrisTNG-100 总氢密度和温度的迁移层参数。所有参数均使用均方误差函数进行优化，样本大小为512个子盒，每个子盒包含 27^3 个体素。

Gas 参数	传递层参数			
	A	B	C	D
密度	1.180	0.657	1.133	0.084
温度	0.898	-0.862	1.34	-0.466

5.4.1. 氢密度

对于氢气密度，图16中橙色的偏差函数显示，一方面，蓝色区域（第1至第90百分位数）的整体偏移略微超出基准偏差区域；另一方面，极高密度下的偏差受到抑制。对于两点统计量，图17显示传递函数振幅整体有所改善，但在$k = 10h$ 兆秒差距$^{-1}$ 的极小尺度上，预测密度在此过程中被过度增强了10%。另一方面，互相关在所有可见空间尺度上均有显著改善。对应于90%相关率的波数增加了 ~ 40%，从6.5h 兆秒差距$^{-1}$ 提高到9h 兆秒差距$^{-1}$。

5.4.2. 温度

应用迁移层后，图 19 中温度的偏差函数有所改善；在第 1 和第 90 百分位数范围内，模拟器预测的平均值稳定在 10% 基准误差以内。从图 20 可以看出，传递函数的振幅大幅改善，几乎是原始 *LyAl-Net* 的两倍。互相关也表明，空间分布精度如今提升至 $k \sim 3h$ 兆秒差距$^{-1}$，表现与 Horizon-noAGN 对应模型相近。

5.5. 莱曼-α 森林吸收率

图 22 展示了 TNG100 中一条归一化莱曼-α 通量的样本视线。上方面板直接对比了蓝色的真实值、橙色的开箱即用 *LyAl-Net* 计算通量，以及配备迁移层的绿色 *LyAl-Net* 。两种配置下 *LyAl-Net* 得出的通量都很好地追踪了真实通量。下方面板的残差图显示，这条视线上的表现相近。然而，原始 *LyAl-Net* 在高吸收区域表现不佳。应用迁移学习层后，该区域得到了显著改善。这主要归功于通过应用迁移层校正温度场，这种做法往往会使吸收变得平滑且范围更宽。
平均透射通量的比较
图23显示，*LyAl-Net* 的相对误差为 ≈ 4%。通过迁移层后，该误差大幅降低。我们可以看到，迁移层提高了吸收精度，这一点从图24顶部面板的传递函数中可以看出。振幅增加至0.9，直至$k \approx 10h$ 兆秒差距$^{-1}$。底部通过互相关函数测得的空间精度仅有微小提升。这表明，当前的迁移层主要改善了吸收量。此外，如图15所示，整体性能与 Horizon-noAGN通量预测几乎完全相同。

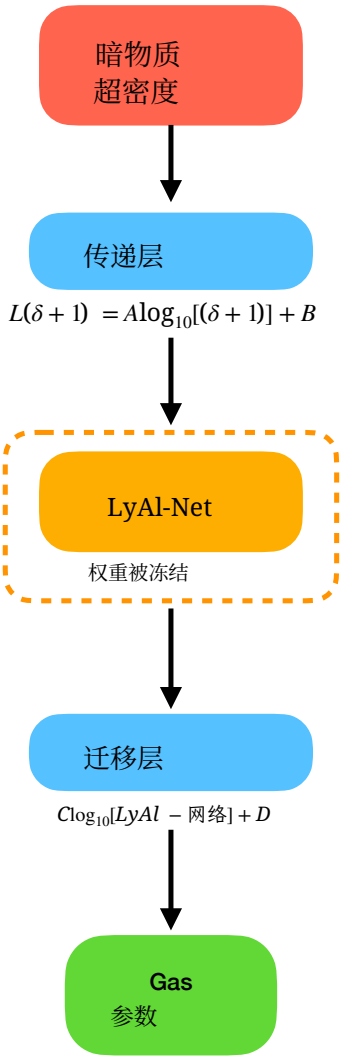

图 21. 用于实施迁移学习过程的改进模拟器的简单示意图。核心网络 *LyAl- Net* 的权重现已冻结，这意味着不再对其进行进一步训练。我们在暗物质超密度与该网络之间引入迁移层，以调整输入来适应新问题。我们仅优化A、B、C和D，以微调预测结果；这也可以视为高阶修正。

这一发现表明，采用实验性迁移层模型的*LyAl-Net* 展现了该框架的可移植性。该模型还可以扩展，用于外推重子反馈的不同配置，以支持未来的宇宙学模拟。此外，在贝叶斯分层模型中，这八个参数（气体密度的四个参数和温度的四个参数）可以与密度场一起，根据Lyman-α森林观测数据进行联合推断，例如Porqueres 等（2020）所示。

6. 讨论与结论

我们介绍了*LyAl-Net*，这是一种神经网络架构和训练流程，可根据宇宙学尺度上的暗物质密度和速度场推导氢气场量。我们还展示了多项基准测试，评估其稳健性、可移植性，以及生成模拟 Lyman-α 森林观测数据的能力。该网络最重要的特点之一，是能够移植到其他气体模型中。在现有文献中，模拟得到的平均透射通量一直通过一个简单的重新缩放因子进行校正，以匹配观测数据

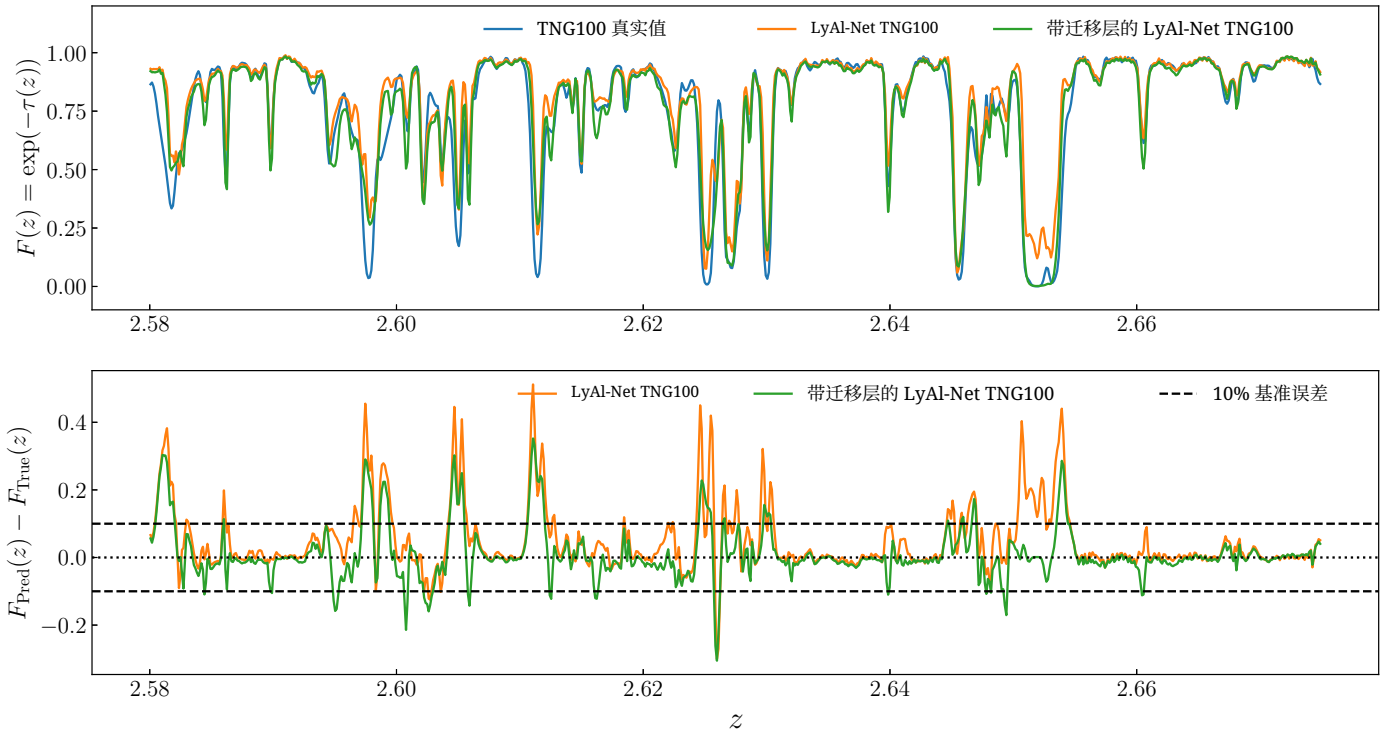

图 22。上图：比较来自同一样本视线的真实值和模拟通量的莱曼-α 森林归一化通量，分别标记为真实值、*LyAl-Net*以及带有*LyAl-Net* 迁移层的结果；数据来自 IllustrisTNG100 快照 28，红移为$z = 2.58$。带迁移层的*LyAl-Net* 显示出改进，尤其是在强吸收区域内。**下图：**通量残差以及以虚线绘制的 10% 基准误差表明，两种配置的残差值大多保持在基准区域内

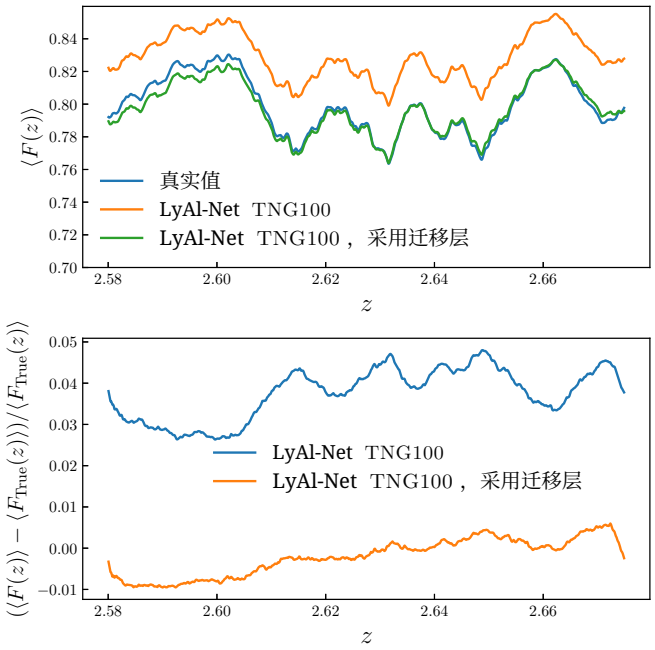

图 23。上图：比较 IllustrisTNG-100 的平均透射通量随红移的变化。真实通量以真实值 表示，并用实心蓝线绘制；另外还比较了根据*LyAl-Net*模拟的星系际介质计算得到的两种模拟通量：1）以橙色表示的经过训练的*LyAl-Net*，以及 2）以绿色表示的带迁移层的经过训练的*LyAl-Net*。**下图：**比较使用经过训练的*LyAl-Net* 和带迁移层的*LyAl-Net* 得到的模拟平均透射通量的相对误差图。结果显示，应用迁移层后，相对误差大幅降低。

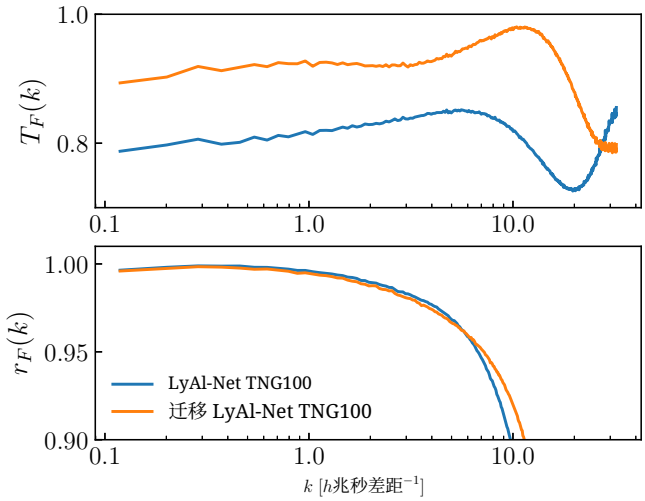

图 24。使用由 *LyAl-Net*（蓝色标示）和带有迁移层的*LyAl-Net*（橙色标示）根据 IllustrisTNG-100 在$z = 2.58$ 暗物质超密度下模拟的透射通量的二点互相关函数进行性能评估，其中上方面板为传递函数比较，下方面板为互相关函数比较。

（Regan 等，2007；Luki´c 等，2014；Borde 等，2014）。这种方法等价于根据观测数据重新缩放紫外背景的强度和光电离率（Meiksin，2009）。然而，这种重新缩放并不能深入揭示星系际介质层面发生的情况。相比之下，近期

这些研究致力于提高该领域的准确性，从而更好地仿真通量吸收。这些研究还采用神经网络方法，根据暗物质的N体模拟生成流体动力学模拟。例如，Horowitz等人（2022年）使用了类似于U-Net架构的全卷积变分自编码器。后续论文中，Wadekar等人（2020年）使用特殊的*U-Net*架构，根据暗物质模拟预测中性氢密度。然而，为了捕捉低密度区域和高密度区域中的物理过程，他们必须结合两个独立的神经网络，并根据局部暗物质密度的振幅选择其输出。Sinigaglia等人（2022年）采用分层的特定领域机器学习方法，从暗物质超密度中获得Lyman-α 森林通量（昵称为BAM），但还需要重建HI和HII。

在这项工作中，我们对现有科学文献中的若干方面作出了改进：针对不同流体动力学物理过程的实现方式，刻画模拟场相对于仿真场的随机性质；增加一个灵活的层，以便将我们的结果迁移到未知的新物理过程；简化整体 U-net 架构，以模拟 Lyman-α；并采用 LyMAS 中使用的自适应滤波器，对暗物质进行滤波，以减轻空洞中的粒子散粒噪声效应（Peirani等人，2014；Peirani 等（2022））。第一点对于将我们的模型用于贝叶斯分析至关重要。我们需要了解模型的局限性，至少要了解其单点统计层面的局限性，这样才能将模型拟合到观测数据，例如将其用于场级分析（如 Porqueres 等（2020）中的 BORG）、基于模拟的方法（如 SELFI，Leclercq等人，2019），或标准的相关函数分析（Slosar等人，2011）。

上述列表中的第二点最为重要，因为我们无法保证控制仿真场（气体密度、温度）的参数（例如恒星形成、活动星系核活动）与观测数据一致。可移植模型让我们能够直接对此进行探查，并根据观测数据了解模型的局限性。Lee等人（2015）；Palanque-Delabrouille等人（2015）采用了类似的方法，但仅限于为 FGPA 模型匹配两点统计量。*LyAl-Net* 模型在观测数据上完成校准后，还能间接让我们将这些模拟场作为副产品加以利用。模拟器在场层面预测流体动力学模拟，这使 *LyAl-Net* 能够利用任何N体模拟代码预测星系际介质的性质。与 LyMAS-2（Peirani等人，2014）相比，我们做出了若干改进。*LyAl-Net* 无需平滑滤波器，但模拟场的分辨率是固定的，取决于训练集。因此，该框架可应用于任意所需的分辨率。为检验*LyAl-Net*的可移植性，我们首先将预训练网络直接应用于 IllustrisTNG-100 暗物质。其重子反馈的物理过程存在于该模拟中，而训练集所源自的 Horizon-noAGN 中则不存在。图 5.5 所示结果表明，$\langle F(z)\rangle$ 的相对误差增大，这正如预期，是由氢密度预测值偏低直接导致的。我们通过使用迁移层进行简单的数值处理解决了这一问题，大幅改善了模拟通量，使其性能与 Horizon-noAGN 的模拟结果相当。我们引入的迁移层可以简单地理解为对状态方程进行调节，将n_H和T的中位数移向更接近 TNG100 真实值的位置。当然，我们所用方程的形式只有少数几个参数，只能对数值进行一维重映射，预计无法模拟高阶修正。简言之，我们可以在不任意缩放通量的情况下改善平均透射率，并应用迁移层微调星系际介质的状态方程。对于给定的重子物理过程，这种方法在提供额外信息和估算的同时，还能将流体动力学模拟的各个场作为副产品获得。

我们提醒读者注意图14中平均透射通量随红移的变化。使用Horizon-noAGN和Horizon-DM时，相对误差表现相同，而Horizon-DM的相对误差更高。根据这一经验性观察，我们可以有把握地认为，*LyAl-Net* 能够基于过密度涨落，将暗物质单元转换为星系际介质属性。然而，由于重子反馈过程具有随机性，它在仿真小尺度效应方面存在一些局限性；在这些反馈过程中，*LyAl-Net* 可能需要被正确捕捉，或者只能提供平均响应，并且精度可能较低。本研究采用的方法仍存在一些局限性。值得注意的是，我们在固定红移下，以共形时间为尺度测试了仿真效果，因此无法保证*LyAl-Net* 能够迁移到其他模拟时间。我们尚未充分探索降低U-Net复杂性的可能性，而更小的版本或许能够达到相同的精度水平。未来的工作中，我们还希望改进模型的可解释性。例如，在保持紧凑性的同时，传递层或许可以采用比方程38更优的形式。我们还希望减少仿真不同环境和反馈所需的自由参数，使其足以匹配宇宙学巡天所需的尺度和分辨率。

我们开发了一种快速的莱曼-α 森林模拟器，利用神经网络模拟星系际介质的流体动力学模拟。我们使用 Horizon-noAGN 作为训练集。当前模型在$z \sim 2.5$ 的模拟中表现良好，同时将整个模拟视为与红移无关的独立对象。我们使用 IllustrisTNG-100 在$z \sim 2.4$ 处的暗物质过密度测试了模拟器的敏感性。该模拟包含不同于训练集的重子反馈，因此预测偏差如预期般出现。为了对模拟生成的场进行后处理，我们引入了一种称为线性层 的方法，用于校准星系际介质状态方程。我们发现，该方法使莱曼-α 吸收通量更接近完整的流体动力学模拟。这一改进表明，对于固定宇宙学参数和重子反馈，可以通过线性层的形式对经过训练的*LyAl-Net* 进行校正，而相应的参数仍需具备可解释性。

此版本的 *LyAl-Net* 假设对于给定训练集红移不变，这意味着它仅限于覆盖特定体积的单一红移。尽管如 IllustrisTNG 测试所示，即使使用具有不同红移的宇宙学模拟，迁移层也可以微调星系际介质状态方程。然而，随着红移升高，星系际介质密度-温度关系变得不太可靠，因为暗物质分布的演化程度较低。此外，随着越来越多高红移莱曼-α 森林光谱观测数据的获得，未来的工作将研究 *LyAl-Net* 对红移的依赖性，使其更加灵活稳健。

LyAl-Net 的另一项发展方向是将其集成到 DELFI（Alsing 等人，2019）、BOLFI（Leclercq，2018）和 SELFI（Leclercq 等人，2019）等推断模型中。这类模型同样需要仿真速度，而这正是本工作的首要重点，因为此类模型需要海量数据才能进行有效分析。LyAl-Net 还能够改进大型宇宙学巡天的经典互相关分析（Seljak 等人，2006；Amendola 等人，2013；Ivezi´c 等人，2019）。此外，LyAl-Net 通过结合高红移处的 Lyman-α 示踪物与星系聚类，为开展场级大尺度结

构推断铺平了道路（Jasche & Wandelt，2013；Porqueres 等人，2019；Ivezi´c 等人，2019；Tsaprazi 等人，2023）。与星系聚类不同，Lyman-α 森林对欠密度区域较为敏感，并以更高的分辨率探测大尺度结构。因此，预计将二者结合能够提高场级结构形成推断的约束能力，同时无需承担流体动力学模拟的巨大计算成本，便可仿真星系际介质的物理过程。

致谢。 我们感谢 Francisco Villaescusa- Navarro、Shy Genel、Simon Prunet 和 Benjamin D. Wandelt 提供的宝贵讨论。本工作得到 ANR BIG4项目（法国国家研究署资助项目 ANR-16-CE23-0002）的支持。本工作得到西蒙斯“学习宇宙”合作项目的支持。CB 感谢索邦人工智能中心（SCAI）提供的经费支持。本工作得益于“理解动态宇宙”研究项目资助，该项目由克努特与爱丽丝·瓦伦堡基金会根据 Dnr KAW 2018.0067 资助。本工作使用了由巴黎天体物理研究所托管的 Horizon集群，并在其中对宇宙学模拟进行了后处理。我们感谢 Stéphane Rouberol 为我们顺利运行该集群。本工作在 Aquila联盟[8]框架内开展。我们感谢以下软件包在本分析中的应用：Numpy（Harris 等，2020年）、Tensor- flow（Abadi 等，2015年）、JAX（Bradbury 等，2018年）、IPython（Pérez 与 Granger，2007年）、Matplotlib（Hunter，2007年）、Numba（Lam 等，2015年）和 SciPy（Jones 等，2001年）。

[8] https://www.aquila-consortium.org/

附录A：沃伊特剖面计算

Lyman-α 的截面采用沃伊特剖面进行建模。该函数的计算成本非常高。我们依赖 scipy.special.wofz 函数来高效获得结果。该软件实现了 Faddeeva算法来计算它（Johnson 2012），即一种复缩放互补误差函数（Faddeeva与Terent'ev 1961）。我们注意到，这是整个流程中计算成本最高的部分。

附录B：核密度估计实现

我们使用 scipy.stats.gaussian_kde 估计概率密度函数，该方法采用高斯核，并使用 Scott规则进行带宽选择（Scott 2015年）。然而，计算时间并不会随样本数量线性增加。因此，实际可行的方法是使用数量有限的体素。在我们的测试中，Horizon-noAGN 的完整模拟体积包含 $1024 \times 1024 \times 1024$ 个体素。我们使用 $1024 \times 1024 \times 500$ 个体素进行分析，约占总体积的 50%。对于这一减少后的点数，在 AMD EPYC 7702（128 个核心）上生成密度估计所需的实际运行时间总计约为 38 小时。

附录C：功率谱掩膜过程

该 $n_{\rm HI}$ 预测在高密度区域受到的影响要大得多，因为模拟器错误地产生了无效预测。我们推测，高密度带来的偏差是正文中记载的使用两点统计量的诊断出现异常的罪魁祸首。为检验这一假设，我们选取了三个不同保真度范围以下的密度像素（见表 3），并将该范围之外的密度掩蔽为零。图 C.1 展示了高保真范围内的密度功率谱比较结果；此时结果与 Horizon-noAGN 吻合良好，证实了前文提到的观点。

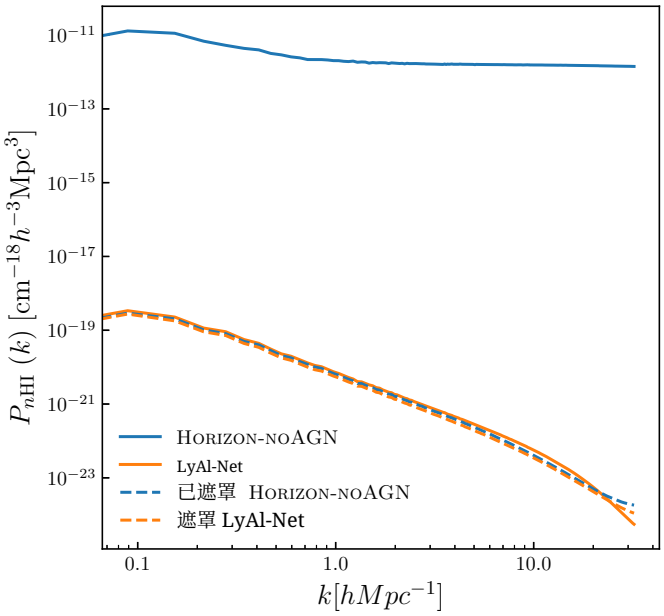

图 C.1。 真实值与模拟中性氢密度之间的功率谱比较。实线表示未遮罩功率谱，虚线表示高保真范围内的遮罩功率谱，二者在所有尺度上都显示出良好的一致性。

参考文献

Abadi, M., Agarwal, A., Barham, P. 等，2015年，《TensorFlow：异构系统上的大规模机器学习》，软件可从 tensor-flow.org 获取；Alsing, J., Charnock, T., Feeney, S. 与 Wandelt, B.，2019年，《皇家天文学会月刊》；Amendola, L., Appleby, S., Bacon, D. 等，2013年，《相对论生活评论》，16，6；Bardeen, J. M., Bond, J. R., Kaiser, N. 与 Szalay, A. S.，1986年，《天体物理学杂志》，304，15；Bautista, J. E., Bailey, S., Font-Ribera, A. 等，2015年，《宇宙学与天体粒子物理学杂志》，2015，060

Becker, G. D. & Bolton, J. S. 2013，皇家天文学会月刊，436，1023 Bernardini, M., Feldmann, R., Anglé s-Alcázar, D., 等 2021，皇家天文学会月刊，509，1323 Bi, H. & Davidsen, A. F. 1997年，《天体物理学杂志》，479，523 Borde, A., Palanque-Delabrouille, N., Rossi, G., 等 2014，宇宙学与天体粒子物理学杂志，2014，005 Bradbury, J., Frostig, R., Hawkins, P., 等 2018，JAX：适用于 Python+NumPy 程序的可组合变换 Colombi, S., Chodorowski, M. J. & Teyssier, R. 2007，皇家天文学会月刊，375，348 Croft, R. A. C., Weinberg, D. H., Katz, N. & Hernquist, L. 1998年，ApJ，495，44 Dai, B. & Seljak, U. 2021，美国国家科学院院刊，118，e2020324118 Dalton, G., Trager, S., Abrams, D. C., 等 2014，第五届地基与机载天文仪器会议

du Mas des Bourboux, H.、Rich, J.、Font-Ribera, A. 等，2020年，《天体物理学杂志》，901，153 Dubois, Y.、Pichon, C.、Welker, C. 等，2014年，皇家天文学会月刊，444，1453 Faddeeva, V. N. 与 Terent'ev, N. M.，1961年，《函数 w(z) 的值表》 = exp ((-z**2)* (1+2l/sqr root pi) 积分 (et**2) dt)，复数自变量 / V.N.Faddeyeva 与 N.M.Terent'ev 编，V.A.Fok 编辑；D.G.Fry 从俄文翻译（培格曼出版社，牛津）Font-Ribera, A.、Kirkby, D.、Busca, N. 等，2014年，《宇宙学与粒子天体物理学杂志》，2014，027 Goodfellow, I. J.、Bengio, Y. 与 Courville, A.，2016年，《深度学习》（美国马萨诸塞州剑桥：麻省理工学院出版社），http://www.deeplearningbook.org Gunn, J. E. 与 Peterson, B. A.，1965年，《天体物理学杂志》，142，1633 Haardt, F. 与 Madau, P.，1996年，《天体物理学杂志》，461，20 Hand, N.、Feng, Y.、Beutler, F. 等，2018年，《天文学杂志》，156，160 Harrington, P.、Mustafa, M.、Dornfest, M.、Horowitz, B. 与 Luki´c, Z.，2021年，arXiv 电子预印本，arXiv:2106.12662 Harris, C. R.、Millman, K. J.、van der Walt, S. J. 等，2020年，《自然》，585，357 He, S.、Li, Y.、Feng, Y. 等，2019年，美国国家科学院院刊，116，13825 Hockney, R. W. 与 Eastwood, J. W.，1988年，《粒子计算机模拟》（美国：泰勒与弗朗西斯出版公司）Horowitz, B.、Dornfest, M.、Luki´c、Z. 与 Harrington, P.，2022年，《天体物理学杂志》，941，42 Horowitz, B.、Lee, K.-G.、White, M.、Krolewski, A. 与 Ata, M.，2019年，《天体物理学杂志》，887，61 Hui, L. 与 Gnedin, N. Y.，1997年，皇家天文学会月刊，292，27 Hunter, J. D.，2007年，《科学与工程计算》，9，90 Iršič, V.、Dio, E. D. 与 Viel, M.，2016年，《宇宙学与粒子天体物理学杂志》，2016，051 Ivezi´c, Z.、Kahn, S. M.、Tyson, J. A. 等，2019年，《天体物理学杂志》，873，111 Jarosik, N.、Bennett, C. L.、Dunkley, J. 等，2011年，《天体物理学杂志增刊》，192，14 Jasche, J. 与 Wandelt, B. D.，2013年，皇家天文学会月刊，432，894 Johnson, S. G.，2012年，Faddeeva 程序包 - AbInitio — ab-initio.mit.edu，http://ab-initio.mit.edu/wiki/index.php/Faddeeva_Package，[访问日期：16-Jan-2023] Jones, E.、Oliphant, T.、Peterson, P. 等，2001年，SciPy：面向 Python 的开源科学工具 Kaushal, N.、Villaescusa-Navarro, F.、Giusarma, E. 等，2022年，《天体物理学杂志》，930，115 Kim, J.、Golwala, S.、Bartlett, J. G. 等，2022年，《天体物理学杂志》，926，179 Kingma, D. P. 与 Ba, J.，2014年，Adam：一种用于随机优化的方法

算法1 用于模拟基本立方体的平铺_立方体 的算法，其中模拟基本立方体是 27^3 模拟基本立方体的输入数组
立方体，*X*、*Y* 和 *Z* 分别表示平铺体积中的行数、列数和层数，*size* 表示每个模拟立方体的大小，而
trim_size 表示为获得所需体积大小 1024^3 而需裁去的多余体素数量。

函数 平铺立方体(模拟_立方体，*X*，*Y*，*Z*，*size*，*trim_size*) ▷平铺模拟立方体，以创建更大的体积
    平铺_立方体 ← 大小为（*X* ∗ 大小，*Y* ∗ 大小，*Z* ∗ 大小）的空三维数组 ▷为平铺立方体创建空数组
    *row*_开始 ← 0 ▷初始化行起始索引
    对于行 ← 0 到*X* − 1 **执行**
        *col*_开始 ← 0 ▷初始化列起始索引
        对 列 ← 0 至 *Y* − 1 **执行**
            层_起始 ← 0 ▷初始化层起始索引
            对 层 ← 0 至 *Z* − 1 **执行**
                模拟_立方体 ← 模拟_立方体[行][列][层] ▷从输入获取仿真立方体
                行_结束 ← 行_起始 + 大小 ▷计算行结束索引
                列_结束 ← 列_起始 + 大小 ▷计算列结束索引
                层_结束 ← 层_起始 + 大小 ▷计算层结束索引
                平铺_立方体[行_起始： 行_结束， 列_起始： 列_结束，层_起始： 层_结束] ← 仿真_立方体 ▷ 插入仿真立方体到平铺体积中
                层_起始 ←层_结束 ▷更新层起始索引
            循环结束
            列起始 ←列结束_ _ ▷更新列起始索引
        循环结束
        行起始 ←行结束_ _ ▷更新行起始索引
    结束循环
    裁剪后的_立方体 ← 每个维度上的前裁剪_大小元素，来自平铺的_立方体 ▷裁剪多余体素以获得所需的体积大小
    返回裁剪后的 _立方体 ▷返回裁剪后的体积
结束函数

Komatsu, E.、Smith, K. M.、Dunkley, J. 等，2011年，《天体物理学杂志增刊》，192，18 Kooistra, R.、Lee, K.-G. 与 Horowitz, B.，2022年，《天体物理学杂志》，938，123 Kraljic, K.、Laigle, C.、Pichon, C. 等，2022年，皇家天文学会月刊，514，1359 Lam, S. K.、Pitrou, A. 与 Seibert, S.，2015年，载于《LLVM 编译器基础设施在高性能计算中的第二届研讨会论文集》- LLVM '15（美国得克萨斯州奥斯汀：ACM 出版社），1–6 Leclercq, F.，2018年，《物理评论D》，98 Leclercq, F.、Enzi, W.、Jasche, J. 与 Heavens, A.，2019年，皇家天文学会月刊，490，4237 Leclercq, F.、Jasche, J.、Gil-Marín, H. 与 Wandelt, B.，2013年，《宇宙学与粒子天体物理学杂志》，2013，048 Lee,K.-G.、Bailey, S.、Bartsch,L.E. 等，2013年，《天文学杂志》，145，69 Lee, K.-G. 等，2015年，《天体物理学杂志》，799，196 Levi, M. E.、Allen, L. E.、Raichoor, A. 等，2019年，《暗能量光谱仪器》（DESI）Liszt, H.，2001年，《天文学与天体物理学》，371，698 Luki´c、Z.、Stark, C. W.、Nugent, P. 等，2014年，皇家天文学会月刊，446，3697 Lynds, R.，1971年，《天体物理学杂志》，164，L73 Martizzi, D.、Vogelsberger, M.、Artale, M. C. 等，2019年，皇家天文学会月刊，486，3766 Meiksin, A. A.，2009年，《现代物理评论》，81，1405 Monaghan, J. J.，1992年，《天文学与天体物理学年评》，30，543 Nelson, D.、Springel, V.、Pillepich, A. 等，2019年，《计算天体物理学与宇宙学》，6，2 Nguyen, C. V.、Cao, T.-D.、Truong-Huu, T.、Pham, K. N. 与 Nguyen, B. T.，2021年，关于采用间隔余弦损失和相对判别器的生成对抗网络的实证研究 O'Shea, K. 与 Nash, R.，2015年，卷积神经网络简介 Palanque-Delabrouille, N.、Yè che, C.、Baur, J. 等，2015年，《宇宙学与粒子天体物理学杂志》，2015，011 Pando, J.、Feng, L.-L. 与 Fang, L.-Z.，2004年，《天体物理学杂志增刊》，154，475 Peirani, S.、Dubois, Y.、Volonteri, M. 等，2017年，皇家天文学会月刊，472，2153 Peirani, S.、Prunet, S.、Colombi, S. 等，2022年，皇家天文学会月刊，514，3222 Peirani, S.、Weinberg, D. H.、Colombi, S. 等，2014年，《天体物理学杂志》，784，11 Pérez, F. 与 Granger, B. E.，2007年，《科学与工程计算》，9，21 Petitjean, P.、Mücket, J. P. 与 Kates, R.，1995年，低红移处的莱曼-α森林：追踪暗物质丝状结构

Pichon, C.、Vergely, J.、Rollinde, E.、Colombi, S. 和 Petitjean, P. 2001年，《皇家天文学会月刊》，326，597 Pieri, M. M. M.、Bonoli, S.、Chaves-Montero, J. 等 2016年，载于《法国天文学与天体物理学周》，法国天文学与天体物理学会（SF2A），法国里昂，8页，2幅图，SF2A会议论文集，里昂，2016年 Planck Collaboration VI。2020年，《天文学与天体物理学》，641，A6 Porqueres, N.、Hahn, O.、Jasche, J. 和 Lavaux, G. 2020年，《天文学与天体物理学》，642，A139 Porqueres, N.、Jasche, J.、Lavaux, G. 和 Enßlin, T. 2019年，《天文学与天体物理学》，630，A151 Ramanah, D. K.、Charnock, T. 和 Lavaux, G. 2019年，《物理评论D》，100 Regan, J. A.、Haehnelt, M. G. 和 Viel, M. 2007年，《皇家天文学会月刊》，374，196 Ronneberger, O.、Fischer, P. 和 Brox, T. 2015年，CoRR，abs/1505.04597[1505。04597] Schaye, J.、Theuns, T.、Leonard, A. 和 Efstathiou, G. 1999年，《皇家天文学会月刊》，310，57 Scott, D. W. 2015年，《多元密度估计》，第2版，威利概率与统计丛书（英格兰奇切斯特：威利-布莱克威尔）Seljak, U.、Slosar, A. 和 McDonald, P. 2006年，《宇宙学与粒子天体物理学杂志》，2006，014 Silverman, B. W. 1986年，《统计与数据分析中的密度估计》，Chapman & Hall/CRC统计与应用概率专著丛书（宾夕法尼亚州费城：Chapman & Hall/CRC）Sinigaglia, F.、Kitaura, F.-S.、Balaguera-Antolínez, A. 等 2022年，《天体物理学杂志》，927，230 Slosar, A.、Font-Ribera, A.、Pieri, M. M. 等 2011年，《宇宙学与粒子天体物理学杂志》，2011，001 Sorini, D.、Oñorbe, J.、Luki´c, Z. 和 Hennawi, J. F. 2016年，《天体物理学杂志》，827，97 Springel, V. 2010年，《皇家天文学会月刊》，401，791–851 Sutherland, R. S. 和 Dopita, M. A. 1993年，《天体物理学杂志增刊》，88，253 Teyssier, R. 2002年，《天文学与天体物理学》，385，337–364 Tsaprazi, E.、Jasche, J.、Lavaux, G. 和 Leclercq, F. 2023年，arXiv电子预印本，arXiv:2301.03581 Villaescusa-Navarro, F.、Vogelsberger, M.、Viel, M. 和 Loeb, A. 2013年，《皇家天文学会月刊》，431，3670 Villasenor, B.、Robertson, B.、Madau, P. 和 Schneider, E. 2022年，《天体物理学杂志》，933，59 Vlah, Z.、Seljak, U.、Chu, M. Y. 和 Feng, Y. 2016年，《宇宙学与粒子天体物理学杂志》，2016，057–057

Wadekar, D.、Villaescusa-Navarro, F.、Ho, S. 和 Perreault-Levasseur, L. 2020年, 《HInet：利用神经网络从暗物质生成中性氢》Weinberg, D. H.、Hernquist, L.、Katz, N.、Croft, R. 和 Miralda-Escudé, J. 1997年, 《莱曼-α森林的哈勃流展宽及其影响》Yèche, C.、Palanque-Delabrouille, N.、Baur, J. 和 du Mas des Bourboux, H. 2017年, 《宇宙学与粒子天体物理学杂志》, 2017，047