Supporting information for ‘Augmenting astrophysical scaling relations with machine learning’

Digvijay Wadekar\textsuperscript{a,b,1}, Leander Thiele\textsuperscript{c}, Francisco Villaescusa-Navarro\textsuperscript{d,e}, J. Colin Hill\textsuperscript{f,d}, Miles Cranmer\textsuperscript{e}, David N. Spergel\textsuperscript{d,e}, Nicholas Battaglia\textsuperscript{g}, Daniel Anglés-Alcázar\textsuperscript{h,d}, Lars Hernquist\textsuperscript{i}, and Shirley Ho\textsuperscript{d,e,j}

\textsuperscript{a}School of Natural Sciences, Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ 08540, USA; \textsuperscript{b}Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003; \textsuperscript{c}Department of Physics, Princeton University, Jadwin Hall, Princeton NJ 08544, USA; \textsuperscript{d}Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY, 10010, USA; \textsuperscript{e}Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton NJ 08544-0010; \textsuperscript{f}Department of Physics, Columbia University, New York, New York 10027, USA; \textsuperscript{g}Department of Astronomy, Cornell University, Ithaca, NY 14853, USA; \textsuperscript{h}Department of Physics, University of Connecticut, 196 Auditorium Road, Storrs, CT, 06269, USA; \textsuperscript{i}Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA; \textsuperscript{j}Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15217

In Fig. S1, we show the mass function of the clusters in two different snapshots of the IllustrisTNG simulation. We also show the relation of different observable properties of clusters to their mass in Fig. S2.

In Fig. 2 of the main text, we showed the performance of $Y_{200c}$ as a mass proxy. Fig. S2 shows the comparison of other mass proxies. We see that the cluster richness has a much larger scatter than $Y_{200c}$, which makes the cluster mass estimation relatively less accurate. One additional issue in using cluster richness from galaxy photometric observations is that, because of the presence of background galaxies, it is not possible to state with absolute confidence that any given galaxy belongs to a given cluster. $M_{\text{gas}}$, on the other hand, also has a low scatter similar to $Y_{200c}$. However, the deviation from a power law relation is much larger than $Y_{200c}$, and the break from power law occurs at a higher halo mass.

We have primarily used the random forest regressor (RF) as feature selection tool in the main text. We show the relative importance of different input features for the RF prediction in Fig. S3. We indeed see that $c_{\text{gas}}, M_\ast/M_{\text{gas}}$ and $c_{\text{NFW}}$ are the most important features.

In the final list of formulae obtained from PySR, we choose the simplest ones to compare in Fig. 4 of the main text. We show a few additional results and compare their performance in reducing the scatter in Fig. S4.

In Fig. 6 of the main text, we show the reduction in the $Y - M$ scatter when cores of clusters in the TNG300 simulation are excised. We show a similar plot for the CAMELS simulations in Fig. S5.

Finally, in Fig. S6, we show the comparison of results from RF on adding two extra parameters (which were used in the previous literature to augment the $Y - M$ relation). Ref. (1) proposed using $R_{500c}/R_{200c}$ as an analogue of the NFW concentration. Ref. (2) proposed using half-light radius for the SZ flux, which is defined as the radius of the sphere that contains half of the total SZ flux. We only collect data for the clusters until $1.5 \times R_{200c}$ and find that the integrated $Y$ has not converged until this radius. We therefore use a different version of the half-light radius than the one proposed in (2); in our case, $R_{SZ,2}$ is obtained from the condition: $Y(r < R_{SZ,2}) = Y_{200c}/2$.

1. Yang HYK, Bhattacharya S, Ricker PM (2010) The Impact of Cluster Structure and Dynamical State on Scatter in the Sunyaev-Zel’dovich Flux-mass Relation. \textit{ApJ} 725(1):1124–1136.
2. Afshordi N (2008) Fundamental Plane of Sunyaev-Zel’dovich Clusters. \textit{ApJ} 686(1):201–205.
Fig. S2. Similar to Fig. 2 in the main text, but for scaling relations for other proxies of halo mass: \(M_{\text{gas}}\) from X-ray surveys, \(M_*\) and richness (i.e., galaxy number counts) from galaxy surveys. The cluster data is from the TNG300 simulation at \(z = 0\). The power-law scaling relation normalized to the most massive halos is shown by the dotted green line.

Fig. S3. The importance of different input variables for the random forest (RF) prediction. The predictions of RF corresponding to the three most important variables is shown in Fig. 3 of the main text.

Fig. S4. Same as Fig. 4 in the main text, but showing the performance of a few additional equations obtained from the symbolic regressor.
**Fig. S5.** Same as Fig. 6 in the main text, but for clusters in the CAMELS simulation suite instead of TNG300. We see a similar reduction in scatter once the cores of these clusters are excised.

**Fig. S6.** Same as Fig. 3 in the main text but adding to the RF training set two additional parameters which have been proposed in the previous literature to augment the $Y-M$ relation. $R_{500c}/R_{200c}$ corresponds to an analogue of the halo concentration and $R_{SZ,2}$ corresponds to an analogue of the SZ half-light radius (see the text for further details).