A deep learning approach to infer galaxy cluster masses from Planck Compton-\( y \) parameter maps

Galaxy clusters are useful laboratories to investigate the evolution of the Universe, and accurate measurement of their total masses allows us to constrain important cosmological parameters. However, estimating mass from observations that use different methods and spectral bands introduces various systematic errors. Here we evaluate the use of a convolutional neural network (CNN) to reliably and accurately infer the masses of galaxy clusters from the Compton-\( y \) parameter maps provided by the Planck satellite. The CNN is trained with mock images generated from hydrodynamic simulations of galaxy clusters, with Planck’s observational limitations taken into account. We observe that the CNN approach is not subject to the usual observational assumptions, and therefore is not affected by the same biases. By applying the trained CNNs to the real Planck maps, we find cluster masses compatible with Planck measurements within a 15% bias. Finally, we show that this mass bias can be explained by the well-known hydrostatic equilibrium assumption in Planck masses, and the different parameters in the integrated Compton-\( y \) signal and the mass scaling laws. This work highlights that CNNs, supported by hydrodynamic simulations, are a promising and independent tool for estimating cluster masses with high accuracy, which can be extended to other surveys as well as to observations in other bands.
among clusters’ physical quantities strictly related to the mass of the object under the self-similarity assumption. Nevertheless, in all the listed methods to infer the mass, we have to face with the problem of mass bias—the derived mass is systematically different from the real cluster mass owing to the assumed approximations in each approach. The presence of the mass bias has an impact on the inference of cosmological parameters and, in particular, the cosmic matter density \( \Omega_m \) and the normalization of the matter power spectrum \( \sigma_8 \). Currently, the value of the bias, \( b = \Delta M / M \) where \( \Delta M \) is the mass difference between the estimated mass and the real mass, needed to reconcile cosmic microwave background (CMB) constraints with thermal SZ (tSZ) cluster counts is \( (1 - b) = 0.58 \pm 0.04 \) (ref. 7). Such a large value for the bias is not consistent with almost any of the estimates based on X-ray, SZ and weak gravitational lensing observations; on average, all are around \( (1 - b) = 0.80 \pm 0.08 \) (ref. 7). Large-scale hydrodynamic simulations play an important role regarding the determination and calibration of the mass bias. However, results from several datasets with different physical processes are included; particle resolutions and number of objects are inconsistent with the Planck bias requirement (see, for example, ref. 8 for a recent review).

Machine learning (ML) (for example, ref. 9) algorithms allow us to analyse data and make predictions without assuming any previous known behaviour, that is, data-driven science. The rapid growth in data complexity in astronomy encourages the development of these techniques, where a wide variety of ML models have been studied so far (for example, see refs. 10,11). Deep learning (for example, refs. 12,13) is an ML tool that makes use of multilayer perceptrons, also known as feedforward neural networks, with numerous ‘deep’ hidden layers. In particular, convolutional neural networks (CNNs)13 are a type of neural network that use convolutions for processing data that show a known grid-like topology. Moreover, recent studies have shown that deep learning methods can be used for inferring galaxy cluster masses directly from mock X-ray images14, mock SZ images15; CMB cluster lensing16; a combination of X-ray, SZ and optical mock images17; and from galaxy member dynamics18,19. These techniques do not rely on any assumption on the dynamics or the spherical symmetry of the CMB, but rather on the quality of the dataset. These theoretical studies further suggested a bias-free estimation of the cluster mass.

In this Article, we make another step by applying these theoretical works to predict the masses of real galaxy clusters observed through the SZ effect. In particular, we apply the trained CNNs to the publicly available second Planck catalogue of SZ sources, that is, the PSZ2 catalogue20, to derive the cluster masses. To do that, we analyse a sample of 6,765 clusters from The Three Hundred (The300)20 simulation with the same redshift and mass ranges as the PSZ2 clusters. In particular, we train our CNNs using simulated tSZ images, aiming at predicting the masses from real Compton-Y parameter Planck images. We also compare our results with masses estimated by Planck and determine that the masses inferred with our CNN are overall in agreement but show some discrepancies that might be attributed to the general assumptions used in Planck, such as the hydrostatic equilibrium and the integrated Compton-y signal and the mass (\( Y - M \)) scaling relation.

To this end, the mock SZ maps have the same noise and beam convolution as the corresponding Planck observations. A summary of the characteristics of each dataset used in this work is presented in Table 1. The interested reader can find further information and technical details regarding the generation of these mock observations in Methods. For information concerning the CNN model, training and validation procedure, the choice of redshift bins and error estimations, we refer the reader to Supplementary Information.

### Results

#### Verifying the CNN models

Our estimated CNN masses, \( M_{\text{CNN}} \), compared with the real cluster mass \( M_{\text{true}} \) (quantified with \( M_{500} \), that is, the mass contained within a sphere of radius \( R_{500} \) corresponding to a spherical overdensity of 500 times the critical density of the Universe) are shown in Fig. 1 for simulated clusters using both the Clean mock dataset and the Planck mock dataset. Figure 1 shows the relative error as a function of the predicted CNN mass \( M_{\text{CNN}} \):

\[
\text{err}(M_{\text{CNN}}, M_{\text{true}}) = \frac{M_{\text{CNN}} - M_{\text{true}}}{M_{\text{CNN}}} = b_t, \tag{1}
\]

where \( b_t \) can be thought as the bias of the true mass with respect to the \( M_{\text{CNN}} \) mass. As shown in Fig. 1 (top), \( b_t \) has median values at around \(-0.02\), at the 4 redshift bins and within the whole mass range. The scatter in this CNN-estimated mass is within 20%. Even when training with and applying to the Planck mock dataset, it is clear (Fig. 1, bottom) that the CNN mass is only slightly biased towards a negative value (\( \leq 5\% \); see

---

**Table 1** | Detailed datasets used in this study

| Dataset       | Mock/real | Beam smoothing (FWHM 10 arcmin) | Instrumental noise | Point-source contaminants |
|---------------|-----------|---------------------------------|--------------------|--------------------------|
| Original mock dataset | Mock      | No (5°)                         | No                 | No                       |
| Clean mock dataset | Mock      | Yes                             | No                 | No                       |
| Planck mock dataset | Mock      | Yes                             | Yes                | No                       |
| Planck real dataset | Real      | Yes                             | Yes                | No                       |
| Golden sample | Real       | Yes                             | Yes                | No                       |

Data properties of the simulated datasets: clean mock dataset and Planck mock dataset and the observations Planck real dataset and Golden sample.

**Fig. 1** | Verifying CNN with mock maps. The relative error \( M_{\text{CNN}} - M_{\text{true}} / M_{\text{CNN}} \) for the Clean mock dataset (top) and the Planck mock dataset (bottom) as a function of the predicted mass \( M_{\text{true}} \). Black squares (bin centre) represent median values, and the shaded region is the 16th–84th percentiles. The red dashed lines correspond to the perfect prediction (0 error) and \( \pm 20\% \) error and different colour points depict different redshift ranges as shown in the legend. A random sample of 200 points per redshift bin is shown but the statistics (median and 16th–84th percentiles) are computed using the whole test set. The data are binned along \( M_{500} \); such that every bin has \( n = 962 \) y maps.
Supplementary Section D for more discussions). However, the shaded region increases from $-0.02^{+0.05}_{-0.05}$ (standard error of ±0.001) to $-0.03^{+0.04}_{-0.14}$ (standard error of ±0.002), which indicates the impact of the instrumental Planck noise in the CNN predictions. We note here that the scatter is comparable to the results from ref. 15 (see their Fig. 7). Furthermore, we also trained our model to estimate the cluster mass $M_{200c}$ for the Golden sample. The results were poor, in the sense of relatively larger scatter in the bias $b_M$, due to the fact that the signal becomes weak at $R > R_{500}$ for the Planck mock dataset. However, $M_{200c}$ can be estimated using the Clean mock dataset with a similar accuracy.

**Predicting the Planck cluster masses**

We simply apply the CNNs trained with the Planck mock dataset to the Planck real dataset for predicting their masses (see Methods for a detailed description of the datasets). The results are shown in Fig. 2 by presenting the relative errors between our CNN masses and the cluster masses estimated by Planck, $M_{\text{Planck}}$, as a function of the predicted mass, $M_{\text{CNN}}$. Similarly to equation (1), we define

$$\text{err}(M_{\text{CNN}}, M_{\text{Planck}}) = \frac{M_{\text{CNN}} - M_{\text{Planck}}}{M_{\text{Planck}}} = b_y,$$

where $b_y$ is the bias of Planck masses with respect to the $M_{\text{CNN}}$. It is noted that the cluster masses estimation from Planck is based on the HE assumption. Reference 14 has predicted an average mass bias of $1 - b = 0.8$. Different to the results in the case of the Planck mock dataset, the median value of $b_y$ is clearly biased towards a positive value, $b_y = 0.11^{+0.05}_{-0.15}$ (standard error of ±0.005), for massive clusters ($M_{\text{CNN}}/M_\odot \geq 4 \times 10^{14}$). At lower cluster mass, $b_y \approx -0.03^{+0.05}_{-0.15}$ (standard error of ±0.02), which means a consistent cluster mass between our CNN and the Planck estimations. We note that the scatter shown by the shaded region in Fig. 2 is also in line with the results in the bottom panel of Fig. 1. It is clear that there is about 0.1 difference between this bias and the Planck estimated bias.

In the full Planck real dataset, roughly two-thirds of the clusters have contamination by point-like sources near their centre, which is not present in the simulated maps. To verify whether this is the case of the mass bias difference between our CNN method and the Planck result, we select a subsample of the Planck real dataset that does not have any relevant radio source or other contaminants in the cluster centre or the vicinity (within 10 arcmin). Furthermore, radio emission contamination outside of the main halo is substituted with a signal intensity that is compatible with instrumental noise. This subsample is named the Golden sample and its result is shown in Fig. 2 (bottom). Although this Golden sample contains a smaller number of objects, its median $b_y$ is in good agreement with the result from the full Planck real dataset. For the exact values of the biases of the Golden sample and the full Planck real dataset, we also refer to Supplementary Section D (see Supplementary Table 2). Clearly, this bias is not caused by the detected point-like contaminants. Therefore, we investigate other possibilities in the following section to explain the difference between $M_{\text{CNN}}$ and $M_{\text{Planck}}$.

**Understanding the mass bias**

Limited to our knowledge on the detailed processes of estimating the $M_{\text{Planck}}$, we perform a simple inference of the cluster mass with the mock $y$ maps to compare with its CNN mass. It is well known that the relation between the integrated Compton-$y$ parameter $Y$, which is proportional to the thermal energy in the ICM, over an aperture of radius $R$, and the mass inside the same aperture, $M$, is a power law. Accordingly, $Y$ is defined as an integral over an aperture subtended by a solid angle $\Omega$:

$$Y = \int_{\Omega} y \, d\Omega \approx \sum_{i} y_i \Omega_i,$$

where $\Omega_i$ is the area of the pixel $i$ and the sum is performed over the image pixels inside $R$. Here we focus on quantities integrated inside $R_{500}$ to compare our estimation with other masses, for example, $M_{\text{Planck}}$. This $Y - M$ scaling law can be written as

$$E(z)^{-2/3} \left[ \frac{D_A(z)Y}{10^{-4} \, \text{Mpc}^{-3}} \right] = B \left[ \frac{h}{70} \right]^{-2+\alpha} \left[ \frac{M_{\text{Planck}}}{6 \times 10^{14} \, M_\odot} \right]^{\alpha},$$

where $D_A(z)$ is the angular distance at redshift $z$ and $E(z) = H(z)/H_0$ is the redshift distortion of the Hubble $H(z)$ parameter where $h$ is its dimensionless value, that is, $H = 100 \, h \, \text{km s}^{-1} \, \text{Mpc}^{-1}$. In particular, the fitted parameters slope $\alpha$ and normalization $B$, for ref. 22 are $\alpha = 1.79 \pm 0.08$ and $\log(B) = -0.19 \pm 0.02$ at $R_{500}$. The estimation of these parameters is based on the cluster masses from a mass-proxy relation from ref. 22. The normalization $B$ parameter is similar between the 300 clusters and the Planck result. However, the slope of this relation in The300 is $\alpha = 1.63 \pm 0.29$, which is compatible with a self-similar relation with $\alpha = 5/3$ (ref. 23). It is noted that the large error in the slope from The300 is due to a mass-complete fitting process; interested readers are referred to ref. 23 for details.

To examine whether the difference in the slope of the $Y - M$ scaling relation is the cause of the bias, we derive the $Y_{500}$ from the original mock dataset. It is noted that the $R_{500}$ estimated in the Amiga Halo Finder (AHF) catalogue is used here. $M_{\text{SZ}}$ is then converted from $Y$ using equation (4) with two slopes: $\alpha = 1.63$ (The300) and $\alpha = 1.79$ (Planck) based on equation (4).

In addition, to meet the Planck results, we applied the same correction factor 1.2 (from $Y_{500}$ to $Y_{200}$, ref. 23) to the $Y$ from the original mock maps (blue line), whereas for the red line, we simply adopt the fitting parameter from ref. 22, which used $Y_{500}$. Here, $Y_{500}$ and $Y_{200}$ are the integrated Compton-$y$ parameter over a cylindrical region and a spherical region, respectively.

In Fig. 3, we show the relative errors between $M_{\text{CNN}}$ and $M_{\text{Planck}}$ as a function of $M_{\text{CNN}}$ for $M_{\text{Planck}}$ masses estimated through the two different...
scaling relations. For an easy comparison, the same error for $M_{\text{SZ}}^\text{Planck}$ in Fig. 2 is included as error bars. It is not surprising to see a larger difference between the two results at higher cluster masses as $M_{\text{c}}$ is normalized to $6 \times 10^{14} h^{-1} M_\odot$. It is clear that the blue dotted-dashed line follows the Planck data points very well, and the red dashed line follows the distribution whose mean and $1 - \sigma$ values are around $0.04 \pm 0.05$ (standard error of $0.0007$) at all mass ranges, which is in agreement with the result from Fig. 1. Although we use the original high-resolution $y$ maps to calculate $Y$, a similar result is obtained using the Planck mock dataset for resolved clusters. In practice, we do not find any noticeable difference for clusters whose $R_{200}$ is greater than 5 pixels. Therefore, a possible explanation for the fact that $b_3$ is different from 0 lies in the intrinsic difference in the assumed $Y$–$M$ relation between the The300 simulated clusters and the Planck clusters. 

As the $Y$–$M$ relation imposes the difference between Planck and The300 simulation and suggests that the root of the bias shown in Fig. 2 lies in this, we discuss the possible reasons for the differences here. From an observational point of view, the uncertainties may come from a couple of sources. (1) The calibration of the $Y$–$M$ relation (where the cluster mass $M_{\text{c}}$ is estimated from X-ray data under the HE assumption). Therefore, an HE mass bias $b_{\text{HE}}$ with a value of about 0.1–0.2 (see ref. 37 for discussions on this value difference) will be inherited. Furthermore, as shown in ref. 13, the Planck $Y$–$M$ slope is steeper than several simulations (see Fig. A2 in ref. 37), which have slopes closer to a self-similar relation. (2) The $Y_{500}$ values derived from the $y$ map are integrated out to $5 < R_{200}$ owing to the large angular resolution of Planck. The angular resolution impact on the $Y$–$M$ relation can be found in ref. 38. Furthermore, the uncertainty in estimating the $R_{200}$ in observations may also play a role; the mis-centre problems may bias the $Y_{500}$ values as well as $M_{\text{c}}$. On the simulation side, the uncertainties in the simulated $Y$–$M$ relation mainly come from the implemented baryon models. However, as indicated in refs. 22,31, the same clusters run with three different baryon models, such as GADGET-MUSIC, without active galactic nuclei (AGN) feedback, GADGET-X, with AGN feedback, and GIZMO-SIMBA, with strong AGN feedback, show consistent fitting results on the $Y$–$M$ relation, especially at the massive halo mass end. However, it is worth noting that ref. 32 showed that including the low-mass halo will increase the slope (see references therein for more discussions); meanwhile, ref. 32 also suggested that the angular resolution plays a critical role in this relation. Lastly, although this scaling relation from The300 seems almost independent of the implemented gas physics (see also ref. 1), refs. 34,35 suggested that different baryon models can violate this self-similarity. Nevertheless, the weak or no redshift evolution of the $Y$–$M$ relation up to $z = 1$ is generally in agreement with other works (for example, refs. 32,34).

In addition, it is also worth noting that $M_{\text{SZ}}^\text{Planck}$ and $M_{\text{c}}$ are intrinsically different: CNN predictions target the true three-dimensional $M_{500}$ based on the physical identified halos in simulation, whereas $M_{\text{c}}$ is a mass estimated through a calibrated $Y$–$M$ scaling relation with the integrated $Y$ from observed clusters within $R_{200}$ from two-dimensional images. However, as indicated in Fig. A3 in ref. 31, the bias depends on the cluster mass—smaller (0.1) at low cluster mass and larger (0.2) at the massive end. This trend is in agreement with the bias shown in Fig. 2, albeit about 0.1% lower (it is noted that $M_{\text{SZ}}^\text{Planck}$ used in this work is not bias corrected). Furthermore, the $Y$–$M$ relation from The300 simulation is in a better agreement with the Planck data at $10^{14} M_\odot \lesssim M_{500} \lesssim 4 \times 10^{15} M_\odot$ (see Fig. 10 in ref. 1). Larger deviation is found at the more massive cluster end. Lastly, we also tried cross-model checks with our CNN, that is, we trained the model with only mock $y$ maps of GADGET-X and applied it to GIZMO-SIMBA or GADGET-MUSIC mock images (Supplementary Section G). Our results are qualitatively in agreement with ref. 1 for a similar approach but to infer cosmological parameters. It suggests that different baryon physics models have a weak impact on our predictions of $M_{500}$ at the cluster-mass scale. In conclusion, we think that the differences between $M_{\text{c}}^\text{Planck}$ and $M_{\text{c}}$ may mainly result from the $Y$–$M$ relation. If we trust the $M_{\text{c}}$ as the true three-dimensional mass of the clusters, the bias in ref. 31 may be just slightly overestimated.

Conclusions

CNN is a powerful tool that allows us to directly apply theoretical models or simulation predictions to raw observational data to derive quantities that we are interested in. By training four CNNs with mock Planck-like $S_{\alpha}$ maps and then applying them to real Planck $y$ maps, we evaluate their relevance and provide CNN-estimated masses of the PSZ2 clusters. We use synthetic clusters selected from The300 simulation to match the PSZ2 clusters in both redshift and mass ranges. The mock $S_{\alpha}$ $y$ maps constitute the Clean mock dataset sharing the same beam size smoothing as in the real Planck cluster maps, whereas the Planck mock dataset further takes the Planck instrumental noise into account. Four CNNs are trained independently by separating the full sample (~200,000 images) into 4 different redshift ranges: $z \leq 0.1$, $0.1 < z \leq 0.2$, $0.2 < z \leq 0.4$ and $z > 0.4$. We show that there are very small biases between the CNN masses and the real three-dimensional cluster masses $M_{500}$ for both the Clean mock dataset and the Planck mock dataset, and the scatter in the CNN masses is also very low (an intrinsic scatter—16th–84th percentiles—of 0.1% and of 0.2%, respectively). By applying these CNNs trained with the Planck mock dataset to the Planck real dataset cluster maps, we provide newly independent CNN-estimated cluster masses with the posterior uncertainties from the simulation-based inference method. Comparing with the cluster mass estimated by Planck mainly with the HE assumption, we find a relevant non-null bias $b_3$ at higher cluster masses, while $M_{\text{c}}^\text{Planck}$ and $M_{\text{c}}$ are in agreement for low-mass clusters. After performing an experiment, the fact that the bias between $M_{\text{c}}$ and $M_{\text{c}}^\text{Planck}$ is not zero might be caused by the
different slopes of the $Y-M$ scaling law between The300 simulations and the Planck one. If the cluster masses estimated by CNN target their true $M_{200}$, this work suggests that the bias for the PSZ2 clusters of $b_\gamma = 0.11$ should mostly be due to the HE bias in Planck. This small bias makes the reconciliation with the CMB constraints even harder.

By training CNNs with mock maps and applying them to real cluster maps, this work establishes that ML models can directly link hydrodynamic simulations with observations. Our approach depends less on some theoretical model assumptions and almost does not require estimations on redshift, $R_{200}$ and so on. Furthermore, it provides the true simulated analogue physical properties of real observations. To this end, this work is only a starting step towards accurate mass estimations that can potentially be extended to other observations as well.

Methods

The Three Hundred simulations

The Three Hundred project is based on hydrodynamic zoomed re-simulations of spherical regions centred on the 324 most massive clusters at $z = 0$, identified in the MultiDark dark-matter (DM)-only simulation (MDPL2; ref. 38). It utilizes the cosmological parameters from the Planck mission 37, and simulates a periodic cubic box of comoving length $1\,h^{-1}\,\text{Mpc}$ containing $3,840^3$ DM particles with mass of $1.5 \times 10^5 \, h^{-1}\,\text{M}_\odot$.

A large region around each cluster of $15\,h^{-1}\,\text{Mpc}$ (over $5 \times R_{200}$) is used for re-simulation with different baryonic physics models: GADGET-MUSIC ref. 49, GADGET-X ref. 41,42 and GIZMO-SIMBA ref. 31,43. In this study, we mostly focus on the galaxy clusters simulated by GADGET-X.

Halos are identified with the AHF package 50. For this work, we select out halos with $M_{200,0} > 10^{14}\, h^{-1}\,\text{M}_\odot$ and free of contamination (that is, the halos do not contain low-resolution DM particles) out to $z = 1$. The mass and redshift distributions of the selected clusters together with the 1,094 PSZ2 clusters are shown in Supplementary Fig. 1. The masses of the Planck clusters are extracted from the PSZ2 catalogue and are further divided (only in the figure) by the expected average hydrostatic bias $b = 0.2$ to be compared with the total mass of the synthetic clusters from The300 simulations.

For each cluster in the sample, the mock map is generated with the PYMSZ package 22,44 with 27 different line-of-sight projections by rotating the cluster around its centre—the maximum mass density peak. The Compton-y parameter maps are estimated as follows:

$$y = \frac{\sigma_T b_\gamma}{m_e c^2} \int n_e T_e d\ell,$$

where $\sigma_T$ is the Thomson cross-section, $b_\gamma$ is the Boltzmann constant, $c$ is the speed of light, $n_e$ is the electron rest-mass, $T_e$ is the electron number density, $T_i$ is the electron temperature, and the integration is done along the observer’s line of sight ($\ell$). Equation (5) is discretized in our simulated data as in refs. 31,46:

$$y = \frac{\sigma_T b_\gamma}{m_e c^2 dA} \sum_r T_{e,i} N_{e,i} W(r, h_i),$$

where $d\ell$ is substituted by $dV/dA$ and $N_{e,i}$ is the volume ($V$) times $n_{e,i}$ and $dA$ is the differential area orthogonal to the line of sight $l$. Moreover, $W(r, h_i)$ is the same sph kernel as in the hydrodynamic simulation with smoothing length $h_i$ at the distance $r$ from the kernel centre. The summation is done over all gas particles $i$.

Originally, each mock image has $1,920 \times 1,920$ pixels with a fixed angular resolution of $5^\circ$ to at least $R_{200}$ in all the clusters. Moreover, the redshift associated with the mock images is the same as the simulation redshift, that is, the snapshot at which the clusters are selected. In real Planck maps, cluster signals can be affected by foreground or background sources, such as radio sources, submillimetre galaxies or other clusters. However, we do not take this contamination into account when generating these mock maps for two reasons: (1) the contamination level is still unclear (see refs. 45,46 for different contamination fractions); (2) the $Y$ signals from different clusters are indistinguishable with Planck’s beam size, especially close to the cluster centre. Therefore, the integrated $Y$ value in Planck might also include these foreground and background clusters, thus, their masses. With these considerations, we do not add contaminating clusters in the mock catalogue, which is a limitation of the current analysis.

Mock Planck observations

To apply our trained CNN models to real Planck maps, mock Planck maps must have the same observational limitations, mainly the same angular resolution and noise. The original maps are post-processed using a procedure similar to the one detailed in refs. 42,43. We detail it here for the reader’s convenience.

Our goal is to create realistic simulations of Planck Compton parameter maps to train our CNN so that it can eventually be applied to cut-outs obtained from Gnomonic projections of the publicly available Planck full-sky y map computed with the modified internal linear combination algorithm (MILCA). To this end, we need to first smooth the The300 images (see The Three Hundred simulations in Methods) by applying a Gaussian kernel with a 10-arcmin full-width at half-maximum (FWHM), filtering the small scales as with the Planck beam. We assume the filtering of large scales by Planck to be negligible. We further process the smoothed simulated y maps by re-gridding them on a grid with 1.7-arcmin-pixel resolution to match the map resolution of Gnomonic projections of the Planck full-sky y map at HEALPix resolution Nside = 2048. This set of maps constitutes what we call the Clean mock dataset. It will be used for characterizing the impact of instrumental noise in the CNN predictions.

Then, we generate a full-sky realization of the Planck instrumental noise based on the publicly available map of the standard deviation of the Compton parameter at HEALPix resolution Nside = 2048 and the noise power spectrum of the Planck full-sky y map. Thus, this realization includes the noise with spatial distribution as observed in the Planck y maps. We extract cut-outs of this noise map using Gnomonic projections centred on cluster locations drawn randomly from the PSZ2 catalogue to match the noise properties of the detected clusters. These cut-outs are generated with the same number of pixels as the maps in the Clean mock dataset.

We generate the Planck mock dataset by adding a noise map cut-out to each map from the Clean mock dataset. The maps in this new dataset are realistically simulated Planck observations of the synthetic clusters from The300 simulation. We note, however, that we did not include the contamination induced by point sources in these simulated maps. This is examined by using real Planck maps without point-source contamination. As shown in the Results section, this should not impact the CNN predictions. We further only select the maps with a higher signal-to-noise ratio (S/N) and a similar cut-off as in the PSZ2 catalogue. However, this selection is performed with cluster mass cut instead of an S/N limit for two reasons. (1) Owing to different estimation procedures of the S/N, the S/N for these mock maps shows much larger scatter at lower Planck S/N. With a simply S/N cut, even with a higher value, we still found many low-mass halos contaminating our sample. Therefore, the corresponding mass cut instead of an S/N cut gives more reliable maps with higher signal. (2) Using an S/N cut will produce an uneven separation between training, validation and test, which is much more complex. This is because each cluster has 27 random projections and the sample is split by cluster to avoid using the same cluster for training and testing/validating, not by maps.

In summary, the Planck mock dataset is composed of the same simulated clusters as the Clean mock dataset but with the addition of Planck noise. Each simulated dataset is composed of $6,765$ different clusters with 27 rotations amounting to a total of $182,655$ maps.
Furthermore, these maps are generated from objects extracted from The300 simulation to cover the Planck sample PSZ2 in the mass range $10^{14} M_\odot < M < 10^{15} M_\odot$, and the redshift range $0 < z < 1$, which is presented in Supplementary Fig. 1. It is noted here that the CNN model is trained and applied in four different redshift bins (see Supplementary Section E for the reasons), and the selected clusters have a different mass range in each redshift bin. This mass cut is based on our S/N estimation as presented in the previous section. With this mass cut, our sample overlaps with the distribution of Planck clusters well as shown in Supplementary Section B (see Supplementary Fig. 1).

**Planck full-sky maps**

The work presented here aims to provide new estimates of the total mass of the PSZ2 clusters resulting from the processing of their y maps by a set of trained CNNs. To this end, we extract the map associated with each PSZ2 cluster with known redshift using a Gnomonic projection of the publicly available Planck full-sky y map centred on the Galactic coordinates provided in the PSZ2 catalogue. We use the gnomview tool provided by the healpy Python library using 96 x 96 pixels of $1.7 \times 1.7$ arcmin$^2$ to match the size of the maps in the Planck mock dataset. This set of maps forms what we call the Planck real dataset.

We further investigate the impact of contamination from the astrophysical signal induced by the Galactic plane and point sources. We perform the extraction of the PSZ2 cluster y maps again by applying the publicly available masks of the Galactic plane and point sources on the full-sky y map before the Gnomonic projections. This allows us to discriminate PSZ2 clusters without any known contamination of the SZ signal within 10 arcmin from the cluster centre defined in the PSZ2 catalogue. In particular, clusters with radio AGN contamination biasing the SZ signal low are excluded with this selection procedure. We find 395 PSZ2 clusters that satisfy this condition. These clusters form what we call the Golden sample. We include the Golden sample to verify that our results based on the training with mock maps, which do not consider point sources, are not affected by that.

We present representative examples of y maps from our three datasets with different masses and redshifts in Supplementary Section B (see Supplementary Fig. 2).

**Data availability**

The catalogue of CNN-estimated masses for Planck clusters can be downloaded from https://github.com/The300th/DeepPlanck. The mock y maps used to train the different CNN models can be accessed upon request to the corresponding authors.

**Code availability**

CNN trained weights and data products are available at https://github.com/The300th/DeepPlanck.

**References**

1. Kravtsov, A. V. & Borgani, S. Formation of galaxy clusters. *Annu. Rev. Astron. Astrophys.* **50**, 353–409 (2012).
2. Planck Collaboration et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **641**, A6 (2020).
3. Biviano, A. et al. On the efficiency and reliability of cluster mass estimates based on member galaxies. *Astron. Astrophys.* **456**, 23–36 (2006).
4. Becker, M. R. & Kravtsov, A. V. On the accuracy of weak-lensing cluster mass reconstructions. *Astrophys. J.* **740**, 25 (2011).
5. Bryan, G. L. & Norman, M. L. Statistical properties of X-ray clusters: analytic and numerical comparisons. *Astrophys. J.* **495**, 80–99 (1998).
6. Planck Collaboration Planck 2015 results. XXIV. Cosmology from Sunyaev–Zeldovich cluster counts. *Astron. Astrophys.* **594**, A24 (2016).
7. Salvati, L., Douspis, M. & Aghanim, N. Constraints from thermal Sunyaev–Zeldovich cluster counts and power spectrum combined with CMB. *Astron. Astrophys.* **614**, A13 (2018).
8. Gianfagna, G. et al. Exploring the hydrostatic mass bias in MUSIC clusters: application to the NIKA2 mock sample. *Mon. Not. R. Astron. Soc.* **502**, 5115–5133 (2021).
9. Bishop, C. M., Nasrabad, N. M. (2006) Pattern recognition and machine learning. Springer, New York.
10. Baron, D. Machine learning in astronomy: a practical overview. Preprint at https://arxiv.org/abs/1904.07248 (2019).
11. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
12. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016); http://www.deeplearningbook.org
13. LeCun, Y. et al. Generalization and network design strategies. *Connectionism Perspect.* **19**, 143–155 (1989).
14. Ntampaka, M. et al. A deep learning approach to galaxy cluster Xray masses. *Astrophys. J.* **876**, 82 (2019).
15. Gupta, N. & Reichardt, C. L. Mass estimation of galaxy clusters with deep learning. I. Sunyaev–Zel'dovich effect. *Astrophys. J.* **900**, 110 (2020).
16. Gupta, N. & Reichardt, C. L. Mass estimation of galaxy clusters with Deep Learning II. Cosmic microwave background cluster lensing. *Astrophys. J.* **923**, 96 (2021).
17. Yan, Z., Mead, A., Van Waerbeke, L., Hinshaw, G. & McCarthy, I. Galaxy cluster mass estimation with deep learning and hydrodynamical simulations. *Mon. Not. R. Astron. Soc.* **499**, 3445–3458 (2020).
18. Ho, M. et al. A robust and efficient deep learning method for dynamical mass measurements of galaxy clusters. *Astrophys. J.* **887**, 25 (2019).
19. Koda, Ramanah, D., Wojtak, R. & Arendse, N. Simulation-based inference of dynamical galaxy cluster masses with 3D convolutional neural networks. *Mon. Not. R. Astron. Soc.* **501**, 4080–4091 (2021).
20. Ho, M., Farahi, A., Rau, M. M. & Trac, H. Approximate Bayesian uncertainties on deep learning dynamical mass estimates of galaxy clusters. *Astrophys. J.* **908**, 204 (2021).
21. Planck Collaboration Planck 2015 results. XXVII. The second Planck catalogue of Sunyaev–Zeldovich sources. *Astron. Astrophys.* **594**, A27 (2016).
22. Cui, W. et al. The Three Hundred project: a large catalogue of theoretically modelled galaxy clusters for cosmological and astrophysical applications. *Mon. Not. R. Astron. Soc.* **480**, 2898–2915 (2018).
23. Planck Collaboration Planck 2013 results. XX. Cosmology from Sunyaev–Zeldovich cluster counts. *Astron. Astrophys.* **571**, A20 (2014).
24. Sunyaev, R. A. & Zeldovich, Y. B. The observations of relic radiation as a test of the nature of X-ray radiation from the clusters of galaxies. *Comments Astrophys. Space Phys.* **4**, 173 (1972).
25. Kravtsov, A. V., Vikhlinin, A. & Nagai, D. A new robust low-scatter X-ray mass indicator for clusters of galaxies. *Astrophys. J.* **650**, 128–136 (2006).
26. Knollmann, S. R. & Knebe, A.; AFI: Amiga's Halo Finder. *Astron. J.* Suppl. Ser. **182**, 608–624 (2009).
27. Gianfagna, G., Rasia, E., Cui, W., De Petris, M. & Yepes, G. The hydrostatic mass bias in the Three Hundred clusters. *EPJ Web Conf.* **257**, 00020 (2022).
28. Yang, T. et al. Understanding the relation between thermal Sunyaev-Zeldovich decrement and halo mass using the Simba and TNG simulations. *Mon. Not. R. Astron. Soc.* (2022).
29. Ferragamo, A. et al. Comparison of hydrostatic and lensing cluster mass estimates: a pilot study in MACS J0647.7+015. *Astron. Astrophys.* **661**, A65 (2022).
30. Cui, W. et al. niFTy galaxy cluster simulations—IV. Quantifying the influence of baryons on halo properties. Mon. Not. R. Astron. Soc. 458, 4052–4073 (2016).
31. Cui, W. et al. The Three Hundred project: the GIZMO-SIMBA run. Mon. Not. R. Astron. Soc. 514, 977–996 (2022).
32. Henden, N. A., Puchwein, E. & Sijacki, D. The redshift evolution of X-ray and Sunyaev–Zel’dovich scaling relations in the FABLE simulations. Mon. Not. R. Astron. Soc. 489, 2439–2470 (2019).
33. Le Brun, A. M. C., McCarthy, I. G. & Melin, J.-B. Testing Sunyaev–Zel’dovich measurements of the hot gas content of dark matter haloes using synthetic skies. Mon. Not. R. Astron. Soc. 451, 3868–3881 (2015).
34. Le Brun, A. M. C., McCarthy, I. G., Schaye, J. & Ponman, T. J. The scatter and evolution of the global hot gas properties of simulated galaxy cluster populations. Mon. Not. R. Astron. Soc. 466, 4442–4469 (2017).
35. Barnes, D. J. et al. The redshift evolution of massive galaxy clusters in the MACSIS simulations. Mon. Not. R. Astron. Soc. 465, 213–233 (2017).
36. de Andres, D. et al. Machine learning methods to estimate observational properties of galaxy clusters in large volume cosmological N-body simulations. Preprint at https://arxiv.org/abs/2204.10751 (2022).
37. Villaescusa-Navarro, F. et al. Robust marginalization of baryonic effects for cosmological inference at the field level. Preprint at https://arxiv.org/abs/2109.10360 (2021).
38. Klypin, A., Yepes, G., Gottlöber, S., Prada, F. & Heß, S. MultiDark simulations: the story of dark matter halo concentrations and density profiles. Mon. Not. R. Astron. Soc. 457, 4340–4359 (2016).
39. Planck Collaboration Planck 2015 results. XIII. Cosmological parameters. Astron. Astrophys. 594, A13 (2016).
40. Semololini, F. et al. The MUSIC of galaxy clusters—I. Baryon properties and scaling relations of the thermal Sunyaev–Zel’dovich effect. Mon. Not. R. Astron. Soc. 429, 323–343 (2013).
41. Murante, G., Monaco, P., Giovalli, M., Borgani, S. & Diacerio, A. A subresolution multiphase interstellar medium model of star formation and supernova energy feedback. Mon. Not. R. Astron. Soc. 405, 1491–1512 (2010).
42. Rasia, E. et al. Cool core clusters from cosmological simulations. Astrophys. J. Lett. 813, L17 (2015).
43. Davé, R. et al. SIMBA: cosmological simulations with black hole growth and feedback. Mon. Not. R. Astron. Soc. 486, 2827–2849 (2019).
44. Baldi, A. S. et al. Kinetic Sunyaev–Zel’dovich effect in rotating galaxy clusters from MUSIC simulations. Mon. Not. R. Astron. Soc. 479, 4028–4040 (2018).
45. Aguado-Barahona, A. et al. Optical validation and characterization of Planck PSZ2 sources at the Canary Islands observatories. II. Second year of LPI5 observations. Astron. Astrophys. 631, A148 (2019).
46. Wen, Z. L. & Han, J. L. Clusters of galaxies up to z = 1.5 identified from photometric data of the Dark Energy Survey and unWISE. Mon. Not. R. Astron. Soc. 513, 3946–3959 (2022).
47. Ruppin, F. et al. Impact of ICM disturbances on the mean pressure profile of galaxy clusters: a prospective study of the NIKA2 SZ large program with MUSIC synthetic clusters. Astron. Astrophys. 631, A21 (2019).
48. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://arxiv.org/abs/1409.1556 (2014).

Acknowledgements
We acknowledge helpful discussions with A. Ferragamo, F. De Luca and F. Semololini. D.d.A., W.C. and G.Y. thank financial support from Ministerio de Ciencia e Innovación (Spain) under project grant PID2021-122603NB-C21. W.C. is supported by the STFC AGP Grant ST/V005594/1 and by the Attrazione di Talento Contract no. 2020-T1/TIC-19882 granted by the Comunidad de Madrid in Spain. W.C. further acknowledges the science research grants from the China Manned Space Project with no. CMS-CSST-2021-A01 and no. CMS-CSST-2021-B01. M.D.P. acknowledges support from Sapienza Università di Roma thanks to Progetti di Ricerca Med 2019, RM1191687540DD8D and Progetti di Ricerca Med 2020, RM120172B32D5BE2.

Author contributions
D.d.A. led the project, wrote and run the machine learning codes and contributed to most of the writing of the paper. W.C. developed, ran The300 simulation, prepared the mock observation images with PYMSZ and contributed to writing most of the paper. F.R. wrote and run the code pipeline to introduce Planck-like limitations into clean mock observations and assisted with the writing of the paper. M.D.P. and G.Y. assisted with interpretation, manuscript preparation and revision. G.G., I.L., G.A., R.D., M.J. and J.V.-F. contributed to this work with the writing of the project and with machine learning technicalities.

Competing interests
The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41550-022-01784-y.

Correspondence and requests for materials should be addressed to Daniel de Andres or Weiguang Cui.

Peer review information Nature Astronomy and the authors thank Ziang Yan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2022