Study of Cosmological Large Scale Structure with Observations and Simulations

OTKA NN 129148 Closing report

Short summary

The standard LCDM model of cosmology has proven very successful in explaining a variety of observations from the early cosmic microwave background (CMB) to the formation of the Large Scale Structure (LSS) of galaxies and the local Universe. In recent years, increasingly precise observations have resulted in some tensions between different measurements of key cosmological parameters. In this research project, we have developed methods to better evaluate observational data, analyzed new observational data and created cosmological simulations. In particular, during the project we improved photometric redshift estimation techniques using advanced machine learning algorithms. Using these methods we have created full sky photometric redshift catalogs, by cross-joining several major surveys. We have developed a multi-resolution N-body simulation method, that avoids problems related to periodic boundary conditions and makes it possible to investigate theories beyond the standard cosmology. We have investigated CMB anomalies and the integrated Sachs-Wolfe effect to constrain cosmological parameters and test models beyond the standard cosmological model. We have involved new students in the research field, strengthened the international collaboration with our partners and extended it with new partners.

Machine learning methods for galaxy spectra, morphology and large-scale structure analysis

Observed photometric properties of galaxies depend both on their redshift and on their spectral type. The disentanglement of these components is key for photometric redshift estimation but also important for understanding stellar population composition and spectral evolution of galaxies. The mechanism of the galaxy evolution is strongly interconnected with cosmology and an active research area where the analysis of the bulge-disk relation and stellar population properties are crucial. Spiral galaxies have significant bimodality in their stellar composition. The disk and the spiral arms are composed of a young, blue stellar population and the central region, the bulge contains mostly old, red stars. The proper separation of the two components in multiband photometric surveys can help to understand the stellar evolution inside spiral galaxies. It is often called a blind source separation problem since we do not know the exact broadband spectral energy distribution (SED) and the morphology of the sources. The state-of-the-art frameworks are based on a generalization of non-negative matrix factorization and although able to simultaneously fit the SEDs and the morphology, but their main focus is to disentangle stars and galaxies or overlapping galaxies in the busy fields of upcoming surveys like LSST and Euclid. We have developed a new morphology-independent method, `sedNN`, to estimate the normalized broadband spectral energy distributions of the two main stellar populations in simulated disk galaxies, where we exploit the relation between the observable color distribution of the pixels and the real spectral energy distributions. The method is based on neural network approach and utilizes the autograd feature of the Tensorflow framework. We trained and tested a neural network on a subset of the recently published CosmoDC2 simulated galaxy catalog containing about 3,600 galaxies. From the color distribution of the galaxy-related pixels in simulated broadband images the algorithm is able to accurately predict the real spectral energy distributions of the young blue and old red stellar population even if the flux contribution of them is very different. The model performance was compared to the results of SCARLET (a state-of-the-art non-parametric multiband source separation algorithm) and we found that our model can predict the spectral energy distributions with ~5% error rate on average, which is about half of the error rate than that of SCARLET. This research was part of the PhD thesis work of Sándor Kunsági-Máté [32]. It was presented at the XXXI ADASS Conference [16] and finally published in MNRAS [17].

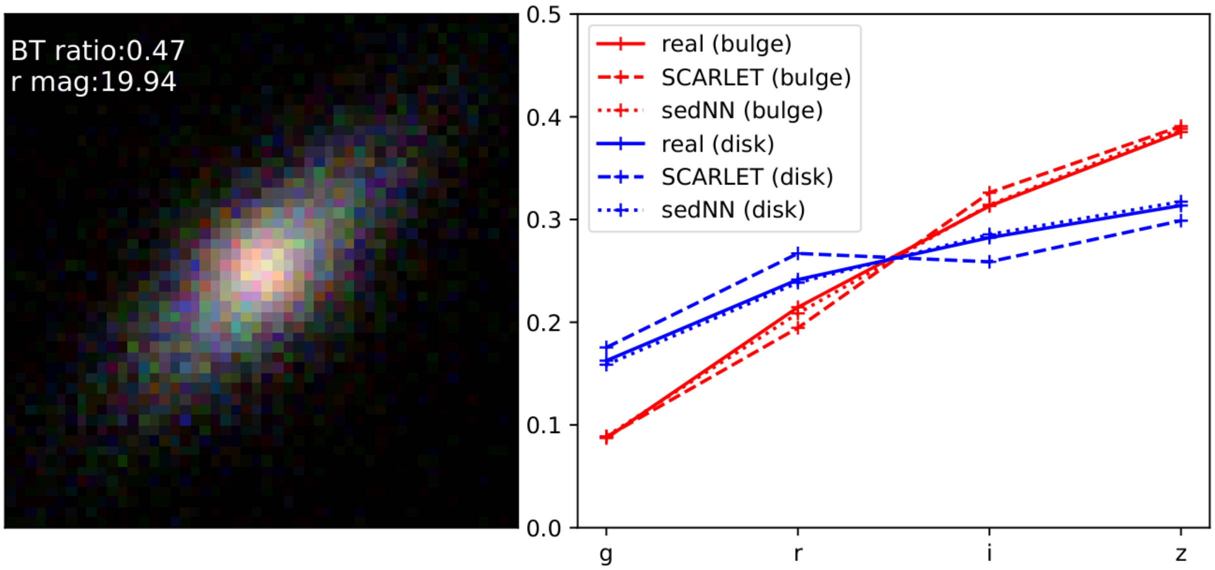

Fig. 1. *Illustration of the SED estimation performance using* `sedNN` *and* `SCARLET` *[17]*

Light from distant galaxies is deflected by the tidal fields of inhomogeneous matter along the line of sight, distorting the shapes of sources, a phenomenon called gravitation lensing. Cosmic shear is the weak gravitational lensing effect arising from no obvious foreground mass apart from the large-scale structure of the Universe. Detecting the effects of the foreground mass density field on distant galaxy images allows us to indirectly map the distribution of the elusive, and apparently very abundant dark matter. Furthermore, characterizing matter distribution at different redshifts through the lensing signal offers a unique window to the evolution of the dark energy dominated late Universe, complementary to other observations. Cosmic shear only distorts the shapes of galaxies at per cent levels, and the signal is dominated by noise from intrinsic shape variation of galaxies, atmospheric, and instrumental distortions. The bulk of the cosmic shear signal is carried by shapes of small and faint galaxies typically those with sizes of a few arcsecs and 24 mag. Weak lensing surveys therefore need to accurately estimate the shapes of galaxies that only cover a few pixels and have very low signal-to-noise ratios (S/N).

We have developed a method based on convolutional neural network (CNN) to predict galaxy shapes from wide-field survey images and trained and evaluated it on the first data release of the Dark Energy Survey (DES DR1). We used conventional shape measurements as ‘ground truth’ from an overlapping, deeper survey with less sky coverage, the Canada–France–Hawaii Telescope Lensing Survey (CFHTLenS). We demonstrated that CNN predictions from single band DES images reproduce the results of CFHTLenS at bright magnitudes and show higher correlation with CFHTLenS at fainter magnitudes than maximum likelihood model fitting estimates in the DES Y1 IM3SHAPE catalogue. Prediction of shape parameters with our CNN is also extremely fast, it takes only 0.2 ms per galaxy, improving more than 4 orders of magnitudes over forward model fitting. The CNN can also accurately predict shapes when using multiple images of the same galaxy, even in different colour bands, with no additional computational overhead. The CNN is again more precise for faint objects, and the advantage of the CNN is more pronounced for blue galaxies than red ones when compared to the DES Y1 METACALIBRATION catalogue, which fits a single Gaussian profile using ‘riz’ band images. We demonstrated that CNN shape predictions within the METACALIBRATION self-calibrating framework yield shear estimates with negligible multiplicative bias, m < 10−3, and no significant point spread function (PSF) leakage. Our proposed set-up is applicable to current and next-generation weak lensing surveys where higher quality ‘ground truth’ shapes can be measured in dedicated deep fields. The results were part of PhD thesis of Dezső Ribli [31] and published in MNRAS [13]

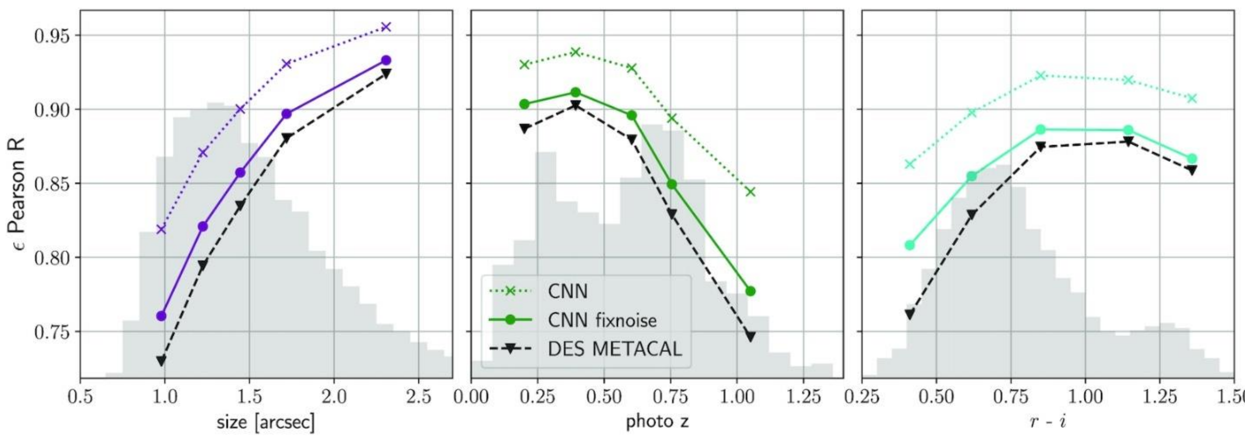

Fig 2. *The CNN estimates galaxy shapes more accurately than the DES METACALIBRATION catalogue regardless of galaxy size (left)and redshift (centre). The advantage is larger for the blue galaxies than the red ones (right). [13]*

Dark matter cannot be observed directly, but its weak gravitational lensing slightly distorts the apparent shapes of background galaxies, making weak lensing one of the most promising probes of cosmology. Several observational studies have measured the effect, and there are currently running and planned efforts to provide even larger and higher-resolution weak lensing maps. Owing to nonlinearities on small scales, the traditional analysis with two-point statistics does not fully capture all of the underlying information. Multiple inference methods have been proposed to extract more details based on higher-order statistics, peak statistics, Minkowski functionals and recently convolutional neural networks. We have developed an improved convolutional neural network that gives significantly better estimates of the Ω_m and σ_8 cosmological parameters from simulated weak lensing convergence maps than state-of-the-art methods and that is also free of systematic bias. We have shown that the network exploits information in the gradients around peaks, and with this insight we have constructed an easy-to-understand and robust peak-counting algorithm based on the steepness of peaks, instead of their heights. The proposed scheme is even more accurate than the neural network on high-resolution noiseless maps. With shape noise and lower resolution, its relative advantage deteriorates, but it remains more accurate than peak counting. With a further development we have show that our improved CNN is able to yield significantly stricter constraints of (Ω_m,σ_8) cosmological parameters than the power spectrum using convergence maps generated by full N-body simulations and ray-tracing, at angular scales and shape noise levels relevant for future observations. In a scenario mimicking LSST or Euclid, our CNN yields 2.4–2.8 times smaller credible contours than the power spectrum, and 3.5–4.2 times smaller at noise levels corresponding to a deep space survey such as WFIRST. We have shown that at shape noise levels achievable in future space surveys our CNN yields 1.4–2.1 times smaller contours than peak counts, a higher order statistic capable of extracting non-Gaussian information from weak lensing maps. The research was done in collaboration with Columbia University and the results were part of PhD thesis of Dezső Ribli [31] and published in Nature Astronomy [14] and MNRAS [15]. Based on these excellent results Dezső Ribli was awarded a prestigious and highly competitive Simons Foundation post-doc position in New York, where he could have continued this research, but finally he has chosen not to accept it and started to work at a leading self-driving car company. We are sorry that he has left academic career in astronomy but on the other hand this case exemplifies that our research approach builds readily applicable skills beyond academic track.

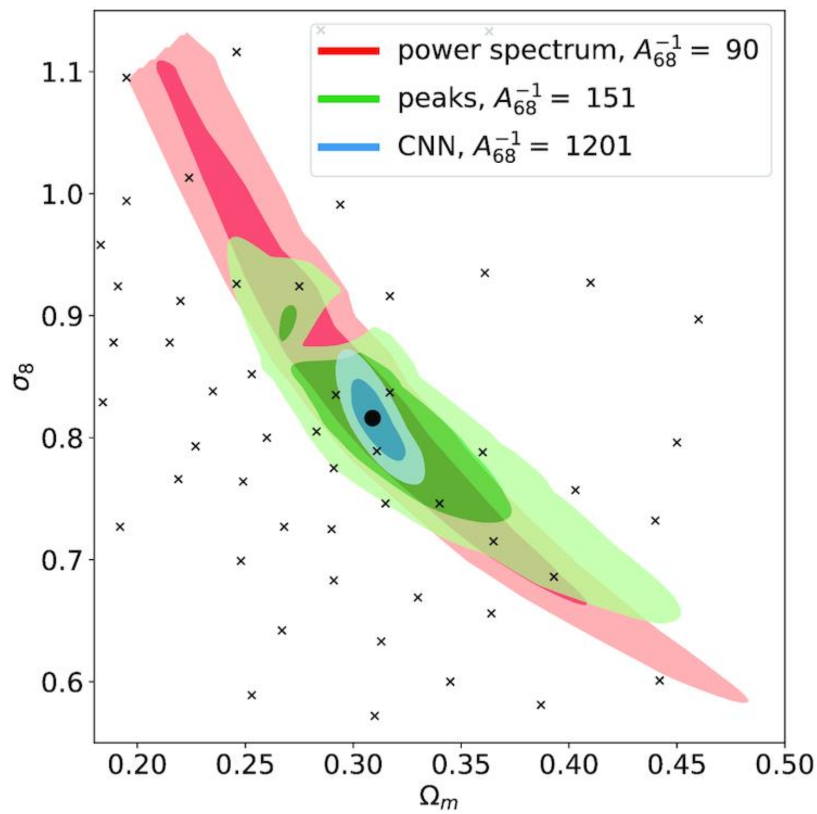

Fig. 3. *The CNN yields significantly smaller credible cosmological parameter contours than the power spectrum or peak counts on noiseless convergence maps. The lighter and darker regions show 95 per cent and 68 per cent confidence areas, respectively. The black crosses represent the parameter pairs of our underlying simulation suite and the larger black dot marks the mock observation. [15]*

Although they are part of the nearby universe, the local dwarf galaxies are unique objects in which the imprint of the conditions that shaped the early structure formation can be studied today at high resolution. In collaboration with project partners at The Johns Hopkins University we have explored how the precise analysis of stellar populations of dwarf galaxies can provide us valuable information about the large-scale and long-term characteristics of both ordinary matter and dark matter. Galactic Archeology focuses on deciphering the formation history of the Milky Way and the Local Group of galaxies via measuring the peculiar motions and chemical abundances of individual stars. The Prime Focus Spectrograph, built for the Subaru telescope is a multi-fiber system of four spectrographs capable of observing approximately 2000 targets at the same time over an area of 1.25 square degrees, or about 1.1 degrees in diameter. This large field of view, combined with the 8.2 mirror, makes this instrument an excellent tool for observing individual red giant stars in nearby satellite galaxies as well in tidal streams around the Milky Way and stars in the nearest galaxies such as M31 and M33. In the frame of this project, we have been developing novel algorithms to deal with relatively low signal-to-noise, low and medium resolution optical spectra to derive physical quantities with reliable error estimates. To measure the line-of-sight velocity of red giants, we combined a state of the art, per pixel cross-correlation algorithm with a precise theoretical derivation of the asymptotic uncertainties in case of fluxing error are present in the data and are accounted for. We ran a large ensemble of simulations of spectroscopic observations to demonstrate that the uncertainty estimates are reliable at S/N > 5 and the uncertainties depend significantly on spectral type. MCMC sampling of the posterior probability distribution of the line-of-sight velocity, as well as the stellar atmospheric parameters indicates that the asymptotic error estimates become unreliable due to non-Gaussian behavior of the posterior at low S/N. Our findings will be published in Dobos et al. [19].

Targeting red giant stars in nearby dwarf spheroidal galaxies is challenging because faint, distant giant stars are often indistinguishable from foreground red dwarf stars in photometric observations. Targeting stars toward the tip of the red giant branch is especially challenging because of the small number of such stars, the uncertainties of stellar population age and chemical abundances, as well as the high number of foreground interlopers. One way to avoid wasting telescope time on foreground stars is to identify dwarf stars from low signal-to-noise, short exposure time spectra. Low S/N stellar spectra, however, pose a problem when fitting stellar parameters and finding the value of surface gravity by template fitting is challenging. We experimented with training artificial neural networks, in the form of a denoiser autoencoder, to approximate the original spectrum without spectrophotometric noise. The proposed network consists of fully connected layers in an hourglass shape with a bottleneck of 256 neurons in the middle layer. The denoiser is trained on simulated spectroscopic observations of stars with a random noise realization on the input and the original, noiseless spectrum on the output. Results indicate that the performance of such a network is not yet meets the low signal-to-noise requirements but reliable denoising is possible above approximately S/N > 5 per pixel. This work was done by Balázs Pál, PhD student at ELTE who has spent 6 months at our research partner at The Johns Hopkins University where he worked together with László Dobos and Alex Szalay. Details of this research will be part of his PhD thesis and is submitted for publication [18].

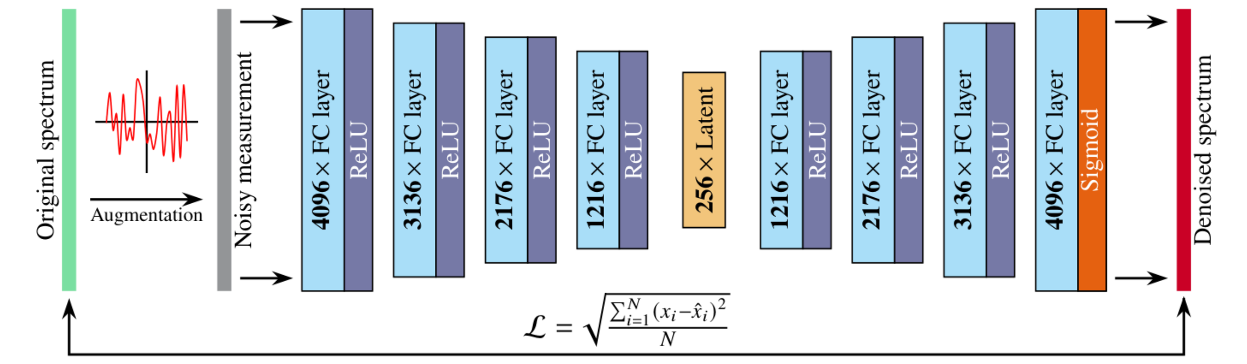

Fig. 4. *Schematic overview of our denoiser training pipeline including the augmentation step*. *[18]*

Photometric redshift estimation and large scale structure catalogs

The redshift of a galaxy uniquely determines its distance in our expanding universe. The colors of a galaxy measured in different bands constrain the approximate photometric redshift. Together with partners at JHU our team was among the first research groups in developing photometric redshift estimation methods and since we have a several decade long experience in improving the models. The prediction models are still imprecise and improving these estimates is a hot topic in cosmology. In a new approach we used machine learning models to refine the photometric redshifts of galaxies using their spatial structure organized into clusters, filaments, and walls. In particular, we tested the hypothesis that additional information from the "neighborhood" of a galaxy sharpens our photometric redshift estimate. We demonstrated that a graph convolutional neural network - trained on a data set of high-resolution redshift observations on a small region of the sky - captures this information by learning to predict the redshift of all the galaxies in a viewing region simultaneously, improving the performance over single-galaxy redshift prediction by 10% median absolute deviation on a held-out region of the sky. The results were presented by Róbert Beck at the NeurIPS machine learning conference [1]. Bendegúz Horváth MSc student performed a similar analysis using another type of machine learning, boosted decision trees in his Master's thesis [29]. Mirkó Mocskonyi BSc student tested another extension of photometric redshift estimation method where galaxy image morphologies provided additional information [28].

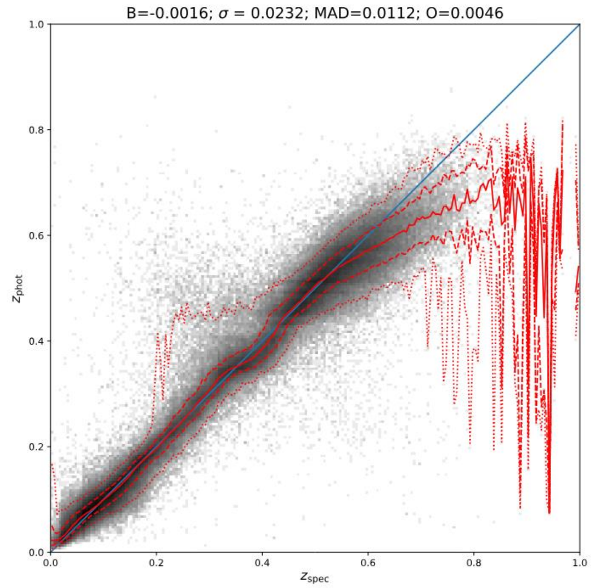

Fig. 5. *Predicted redshift from photometric observations using graph convolution network model (zphot) vs. ground truth measurement via spectroscopy (zspec) of the validation set [1].*

Due to the well-known limitations of spectroscopic observations, ongoing and upcoming wide area cosmological surveys, such as DES, HSC, KiDS, LSST, Euclid, depend highly on reliable photometric redshifts to recover the cosmological parameters with sufficient accuracy. The photometric accuracy of these new generation surveys may be sensitive not just for the shape of the general continuum, but for finer details of the spectra, like stronger spectral lines. So, we investigated the effect of strong emission line galaxies on the performance of empirical photometric redshift estimation methods. In order to artificially control the contribution of photometric error and emission lines to total flux, we developed a PCA-based stochastic mock catalogue generation technique that allows for generating infinite signal-to-noise ratio model spectra with realistic emission lines on top of theoretical stellar continua. Instead of running the computationally expensive stellar population synthesis and nebular emission codes, our algorithm generates realistic spectra with a statistical approach, and – as an alternative to attempting to constrain the priors on input model parameters – works by matching output observational parameters. Hence, it can be used to match the luminosity, color, emission line and photometric error distribution of any photometric sample with sufficient flux-calibrated spectroscopic follow-up. We tested three simple empirical photometric estimation methods and compared the results with and without photometric noise and strong emission lines. While photometric noise clearly dominates the uncertainty of photometric redshift estimates, the key findings were that emission lines play a significant role in resolving color space degeneracies and good spectroscopic coverage of the entire color space is necessary to achieve good results with empirical photo-z methods. Template-fitting methods, on the other hand, must use a template set with sufficient variation in emission line strengths and ratios, or even better, first estimate the redshift empirically and fit the colors with templates at the best-fit redshift to calculate the K-correction and various physical parameters. This research was part of Géza Csörnyei former MSc student's (advisor L. Dobos) MSc thesis [26] work that was started during his visit at partner institute JHU in 2019. Based on his excellent work G. Csörnyei got a PhD position at International Max Planck Research School on Astrophysics. The results were published in MNRAS [4]

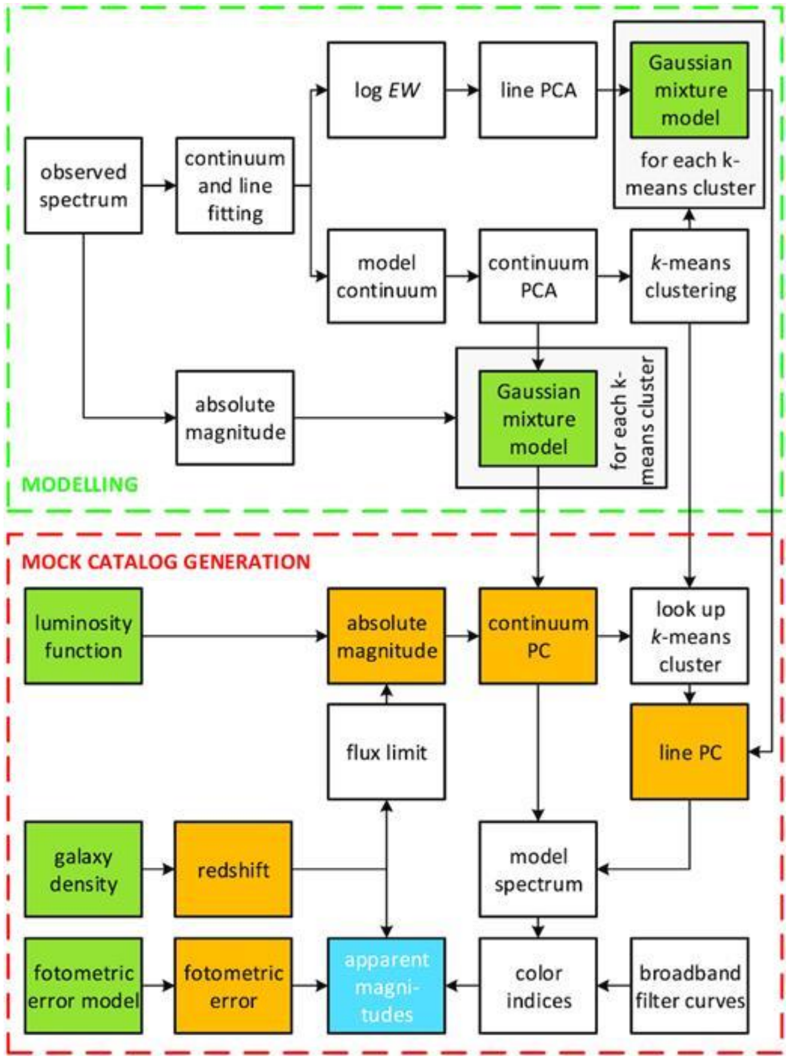

Fig. 6. *The method of mock catalogue generation for galaxy spectral models with emission lines. The upper panel shows how we model the continuum and emission lines of the observed galaxies. The lower panel explains how model spectra are generated by drawing samples from the theoretical distributions and fitted stochastic models [4].*

Based on our previous photometric redshift estimation algorithm and database technology developments we have finalized and published the largest photometric redshift catalog a 3D map of the Universe. Previously, the largest map of the universe was created by the Sloan Digital Sky Survey (SDSS), which covers only one-third of the sky. The new catalog, based on the Pan-STARRS1 (Panoramic Survey Telescope and Rapid Response System) 3π survey which is the world's largest deep multi-color optical survey, spanning three-quarters of the sky. It doubles the area surveyed, has greater statistics, and contains specific areas the SDSS missed. Utilizing a state-of-the-art optimization algorithm, we leveraged the spectroscopic training set of almost 4 million light sources from various catalogs to teach the neural network to predict source types and galaxy distances, while at the same time correcting for light extinction by dust in the Milky Way. The resulting 3D catalog is available as a high-level science product through the Mikulski Archive for Space Telescopes. It is approximately 300 GB in size, and science users can query the catalog through the MAST CasJobs SQL interface, or download the entire collection as a computer-readable table from https://archive.stsci.edu/hlsp/ps1-strm/. The description of the photometric redshift estimation method and details of the catalog was published by Róbert Beck et al. in MNRAS [3].

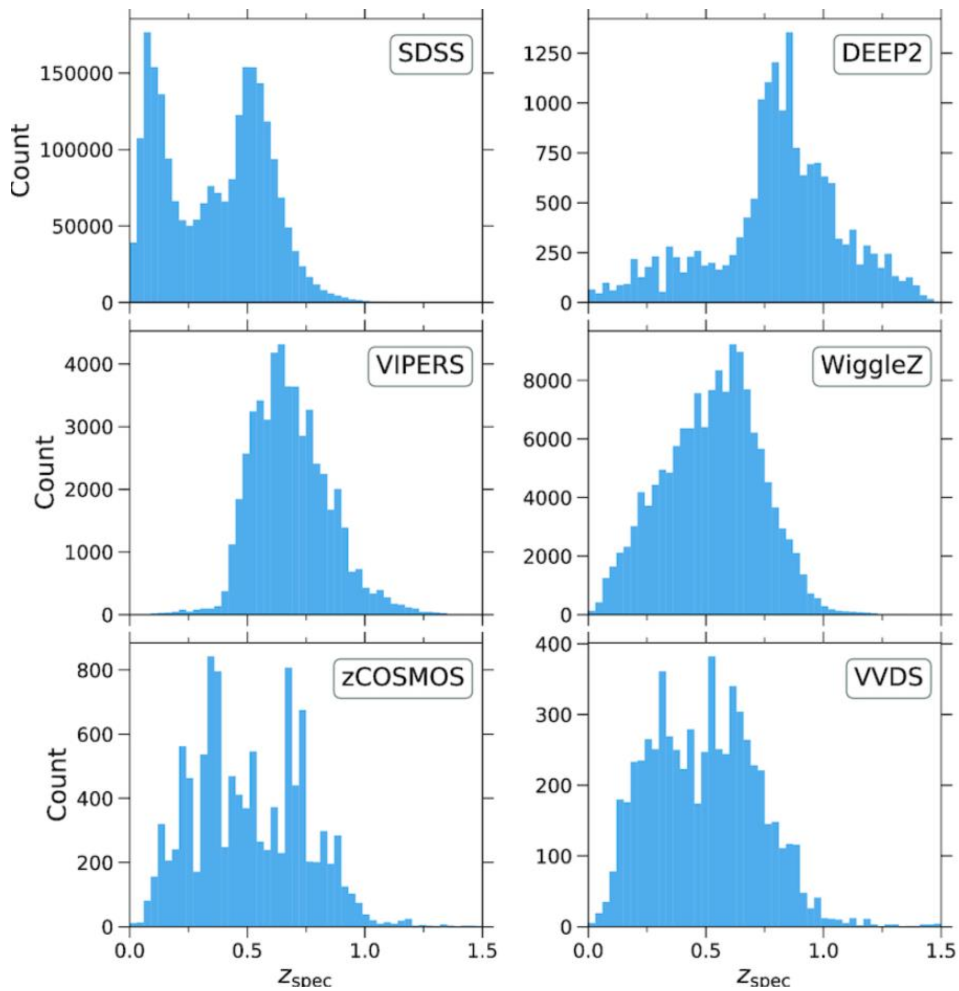

Fig. 6. *The redshift distribution of galaxies with spectroscopic observations that were used to calibrate the photometric redshift estimation in [3].*

We have also published a large cross-match catalogue between the WISE All-Sky survey and Pan-STARRS PS1 3 πDR2 sources. Since Pan-STARRS is a ground-based survey, it spans optical colors, and it is subject to inhomogeneities due to variations in day-to-day operations, weather, etc. The space-based Wide-field Infrared Survey Explorer mission (WISE) imaged the whole sky in four infrared colors. The data set generated by WISE is still the most complete in terms of infrared sky coverage. Combining these two unique datasets with the wider spectral coverage can be the basis of better photometric redshift estimation and several cosmological studies involving the large scale structure of the universe. The resulting catalogue has more than 350 million objects, though significantly fewer than the parent PS1 catalogue, but its combination of optical and infrared colors facilitates both better source classification and photometric redshift estimation. Using these data, we also performed a neural network-based classification of the objects into galaxies, quasars, and stars, then run neural network-based photometric redshift estimation for the galaxies. Thanks to the information richness from the multiple bands, the quasar completeness jumped from 81% to 94 .6% and star purity also increased to 99.5% from 97.1%. The resulting galaxy photo-z's are significantly more accurate in terms of statistical scatter and bias than those calculated from PS1 properties alone. The redshift-dependent bias also decreased substantially, in particular, at lower redshifts. The improvements are partly due to the increased number of colours, extending into the infrared, and partly to our enhanced methodology: we used hyperparameter optimization for our neural networks, and revised the loss function definition. The catalogue will be a basis for future large-scale structure studies and was made available as a high-level science product via the Mikulski Archive for Space Telescopes [6]. The description of the source classification is published in MNRAS [5].

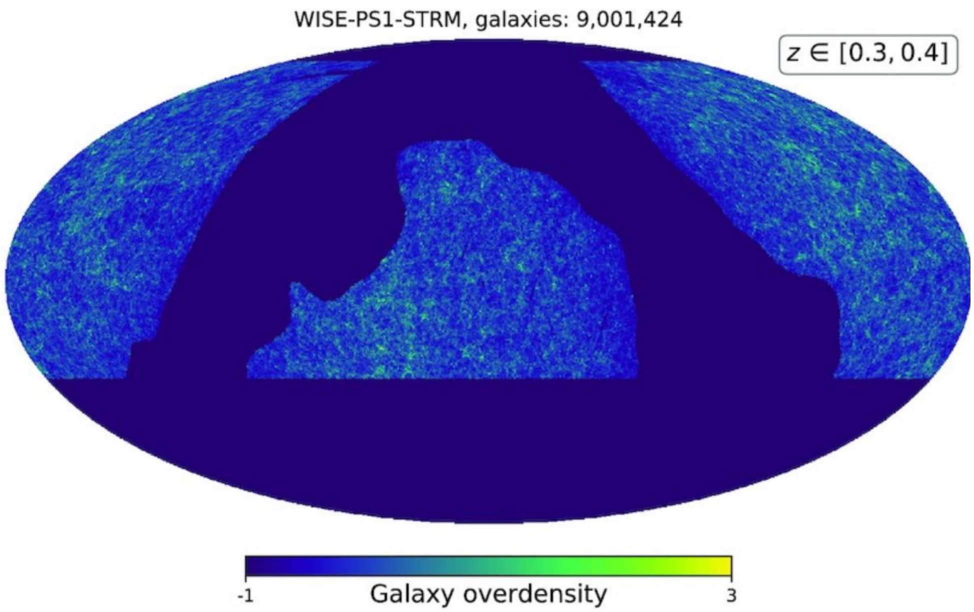

Fig. 7. *Galaxy overdensity in the cross-matched PS1 3π DR2 source catalogue and the WISE All-Sky catalogue [5].*

Three-dimensional deep wide-field galaxy surveys are fundamental for cosmological studies. At higher redshift ranges ($z > 1$) where galaxies are too faint, quasars still trace the large-scale structure of the Universe. The $z>1.3$ redshift range is especially interesting for testing the anomalous integrated Sachs-Wolfe (ISW) effect predictions of our AvERA cosmological model [Beck et al. MNRAS (2018) 479 (3), 3582–3591] and comparing it to other models. Since available telescope time limits spectroscopic surveys, only photometric methods are feasible for estimating redshifts for many high redshift quasars in large areas. Machine learning methods are increasingly successful for quasar photometric redshift estimation also; however, they hinge on the distribution of and possible biases in the training set. Therefore, a rigorous estimation of reliability is critical. We extracted optical and infrared photometric data from the cross-matched catalogue of the WISE All-Sky and PS1 DR2 sky surveys (Beck et al. 2022 MNRAS [5,6], described above). We used an XGBoost machine learning model as a base line method and then a more advanced artificial neural network model for the final predictions. We approximated the effective training set coverage with the K nearest neighbors algorithm, hence selected object for which reliable redshift estimation is feasible. We estimated photometric redshifts for 2.5 million quasars from the WISE-PS1 cross-matched catalogue, validated the derived redshifts with an independent, clustering-based redshift estimation technique. A good accordance was found between the results of the two methods up to z ~ 2.5 therefore the published catalog will be useful for several cosmological large-scale structure studies, including our ISW investigations. A detailed analysis of the results and an explanation for the remaining biases were also performed and the final catalog was made publicly available. The paper on details of the analysis and the catalog were published in MNRAS [7] and the research was part of PhD thesis work of Sándor Kunsági-Máté [32].

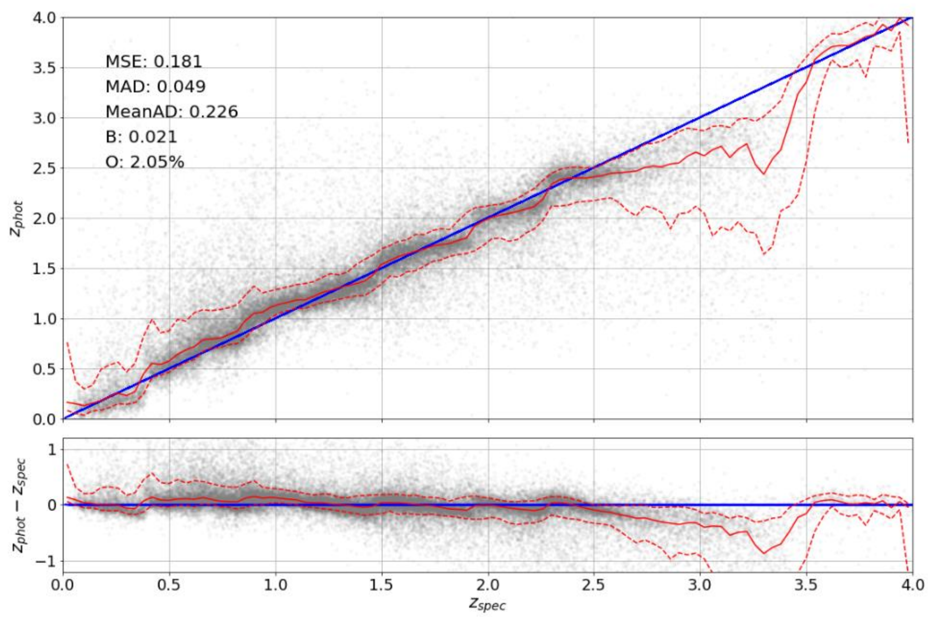

Fig. 8. *Photometric redshift estimation calibration for the WISE-PS1 cross-matched quasar catalogue [7].*

Cosmological N-body simulations

As one of our central goals, we developed a new cosmological N-body simulation code that overcomes the limitation of other state-of-the-art codes that use periodic boundary conditions, limiting the scale of the simulations and also imposing non-physical anisotropies (see later in [10]). Our multi-GPU realization of the StePS (Stereographically Projected Cosmological Simulations) algorithm is built upon MPI–OpenMP–CUDA hybrid parallelization and nearly ideally scales-out to multiple compute nodes. The key novelty of this new zoom-in cosmological direct N-body simulation method is that it simulates the infinite universe with unprecedented dynamic range for a given amount of memory and, in contrast to traditional periodic simulations, its fundamental geometry and topology match observations. By using a spherical geometry instead of periodic boundary conditions, and gradually decreasing the mass resolution with radius, our code is capable of running simulations with a few gigaparsecs in diameter and with a mass resolution of in the center in four days on three compute nodes with four GTX 1080 Ti GPUs in each. The code can also be used to run extremely fast simulations with reasonable resolution for fitting cosmological parameters, we have made extensive benchmarking on the JHU MARCC GPU cluster at our partner institute. These simulations are useful for prediction needs of large surveys and as a demonstration we were able to re-run a Millennium simulation scale sample. The StePS code is publicly available for the research community, the paper summarizing the method was published in Astronomy and Computing [8] and was part of PhD thesis work of Gábor Rácz [25]. Gábor had the opportunity to visit and do research at both of our partner institutes, The Johns Hopkins University and University of Hawaii. After completion of his PhD, based on his experience in N-body simulations and alternative cosmological models he got postdoc position at NASA Jet Propulsion Laboratory, Pasadena. As the acceptance letter says, "*NASA Postdoctoral Program is extremely selective, and this Fellowship is offered to you in recognition of your highly ranked academic and scientific achievement*". We consider this as a positive "performance metric" of our project and fruit of the efforts of building up cosmological simulation expertise in the group. At JPL Gábor leads the alternative cosmology simulation studies for the Euclid survey mission.

For more efficient handling of large number of multidimensional point-cloud data, like the results of cosmological N-body simulations we have developed and evaluated a distributed point-cloud database that is able to perform fast box- and nearest neighbor spatial queries. The method was described in a short paper published in Annales Univ. Sci. Budapest., Sect. Comp. [9]. Another PhD student, János Szalai-Gindl has completed and defended his PhD thesis, titled "Data-intensive methods for managing scientific data with cosmological, GIS and biological applications". Two thesis points were related to optimal multidimensional point-cloud databases (like the snapshots of N-body simulations) and one to hierarchical Bayesian model based CUDA implementation of galaxy luminosity function estimation. [27]

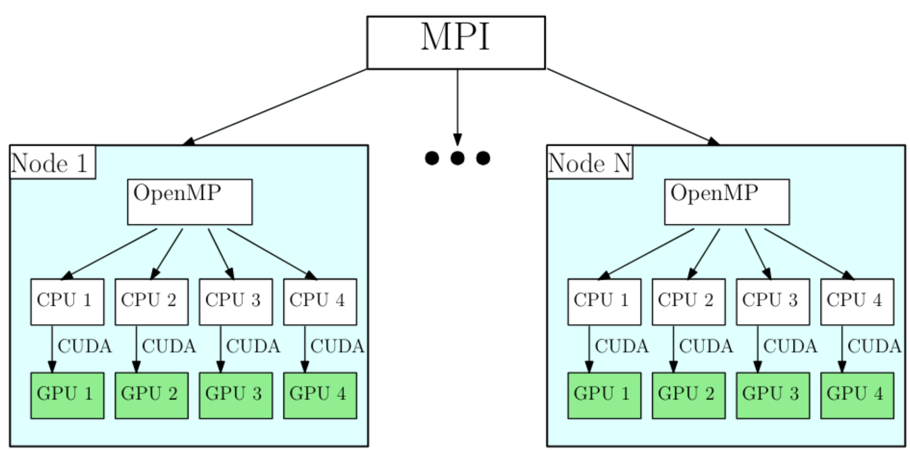

Fig. 9. *Force calculation for the StePS algorithm in a GPU cluser with MPI-OpenMP-CUDA hybrid parallelization [8].*

Related to the above-described StePS code, together with the University of Hawaii group, we have analyzed the anisotropy of the power spectrum in periodic n-body simulations. In short, most of the current cosmological n-body codes use periodic boundary conditions to avoid the finite size edge effects. But the classical gravitational force on a torus topology is anisotropic and always smaller than Newton's $1/r^2$ law. This is a less known effect as this is not present in molecular dynamics simulations where the effective interaction distance is finite due to the quickly decaying Van der Waals force. This contrasts with the effectively infinite range gravitational interaction and ion top of this the matter distribution is inhomogeneous over several scales that further enhance the anisotropy. In the analysis we demonstrated the effects of periodicity in dark matter only n-body simulations of spherical collapse and standard ΛCDM initial conditions. Periodic boundary conditions cause an overall negative and anisotropic bias in cosmological simulations of cosmic structure formation. The lower amplitude of power spectra of small periodic simulations is a consequence of the missing large-scale modes and the equally important smaller periodic forces. The effect is most significant when the largest mildly non-linear scales are comparable to the linear size of the simulation box, as often is the case for high-resolution hydrodynamic simulations. Spherical collapse morphs into a shape similar to an octahedron and the anisotropic growth distorts the large-scale ΛCDM dark matter structures. We introduced a direction-dependent power spectrum invariant under the octahedral group of the simulation volume and have shown that the results break spherical symmetry. The results further emphasize the importance of boundary-free simulations, like our StePS method. The results of the study were published in MNRAS [10] and were part of PhD thesis work of Gábor Rácz [25]. We have also modified an n-body code to be able to handle non-unique mass objects and studied the effects of non-uniformity on large scale structure formation. This was part of the thesis work of a that time MSc, and now PhD student Balázs Pál [30].

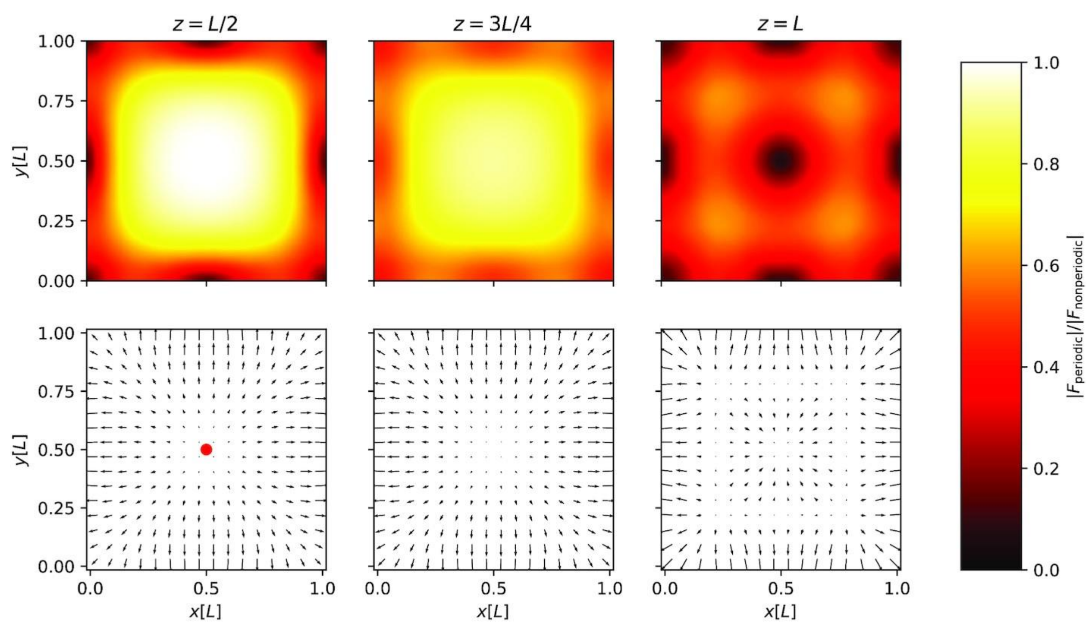

Fig. 10. *Comparison between the free Newtonian and the periodic gravitational force calculated with the Ewald method [25].*

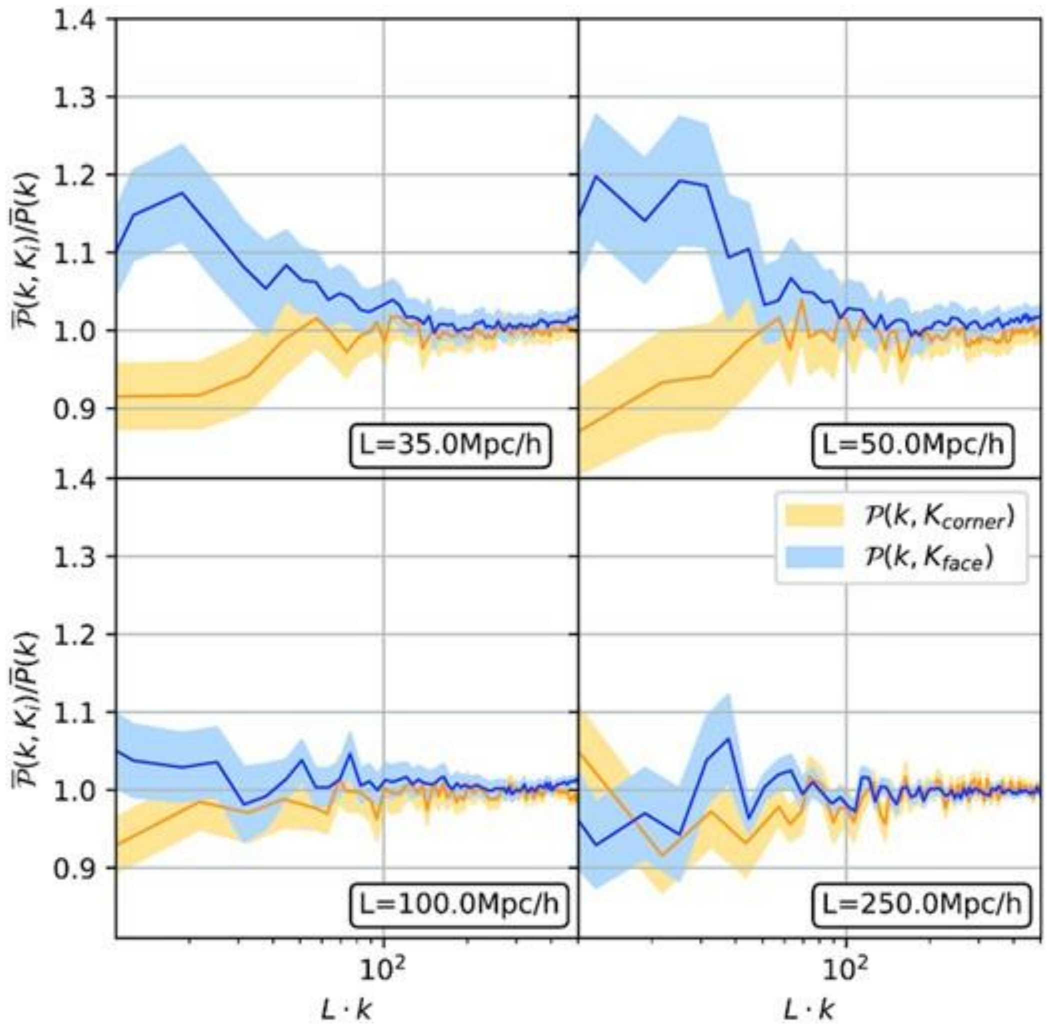

Fig. 11. *The ratios of the direction-dependent power spectra and the direction-independent P(k) spectrum of the same simulation series, for four different linear box sizes at z = 0. The spectra are averaged over directions within regions of the fundamental triangle of the simulation cube, corresponding to the wavenumber vectors pointing towards the corner and face directions [25].*

N-body simulations are done as a realization of cosmological models and their statistical analysis makes it possible to compare them to the observed cosmic density fields and verify the underlying cosmological model. As simulation distributions and observed distributions cannot be matched directly two-point statistics of the cosmic density field are the principal tools for constraining cosmological models. Large-scale galaxy surveys, such as the SDSS, 2MASS, DESI, PanSTARRS or DES provide data for such measurements. In the near future even larger surveys are planned or started (e.g., Euclid, WFIRST/Roman, SPHEREx, LSST/Rubin, Subaru Prime-Focus Spectrograph). While some day we might map the cosmological density field of the observable universe, at present these wide field surveys are complemented with narrower deep surveys. The geometry and volume of the survey window and the super-survey modes modulate the interpretation of any measured two-point statistics. The effects of the survey window are described in detail in several papers and the effects of super-survey modes are usually treated as overdensity and tidal effects. These are based on standard perturbation theory or the halo-model. The biasing effect is noticeable even in large volumes but becomes large when only small volume available. The state-of-the-art methods provide precise results in most cases at the expense of considerable complexity and they usually ignore the effects of the non-Gaussianity of the density distribution. We decided to quantify the non-linear overdensity response for finite volumes beyond linear and second order SPT, to show the effect of the density field distribution, and to provide an easy-to-use use yet accurate fit as a function of cosmological parameters. In our study we measured the dark matter power spectrum amplitude as a function of the overdensity in N-body simulation subsamples. We showed that the response follows a power law form, and we provided a new fit in terms of the variance of a sub-volume. Our fit has similar accuracy and comparable complexity as second order standard perturbation theory on large scales, but it is valid for non-linear (smaller) scales, where perturbation theory needs higher order terms for comparable precision. Furthermore, we have shown that the lognormal nature of the overdensity distribution causes a previously unidentified bias: the power spectrum amplitude for a sub-sample with average density is typically underestimated. Our result opens the road towards precise measurement and interpretation of the power spectrum from smaller surveys and simulation subsamples. In particular, zoom-in simulations, and our recently published compactified multiresolution StePS simulations, will benefit from these analyses. The results are published in A&A [11].

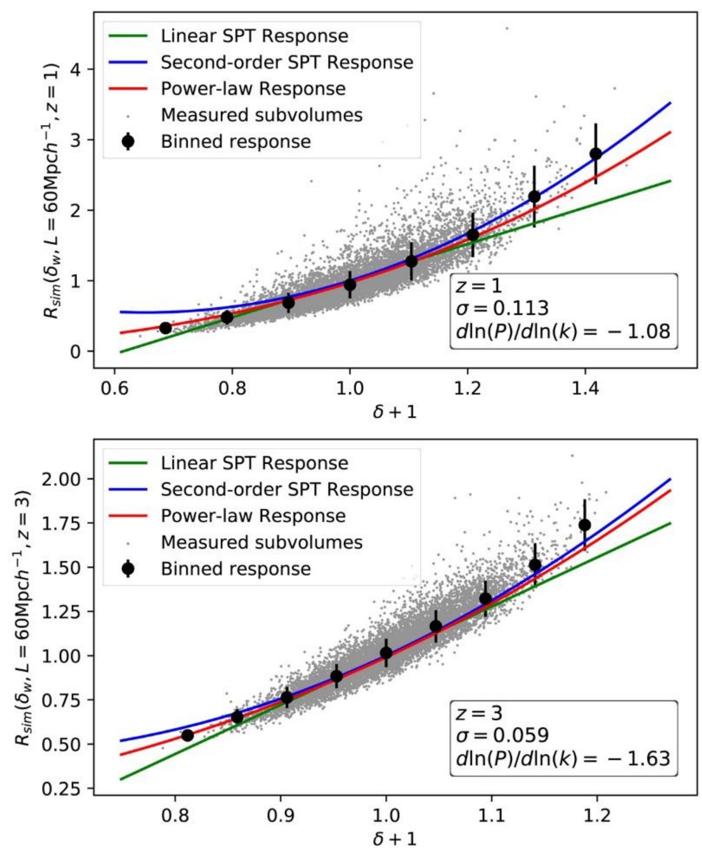

Fig. 12. *Simulated and theoretical* $R_{sim}(\delta_w, \sigma, z)$ *responses in ΛCDM cosmology for the* $L = 60\,\mathrm{Mpc}\,h^{-1}$ *window size at* $z = 1$ *and 3 redshifts. Our new power-law response function fits the simulated responses well, especially in the low-density regions. At higher redshifts, the SPT and power-law responses converge as the density field becomes more Gaussian [11].*

The n-body simulation process is in general computationally expensive since it follows the motion of billions of dark matter particles in very large volumes. Simulation sample variance that is caused by the specific distribution of initial phases of fluctuations, can grow significantly over the evolution of the simulation, and this will introduce bias in important statistics used for comparison to observations. Sample variance is typically mitigated by running hundreds of simulations, or by using extremely large volumes which can be prohibitively expensive computationally. In a study together with University of Hawaii and NASA JPL groups we have developed a new method for reducing the cosmic variance in cosmological N-body simulations. In this method, the simulations are run in matching pairs with initially compensated amplitudes and shifted phases in Fourier space. The main advantage of this method is that a complementary simulation can be generated for any existing simulation. To demonstrate the effectiveness of this novel technique, we ran the complementary pair of the famous Millennium simulation and showed how the baryon acoustic oscillation (BAO) fluctuations evolved over time in the simulations. By itself the original Millennium simulation was not large enough to resolve the baryonic acoustic oscillation (BAO) wiggles in the power spectrum due to the sample variance at the larger scales. Combining it with our new complementary simulation we were able to recover the evolution of these fine details. Using complementary pairs of simulations greatly reduces the computational cost of cosmological predictions and analyses since the evolution of the Universe can be followed using a significantly smaller volume than with a single simulation. Building on this key achievement, we will be able to compare more cosmological scenarios and alternative models with the upcoming large observations while using significantly fewer computational resources. The paper describing the results was published in A&A [12].

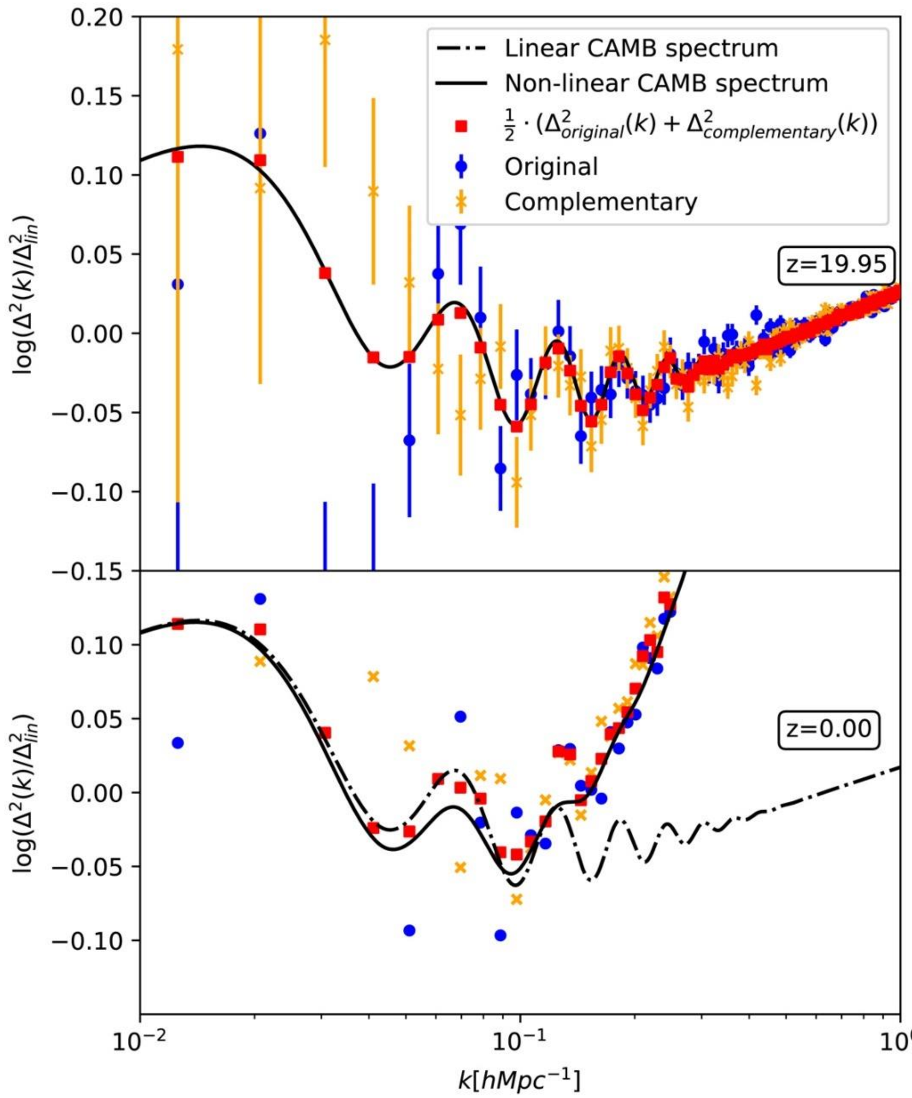

Fig. 13. *Power spectra of the dark matter density field in the Millennium run and its complementary pair at the BAO scales. All data points have been divided by a linear dark matter-only power spectrum. While the original Millennium simulation cannot resolve these scales by itself due to sample variance, the average spectrum of the original and complementary simulations (red squares) can precisely follow the evolution of the BAO features and matches well at the linear scales with the linear CAMB power spectra [12].*

Analyzing cosmological models through large surveys

During the years of the project more and more groups have published studies about the various tensions within the standard ΛCDM model, validating our original plans to investigate this issue in detail. One of the most widely debated issue is the so called “Hubble-tension”, namely that the “early” CMB based estimates (H0=67.5) and the “local” supernova (and other observations) based estimates (H0=74.0) differ by over 4 standard deviations. Another, less studied anomaly is the repeated observation of integrated Sachs-Wolfe (ISW) imprints ~5x stronger than expected in the ΛCDM model and also the related effect of large, >~ 100 Mpc/h super-structures, like the CMB cold spot. As we have previously suggested, our AvERA model, that incorporates the effects of inhomogeneity into the matter density evolution may explain these anomalies. We elaborated on the details and we have shown, that indeed the AvERA evolution is capable of providing a plausible albeit radically different story (w.r.t. the standard ΛCDM model) that explains both observational anomalies without dark energy. The manuscript in collaboration with András Kovács, that describes our results has been recently accepted in MNRAS [21].

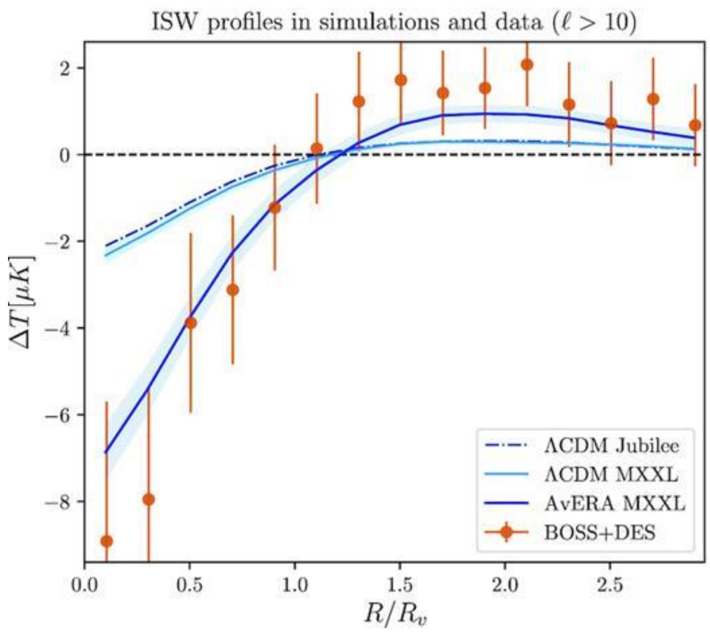

Fig. 14. *Radial profiles of ISW imprints in ΛCDM and in AvERA cosmologies are compared to observed ISW signals from a combined stacking analysis of DES and BOSS supervoids. The shaded bands around simulated results show the standard deviation of profiles measured from random ISW-only stacking measurements. The BOSS + DES data points are shown with their actual observational error bars including CMB noise that dominates the uncertainties. [21]*

Studying the high redshift galaxy clustering is especially challenging. Here due to the lack of large training sets and low signal-to-noise ratio, most photometric redshift estimation methods fail. We have performed an analysis of two-point galaxy clustering and galaxy bias using Subaru Hyper-Suprime Cam (HSC) data taken jointly by the Subaru Strategic Program and the University of Hawaii in the COSMOS field. The depth of the data is similar to the ongoing Hawaii Two-0 (H20) optical galaxy survey; thus the results are indicative of future constraints from tenfold area. We have measured the angular auto-power spectra of the galaxy overdensity in three redshift bins, defined by dropouts from the g-, r- and i-bands, and compared them to the theoretical expectation from concordance cosmology with linear galaxy bias. We have determined the redshift distribution of each bin using a standard template-based photometric redshift method, coupled with a self-organizing map (SOM) to quantify colour space coverage. We also investigated sources of systematic errors to inform the methodology and requirements for Hawaii Two-0. The results were published in MNRAS [20].

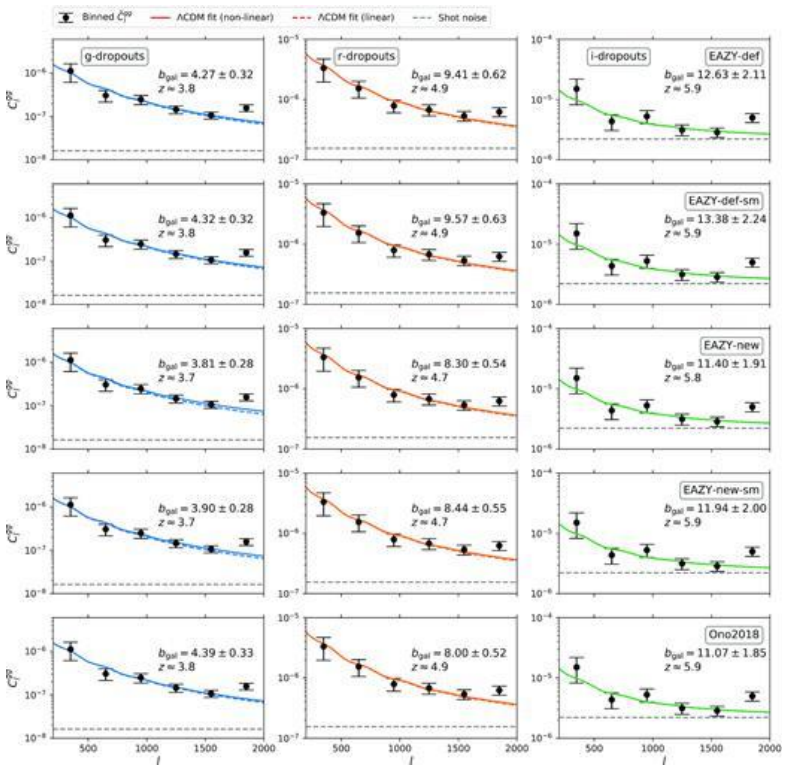

Fig. 15. *Theoretical model fits to the spherical autocorrelation power spectra of the g-, r-, and i-dropout galaxy samples [20].*

As a key study w.r.t. deviations from the standard cosmological scenario we analyzed the un-probed key redshift range $0.8 < z < 2.2$ where the ISW signal is expected to fade in ΛCDM, due to a weakening dark energy component, and eventually become consistent with zero in the matter dominated epoch. On the other hand, one of the key observables of our AvERA model - that incorporates the effects of inhomogeneity into the matter density evolution - is an anomalous integrated Sachs-Wolfe (ISW) effect. As we (Beck et al. MNRAS 479(3), pp.3582-3591. 2018) have shown previously, at low and moderate redshifts ($z < 1$) the AvERA cosmology predicts ISW amplitude that is much larger than for the standard ΛCDM model. Even more strikingly, at larger redshifts ($z >\sim 1.4$) the sign of the ISW, that is always positive in ΛCDM, becomes negative due to the possible growth of large-scale potentials that is absent in the standard model. As described above and in Kovács et al. [21] for moderately low redshift ranges and also for the largest super-void statistics we have found better agreement with the AvERA than with the ΛCDM model. To discriminate the cosmological models at the higher redshift regime, we estimated the high-z ΛCDM ISW signal using the Millennium XXL mock catalogue, and compared it to our measurements from about 800 super-voids identified in the eBOSS DR16 quasar catalogue. At $0.8 < z < 1.2$, we found an excess ISW signal with $A_{ISW} \approx 3.6 \pm 2.1$ amplitude. The signal is then consistent with the ΛCDM expectation (AISW = 1) at $1.2 < z < 1.5$ where the standard and AvERA model predicts similar amplitudes. Most interestingly, we also observed an opposite-sign ISW signal at $1.5 < z < 2.2$ that is in 2.7σ tension with the ΛCDM prediction. Taken at face value, these recurring hints for ISW anomalies suggest an alternative growth rate of structure in low-density environments at ~ 100 Mpc/h scales. To further enhance the significance of the results we need larger cosmic volume that is not available in form of spectroscopic surveys, so quasar photometric redshift catalogs, like the one described above and in Kunsági-Máté et al. [7] will be needed. The high-z ISW study was published in MNRAS [22].

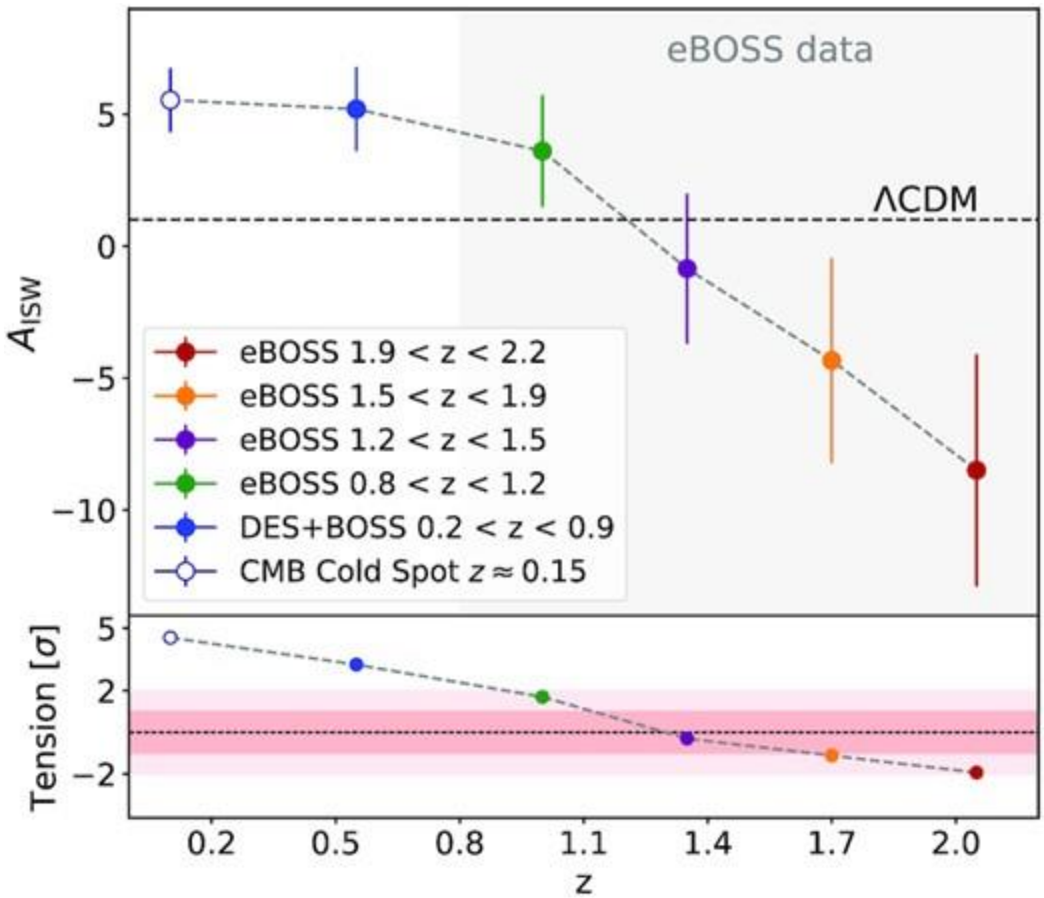

Fig. 16. *Measured ISW amplitudes with emerging trends. At $z \lesssim 1.5$, multiple observational results point to an enhanced positive ISW amplitude. In contrast, our new eBOSS results showed a large negative ISW amplitude at $z \gtrsim 1.5$. The bottom panel shows the estimated tension with the baseline ΛCDM model predictions (shaded bands correspond to 1σ and 2σ). The higher low redshift and negative high redshift values are consistent with the prediction of our AvERA model [22].*

As an extension of our work on large scale structure observations, in collaboration with A. Kovács and a short-term guest student, Gisela Camacho, we performed a cross-correlation of cosmic voids with the CMB lensing signal to probe the cosmological model. We selected luminous red galaxies (LRGs) from the WISE-Pan-STARSS data set, that we have created during the past few years, allowing an extended analysis using 14,200 deg^2 sky area which is larger than in previous works (~ 5,000 deg^2). We created both 2D and 3D void catalogs to cross-correlate their locations with the Planck CMB lensing map and studied their average imprint signal using a stacking methodology. Applying the same procedure, we also generated a mock galaxy catalog from the WebSky simulation to serve as a basis for comparison. The 2D void analysis revealed good agreement with the standard cosmological model with A_κ = 1.05 +/- 0.08 amplitude, i.e. S/N=13.2, showing a higher signal-to-noise than previous studies using DES voids. The 3D void analysis exhibited a lower signal-to-noise ratio and demonstrated worse agreement with our mock catalog compared to the 2D voids. These deviations could be attributed to limitations in the mock catalog, such as imperfections in the LRG selection, as well as a potential asymmetry between the North and South patches. The methodology that we employed can be applied to different catalogs for cosmological measurements, and given the unique nature of cosmic voids, these future results hold great promise. The paper describing the results was accepted for publication in A&A [24].

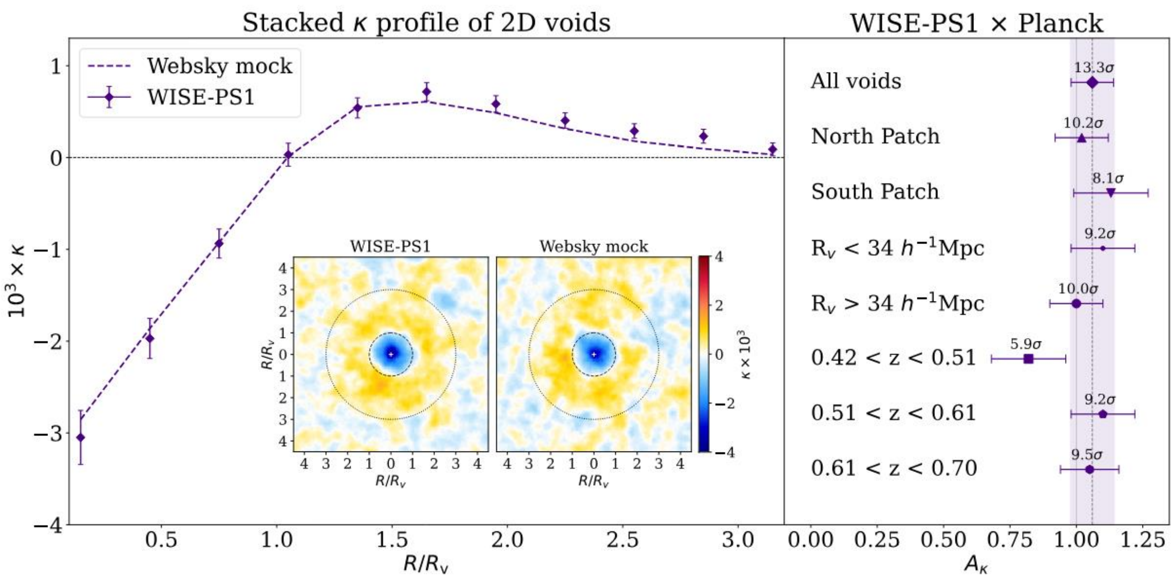

Fig. 17. *Stacked CMB κ radial profile and image of the Websky mock and WISE-PS1 catalogs using all 2D voids. The κ profiles show great agreement, with an 6% higher than expected signal at the center, and a clear negative κ value inside of the voids, followed by a clear detection of the void edges in the lensing signal too [24].*

Regarding the international collaboration, during the project period we had regular weekly Budapest-Baltimore-Honolulu-Pasadena online meetings and a weekly in person meeting with local researchers and students. We extended the collaboration with András Kovács, a former research fellow in Tenerife, who has recently returned to Hungary with a "Lendület" grant. András is participating in several large surveys (DES, J-PAS, Euclid, with priority data access). His students had joined many of our discussions and some joint work resulted in publications. Project partner István Szapudi, from the University of Hawaii, Institute for Astronomy, has spent several months and part of his sabbatical year in Budapest. This helped to deepen our collaboration, explore further research directions, and creates a good opportunity for our students to meet him in person. Our students had the opportunity to visit research partners at The Johns Hopkins University and University of Hawaii. Through Gábor Rácz we have established a new perspective collaboration with NASA JPL where Gábor became key personnel in coordinating the alternative cosmology simulations for Euclid Consortium. He and his EC group developed a new pipeline to generate scientific results from a suite of cosmological N-body simulations of non-standard models including dynamical dark energy, interacting dark energy, modified gravity, massive neutrinos, and primordial non-Gaussianities. During the project period Hungary became a member of the Euclid collaboration that has recently started observation and published the first overview papers [23]. The soon to be collected data will be crucial to our future cosmological analyses and constraining cosmological models.

References

1. Beck, R., Glaser, Y., Sadowski, P. and Szapudi, I.: Refined Redshift Regression in Cosmology with Graph Convolution Networks, 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver Canada, December 2019
2. D. Ribli: Exploring domain adaptation with generative adversarial networks for galaxy images, Kavli Institute Summer School Report, 2019
3. Beck, R., Szapudi, I., Flewelling, H., Holmberg, C., Magnier, E. and Chambers, K.C.: PS1-STRM: Neural network source classification and photometric redshift catalogue for PS1 3π DR1., MNRAS doi:10.1093/mnras/staa2587, 2021
4. Csörnyei, G., Dobos, L. and Csabai, I.: The effect of emission lines on the performance of photometric redshift estimation algorithms, Monthly Notices of the Royal Astronomical Society, 502(4), pp.5762-5778., 2021
5. Beck, Róbert, S. Curt Dodds, and István Szapudi: WISE-PS1-STRM: neural network source classification and photometric redshifts for WISE× PS1, Monthly Notices of the Royal Astronomical Society 515, no. 4 (2022): 4711-4721., 2022
6. István Szapudi, Robert Beck, Curt Dodds: WISE + Pan-STARRS1 Source Types and Redshifts with Machine Learning (WISE-PS1-STRM), MAST archive at Space Telescope Science Institute, 2022
7. Kunsági-Máté, Sándor, Róbert Beck, István Szapudi, and István Csabai: Photometric redshifts for quasars from WISE-PS1-STRM, Monthly Notices of the Royal Astronomical Society, (2022) Volume 516, Issue 2, Pages 2662–2670, 2022
8. Rácz, G., Szapudi, I., Dobos, L., Csabai, I. and Szalay, A.S.: StePS: A multi-GPU cosmological N-body Code for compactified simulations., Astronomy and Computing, 28, p.100303., 2019
9. Csabai, I., Dobos, L., Kiss, A., Szalai-Gindl, JM.: LOAD BALANCING IN DISTRIBUTED POINT CLOUD DATABASES, Annales Univ. Sci. Budapest., Sect. Comp. 49 97–116, 2019
10. Rácz, G., Szapudi, I., Csabai, I. and Dobos, L.: The anisotropy of the power spectrum in periodic cosmological simulations, Monthly Notices of the Royal Astronomical Society, 503(4), pp.5638-5645., 2021
11. Rácz, Gábor, István Szapudi, and István Csabai: An empirical nonlinear power spectrum overdensity response, Astronomy & Astrophysics 661 (2022): A96., 2022
12. G. Rácz, A. Kiessling, I. Csabai, I. Szapudi: Complementary Cosmological Simulations, Astronomy & Astrophysics, 672, p.A59., 2023
13. Ribli, D., Dobos, L., Csabai, I.: Galaxy shape measurement with convolutional neural networks, MNRAS Volume 489, Issue 4, November 2019, Pages 4847–4859, 2019
14. Ribli, D., Pataki, B.Á. and Csabai, I.: An improved cosmological parameter inference scheme motivated by deep learning, Nature Astronomy, 3(1), p.93., 2019
15. Ribli, D., Pataki, B.Á., Matilla, J.M.Z., Hsu, D., Haiman, Z. and Csabai, I.: Weak lensing cosmology with convolutional neural networks on noisy data., MNRAS Volume 490, Issue 2, Pages 1843–1860, 2019

16. Kunsági-Máté Sándor: Decomposition of stellar populations in simulated disk galaxies using the TensorFlow/Keras framework, XXXI ADASS Conference (24-28 October 2021, Cape Town, South Africa), 2021
17. Kunsági-Máté, Sándor, and István Csabai: Decomposition of stellar populations in CosmoDC2 galaxies using SCARLET and Deep Learning., Monthly Notices of the Royal Astronomical Society 512, no. 1 (2022): 1045-1, 2022
18. Balazs Pal, Laszlo Dobos: Denoising low resolution stellar spectra with neural networks, submitted, 2024
19. L. Dobos, A.S. Szalay, T. Budavari, E.N. Kirby, R.H. Lupton and R.F.G. Wyse: Uncertainty of line-of-sight velocity measurement of faint stars from low and medium resolution optical spectra, ApJ submitted, 2024
20. Beck, R., McPartland, C., Repp, A., Sanders, D. and Szapudi, I.: Hawaii Two-0: high-redshift galaxy clustering and bias, MNRAS 493(2), pp.2318-2328., 2020
21. Kovács, A., Beck, R., Szapudi, I., Csabai, I., Rácz, G. and Dobos, L.: A common explanation of the Hubble tension and anomalous cold spots in the CMB, MNRAS 499(1), pp.320-333, 2020
22. Kovács, A., Beck, R., Smith, A., Rácz, G., Csabai, I. and Szapudi, I: Evidence for a high-z ISW signal from supervoids in the distribution of eBOSS quasars, Monthly Notices of the Royal Astronomical Society, 513(1), pp.15-26., 2022
23. Euclid Collaboration: Y. Mellier et al. including A. Kovacs, G. Racz, I. Szapudi: Euclid. I. Overview of the Euclid mission, Astronomy & Astrophysics, submitted, 2024
24. G. Camacho-Ciurana, P. Lee, N. Arsenov, A. Kovács, I. Szapudi, I. Csabai: The CMB lensing imprint of cosmic voids detected in the WISE-Pan-STARRS luminous red galaxy catalog, Astronomy & Astrophysics, accepted, 2024
25. Gábor Rácz: Kozmológiai modellek vizsgálata numerikus szimulációkkal, ELTE PhD thesis, 2020
26. Géza Csörnyei: Generalized multi-wavelength classification of galaxies based on emission lines, ELTE MSc thesis, 2020
27. János M. Szalai-Gindl: Adatintenzív módszerek tudományos adatok kezelésére kozmológiai, térinformatikai és biológiai alkalmazásokkal, ELTE PhD thesis, 2020
28. Mirkó Mocskonyi: Deep learning alapú kozmológiai vöröseltolódás-becslés fotometriai képek alapján, ELTE BSc thesis, 2020
29. Horváth Bendegúz (advisor: Csabai István): Galxiscsoportok kozmológiai vöröseltolódás becslésének javítása gépi tanulási módszerekkel, ELTE TTK MSc thesis, 2021
30. Pál Balázs (advisor: Csabai István): Variábilis részecsketömegű kozmológiai szimulációk elemzése, ELTE TTK MSc thesis, 2021
31. Ribli Dezső (advisor: Csabai István): MESTERSÉGES NEURÁLIS HÁLÓZATOK ALKALMAZÁSA ÉS ELEMZÉSE ADATINTENZÍV TUDOMÁNYOS PROBLÉMÁKBAN, ELTE TTK PhD dissertation, 2021
32. Kunsági-Máté, Sándor: Machine Learning in Astronomy and Computational Finance ELTE PhD thesis, 2023