Teaching dark matter simulations to speak the halo language

Shivam Pandey [1] **Francois Lanusse** [2] **Chirag Modi** [2] **Benjamin Wandelt** [2 3]

Abstract

We develop a transformer-based conditional generative model for discrete point-objects and their properties and use to to build a model for populating cosmological simulations with gravitationally collapsed structures called dark matter halos. Specifically, we condition our model with dark matter distribution obtained from fast, approximate simulations to recover the correct three-dimensional positions and masses of individual halos. This leads to a first model that can recover the statistical properties of the halos at small scales to better than 3% level using an accelerated dark matter simulation. This trained model can then be applied to simulations with significantly larger volume which would otherwise be computationally prohibitive with traditional simulations, and also provides a crucial missing link in making end-to-end differentiable cosmological simulations. The code, named GOTHAM (Generative Conditional Transformer for Halos And their Masses) is made publicly available.

1. Introduction

Transformer-based architecture, that is now a staple in natural language processing (NLP) applications, excels at learning the conditional distribution of data. It scales well, is highly flexible and has a native auto-regressive structure. Through the attention mechanism, it is able to learn the syntactical meaning of a token in relation to other surrounding tokens. This property has remarkable applications beyond NLP. Here we apply the transformer-based architecture to solve one of the long standing cosmological problems: creating a differentiable end-to-end simulations that can be used as a Bayesian forward model to exhaust the information content in the observations. However, the framework and code developed here can in general be applied to any point cloud generation problem that is conditioned on external continuous field.

Cosmological N-body simulations evolve system of more than a billion dark matter particles under gravity from initial Gaussian random fluctuations to present state non-Gaussian large scale structure. After billions of years of evolution, most of the mass in the simulations ends up in the collapsed dark matter structure called halos. These form the hosts of the galaxies we observe and are, to first-order, described by their masses. However, solving particle-particle interactions of billions of particles for many time-steps makes these simulations computationally intensive. Moreover, finding and characterizing the collapsed dark matter structures further exacerbates the computational requirements. Finally, both the traditional N-body simulations and halo finders are non-differentiable processes, making it challenging to use machine learning based methods for their analysis.

Particle mesh (PM) simulations instead put all the particles on a regular grid and solve their equations of motions using fast Fourier transform techniques (Feng et al., 2016). These simulations capture the large scale dark matter distribution accurately, are significantly faster compared to N-body and can also be written in a differentiable form (Modi et al., 2021; Li et al., 2022). However, as the particles are placed on a grid, it lacks the resolution capabilities of a N-body simulation and underestimates the small scale structures. As the dark matter halos form and get their properties from small scale interaction of the particles, this means that the PM simulations severely underestimate the number and properties of dark matter halos. In this study, we learn the mapping between large scale dark matter distribution as obtained from PM simulations and N-body like dark matter halos. This architecture can also be made differentiable (Horowitz et al., 2024), which when combined with PM simulations, leads to an end-to-end differentiable cosmological simulator.

Having an accurate differentiable forward model is required if we hope to extract most of the information in the observed galaxy data. The volume of the galaxy survey (Alam et al., 2015) that ended a decade ago is atleast a factor of 27 larger than available from the current best N-body simulation suites (Villaescusa-Navarro et al., 2020) for forward

[1]Department of Physics, Columbia University, 538 West 120th Street, New York, NY, USA 10027, USA [2]The Flatiron Institute, 162 5th Ave, New York, NY, 10010, USA [3]CNRS & Sorbonne Université, Institut d'Astrophysique de Paris (IAP), Paris, France. Correspondence to: Shivam Pandey <sp4204@columbia.edu>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

modeling the observations. The model developed here describes a method to use fast and approximate dark matter simulations that are easy to scale to obtain N-body like halo catalogs.

2. Related Work

As halos are a set of discrete objects in 3D space, techniques based on point cloud and their transformations are relevant. Several studies have implemented various ways of completing a point cloud by leveraging the features learned from the given partial set of points (Yan et al., 2022; Yu et al., 2022). However, here we are interested in generating full set of point cloud conditioned only on the surrounding dark matter distribution which requires a different architecture. Moreover, usually the point cloud transformers treat each point as unweighted. However, for halos, their masses is an important property to capture accurately as it can significantly impact the properties of galaxies it hosts which is what we measure in the observations. Therefore, here we develop an architecture that can jointly infer the position and masses of the halos.

Several past studies have also aimed at learning a mapping from simulated dark matter distributions to dark matter halos. For example, Kodi Ramanah et al. (2019) trained a Wasserstein GAN to predict counts of halos within four mass bins based on gridded dark matter density from a full N-body simulation. Jamieson et al. (2022) learn the correction to velocities and positions of the particles in PM simulations to replicate N-body like simulations. However, even after the correction, the statistical properties of halos are only reproduced at 20-30% level. Pandey et al. (2023) learn the mapping between PM simulation and halos, but are limited to large scales ($k < 0.2h/\text{Mpc}$) in reproducing the statistical properties of halos. In this work we treat the conditional halo catalog generation like language translation problem and use large language model like architecture to solve it. The flexibility of such a model allows us to model the halo distribution and their masses at the 2-3% level to significantly smaller scales ($k \sim 1h/\text{Mpc}$).

3. Dataset

The halo catalogs for training are obtained from the public Quijote N-body 'high-resolution' simulation suite at a fiducial cosmology (Villaescusa-Navarro et al., 2020) which evolve 1024^3 particles inside a box with a side length of 1000 Mpc/h. We learn the distribution of these catalogs when conditioned on 3D dark matter densities derived from the low-resolution PM simulations.

3.1. Input : Continuous 3D conditional field

We run PM simulations with same initial conditions and volume as the Quijote simulations. However, to significantly reduce the run time, we run them with only 384^3 particles and the forces between particles are calculated in a 768^3 grid. When run on CPUs, each simulation has a runtime of 5 CPU hours (c.f. 5000 CPU hours for N-body simulation), which can be further reduced by using its GPU implementation (Li et al., 2022). We run 11 different PM simulations for different set of initial conditions to capture the stochastic contribution. We sub-divide each parent 1000 Mpc/h box into 32^3 sub-boxes (giving $L_{\rm sub-box} = 31.25\,{\rm Mpc}/h$) and treat each of these sub-boxes as independent. We use the sub-boxes from first three simulations for training (80%) and validation (20%) and use remaining eight simulations as test set.

3.2. Target: Weighted discrete 3D point cloud

We use an accurate definition of halos which uses phase space distribution of dark matter particles to identify collapsed structures and assign them masses (Behroozi et al., 2012). In this study, we only focus on halos with masses above $M_{\rm halo} > 10^{13.5} M_\odot/h$. We aim to learn the three spatial coordinates and mass of each halo, when conditioned on the input density field.

3.3. Tokenization

As the position and masses of the halos are continuous, we tokenize them. First we scale each of 3D coordinate and logarithm of masses in the range $(0, 1)$. Then we divide them into 64 bins and assign each halo four integers corresponding to its 3D coordinate and mass. Therefore, each halo effectively becomes a "word" with four characters. We concatenate the tokens of all the halos in the sub-box by spacing them with a `<SPACE>` token. Finally we pre-pend a `<START>` token and append an `<END>` token to create a "sentence" of halos for each sub-box (giving $n_{\rm vocab} = 67$). As we physically expect the halo with heaviest mass to dominate the structure formation in a sub-box, we concatenate the halos with a descending order of their masses. This is then padded with a `<PAD>` token to reach the maximum sequence length of $N_{\rm seq-dec} = 101$ (corresponding to having a maximum of 20 halos in a sub-box). We then create a right-shifted prediction vector to predict these tokens conditioned on the input field. Therefore the task effectively becomes understanding the "grammatical" structure of these halo "sentences".

4. Methodology

We use the transformer architecture along with residual network in this task. Similar to NLP models (Vaswani et al., 2023), we can divide our architecture into two parts, encoder and decoder. We show the details of the architecture in Fig. 1.

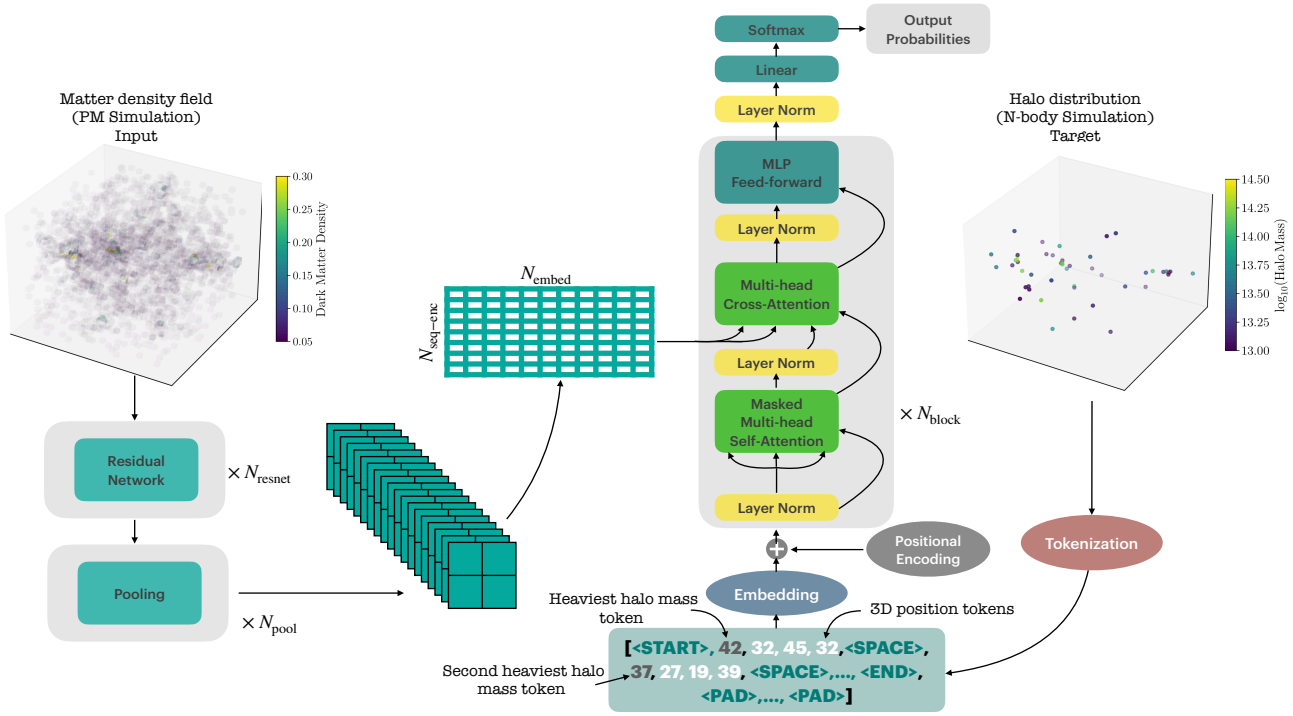

Figure 1. Model architecture: On left we show the input dark matter density distribution for one training sub-box and on right we show the target 3D distribution of halos colored by their masses. These four properties are tokenized and concatenated for all the halos in the sub-box in a way that it forms a 'sentence'. On the encoder side, we use stacks of 3D residual networks to extract the features from the density field and input that to the cross-attention module of the decoder model to learn the conditional log-probability of the tokens. See Sec. 3 and Sec. 4 for more details.

4.1. Encoder

To extract the feature vectors that can inform the halo selection, we run stack of $N_{\rm res-net} = 4$ residual networks. We input three PM density fields to the network, each with a resolution of $32 \times 32 \times 32$, but obtained from a physical volume of 32, 48 and 96 Mpc$/h$ respectively, centered on each sub-box. This captures the information from the surrounding environment that is crucial to capture halo formation physics. After passing through the residual network with a filter of size $n_f = 3$, we obtain an output with shape $16 \times 16 \times 16 \times N_{\rm embed}$, where $N_{\rm embed} = 64$ are the features we extract. We then spatially downsample the output with $N_{\rm pool} = 4$ layers to get the final output with shape of $4 \times 4 \times 4 \times N_{\rm embed}$. This output is spatially flattened to finally obtain a matrix of shape $N_{\rm seq-enc} \times N_{\rm embed}$, where $N_{\rm seq-enc} = 64$ and this is used as input to the cross-attention in the decoder module.

4.2. Decoder

The decoder part mostly follows the architecture introduced in Vaswani et al. (2023). The tokenized halo "sentence" (Sec. 3.3) is first embedded to $N_{\rm embed} = 64$ dimensions. We then add linear positional embedding to this input and pass it to a stack of $N_{\rm block} = 4$ multi-head attention modules, each with $N_{\rm head} = 4$ heads. Note that the cross-attention module gets its key and value from the encoder described above whereas the query is generated from the halo tokens. We use standard multi-class cross-entropy loss over $n_{\rm vocab} = 67$ classes to predict the next token number, conditioned on previous tokens and the features from the PM density fields.

When predicting the mock halo catalog from test simulations, we provide the encoder the dark matter density from the PM simulations and a <START> token to the decoder part. We end the prediction once the <END> token is predicted. The predicted token numbers are then used to convert back to the 3D positions within the sub-box and the masses of the halos.

5. Results

5.1. Local performance

We compare the histogram of the predicted mock halo catalogs to the true N-body catalogs on 8 test simulations in the top panel of Fig. 2. We find that the architecture can correctly predict the number of the halos in each sub-box as well as their masses. We show the mean of the histogram obtained from 8 simulations as well as their standard deviation, finding that the stochastic uncertainties due to varying initial conditions are also correctly captured in the model. This tests the local performance of the model.

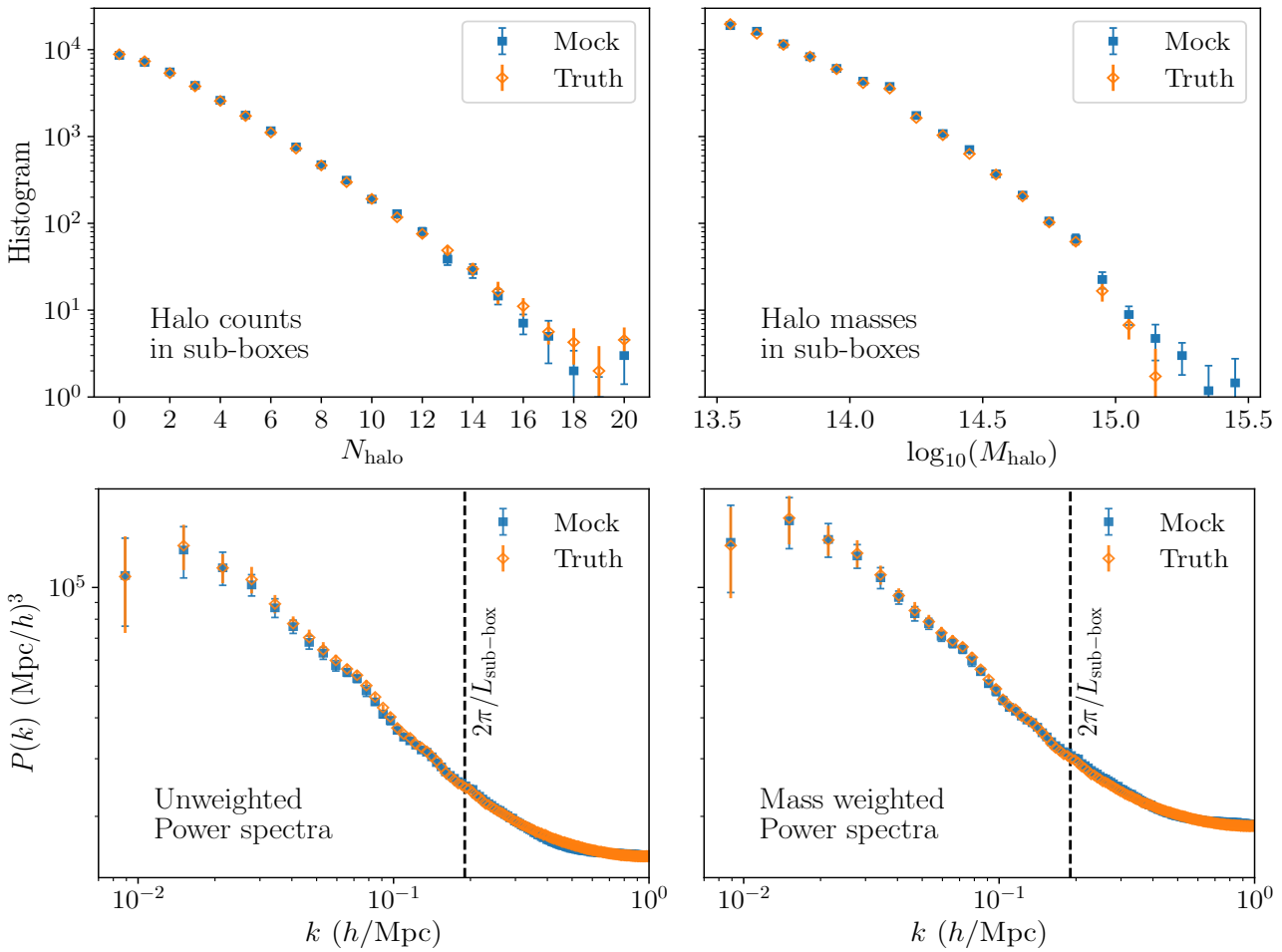

Figure 2. Top row, local performance: Histogram of the predicted and true halo number counts (left) and masses (right) in 8 test simulations. We also show the standard deviation in true and mock halo catalogs, finding that the trained network can accurately capture the mean and their uncertainties. *Bottom row, global performance*: Comparison of the power spectrum of the mock and true halo catalogs, unweighted (left) or weighted by mass (right). We also show the Nyquist frequency of the sub-box ($L_{\rm sub-box} = 31.25\,{\rm Mpc}/h$).

5.2. Global performance

To test the global performance of the model, we take the predicted mock catalogs in each sub-box (of box-length, $L_{\rm sub-box} = 31.25\,{\rm Mpc}/h$) corresponding to one realization and re-create the distribution in the full box of length $L_{\rm full-box} = 1000\,{\rm Mpc}/h$ by stacking the sub-volumes. We then measure the power spectrum of the halos in this full box on a wide range of scales as shown in the bottom left panel of Fig. 2. We find that the resulting mock power spectrum matches to the true power spectrum at 3% level. To test that the model has correctly captured the correlation between halo masses and their spatial distribution, we additionally weight each halo with their predicted mass value using a power-law weighting ($w_i = (M/M_1)^\alpha$, where $M_1 = 10^{14} M_\odot/h$ and $\alpha = 0.66$) and calculate the power spectrum. The comparison is shown in the bottom right panel of Fig. 2 and we again find a 3% agreement between mock and truth catalogs. We show the mean and variance of the prediction and true power spectra calculated from 8 test simulations and we find that the mock catalogs also correctly capture the variance in the power. Note that as the power matches well between true and mock catalogs on both the scales much smaller and larger compared to the size of sub-boxes ($k \sim 2\pi/L_{\rm sub-box} = 0.2\,h/{\rm Mpc}$), the model is capturing the correlation between spatial position and masses correctly.

6. Conclusion

In this study we developed a novel model to generate conditional distribution of weighted discrete point cloud (dark matter halos) when conditioned on a correlated 3D field (dark matter density field). We showed that such a model can accurately reproduce the catalog such that it has correct distribution and statistics. We plan to enhance the network to correctly predict even lower mass halos and include more properties in the inference such as velocity and concentration of the dark matter halo by increasing the size of each "word" in the halo "sentence" (see Sec. 3.3). We also plan to generalize the model for

different cosmologies. The code is publicly available at https://github.com/shivampcosmo/GOTHAM

References

Alam, S., Albareti, F. D., Prieto, C. A., Anders, F., Anderson, S. F., Anderton, T., Andrews, B. H., Armengaud, E., Aubourg, , Bailey, S., Basu, S., Bautista, J. E., Beaton, R. L., Beers, T. C., Bender, C. F., Berlind, A. A., Beutler, F., Bhardwaj, V., Bird, J. C., Bizyaev, D., Blake, C. H., Blanton, M. R., Blomqvist, M., Bochanski, J. J., Bolton, A. S., Bovy, J., Bradley, A. S., Brandt, W. N., Brauer, D. E., Brinkmann, J., Brown, P. J., Brownstein, J. R., Burden, A., Burtin, E., Busca, N. G., Cai, Z., Capozzi, D., Rosell, A. C., Carr, M. A., Carrera, R., Chambers, K. C., Chaplin, W. J., Chen, Y.-C., Chiappini, C., Chojnowski, S. D., Chuang, C.-H., Clerc, N., Comparat, J., Covey, K., Croft, R. A. C., Cuesta, A. J., Cunha, K., Costa, L. N. d., Rio, N. D., Davenport, J. R. A., Dawson, K. S., Lee, N. D., Delubac, T., Deshpande, R., Dhital, S., Dutra-Ferreira, L., Dwelly, T., Ealet, A., Ebelke, G. L., Edmondson, E. M., Eisenstein, D. J., Ellsworth, T., Elsworth, Y., Epstein, C. R., Eracleous, M., Escoffier, S., Esposito, M., Evans, M. L., Fan, X., Fernández-Alvar, E., Feuillet, D., Ak, N. F., Finley, H., Finoguenov, A., Flaherty, K., Fleming, S. W., Font-Ribera, A., Foster, J., Frinchaboy, P. M., Galbraith-Frew, J. G., García, R. A., García-Hernández, D. A., Pérez, A. E. G., Gaulme, P., Ge, J., Génova-Santos, R., Georgakakis, A., Ghezzi, L., Gillespie, B. A., Girardi, L., Goddard, D., Gontcho, S. G. A., Hernández, J. I. G., Grebel, E. K., Green, P. J., Grieb, J. N., Grieves, N., Gunn, J. E., Guo, H., Harding, P., Hasselquist, S., Hawley, S. L., Hayden, M., Hearty, F. R., Hekker, S., Ho, S., Hogg, D. W., Holley-Bockelmann, K., Holtzman, J. A., Honscheid, K., Huber, D., Huehnerhoff, J., Ivans, I. I., Jiang, L., Johnson, J. A., Kinemuchi, K., Kirkby, D., Kitaura, F., Klaene, M. A., Knapp, G. R., Kneib, J.-P., Koenig, X. P., Lam, C. R., Lan, T.-W., Lang, D., Laurent, P., Goff, J.-M. L., Leauthaud, A., Lee, K.-G., Lee, Y. S., Licquia, T. C., Liu, J., Long, D. C., López-Corredoira, M., Lorenzo-Oliveira, D., Lucatello, S., Lundgren, B., Lupton, R. H., III, C. E. M., Mahadevan, S., Maia, M. A. G., Majewski, S. R., Malanushenko, E., Malanushenko, V., Manchado, A., Manera, M., Mao, Q., Maraston, C., Marchwinski, R. C., Margala, D., Martell, S. L., Martig, M., Masters, K. L., Mathur, S., McBride, C. K., McGehee, P. M., McGreer, I. D., McMahon, R. G., Ménard, B., Menzel, M.-L., Merloni, A., Mészáros, S., Miller, A. A., Miralda-Escudé, J., Miyatake, H., Montero-Dorta, A. D., More, S., Morganson, E., Morice-Atkinson, X., Morrison, H. L., Mosser, B., Muna, D., Myers, A. D., Nandra, K., Newman, J. A., Neyrinck, M., Nguyen, D. C., Nichol, R. C., Nidever, D. L., Noterdaeme, P., Nuza, S. E., O'Connell, J. E., O'Connell, R. W., O'Connell, R., Ogando, R. L. C., Olmstead, M. D., Oravetz, A. E., Oravetz, D. J., Osumi, K., Owen, R., Padgett, D. L., Padmanabhan, N., Paegert, M., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Park, C., Pattarakijwanich, P., Pellejero-Ibanez, M., Pepper, J., Percival, W. J., Pérez-Fournon, I., Pe´rez-Ra'fols, I., Petitjean, P., Pieri, M. M., Pinsonneault, M. H., Mello, G. F. P. d., Prada, F., Prakash, A., Price-Whelan, A. M., Protopapas, P., Raddick, M. J., Rahman, M., Reid, B. A., Rich, J., Rix, H.-W., Robin, A. C., Rockosi, C. M., Rodrigues, T. S., Rodríguez-Torres, S., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Ruan, J. J., Rubiño-Martín, J. A., Rykoff, E. S., Salazar-Albornoz, S., Salvato, M., Samushia, L., Sánchez, A. G., Santiago, B., Sayres, C., Schiavon, R. P., Schlegel, D. J., Schmidt, S. J., Schneider, D. P., Schultheis, M., Schwope, A. D., Scóccola, C. G., Scott, C., Sellgren, K., Seo, H.-J., Serenelli, A., Shane, N., Shen, Y., Shetrone, M., Shu, Y., Aguirre, V. S., Sivarani, T., Skrutskie, M. F., Slosar, A., Smith, V. V., Sobreira, F., Souto, D., Stassun, K. G., Steinmetz, M., Stello, D., Strauss, M. A., Streblyanska, A., Suzuki, N., Swanson, M. E. C., Tan, J. C., Tayar, J., Terrien, R. C., Thakar, A. R., Thomas, D., Thomas, N., Thompson, B. A., Tinker, J. L., Tojeiro, R., Troup, N. W., Vargas-Magaña, M., Vazquez, J. A., Verde, L., Viel, M., Vogt, N. P., Wake, D. A., Wang, J., Weaver, B. A., Weinberg, D. H., Weiner, B. J., White, M., Wilson, J. C., Wisniewski, J. P., Wood-Vasey, W. M., Ye'che, C., York, D. G., Zakamska, N. L., Zamora, O., Zasowski, G., Zehavi, I., Zhao, G.-B., Zheng, Z., Zhou (), X., Zhou (), Z., Zou (), H., and Zhu, G. The eleventh and twelfth data releases of the sloan digital sky survey: Final data from sdss-iii. *The Astrophysical Journal Supplement Series*, 219(1):12, July 2015. ISSN 1538-4365. doi: 10.1088/0067-0049/219/1/12. URL http://dx.doi.org/10.1088/0067-0049/219/1/12.

Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. The rockstar phase-space temporal halo finder and the velocity offsets of cluster cores. *The Astrophysical Journal*, 762(2):109, December 2012. ISSN 1538-4357. doi: 10.1088/0004-637x/762/2/109. URL http://dx.doi.org/10.1088/0004-637X/762/2/109.

Feng, Y., Chu, M.-Y., Seljak, U., and McDonald, P. FASTPM: a new scheme for fast simulations of dark matter and haloes. , 463(3):2273–2286, December 2016. doi: 10.1093/mnras/stw2123.

Horowitz, B., Hahn, C., Lanusse, F., Modi, C., and Ferraro, S. Differentiable stochastic halo occupation distribution. , 529(3):2473–2482, April 2024. doi: 10.1093/mnras/stae350.

Jamieson, D., Li, Y., Alves de Oliveira, R., Villaescusa-Navarro, F., Ho, S., and Spergel, D. N. Field Level Neural

Network Emulator for Cosmological N-body Simulations, June 2022.

Kodi Ramanah, D., Charnock, T., and Lavaux, G. Painting halos from cosmic density fields of dark matter with physically motivated neural networks. , 100(4):043515, August 2019. doi: 10.1103/PhysRevD.100.043515.

Li, Y., Modi, C., Jamieson, D., Zhang, Y., Lu, L., Feng, Y., Lanusse, F., and Greengard, L. Differentiable cosmological simulation with adjoint method. *arXiv preprint arXiv:2211.09815*, 2022.

Modi, C., Lanusse, F., and Seljak, U. FlowPM: Distributed TensorFlow implementation of the FastPM cosmological N-body solver. *Astronomy and Computing*, 37:100505, October 2021. doi: 10.1016/j.ascom.2021.100505.

Pandey, S., Modi, C., Wandelt, B., and Lavaux, G. CHARM: Creating halos with auto-regressive multi-stage networks. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL https://openreview.net/forum?id=dz3O7M1QzA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Villaescusa-Navarro, F., Hahn, C., Massara, E., Banerjee, A., Delgado, A. M., Ramanah, D. K., Charnock, T., Giusarma, E., Li, Y., Allys, E., Brochard, A., Uhlemann, C., Chiang, C.-T., He, S., Pisani, A., Obuljen, A., Feng, Y., Castorina, E., Contardo, G., Kreisch, C. D., Nicola, A., Alsing, J., Scoccimarro, R., Verde, L., Viel, M., Ho, S., Mallat, S., Wandelt, B., and Spergel, D. N. The quijote simulations. *The Astrophysical Journal Supplement Series*, 250(1):2, August 2020. ISSN 1538-4365. doi: 10.3847/1538-4365/ab9d82. URL http://dx.doi.org/10.3847/1538-4365/ab9d82.

Yan, X., Lin, L., Mitra, N. J., Lischinski, D., Cohen-Or, D., and Huang, H. Shapeformer: Transformer-based shape completion via sparse representation, 2022.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling, 2022.