Learning to Predict the Cosmological Structure Formation

Siyu He\textsuperscript{a,b,c}, Yin Li\textsuperscript{d,e,f}, Yu Feng\textsuperscript{d,h}, Shirley Ho\textsuperscript{c,e,d,a,b}, Siamak Ravanbakhsh\textsuperscript{g}, Wei Chen\textsuperscript{b}, and Barnabáš Póczos\textsuperscript{h}

\textsuperscript{a}Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, USA; \textsuperscript{b}McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15213, USA; \textsuperscript{c}Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, NY 10010, USA; \textsuperscript{d}Berkeley Center for Cosmological Physics, University of California, Berkeley, CA 94720, USA; \textsuperscript{e}Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; \textsuperscript{f}Kavli Institute for the Physics and Mathematics of the Universe (WPI), UTIAS, The University of Tokyo, Chiba 277-8583, Japan; \textsuperscript{g}Computer Science Department, University of British Columbia, Vancouver, BC V6T1Z4, Canada; \textsuperscript{h}Machine Learning Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh 15213, USA

Matter evolved under influence of gravity from minuscule density fluctuations. Non-perturbative structure formed hierarchically over all scales, and developed non-Gaussian features in the Universe, known as the Cosmic Web (1). To fully understand the structure formation of the Universe is one of the holy grails of modern astrophysics. Astrophysicists survey large volumes of the Universe (2–9) and employ a large ensemble of computer simulations to compare with the observed data in order to extract the full information of our own Universe. However, to evolve trillions of galaxies over billions of years even with the simplest physics is a daunting task. We build a deep neural network, the Deep Density Displacement Model (hereafter D$^3$M), to predict the non-linear structure formation of the Universe from simple linear perturbation theory. Our extensive analysis demonstrates that D$^3$M outperforms the second order perturbation theory (hereafter 2LPT), the commonly used fast approximate simulation method, in point-wise comparison, 2-point correlation, and 3-point correlation. We also show that D$^3$M is able to accurately extrapolate far beyond its training data, and predict structure formation for significantly different cosmological parameters. Our study proves, for the first time, that deep learning is a practical and accurate alternative to approximate simulations of the gravitational structure formation of the Universe.

\textbf{cosmology | deep learning | simulation |}

At its core, modeling structure formation of the Universe is a computationally challenging task; it involves evolving trillions of particles with the correct physical model over a large volume over billions of years. To simplify this task, we either simulate a large volume with simpler physics or a smaller volume with more complex physics. In order to produce the cosmic web in large volume, we select gravity, the most important component of the theory, to simulate at large scales. A gravity-only $N$-body simulation is the most popular; and effective numerical method to predict the full 6D phase space distribution of a large number of massive particles whose positions and velocities evolve over time in the Universe (10). Nonetheless, $N$-body simulations are relatively computationally expensive, thus making the comparison of the $N$-body simulated large-scale structure (of different underlying cosmological parameters) with the observed Universe a challenging task. Here, we propose to learn a deep model that predicts the structure formation as an alternative to expensive $N$-body simulations.

Deep learning (11) is a fast growing branch of machine learning where recent advances have lead to models that reach and sometimes exceed human performance across diverse areas, from analysis and synthesis of images (12–14), sound (15, 16), text (17, 18) and videos (19, 20) to complex control and planning tasks as they appear in robotics and game-play (21–23). This new paradigm is also significantly impacting a variety of domains in the sciences, from biology (24, 25) to chemistry (26, 27) and physics (28, 29). In particular, in astronomy and cosmology, a growing number of recent studies are using deep learning for a variety of tasks, ranging from analysis of cosmic microwave background (30–32), large-scale structure (33, 34), and gravitational lensing effects (35, 36) to classification of different light sources (37–39).

The ability of these models to learn complex functions has motivated many to use them to understand the physics of interacting objects leveraging image, video and relational data (40–50). However, modeling the dynamics of billions of particles in $N$-body simulations poses a distinct challenge.

In this paper we show that a variation on the architecture of a well-known deep learning model (51), can efficiently transform the first order approximations to the displacement field and approximate the exact solution, thereby producing accurate estimates of the large-scale structure. Our objective is to show that this approach is a serious contender in replacing expensive cosmological simulations, and to this end we provide an extensive analysis of the results in the following section.

The outcome of a typical $N$-body simulation depends on both the initial conditions and cosmological parameters. We report a rather surprising observation that our Deep Density Displacement Model (D$^3$M), trained using a “single” set of

\textbf{Fig. 1. (left)} the displacement vector-field produced by D$^3$M and; (right) the resulting density field. The vectors in the left figure are uniformly scaled down for better visualization.
cosmological parameters can generalize to new set of significantly different parameters, minimizing the need for training data on diverse range of parameters in practical settings.

**Setup**

We build a deep neural network, Deep Density Displacement Model (D^3M) with similar input and output variables of an N-body simulation. The input to our D^3M is the displacement field from the so called Zel’dovich Approximation (52) (hereafter ZA). A displacement vector is the difference in particle position between target redshift \( z = 0 \), i.e., the present time, and it’s Lagrangian position on a uniform grid. ZA evolves the particles on linear trajectories along their initial displacements. It is accurate when the displacement is small, therefore is frequently used to construct the initial conditions of N-body simulations (53).

As the ground truth, the target displacement field is produced using FastPM (54), a recent approximate N-body simulation scheme that is based on a particle-mesh (PM) solver. FastPM quickly approaches a full N-body simulation with high accuracy and provides a viable alternative for exact N-body simulations (53).

A significantly faster approximation to N-body simulations, is produced by second-order Lagrangian perturbation theory (hereafter 2LPT), which bends each particle’s trajectory with a quadratic correction (55). Many cosmological analyses use 2LPT to generate a large number of cosmological simulations for comparison of astronomical dataset against the physical model (56, 57) or to compute the covariance of the dataset (58–60). We regard 2LPT as an effective way to generate a relatively correct description of the large-scale structure and therefore, this method serves as a natural competitor to D^3M.

We generate 10,000 pairs of ZA approximations as input and accurate FastPM approximations as target. We use simulations of \( 32^3 \) N-body particles in a volume of 128 Mpc/h (600 million light years) on each side. The particles have a mean separation of 4 Mpc/h per dimension.

**Displacement versus Density Field.** An important choice in our approach is training with displacement field rather than density field. Displacement field \( \Psi \) and density field \( \rho \) are two ways of describing the same distribution of particles, and they are related by the law of mass conservation. The mean value of the density field is \( \bar{\rho} \), relative to which the density fluctuation can be described by the over-density field, \( \delta = \rho / \bar{\rho} - 1 \).

When the displacement field is small and curl-less, the choice of density vs displacement field for the output of the model is irrelevant, as there is a bijective mapping between these two representations. However as the displacement grow into the non-linear regime of structure formation, different displacement fields can produce identical density fields. (e.g. 61) Therefore, providing the model with the target displacement field during the training is significantly more informative as it does not have the ambiguity associated with the density field. Indeed we could not produce comparable results when using the density field as our input and target.

**Results and Analysis**

Figure 1 shows the displacement vector field as predicted by D^3M (left) and the point-cloud representation of the predicted structure formation (right). It is possible to identify structures such as clusters, filaments and voids in this point-cloud representation. In the following we quantify the accuracy of D^3M and compare its performance to 2LPT.

**Point-Wise Comparison.** Let the rank four tensors \( \Psi \in \mathbb{R}^{d \times d \times d \times d} \) denote the displacement field, where \( d \) is the number of spatial resolution elements per side (\( d = 32 \)). A natural error measure is the norm of the error \( e = \hat{\Psi} - \Psi_{t} \), where \( \Psi_{t} \) is the true displacement field (FastPM) and \( \hat{\Psi} \) is the prediction (ZA, 2LPT or D^3M). Figure 2 compares this error for different approximations in a single simulation. We observe that D^3M predictions are very close to the ground truth, with a maximum error of 4.46 Mpc/h over all 1000 simulations. For 2LPT this number is significantly higher at 15.20 Mpc/h.

Since a few large error vectors can dominate the L2 norm, we also measure the relative displacement error \( |\hat{\Psi} - \Psi_{t}|/|\Psi_{t}| \). Using this measure we observe the same pattern: D^3M is almost 3 times more accurate than 2LPT. In particular, D^3M is able to achieve an average of 2.8% in relative displacement error per voxel on test data, while 2LPT can only achieve 9.3%.

**2-Point Correlation Comparison.** As suggested by Figure 2 the more dense regions seem to have a higher error for all methods – that is the error is larger when the structure formation is more non-linear for both D^3M and 2LPT. This correlation of error with scale is captured in 2-point and 3-point correlation analysis.

Cosmologists often employ compressed summary statistics of the density field in their studies. The most widely used of these statistics is 2-point correlation function (2PCF) \( \xi(r) \).
(b) 3-point analysis

Fig. 3. (a) From top to bottom: (top) density and displacement power-spectrum of D^3M, 2LPT, and FastPM; (middle) transfer function – i.e., the square root of the ratio of the predicted power-spectrum and the ground truth, and; (bottom) 1−r^2 where r is the correlation coefficient between the predicted fields and the true fields. Results are the averaged values from 1000 test simulations.
(b) 3-point correlation function measured as binned multipole coefficients $\zeta(r_1,r_2)$. Results are averaged values from 10 test simulations. The top panel shows $\zeta(r_1,r_2)$ for D^3M, FastPM, and 2LPT. In order to amplify the finer features, we have weighted $\zeta(r_1,r_2)$ by $r_1^2r_2^2/(100 \text{ Mpc}/h)^4$ in each spherical shell. On the diagonal elements, the 3PCF is expected to be dominated by squeezed triangles for which perturbation theory is not valid, so we have not computed the diagonal elements (thus blank in the plots here). The bottom panel shows the relative residual of D^3M and 2LPT models compared to FastPM.

and its Fourier transform, the power spectrum $P(k)$:

$$\xi(|r|) = \langle \delta_A(r) \delta_B(r + r) \rangle,$$

where the ensemble average $\langle \rangle$ is taken over all possible realizations of the Universe. Our Universe is observed to be homogeneous and isotropic on large scales, i.e. without any special location or direction. This allows one to drop the dependences on $r'$ and on the direction of $r$, leaving only the amplitude $|r|$ in the final definition of $\xi(r)$. In the second equation, $P(k)$ is simply the Fourier transform of $\xi(r)$, and captures the dispersion of the plane wave amplitudes at different scales in the Fourier space. $k$ is the 3D wavevector of the plane wave, and its amplitude $k$ (the wavenumber) is related to the wavelength $\lambda$ by $k = 2\pi/\lambda$. Due to isotropy of the Universe, we drop the vector form of $r$ and $k$ when we discuss $\xi(r)$ and $P(k)$ from here onwards.

We focus on the Fourier-space representation of the 2-point correlation. We employ the transfer function $T(k)$ and the correlation coefficient $r(k)$ as metrics to quantify the model performance against the ground truth (FastPM) on the 2-point correlation. We define the transfer function $T(k)$ as the square root of the ratio of two power spectra,

$$T(k) = \sqrt{\frac{P_{\text{pred}}(k)}{P_{\text{true}}(k)}}$$

where $P_{\text{pred}}(k)$ is the density or displacement power spectrum from 2LPT or D^3M predictions, and $P_{\text{true}}(k)$ is the equivalent from the ground truth (FastPM). The correlation coefficient $r(k)$ is a form of normalized cross power spectrum,

$$r(k) = \frac{P_{\text{pred} \times \text{true}}(k)}{\sqrt{P_{\text{pred}}(k)P_{\text{true}}(k)}}$$

where $P_{\text{pred} \times \text{true}}(k)$ is the cross power spectrum between 2LPT or D^3M predictions and the ground truth (FastPM) simulation. Transfer function captures the discrepancy in amplitudes,

while the correlation coefficient can indicate the discrepancy in phases as functions of scales. For a perfectly accurate prediction, $T(k)$ and $r(k)$ are both 1. In particular, $1−r^2$ describes stochasticity, the fraction of the variance in the prediction that cannot be explained by the true model.

Figures 3(a) shows the average power spectrum, transfer function $T(k)$ and stochasticity $1−r^2(k)$ of the displacement field and the density field over 1000 simulations. The power spectrum of density from 2LPT prediction is 3% smaller than that of FastPM on large scales ($k \approx 0.05 \text{ h/Mpc}$). This is expected since 2LPT performs accurately only on very large scales ($k < 0.01 \text{ h/Mpc}$). The displacement transfer function of 2LPT increases above 1 at $k \approx 0.35 \text{ h/Mpc}$ and then drops sharply. The increase of 2LPT displacement transfer function is because 2LPT over-estimates the displacement power at small scales (see, e.g. 62). There is a sharp drop of power near the voxel scale, as we smooth over voxel scales in our predictions, which automatically erases power at scales smaller than the voxel size. Now we turn to D^3M predictions: both the density and displacement transfer functions of the D^3M differ from 1 by a mere 0.4% at scale $k \lesssim 0.4 \text{ h/Mpc}$, and this discrepancy only increases to 3% and 6% for density field and displacement field respectively, as $k$ increases to the Nyquist frequency around 0.7 h/Mpc. The stochasticity hovers at approximately $10^{-3}$ and $10^{-2}$ for most scales. In other words, the correlation coefficient of either density or displacement field is over 90% between the D^3M predictions and FastPM simulations all the way down to small scales of $k = 0.7 \text{ h/Mpc}$. The transfer function and correlation coefficient of the D^3M suggest its nearly perfect predicting power from large to intermediate scales — the regimes where traditional perturbation theories like the 2LPT work reasonably well — and even beyond. Only on fairly small scales that the D^3M starts to falter, which is not surprising as the deeply nonlinear evolution at these scales is also the most difficult to solve accurately in simulations and even impossible for any analytical theories to model.
3-Point Correlation Comparison. The 3-point correlation function (3PCF) measures the correlation among the field of interest at 3 locations in the configuration space, which is equivalently defined as bispectrum in Fourier space. Here we concentrate on the 3PCF due to computational convenience:

$$\zeta(r_1, r_2, \theta) = \langle \delta(x) \delta(x + r_1) \delta(x + r_2) \rangle. \tag{4}$$

where $r_1 = |r_1|$ and $r_2 = |r_2|$. Translation invariance guarantees that $\zeta$ is independent of $x$. Rotational symmetry further eliminates all direction dependence but that on $\theta$, the angle between $r_1$ and $r_2$. The multipole moments of $\zeta(r_1, r_2, \theta)$, $\zeta_\ell(r_1, r_2) = (2\ell + 1) \int d\Omega P_\ell(\cos \theta) \zeta(r_1, r_2, \theta)$ where $P_\ell(\cos \theta)$ is the Legendre polynomial of degree $\ell$, can be efficiently estimated with pair counting (63). While the input variables (ZA) does not contain significant correlations beyond the second order (power spectrum level), we expect our $D^3 M$ to be capable of generating densities with a 3PCF that mimics that of ground truth.

We compare the 3PCF calculated from FastPM, 2LPT and $D^3 M$ by analyzing the 3PCF through the multipole moments $\zeta_\ell(r_1, r_2)$. Figure 3(c) shows the binned multipole coefficients $\zeta_\ell$ of the 3PCF. We used 10 radial bins with $\Delta r = 5 \text{ Mpc$/h$}$. The upper panel in Figure 3(c) shows the average $\zeta_\ell(r_1, r_2)$ over 10 simulations. $\ell = 2$ contributes the largest amplitude contribution to the 3PCF, especially for triangles that are not close to squeezed configuration (two sides of very similar amplitude). For $\ell \geq 3$, the amplitude of 3PCF near the diagonal dominates. For higher multipoles, $\zeta_\ell \sim 1/(r_1^2 r_2^2) + 1/((r_1 r_2)^2) + 1/r_1^4 + 1/r_2^4$, thus the amplitude is dominated by a small population of squeezed triangles where two sides are equal and the third side near 0 (63). To compare the performance of $D^3 M$ and 2LPT predictions, we show the relative residual of $D^3 M$ prediction and 2LPT with regard to FastPM in the lower panel in Figure 3(c). The difference between FastPM and $D^3 M$ is notably smaller than that between FastPM and 2LPT at all angular scales. To quantify our comparison further, we calculate the relative 3PCF residual defined as

$$3\text{PCF relative residual} = \frac{1}{9 \times N_r} \sum_{\ell=2}^8 \sum_{r_1, r_2} |\zeta_\ell(r_1, r_2)_{\text{pred}} - \zeta_\ell(r_1, r_2)_{\text{true}}| \bigg/ |\zeta_\ell(r_1, r_2)_{\text{true}}| \tag{5}$$

where $N_r$ is the number of $(r_1, r_2)$ bins. The mean relative 3PCF residual of $D^3 M$ prediction and 2LPT compared to FastPM is 0.79% and 7.82% respectively. The $D^3 M$ performance on 3PCF is also an order of magnitude better than 2LPT, which indicates the $D^3 M$ is far better at capturing the non-Gaussian structure formation.

Generalizing to New Cosmological Parameters

So far, we trained our model using a “single” choice of cosmological parameters $A_s = 2.142 \times 10^{-9}$ (hereafter $A_0 = 2.142 \times 10^{-9}$) and $\Omega_m = 0.3089 \ (64)$. $A_s$ is the primordial amplitude of the scalar perturbation from cosmic inflation, and $\Omega_m$ is the fraction of the total energy density that is matter at the present time, and we will call it matter density parameter for short. The true exact value of these parameters are unknown and different choices of these parameters change the large-scale structure of the Universe; see Figure 4.

Here, we report an interesting observation: $D^3 M$ trained on a single set of parameters, can predict the structure formation for widely different choices of $A_s$ and $\Omega_m$. These results give important insight about structure formation that we will discuss shortly. From the computational point of view this indicates the possibility of producing simulations for a diverse range of parameters, with minimal training data.

Varying Primordial Amplitude of Scalar Perturbations $A_s$. After training the $D^3 M$ using $A_s = A_0$, we change our $A_s$ in the input of our test set by nearly one order of magnitude: $A_s = 1.8A_0$ and $A_s = 0.2A_0$. Again, we use 1000 simulations for analysis of each test case. The average relative displacement error of $D^3 M$ remains less than 4% per voxel (compared to < 3% when train and test data have the same parameters). This is still well below the error for 2LPT, which is 15.5% and 6.3% for respectively smaller and larger values of $A_s$.

Figure 5(a) shows the average power spectrum, transfer function and correlation coefficient for both $D^3 M$ and 2LPT. Overall, $D^3 M$ performs much better than 2LPT for $A_s = 1.8A_0$. For small $A_s = 0.2A_0$, 2LPT does a better job at predicting the density transfer function and correlation coefficient at the largest scales, otherwise $D^3 M$ predictions outperforms 2LPT at scales larger than $k = 0.08 \text{ h/Mpc}$. We observe a similar trend with 3PCF analysis: the 3PCF of $D^3 M$ predictions is notably better than 2LPT ones for larger $A_s$, compared to smaller $A_s$ where it is only slightly better. These
results confirm our expectation that increasing $A_s$ increases the non-linearity of the structure formation process. While 2LPT can predict fairly well in linear regimes, its performance deteriorates as compared to D$^3$M with increased non-linearity. It is interesting to note that D$^3$M prediction maintains its advantage despite being trained on data from more linear regimes.

**Varying matter density parameter $\Omega_m$.** We repeat the same experiments, this time changing $\Omega_m$ to 0.5 and 0.1, where the model is trained on $\Omega_m = 0.3089$, which is quite far from both of the test sets. The relative residual displacement error of the D$^3$M (2LPT) averaged over 1000 simulations is 3.8% (15.2%) and 2.5% (4.3%) respectively. Figures 5(c)(d) show the two-point statistics for different values of $\Omega_m$. For $\Omega_m = 0.5$, the results show that the D$^3$M outperforms 2LPT at all scales, while for smaller $\Omega_m = 0.1$, D$^3$M outperforms 2LPT on smaller scales ($k > 0.1 \ h/Mpc$). As for the 3PCF of different values of $\Omega_m$, the mean relative 3PCF residual of deep learning model (2LPT) for $\Omega_m = 0.5$ and $\Omega_m = 0.1$ are 1.7% (7.6%) and 1.2% (1.7%) respectively. The D$^3$M prediction performs better at $\Omega_m = 0.5$ than $\Omega_m = 0.1$. This is also because the universe is much more non-linear at $\Omega_m = 0.5$ than $\Omega_m = 0.1$. There is much more non-linearity to learn that are not modeled with 2LPT.

**Conclusions**

D$^3$M can predict accurately the large-scale structure of Universe represented by our input simulations (FastPM) at all scales as seen in Table 1. Furthermore, D$^3$M also performs better than our benchmark model (2LPT) at most scales, in terms of point-wise comparison, 2- and 3-point correlations. Finally, to our surprise, our model also generalizes well to test simulations with cosmological parameters ($A_s$ and $\Omega_m$) different from the training set (Table 1).

This indicates the deep learning model has the potential for deployment for producing a broader range of simulations beyond the parameter space covered by the training data. All results demonstrate that the D$^3$M successfully learns the nonlinear mapping from first order perturbation theory to FastPM simulation beyond what higher order perturbation theories can achieve.

We have no reason to believe that replacing FastPM with exact N-body simulations will deteriorate the performance of our method. Moreover, since the complexity of D$^3$M is linear in the number of voxels, we expect to be able to further improve our encouraging results when moving to higher resolutions.

Laws of physics often respects large symmetry groups, which lead to nearly unique sets of equations (65). Similarly, after factoring out equivalence, one would expect that few distinct ML models will describe a physical law at the data level. We look forward to developing further in these directions towards interpretability of our deep models.

**Materials and Methods**

**Dataset.** The full simulation data consists of 10,000 ZA-FastPM simulation boxes as input-output pairs, with an effective volume of 20 (Gpc/h)^3 (190 × 10^9 ly^3), comparable to the volume of a large spectroscopic sky survey like DESI or EUCLID. We split the full simulation data set into 80%, 10%
and 10% for training, validation and test, respectively. We also generated 1000 simulations for 2LPT for each set of tested cosmological parameters.

**Model and Training.** The D³M adopts the U-Net architecture (51) with 15 convolution-deconvolution layers and approximately $8.4 \times 10^6$ trainable parameters. Our D³M generalizes the standard U-Net architecture to work with three-dimensional data.

**Padding and Periodic Boundary.** It is common to use constant or reflective padding in deep models for image processing. However, these approaches are not suitable for our setting. The physical model we are learning is constructed on a spatial volume with a periodic boundary condition. This is sometimes also referred to as a torus geometry, where the boundaries of the simulation box are topologically connected – that is $x_{i+L} = x_i$, where $i$ is the index of the spatial location, and $L$ is the periodicity (size of box). Constant or reflective padding strategies break the connection between the physically nearby points separated across the box, which not only loses information but also introduces noise during the convolution, further aggravated with increased number of layers.

We find that the periodic padding strategy significantly improves the performance and expedites the convergence of our model, comparing to the same network using a constant padding strategy. This is not surprising, as one expects it is easier to train a model that can explain the data than to train a model that does not.

**Loss Function and Training.** We train the D³M to minimize the mean square error on particle displacements

$$
\mathcal{L} = \frac{1}{N} \sum_i (\hat{\psi}_{t,i} - \psi_{t,i})^2,
$$

[6] where $i$ labels the particles and the $N$ is the total number of particles. This loss function is proportional to the integrated squared error, that in the Fourier space can be rewritten using the Parseval’s theorem as

$$
\int (\hat{\psi} - \psi_t)^2 d^3q = \int |\hat{\psi} - \psi_t|^2 d^3k =
$$

$$
\int d^3k \left( |\psi_t|^2 (1 - T)^2 + 2 |\psi_t| |\psi_t| (1 - r) \right),
$$

[7] where $q$ is the Lagrangian space position, and $k$ is its corresponding wavevector. $T$ is the transfer function defined in Eq. 2, and $r$ is the correlation coefficient defined in Eq. 3. They characterize the similarity between the predicted and the true fields, in amplitude and phase respectively. Eq. 7 shows that our simple loss function jointly captures both of these measures; as $T$ and $r$ approach 1, the loss function approaches 0.

In the training phase, we employ the Adam Optimizer (66) with a learning rate of 0.0001, and first and second moment exponential decay rate of 0.9 and 0.999, respectively. We use the L2 regularization with regularization coefficient 0.0001.

**Details of the D³M Architecture.** The contracting path follows the typical architecture of a convolution network. It consists of two blocks, each of which consists of two successive convolutions of stride 1 and a down-sampling convolution with stride 2. The convolution layers use $3 \times 3 \times 3$ filters with a periodic padding of size 1 (see Padding and Periodic Boundary) on both sides of each dimension. The padding ensures that the first two layers keep the shape of each channel. The following down-sampling layer reduces the shape of the layer by half on each spatial dimension. Notice that at each of the two down-sampling steps, we double the number of feature channels. At the bottom of the D³M, another two successive convolutions with stride 1 and the same periodic padding as above are applied. The expansive path of our D³M is an inverted version of the contracting path of the network. (It includes two repeated applications of the expansion block, each of which consists of one upsampling transposed convolution with stride 1/2 and two successive convolution of stride 1. The transposed convolution and the convolution are constructed with $3 \times 3 \times 3$ filters.)

We take special care in the padding and cropping procedure to preserve the shifting and rotation symmetry in the upsampling layer in expansive path. We apply before the transposed convolution a periodic padding of length 1 on the right, down and back sides of the box (padding=(0,1,0,1,0,1) in pytorch), and after the transposed convolution, we discard one column on the left, up and front sides of the box and two columns on the right, down and back sides (crop=(1,2,1,2,1,2)).

A special feature of the D³M is the concatenation procedure, where the up-sampling layer halves the feature channels and then concatenates them with the corresponding feature channels on the contracting path, doubling the number of feature channels.

The expansive building block then follows a $1 \times 1 \times 1$ convolution without padding, which converts the 64 features to the the final 3-D displacement field. All convolutions in the network except the last one are followed by a rectified linear unit (ReLU) activation and batch normalization (BN).

**ACKNOWLEDGMENTS.** We thank Angus Beane, Peter Braam, Gabriella Contardo, David Hogg, Laurence Levasseur, Pascal Ripoche and David Spergel for useful suggestions and comments, Angus Beane for comments on the paper, Nick Carriero for help on CCA computing clusters. The FastPM simulations are generated on the computer cluster Edison at the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. The training of neural network model is performed on the CCA computing facility and CMU AutonLab computing facility. The open source software toolkit nbodykit (67) is employed for the clustering analysis. YL acknowledges support from the Berkeley Center for Cosmological Physics and the Kavli IPMU established by World Premier International Research Center Initiative (WPI) of the MEXT, Japan. S.Ho thanks NASA for their support in grant number: NASA grant 15-WFIRST15-0008 and NASA ROSES grant 12-EUCLID12-0004.

**References**

1. Bond JR, Kolman L, Pogosyan D (1996) How filaments of galaxies are woven into the cosmic web. Nature 380:603–606.

2. Colless, M., , et al. (2001) The 2dF Galaxy Redshift Survey: Spectra and redshifts. Mon. Not. Roy. Astron. Soc. 328:1039.

3. Eisenstein, D.J., , et al. (2011) SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. Astron. J. 142:72.

4. Jones, H.D., , et al. (2009) The 6dF Galaxy Survey: Final Redshift Release (DR3) and Southern Large-Scale Structures. Mon. Not. Roy. Astron. Soc. 399:683.

5. Limke, J., , et al. (2010) Galaxy and mass assembly (gama): end of survey report and data release 2. Monthly Notices of the Royal Astronomical Society 452(2):2087.

6. Scodeggio, M., , et al. (2016) The VIMOS Public Extragalactic Redshift Survey (VIPERS). Full spectroscopic data and auxiliary information release (PDR-2). ArXiv e-prints.
