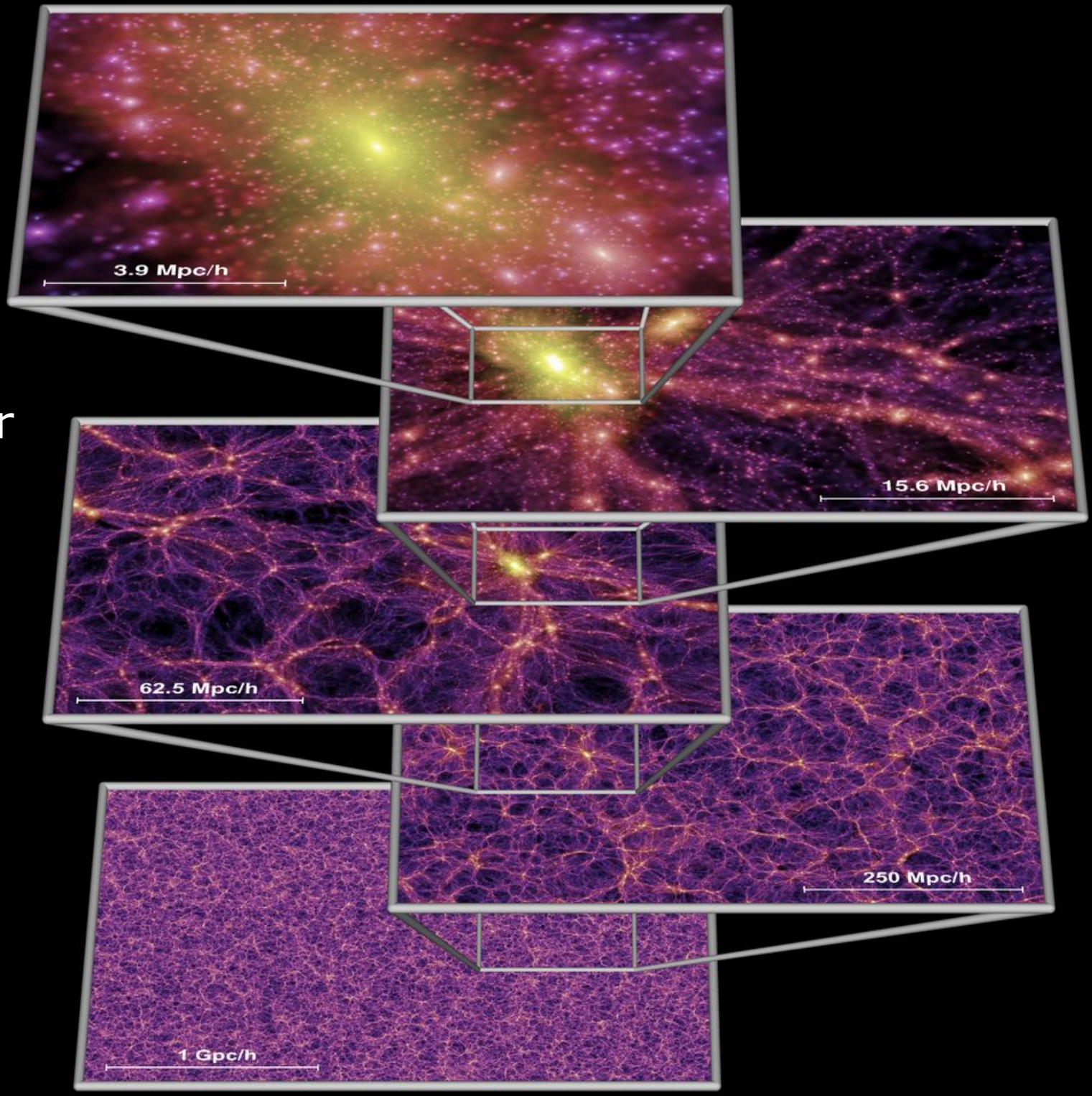
The Universe in a Supercomputer

Volker Springel

Max-Planck-Institute for Astrophysics



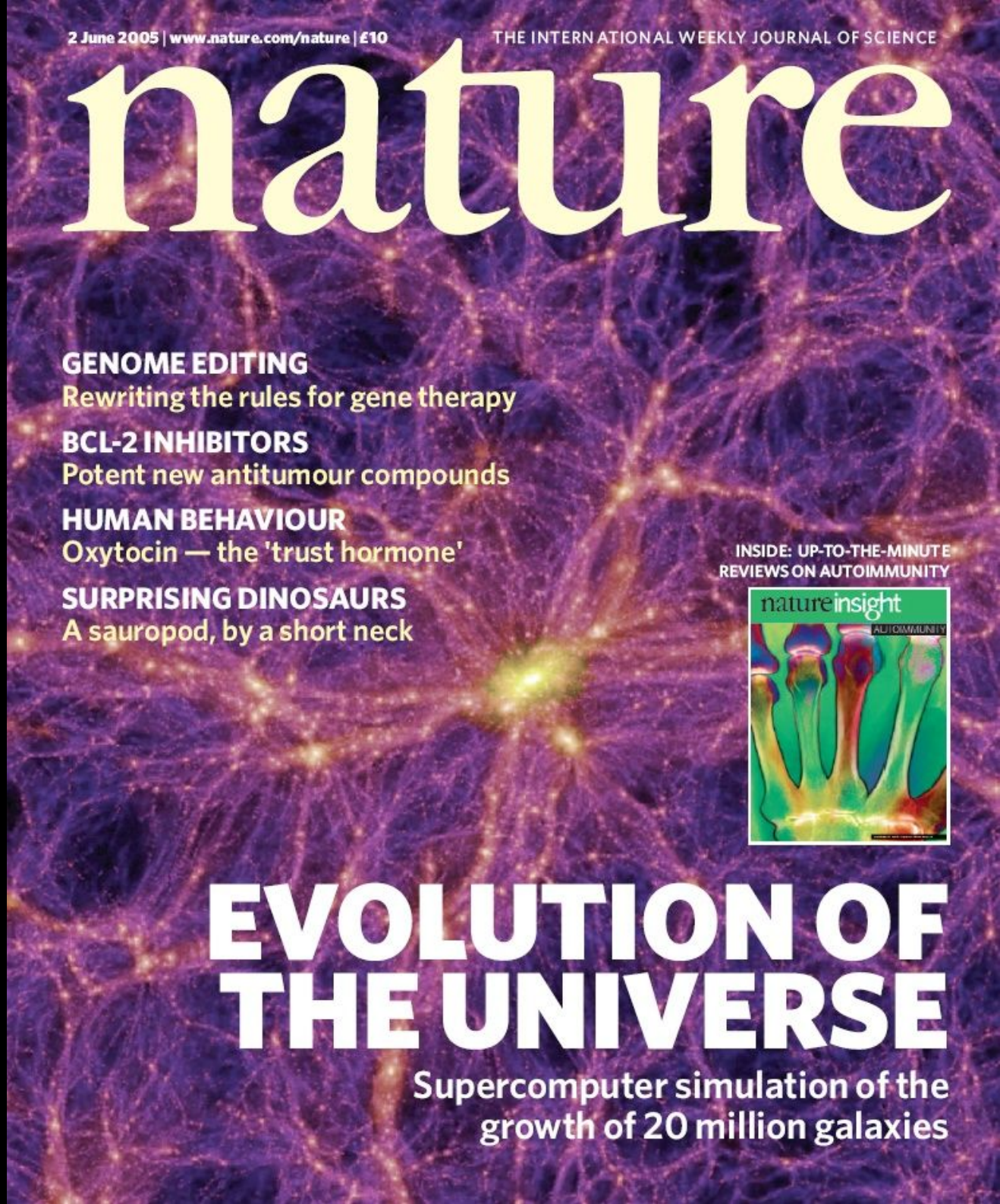
The universe in a supercomputer: From the first quasars to the large-scale galaxy distribution

Volker Springel
Simon White
Adrian Jenkins
Carlos Frenk
Naoki Yoshida
Liang Gao
Julio Navarro
Robert Thacker
Darren Croton
John Helly
Shaun Cole
Peter Thomas
Hugh Couchman
August Evrard
Jörg Colberg
Frazer Pearce
Gabriella Di Lucia
Manfred Kitzbichler
Stefan Hilbert

- **The 'Millennium Run' – Pushing the computational frontier**
- **High-precision measurements of dark matter clustering**
- **Semi-analytic models in representative pieces of the universe**
- **Acoustic oscillations, Rees-Sciama effect**
- **Some code aspects**

The first article on the **Millennium Simulation** will be published in tomorrow's issue of ***Nature***

COMPUTATIONAL COSMOLOGY APPEALS TO A GENERAL AUDIENCE



The initial conditions of cosmological structure formation now **almost unambigously known**

COSMOLOGICAL SIMULATIONS AS INITIAL VALUE PROBLEM

Cosmic Microwave Background

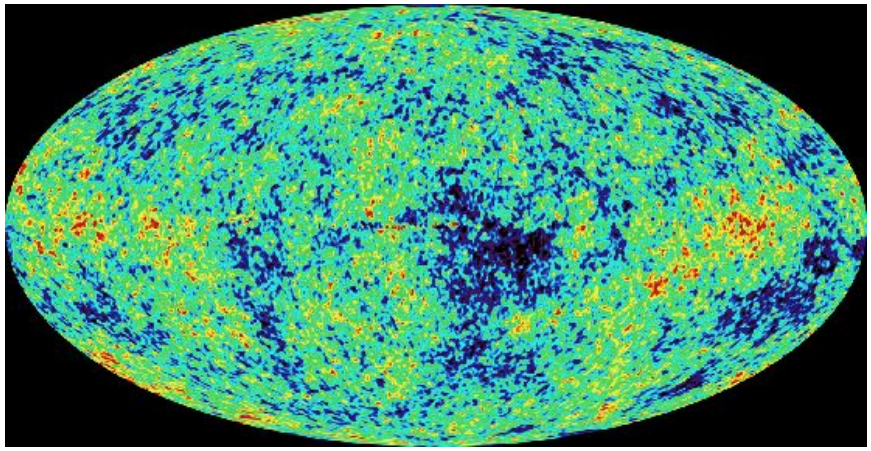
Inflation

Big-bang nucleosynthesis

Distant supernovae

Large-scale structure

Lyman-α forest

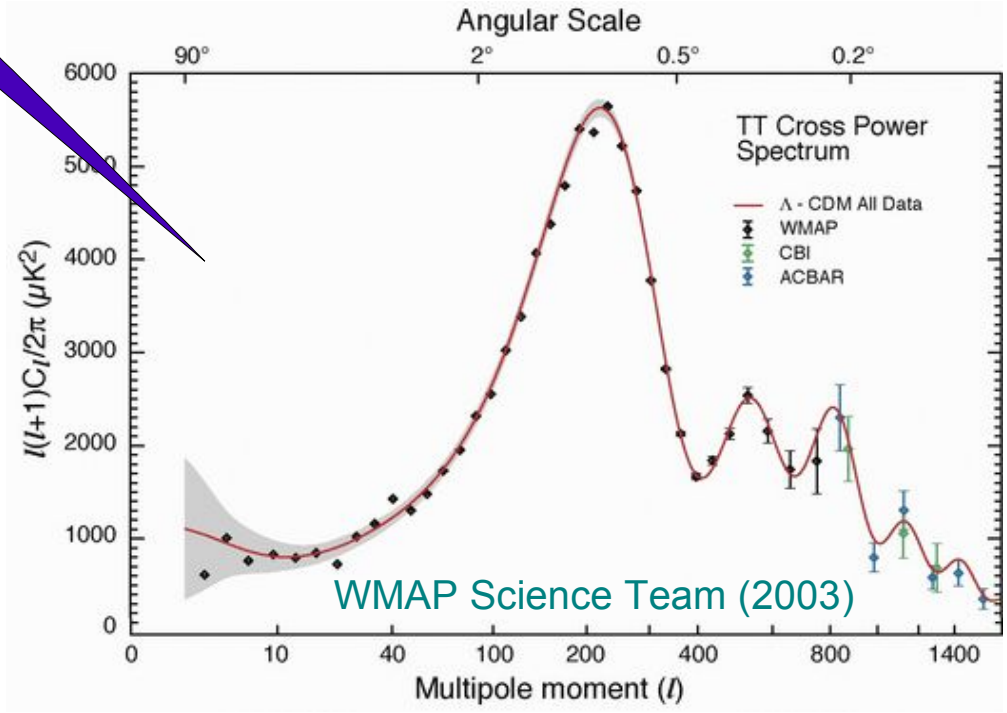


- **Consistent** cosmological parameters:

ΛCDM Modell

$$
\begin{aligned}
H_0 &= 71 \pm 4\ \mathrm{km\,s^{-1}} \\
\Omega_0 &= 0.29 \pm 0.07 \\
\Omega_b &= 0.047 \pm 0.006 \\
\sigma_8 &= 0.9 \pm 0.1 \\
t_0 &= 13.4 \pm 0.3\ \mathrm{Gyr}
\end{aligned}
$$

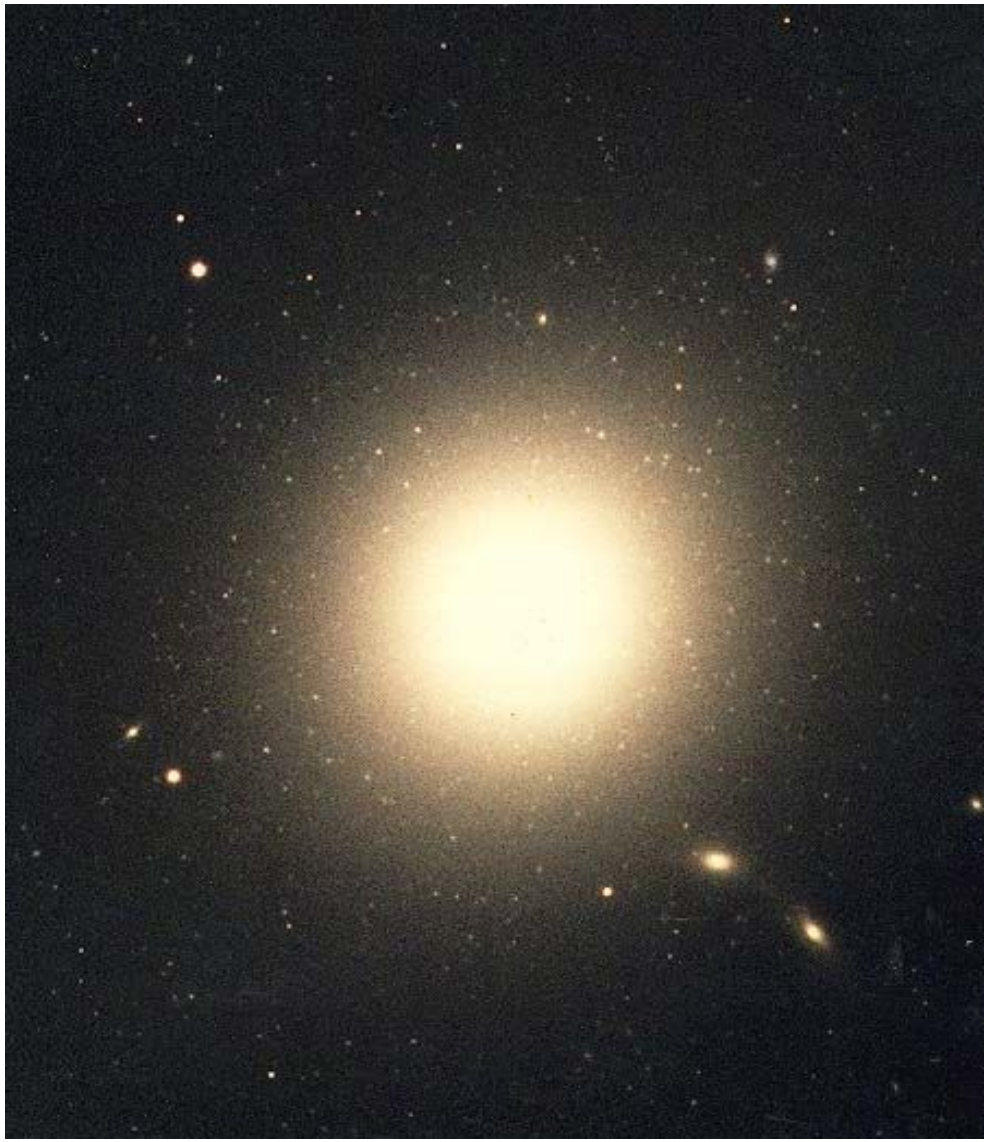
- **Initial conditions** for simulations of structure formation are **known**



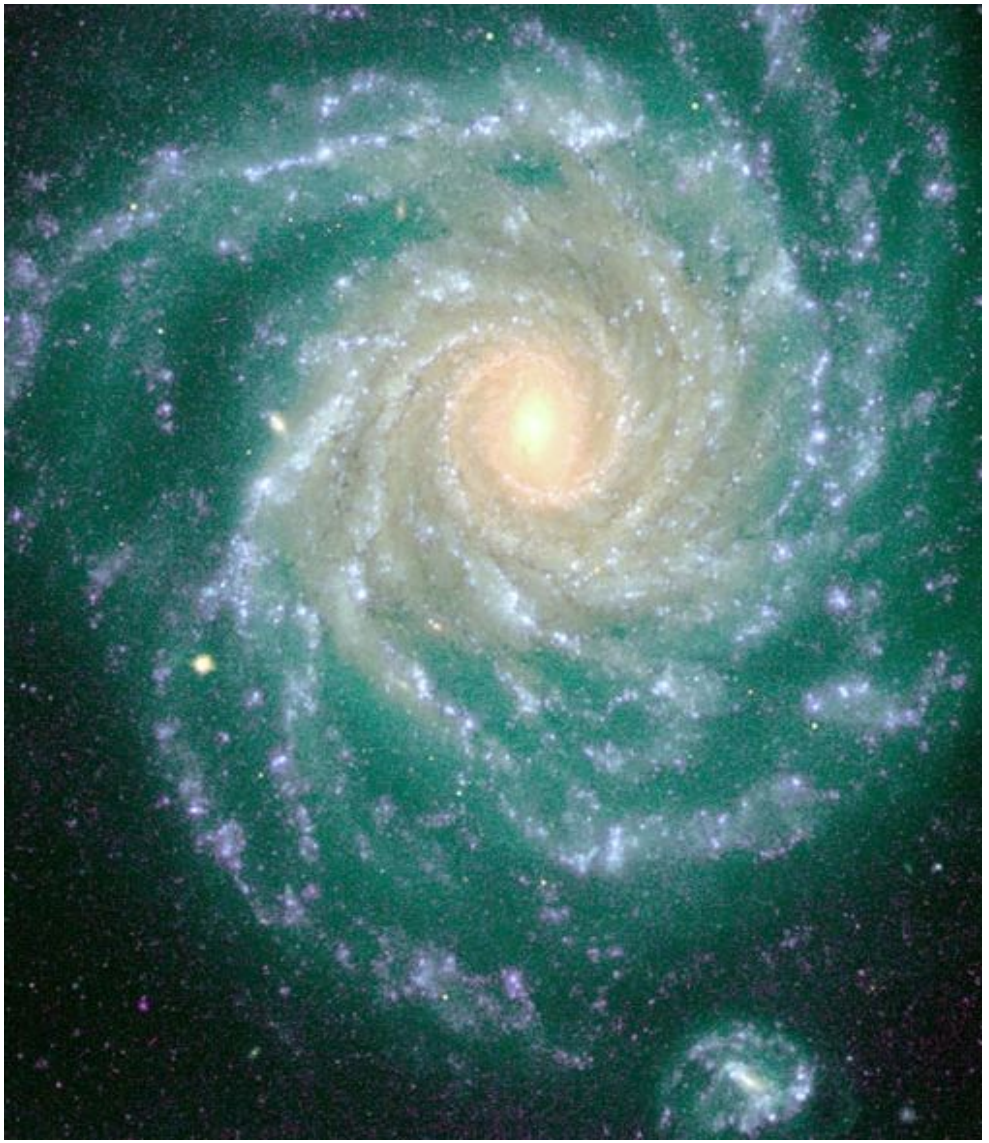
WMAP Science Team (2003)

CDM simulations are still unable to reproduce the bewildering variety in shapes and sizes of observed galaxies in detail

MORPHOLOGY OF OBSERVED GALAXIES



M87 – Anglo-Australian Observatory

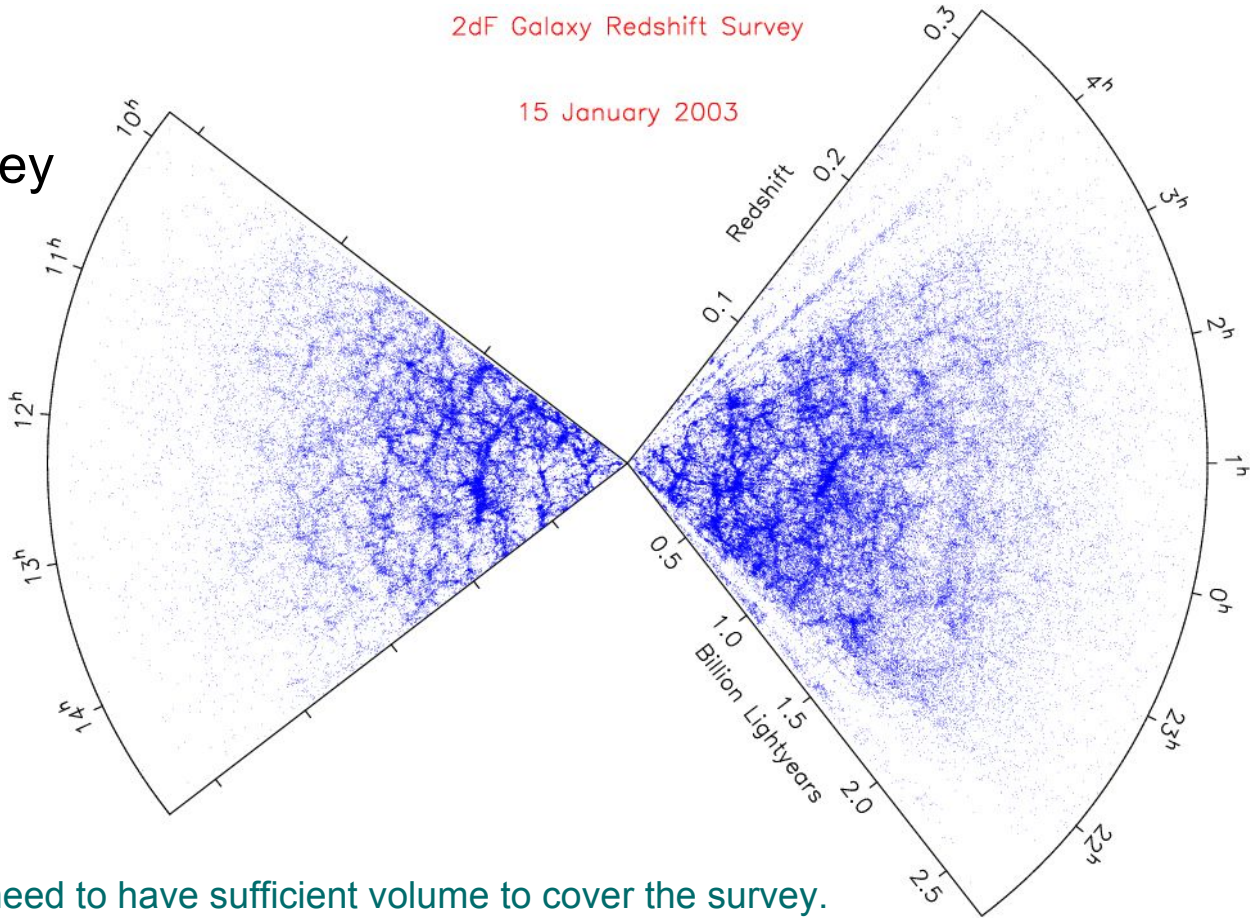


NGC1332 – ESO, VLT

Full exploitation of large observational galaxy surveys relies on theoretical mock catalogues of adequate size

EXAMPLES OF CURRENT OBSERVATIONAL SURVEYS

- Sloan Digital Sky Survey
- 2dF Galaxy Redshift Survey



Simulations and mock catalogues are needed for:

- Investigating systematic effects of the survey (selection effects, biases, etc.)
- Interpreting the observations in terms of the physics of galaxies
- Constraining the theory of galaxy formation

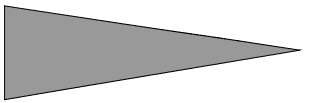
But:
- Theoretical mock catalogues need to have sufficient volume to cover the survey.
- Simultaneously, the particle mass must be small enough to resolve even the faintest galaxies that enter the survey.

Thanks to its extremely large dynamic range, the Millennium Run can meet the opposing requirements of large volume and small particle mass like no other simulation before.

Two conflicting requirements complicate the study of **hierarchical** structure formation

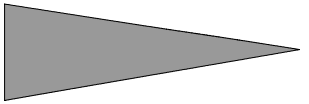
DYNAMIC RANGE PROBLEM FACED BY COSMOLOGICAL SIMULATIONS

Want **small particle mass** to resolve internal structure of halos



Either small volume, or many particles

Want **large volume** to obtain respresentative sample of universe



Either large particle mass, or many particles

$$N$$

Problems due to a small box size:

- Fundamental mode goes non-linear soon after the first halos form. ⇨ Simulation cannot be meaningfully continued beyond this point.
- No rare objects (the first halo, **rich** galaxy clusters, etc.)

Problems due to a large particle mass:

- Physics cannot be resolved.
- Small galaxies are missed.

At any given time, halos exist on a large range of mass-scales !

The recent progress in computer technology and simulation methods allows extremely large simulations

DESIGN GOALS OF A NEXT GENERATION "HUBBLE" SIMULATION

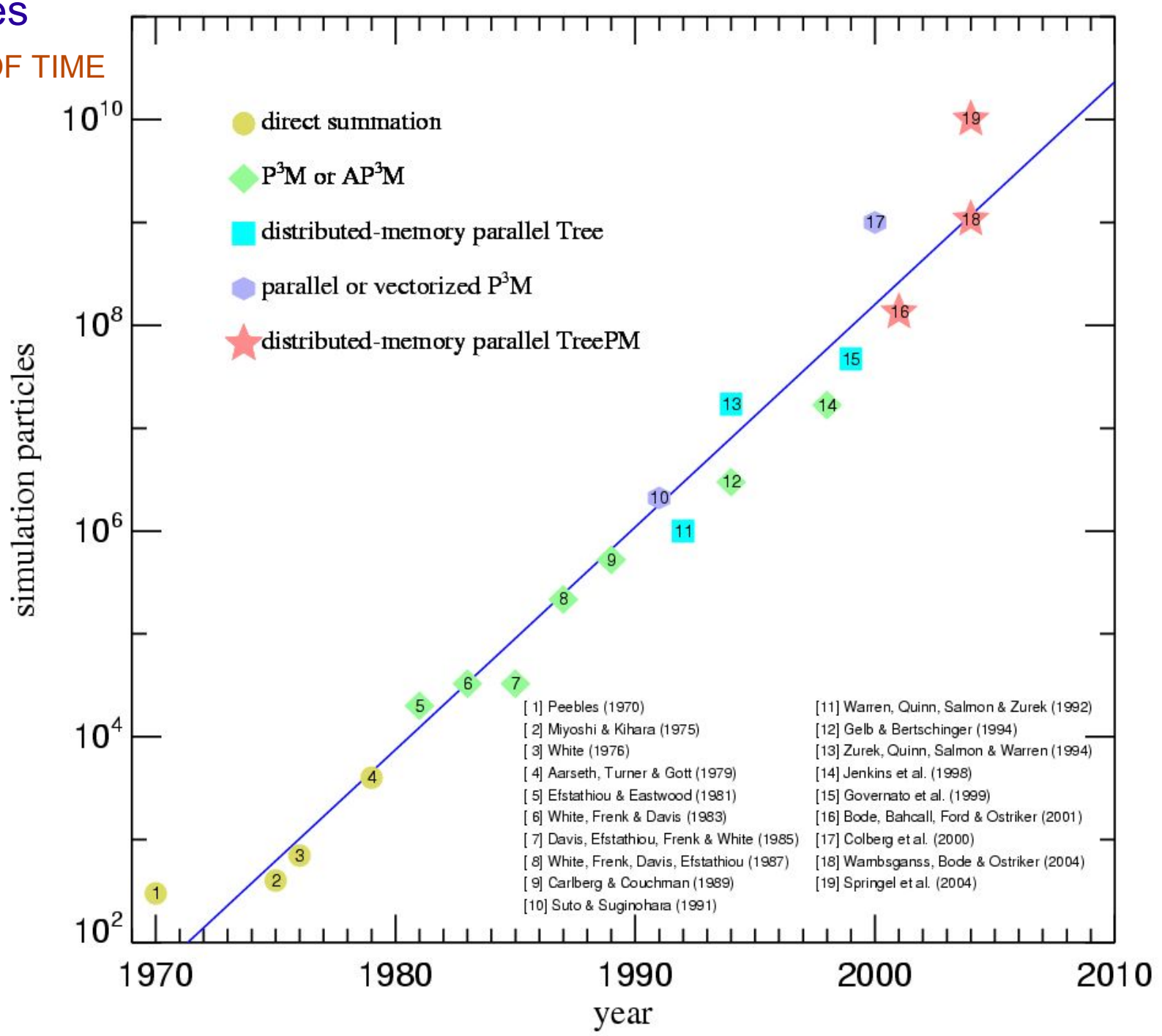
- Large enough volume to cover the big surveys like Sloan and 2dF
- Sufficient mass resolution to predict detailed properties of galaxies substantially fainter than L_*
- If feasible, up to an order of magnitude larger in size than what has been possible so far - that would mean up to 10^{10} particles

→ Will require an extremely efficient code on the largest available machines, and result in a cosmological simulation with extremely large dynamic range.

Cosmological N-body simulations have grown rapidly in size over the last three decades

"**N**" AS A FUNCTION OF TIME

- Computers double their speed every 18 months (Moore's law)
- N-body simulations have doubled their size every 16-17 months
- Recently, growth has accelerated further.
  The Millennium Run should have become possible in 2010 – we have done it in 2004 !



The simulation was run on the *Regatta* supercomputer of the RZG

REQUIRED RESSOURCES

1 TByte RAM needed

16 x 32-way Regatta Node
64 GByte RAM
512 CPU total

CPU time consumed

350.000 processor hours

- 28 days on 512 CPUs/16 nodes
- 38 years in serial
- ~ 6% of annual time on total Regatta system
- sustained average code performance (hardware counters) 400 Mflops/cpu
- 5 x 10^{17} floating point ops
- 11000 (adaptive) timesteps

We have developed a new, memory-efficient cosmological code: GADGET-II

NEW FEATURES OF GADGET-II

- New symplectic integration method
- Higher speed of the tree algorithm
- Less memory consumption for tree and particle storage (~factor 2 saving)

Key feature for Millenium Run

- **Code may be run optionally as a TreePM hybrid code**
- SPH neighbour search faster
- Conservative SPH formulation
- Fully consistent dynamic tree updates
- Additional types of runs possible (e.g. hydrodynamics-only, long periodic boxes)
- Efficient and clean formation of star particles
- More physics (B-fields, thermal conduction, chemical enrichment)
- More output options, including HDF5
- Still fully standard C & standard MPI. The FFTW and GSL libraries are needed.
- Reduced communication overhead and better scalability

The new code is quite a bit better than the old version...

The maximum size of a TreePM simulation with *Lean*-GADGET-II is essentially memory bound

A HIGHLY MEMORY EFFICIENT VERSION OF GADGET-II

Particle Data	Tree storage	FFT workspace
48 bytes / particle	40 bytes / particle	24 bytes / mesh-cell

Not needed concurently!

Special code version ***Lean*-GADGET-II** needs: **88 bytes / particle** (Assuming 1.5 mesh-cells/particle)

Preliminary Simulation Set-up

- Particle number: $2160^3 = 10.077.696.000 = \sim 10^{10}$ particles
- Boxsize: $L = 500\ h^{-1}$ Mpc
- Particle mass: $m_p = 8.6 \times 10^8\ h^{-1}\ M_\odot$
- Spatial resolution: $5\ h^{-1}$ kpc
- Size of FFT: $2560^3 = 16.777.216.000 = \sim 17$ billion cells

Memory requirement of simulation code **880 GByte**

Compared to Hubble-Volume simulation:
- 2000 times better mass resolution
- 10 times larger particle number
- 13 better spatial resolution

The simulation produced a multi-TByte data set

RAW SIMULATION OUTPUTS

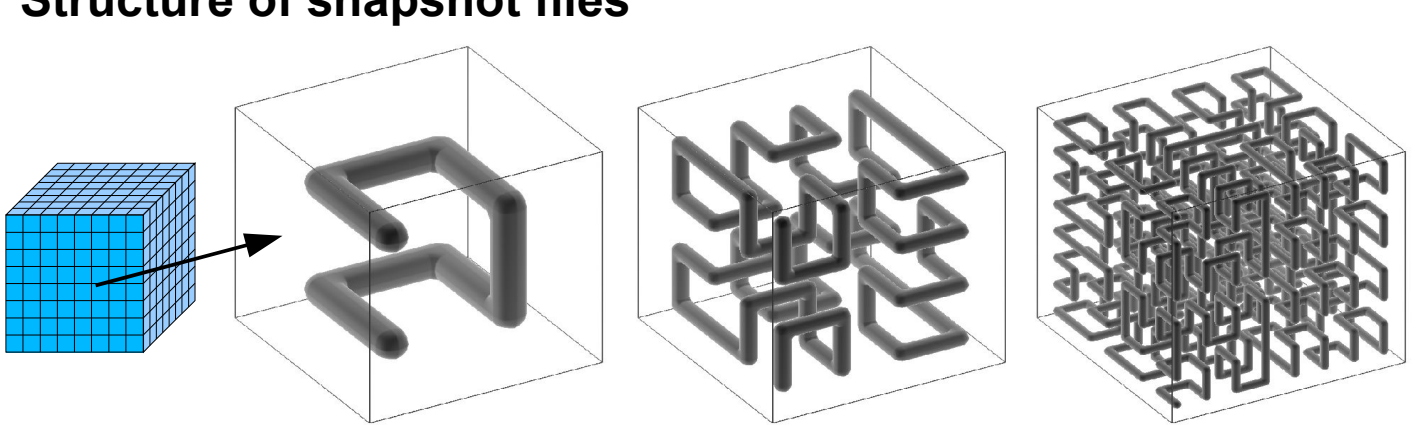
Data size

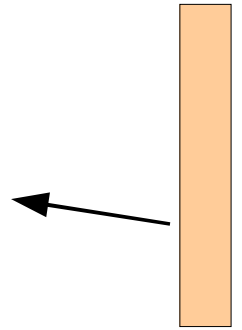
One simulation timeslice
360 GByte

we have stored
64 outputs

Raw data volume
23 TByte

Structure of snapshot files



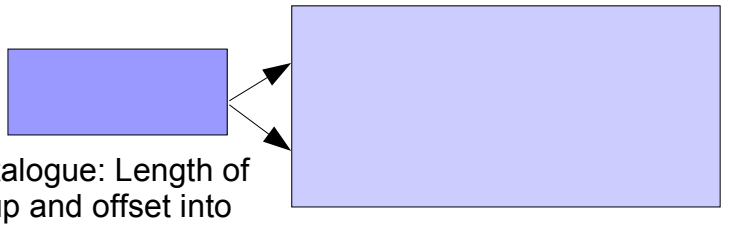


- The particles are stored in the sequence of a 256^3 Peano-Hilbert grid that covers the volume. On average, 600 particles per grid-cell.
- Each output is split into $8^3 = 512$ files which roughly map to subcubes in the simulation volume.
- Each file has ~ 20 million particles, 600 MB.

- A hash-table is produced for each file of a snapshot. Each entry gives the offset to the first particle in corresponding cell of the 256^3 grid, relative to the beginning of the file.
  Size of tables:
  512 x 128 Kb = 64 MB
- Allows random access to particle data of subvolumes.

FoF group catalogues

Are computed on the fly



- Group catalogue: Length of each group and offset into particle list
- Long list of particle keys (64 bit) that make up each group

- Structure of 64-bit particle key

15 bit Hash-Key | 9 bit File-Key | 34 bit Particle-ID

- Allows fast selective access to all particles of a given group

Postprocessing of the simulation data requires effcient analysis codes

VARIOUS POSTPROCESSING-TASKS

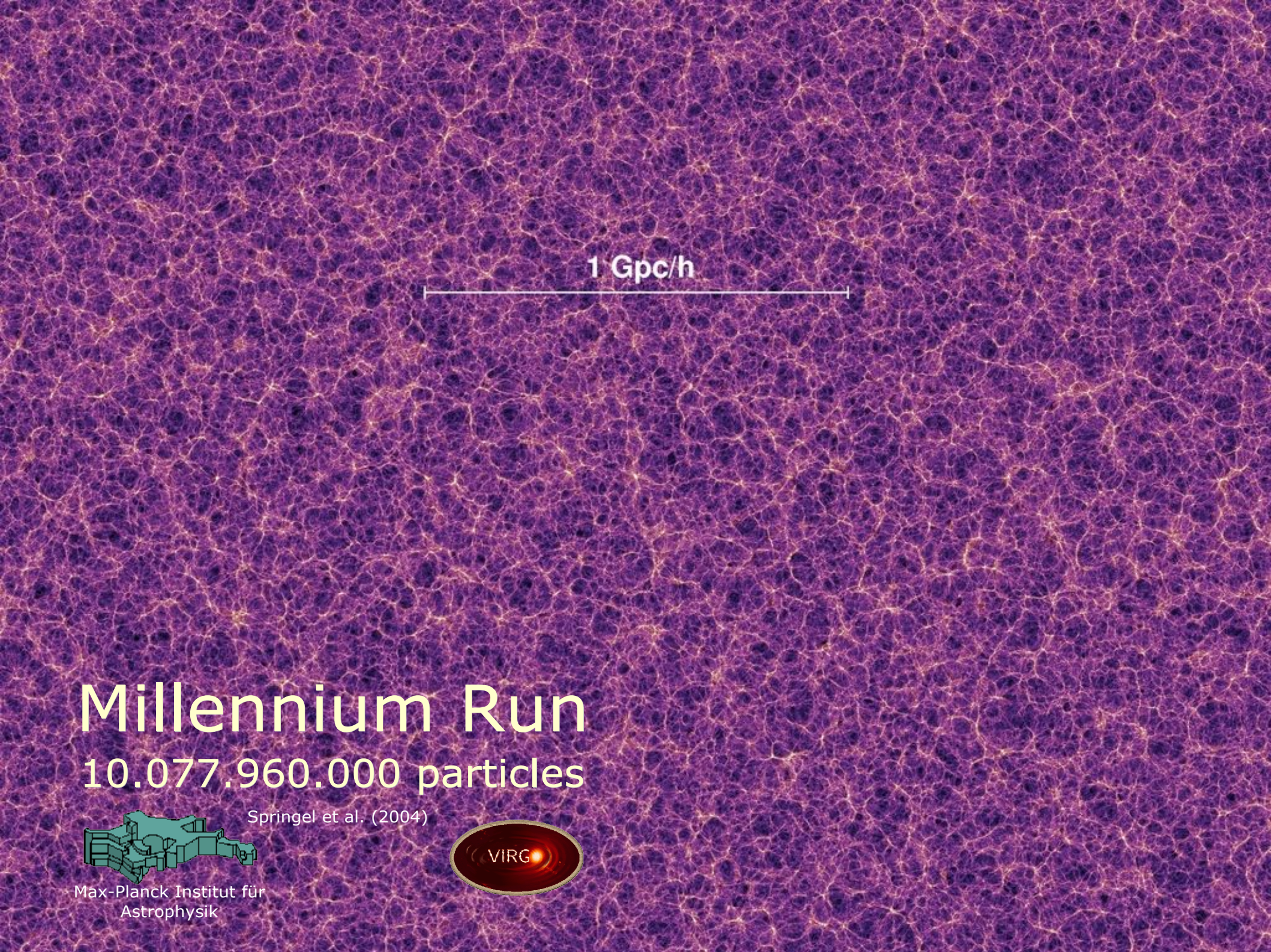
Things done on the fly by the simulation code

- **FoF group finding**
- **Power spectrum and correlation function measurement**

Tasks carried out as true postprocessing

- **Substructure finding and halo/subhalo properties**
  - Done by L-SubFind in massiv parallel mode
  - With 32 CPU/256 GB (chubby queue) can process one clustered snapshot in ~4-5 hours
- **Construction of merger history trees**
  - Two step procedure. L-BaseTree finds halos descendants in future snapshots, thereby providing horizontal links in the merger tree. Serial/OpenMP-parallel, requires ~200 GB shared RAM, fast.
  - In a second step, L-HaloTrees builds up fully threaded vertical trees for each halo. These are the input objects for the semi-analytic code.
- **Semi-analytic galaxy formation**
  - New semi-analytic code L-Galaxies, can be run in massively parallel mode on the merger trees generated for the Millennium Run.
  - Interfacing with VO databases is in preparation.
- **Data visualization**
  - Challenging because of the data volume. L-HsmlFind (massively parallel) determines dark matter adaptive smoothing lengths, while L-Picture (serial) makes a picture for an arbitrarily inclinded and arbitrarily large slice through the periodic simulation.

	Lines of C-Code
L-GenIC	1900
L-Gadget2	12000
L-SubFind	3000
L-BaseTree	700
L-HaloTrees	900
L-Galaxies	2600
L-HsmlFind	1800
L-Picture	1600
Lines	24500
Characters	580000



Millennium Run

10.077.960.000 particles

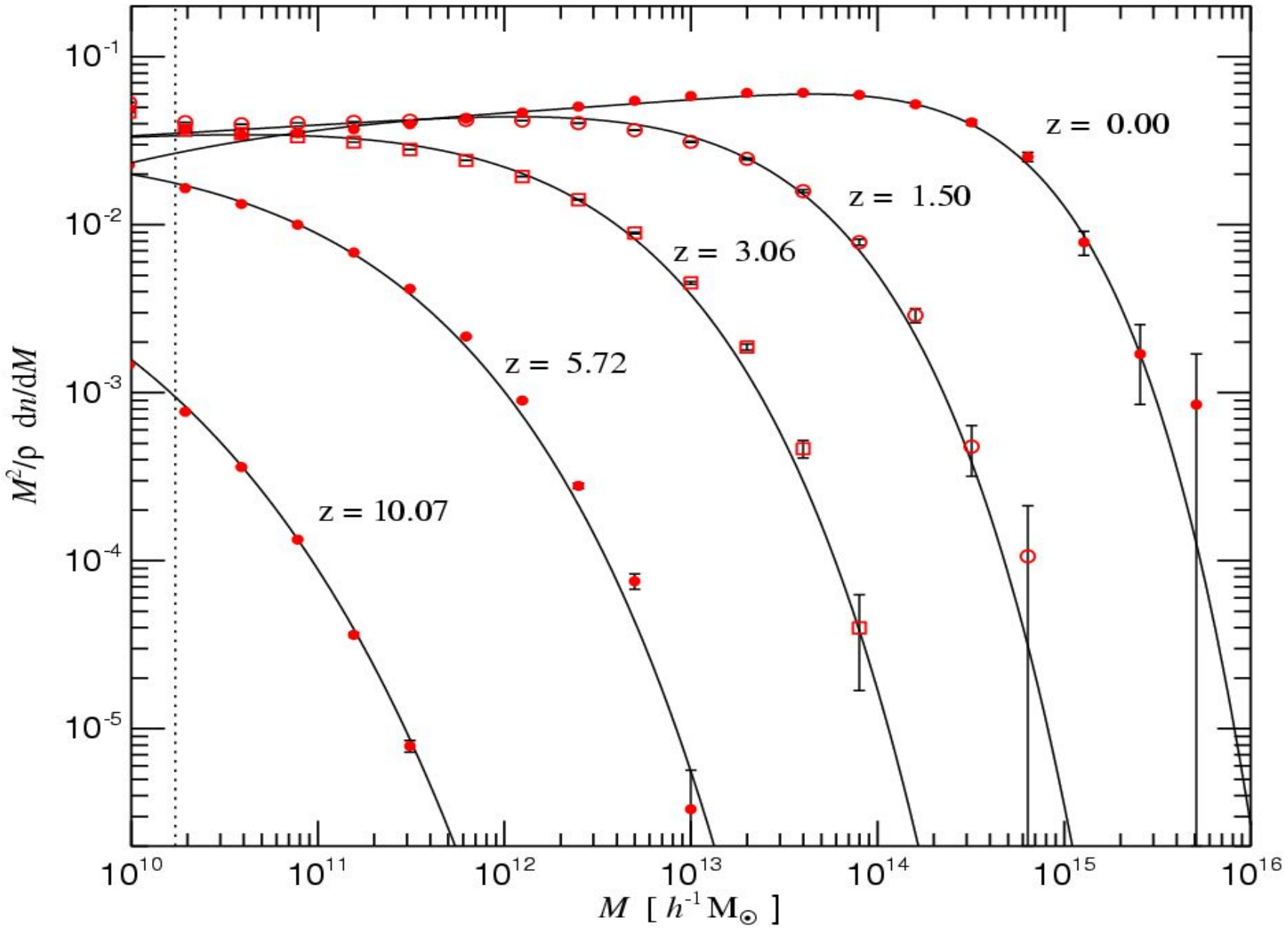
Springel et al. (2004)

Max-Planck Institut für Astrophysik

VIRGO

The halo mass function is very well fit by the model of Jenkins et al. (2001)

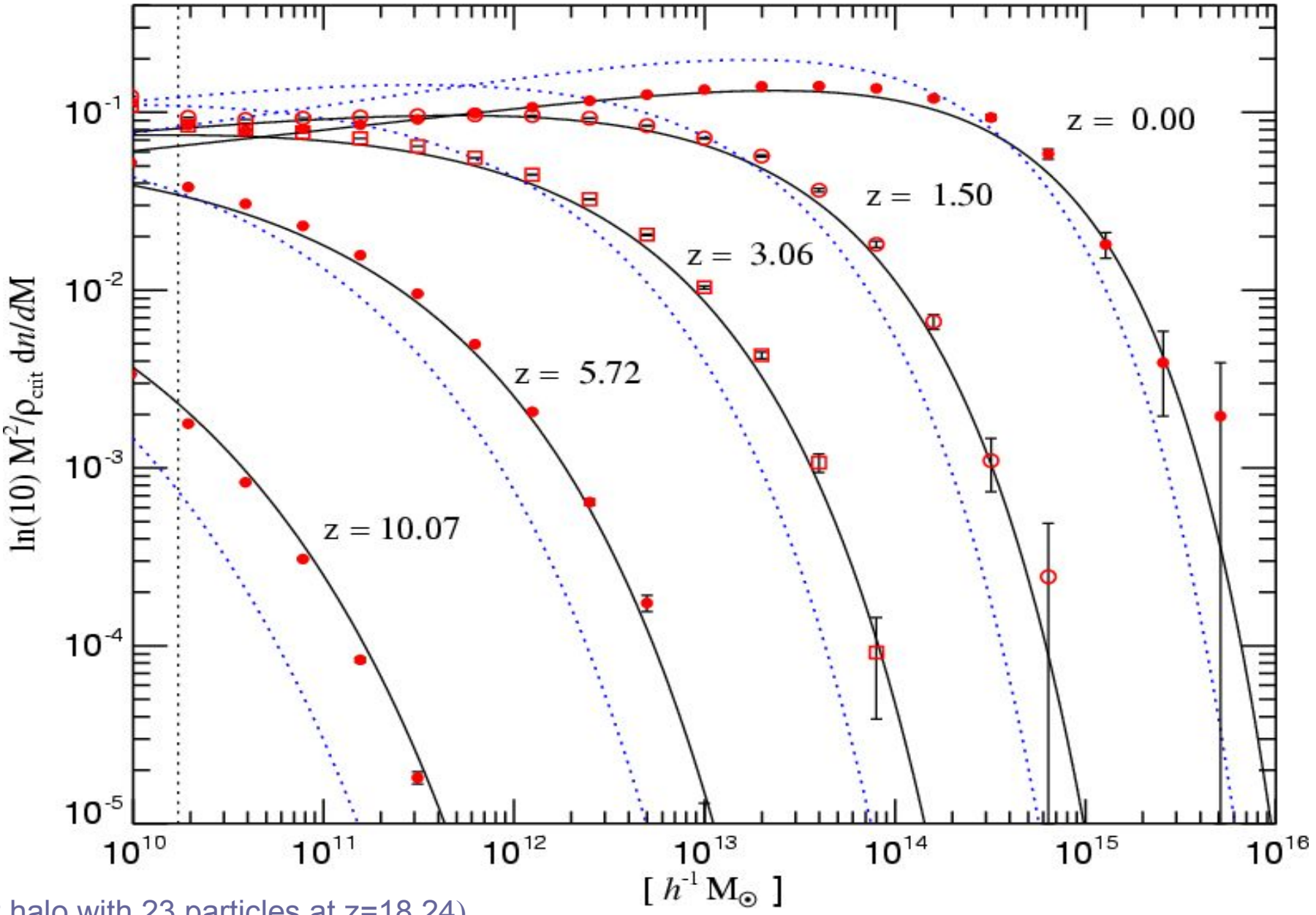
MASS MULTIPLICITY FUNCTION



(First halo with 23 particles at z=18.24)

The Sheth & Tormen mass function provides still an acceptable description, but Press & Schechter is quite discrepant

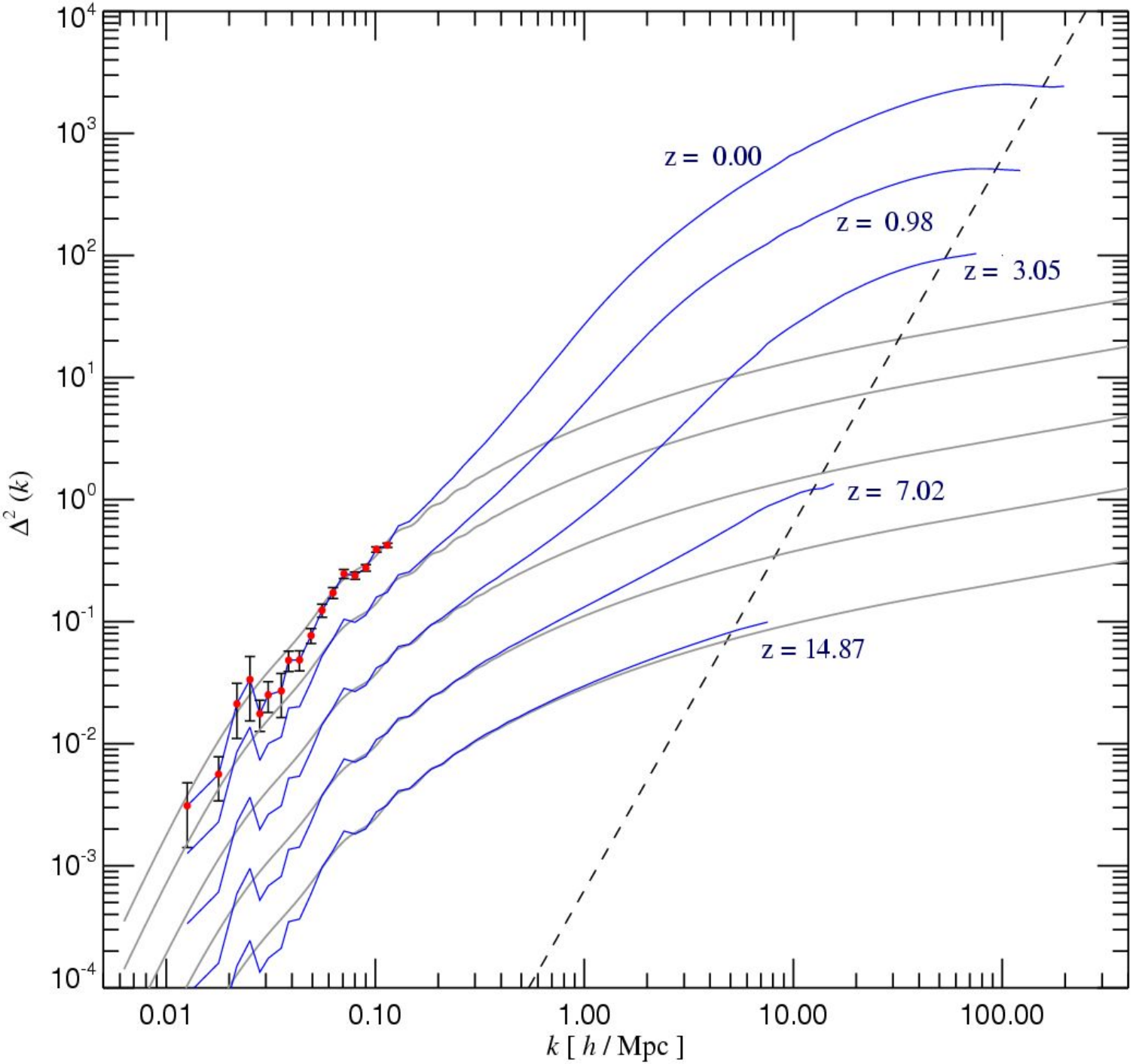
MASS MULTIPLICITY FUNCTION



(First halo with 23 particles at z=18.24)

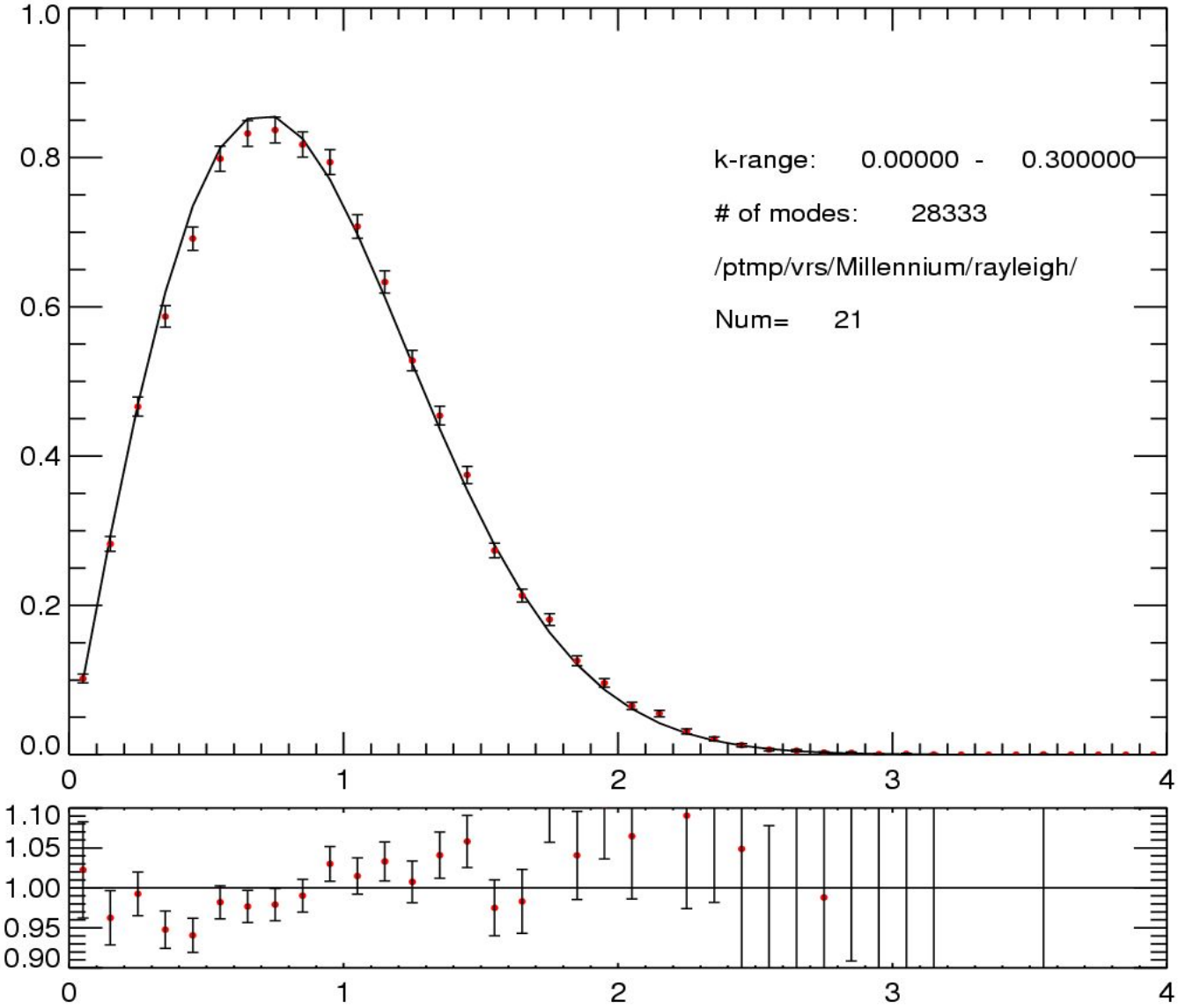
The non-linear evolution of the mass power spectrum is accurately determined by the Millennium Run over a large range of scales

POWER SPECTRUM OF MASS FLUCTUATIONS



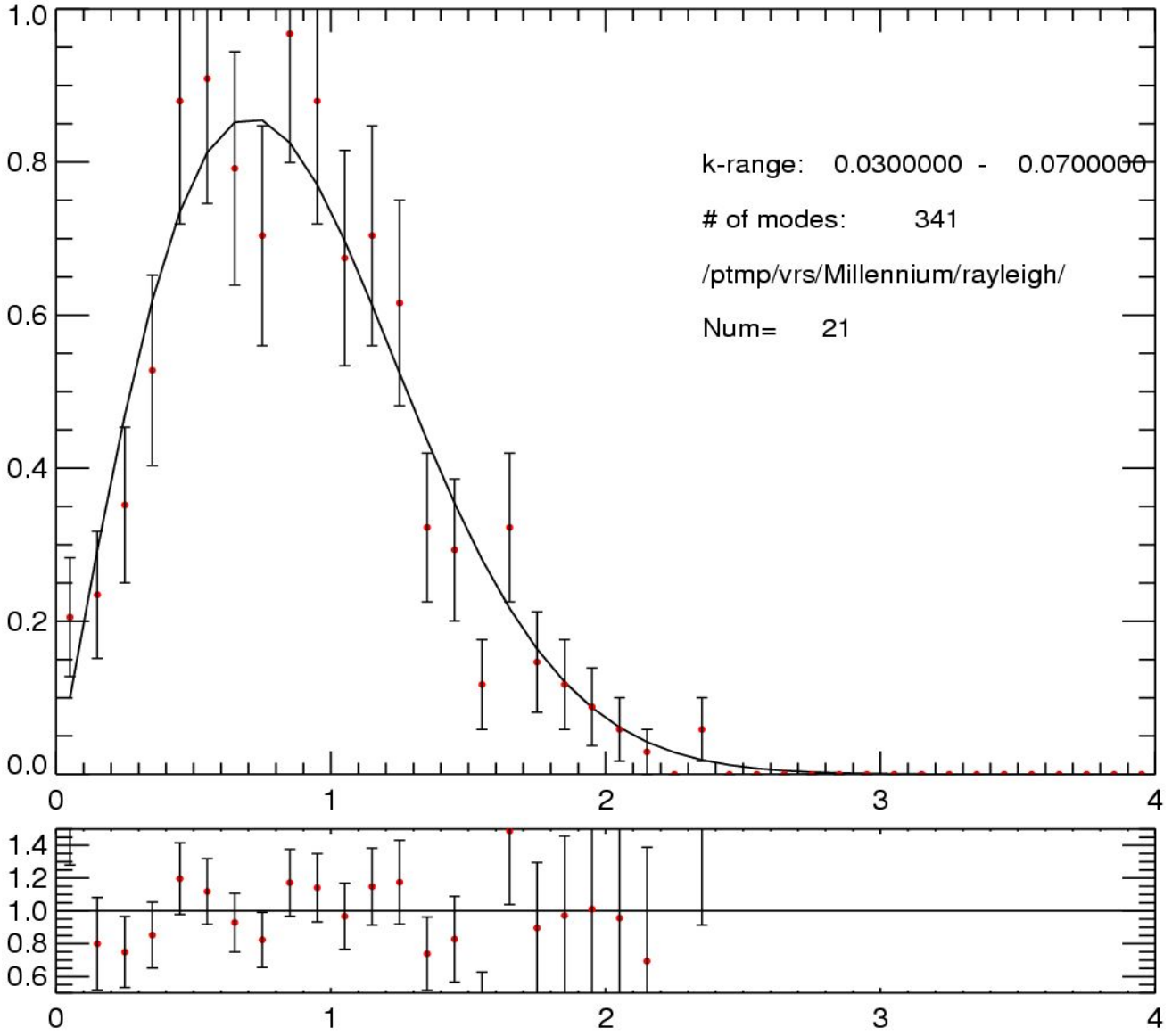
The power in individual modes is Rayleigh distributed around the mean

DISTRIBUTION OF MODE AMPLITUDES RELATIVE TO THE MEAN



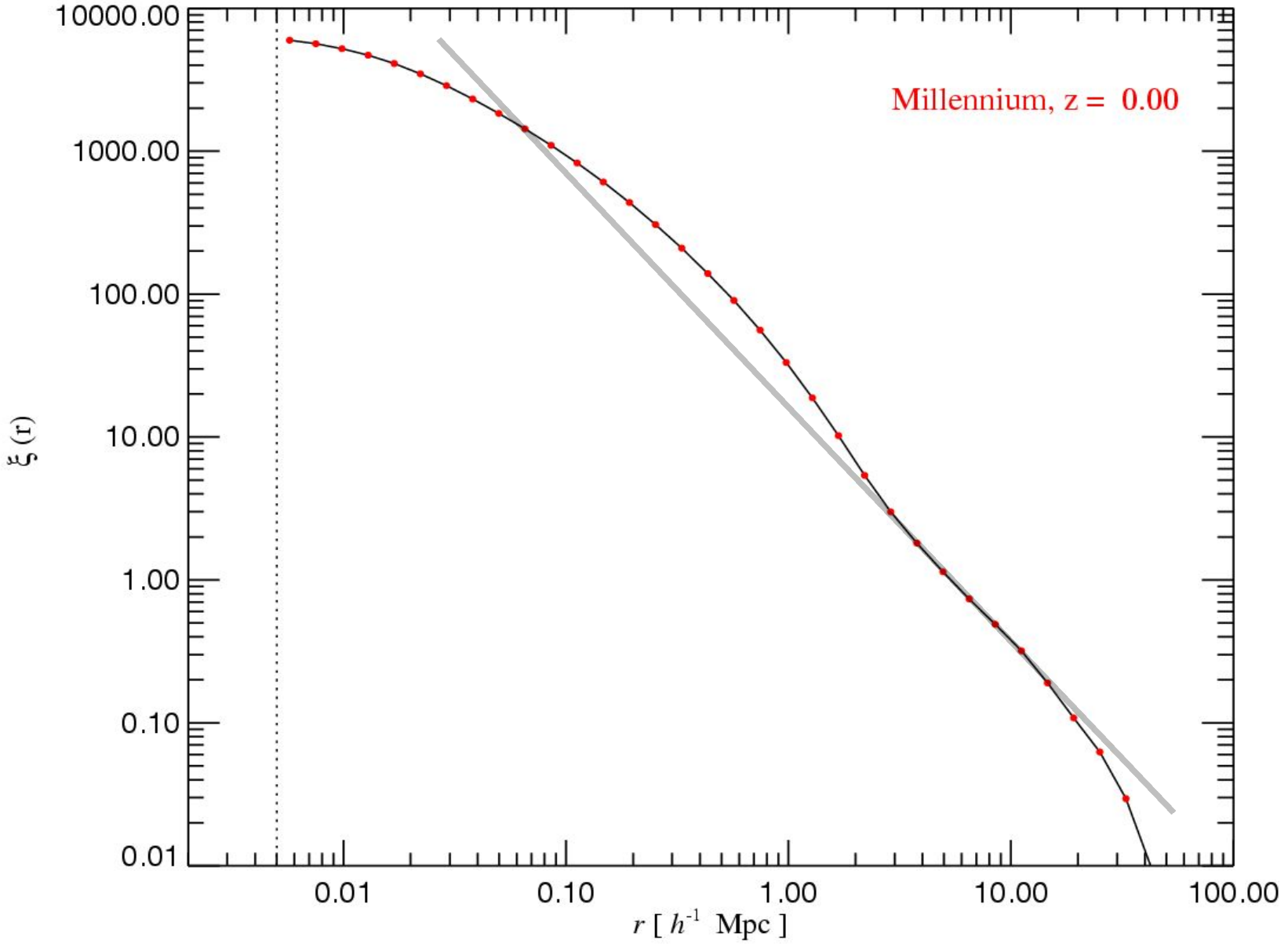
The power in individual modes is Rayleigh distributed around the mean

DISTRIBUTION OF MODE AMPLITUDES RELATIVE TO THE MEAN IN A NARROW k-RANGE



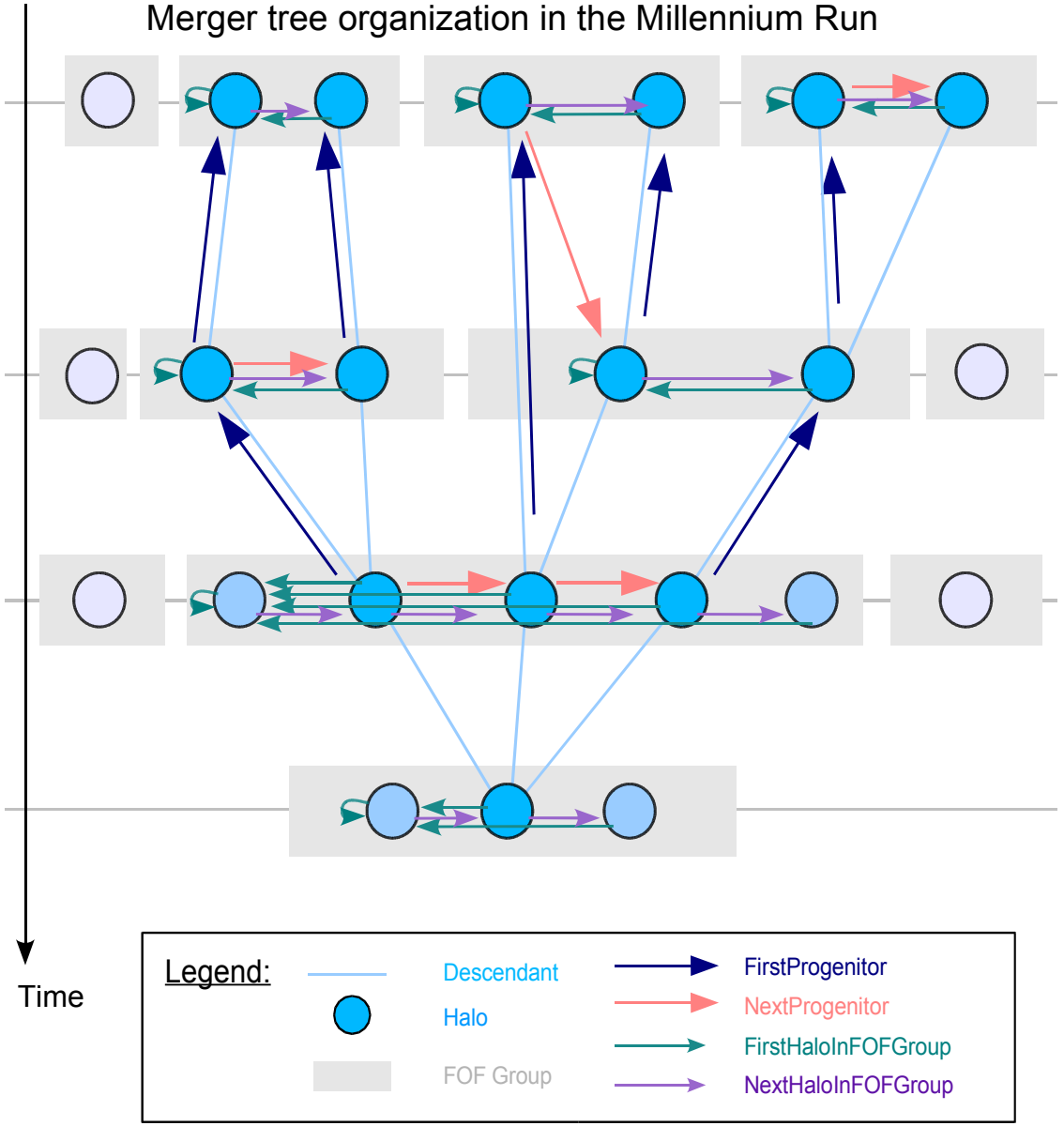
The dark matter autocorrelation function of the dark matter is measured with high precision and deviates strongly from a power-law

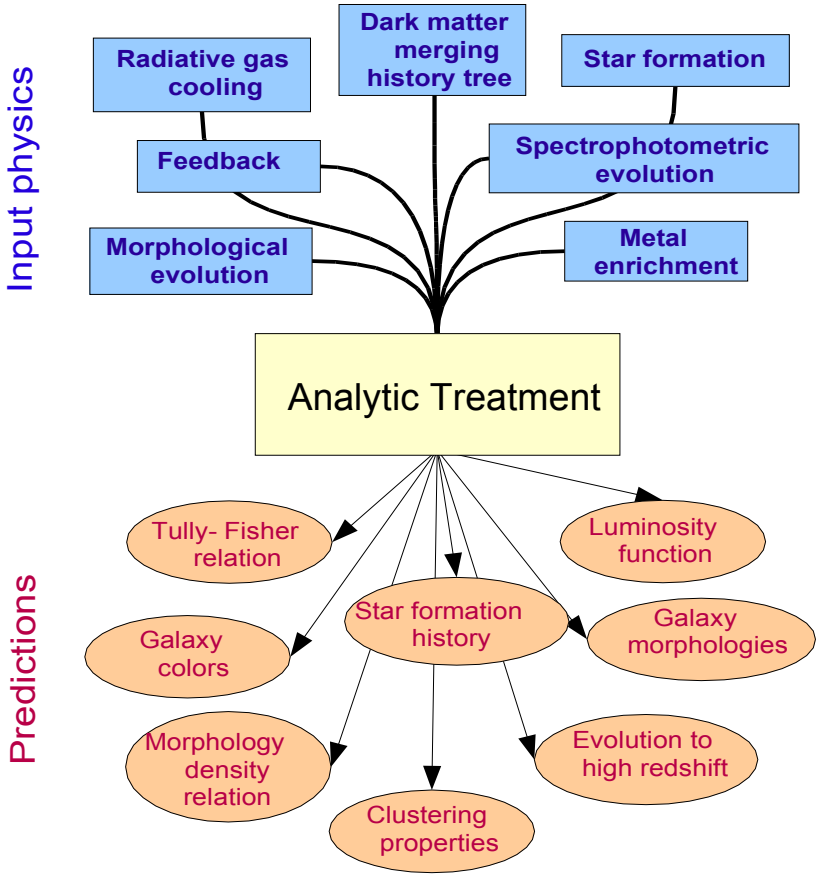
DARK MATTER TWO-POINT FUNCTION



The semi-analytic model follows *post hoc* the most important physics of galaxy formation

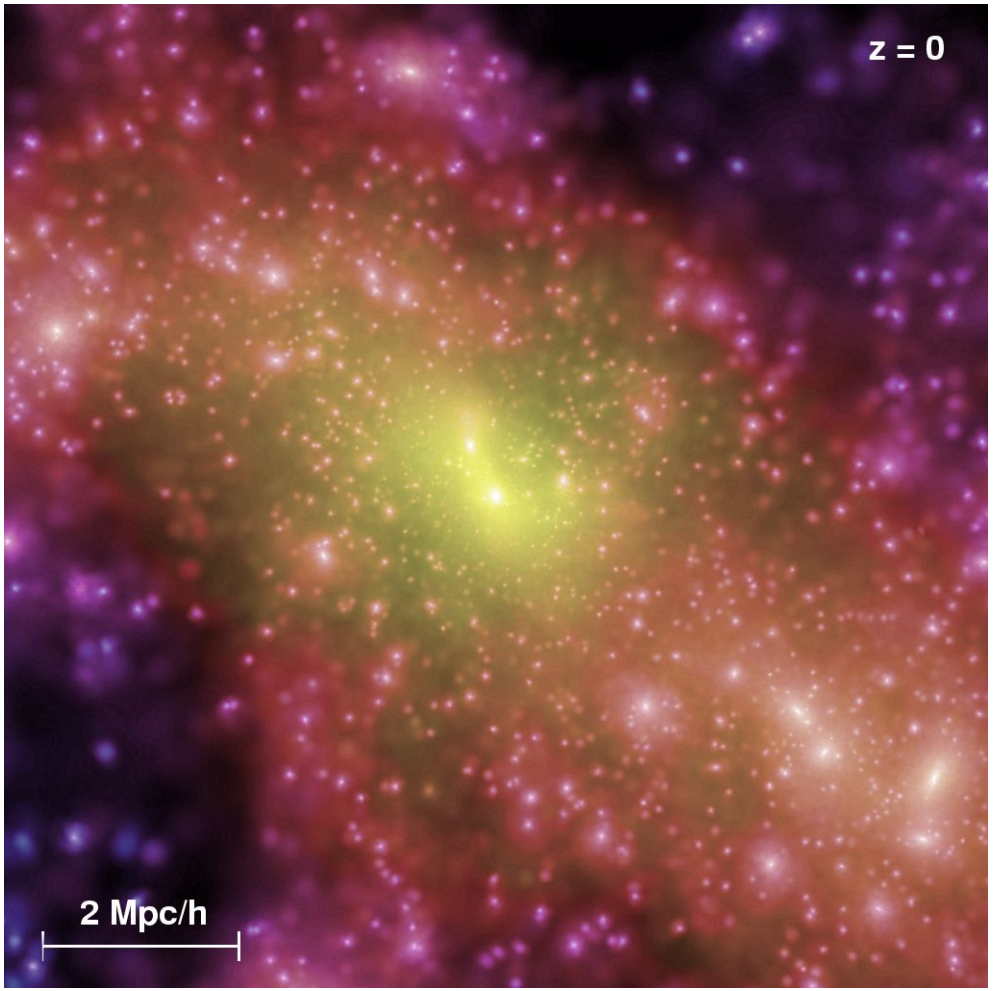
SCHEMATIC MERGER TREE AND SEMI-ANALYTIC MODEL

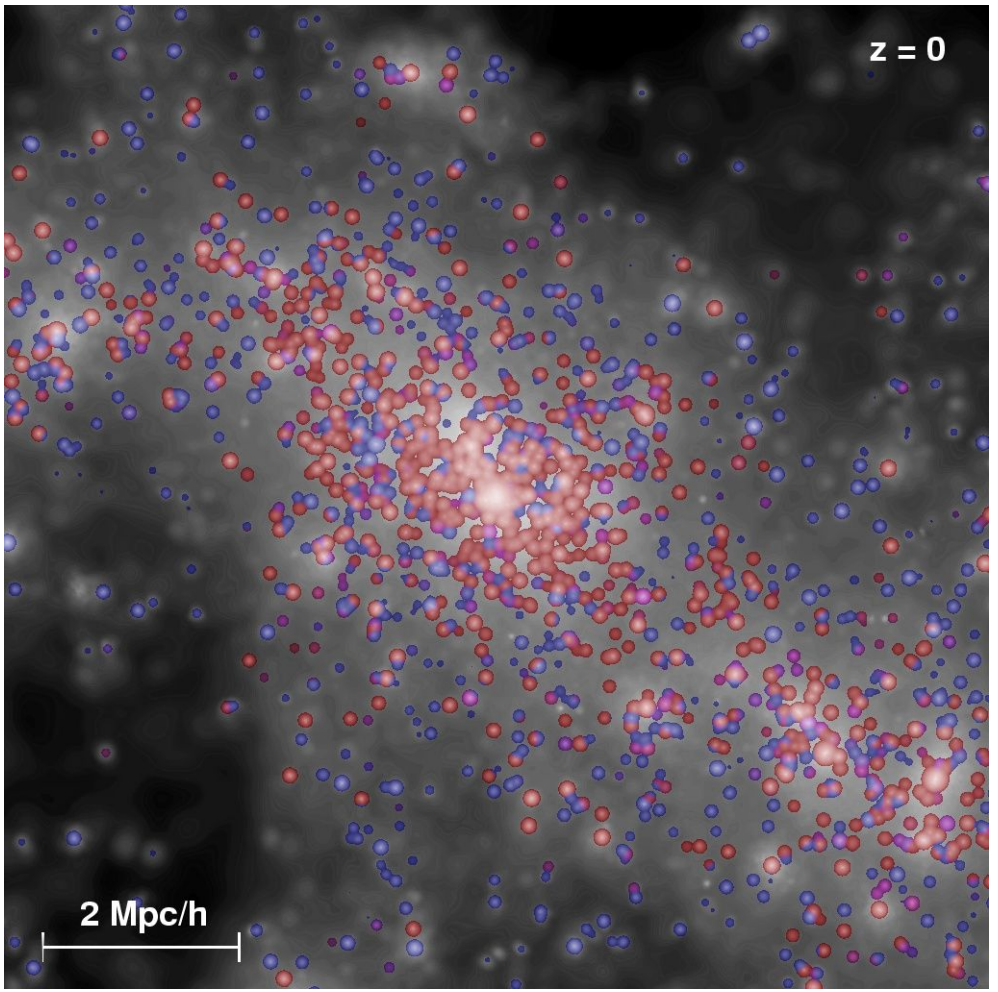




A merger tree containing 800 million dark matter (sub)halos is used to compute semi-analytic models of galaxy formation

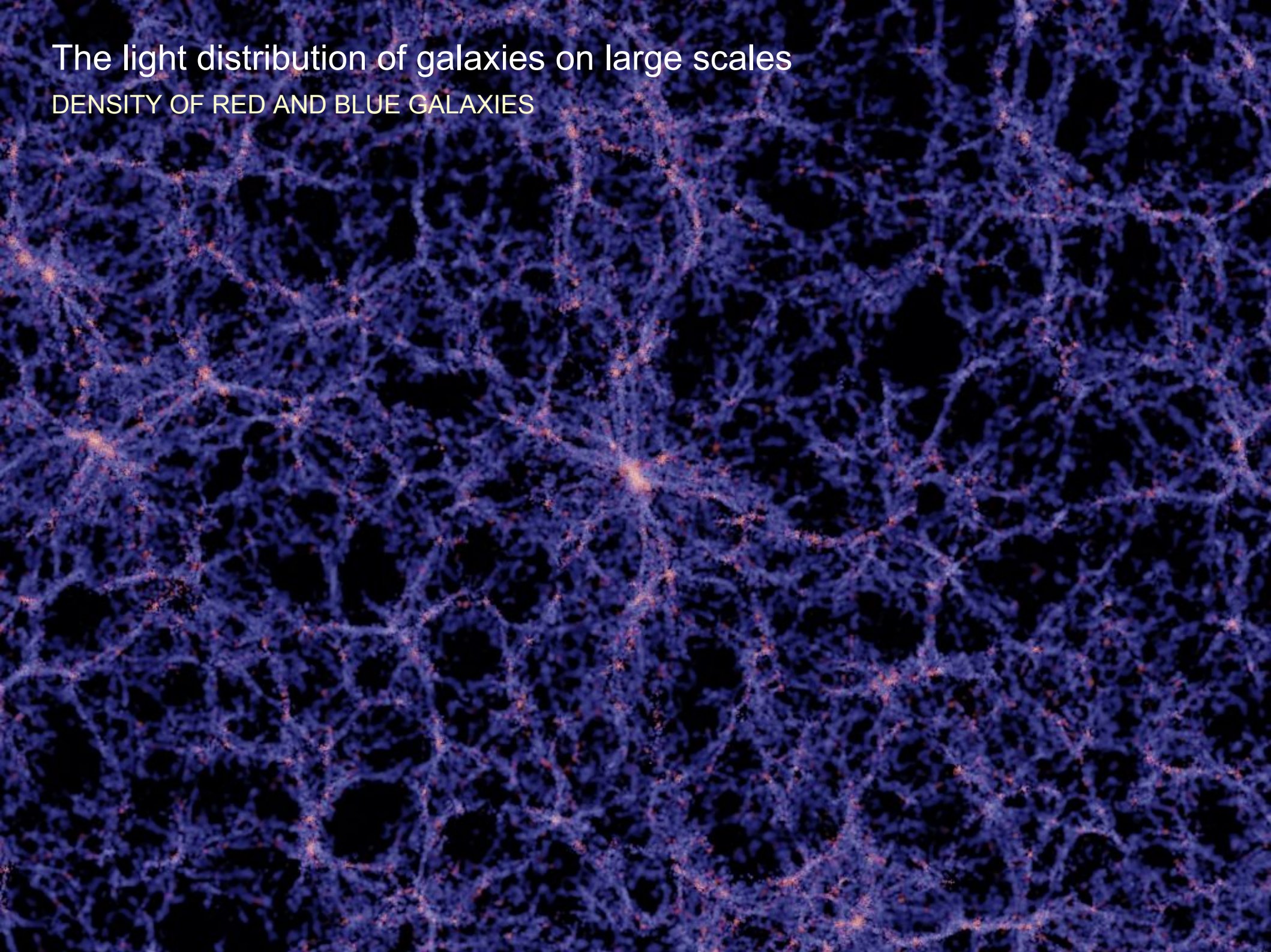
DARK MATTER AND GALAXY DISTRIBUTION IN A CLUSTER OF GALAXIES





The light distribution of galaxies on large scales

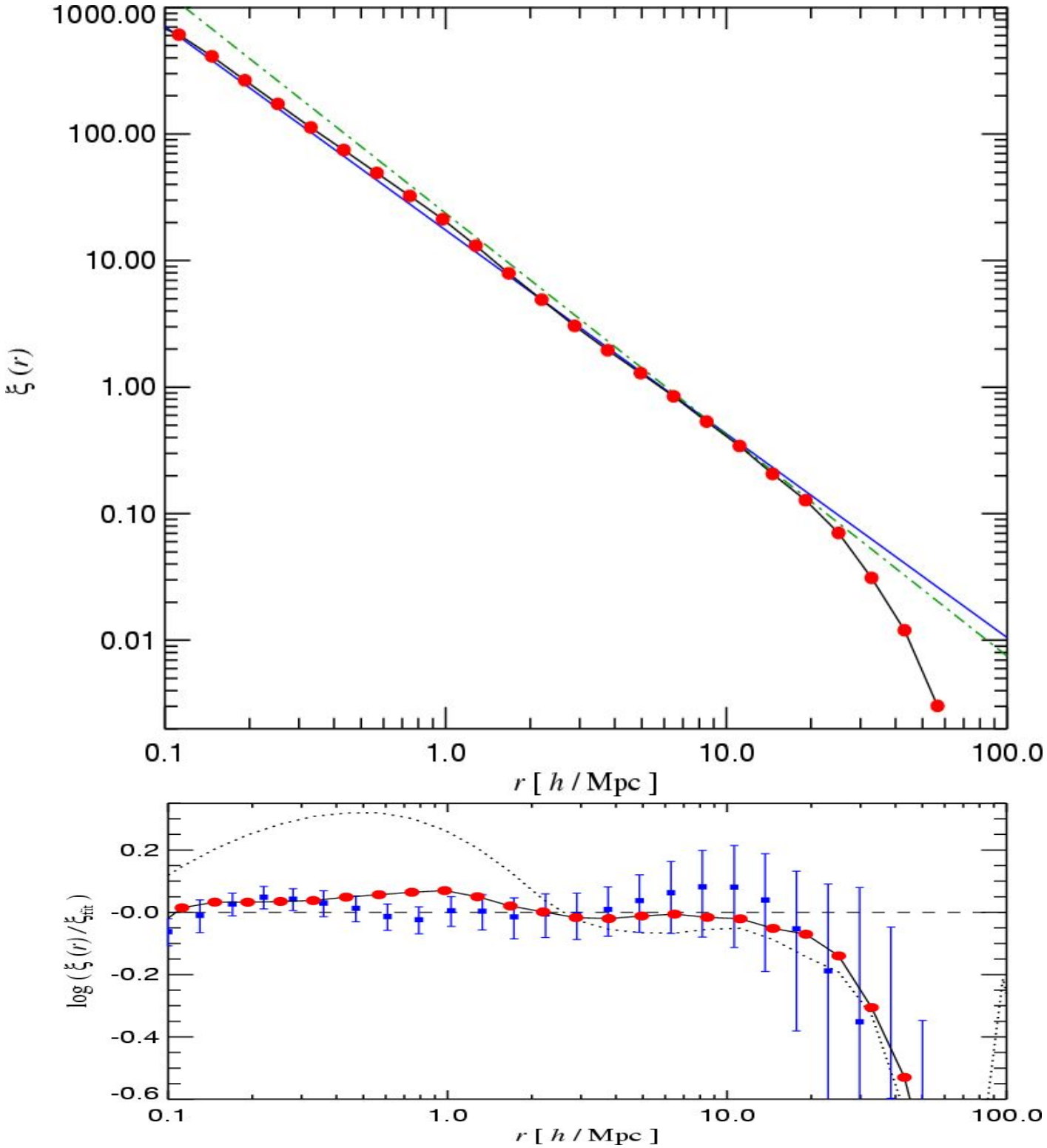
DENSITY OF RED AND BLUE GALAXIES




The distribution of dark matter on large scales
DARK MATTER DENSITY, COLOR-CODED BY DENSITY AND VELOCITY DISPERSION
125 Mpc/h

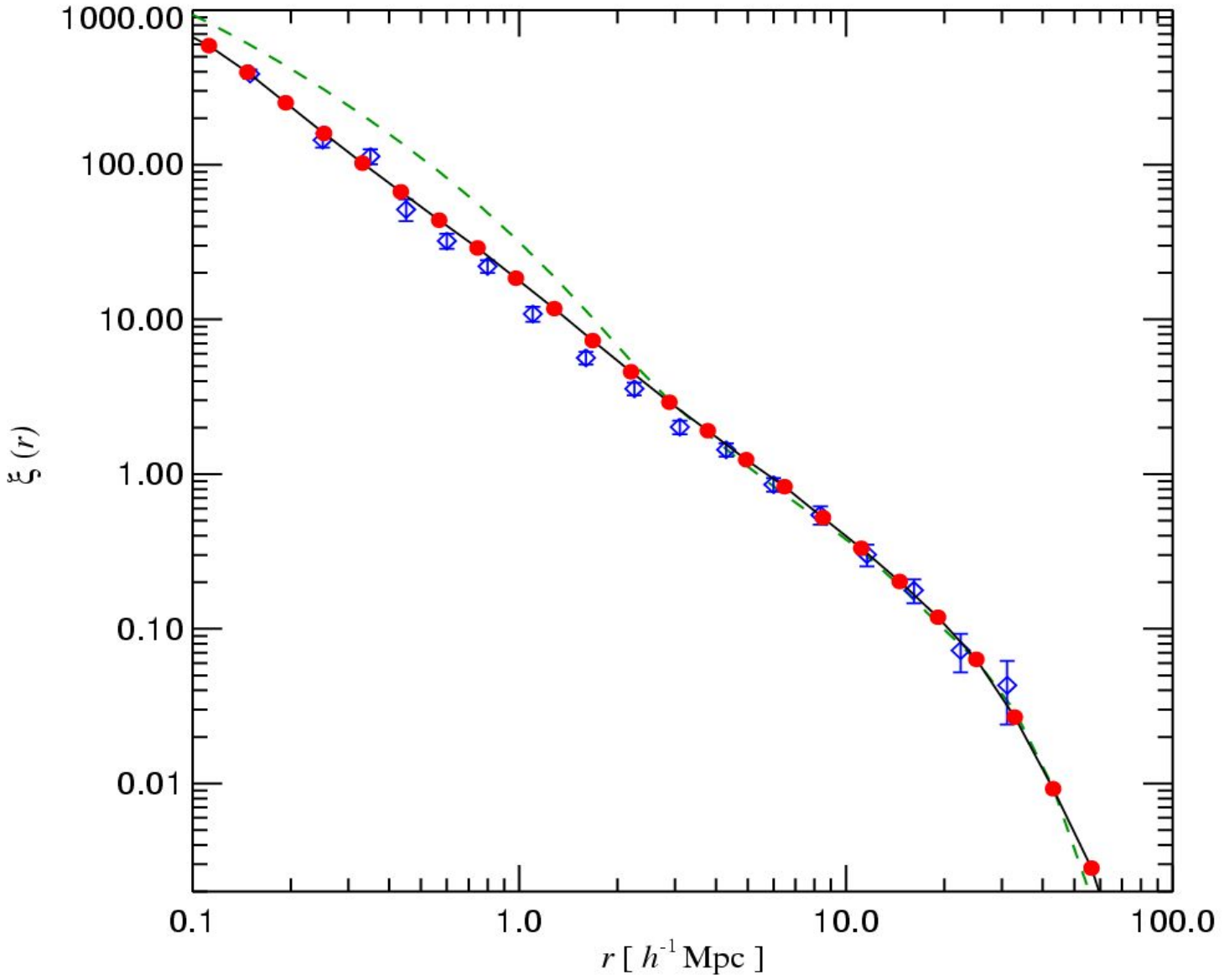
The two-point correlation function of galaxies in the Millennium run is a very good power law

GALAXY TWO-POINT FUNCTION COMPARED WITH APM AND SDSS



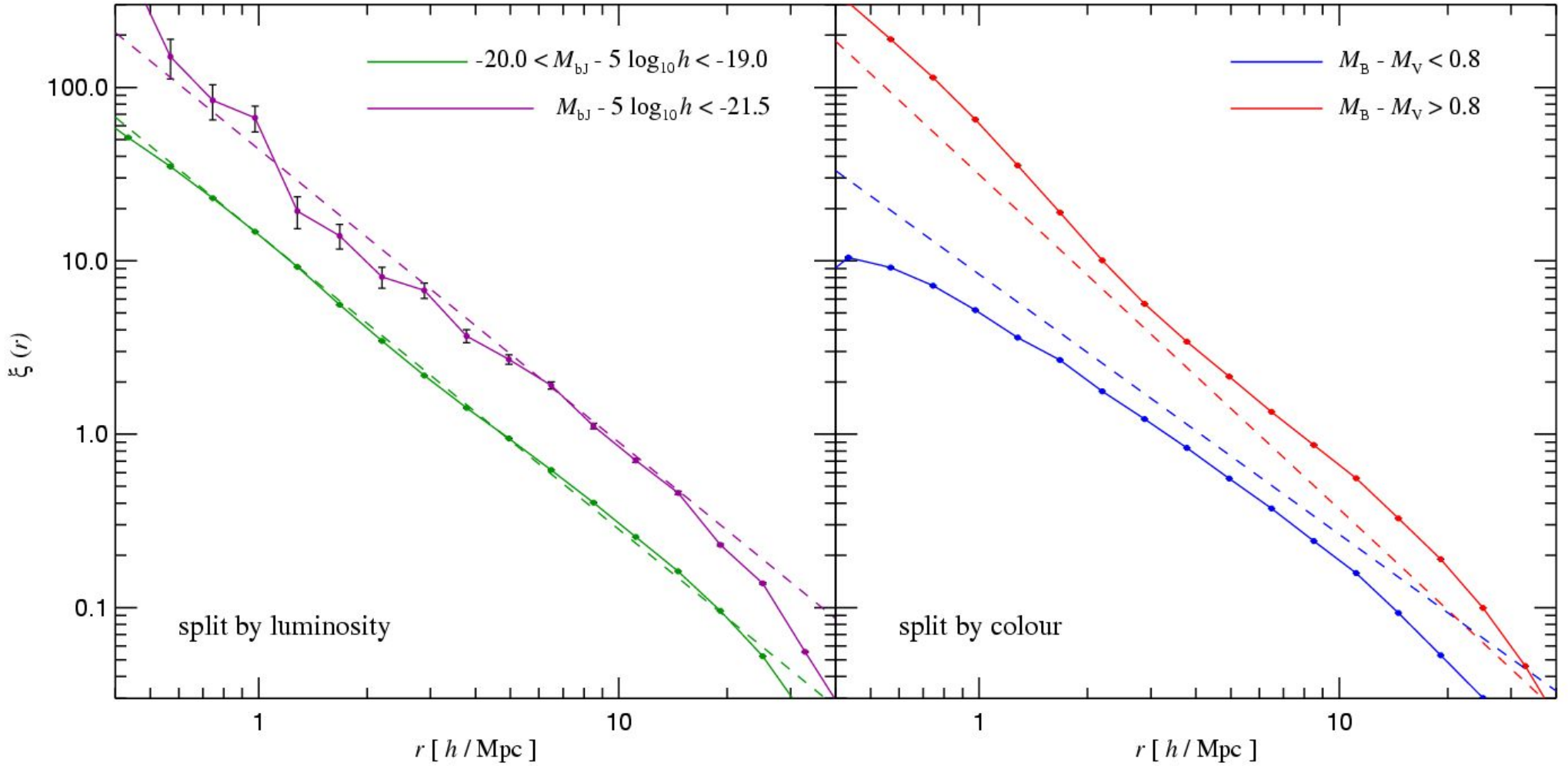
The two-point correlation function of galaxies in the Millennium run is a very good power law

GALAXY TWO-POINT FUNCTION COMPARED WITH **2dFGRS**



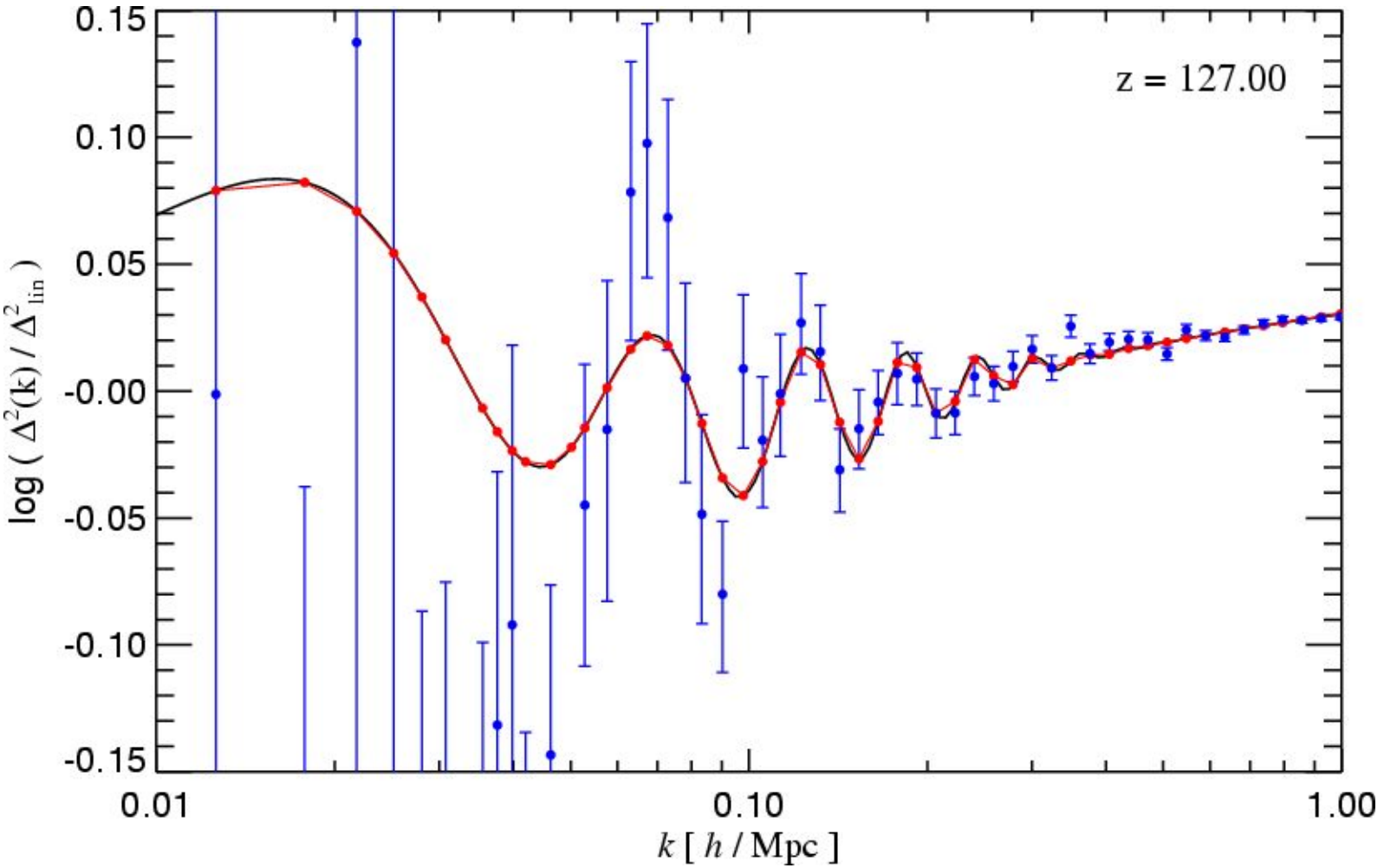
The semi-analytic model fits a multitude of observational data

CLUSTERING BY MAGNITUDE AND COLOR



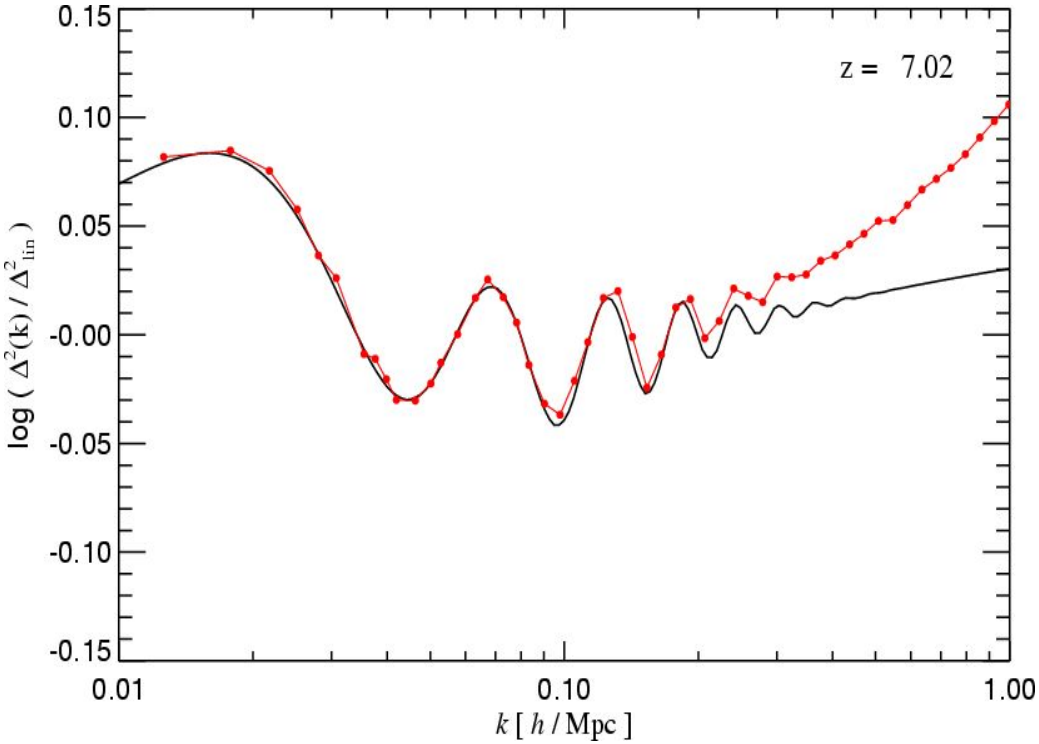
The boxsize of the Millennium Run is large enough to resolve the baryonic wiggles in the matter power spectrum

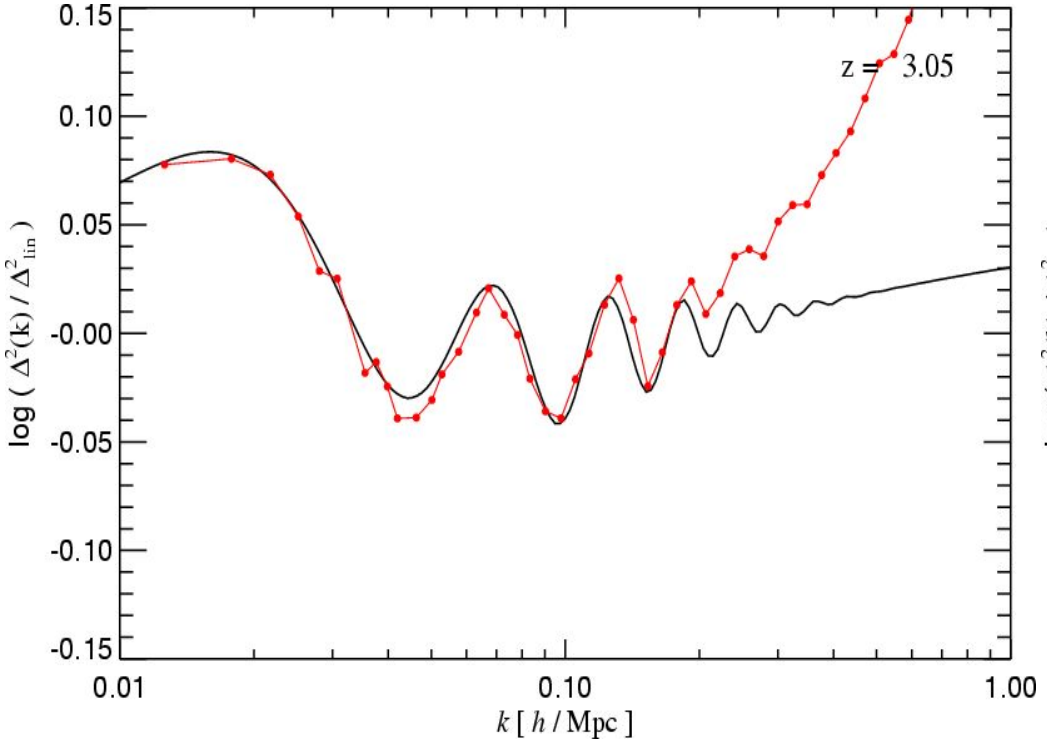
LINEAR MASS POWER SPECTRUM AT HIGH REDSHIFT

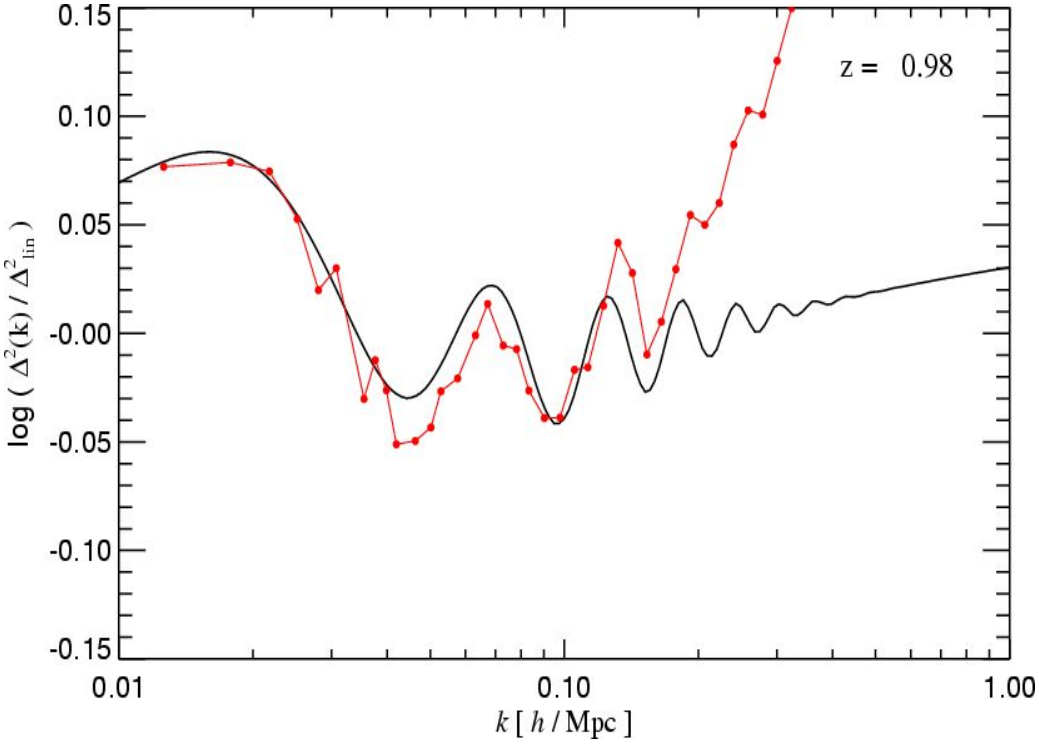


Non-linear evolution accelerates the growth of power and eliminates structure in the spectrum by mode-coupling

TIME EVOLUTION OF THE DARK MATTER POWER SPECTRUM IN THE "WIGGLE" REGION

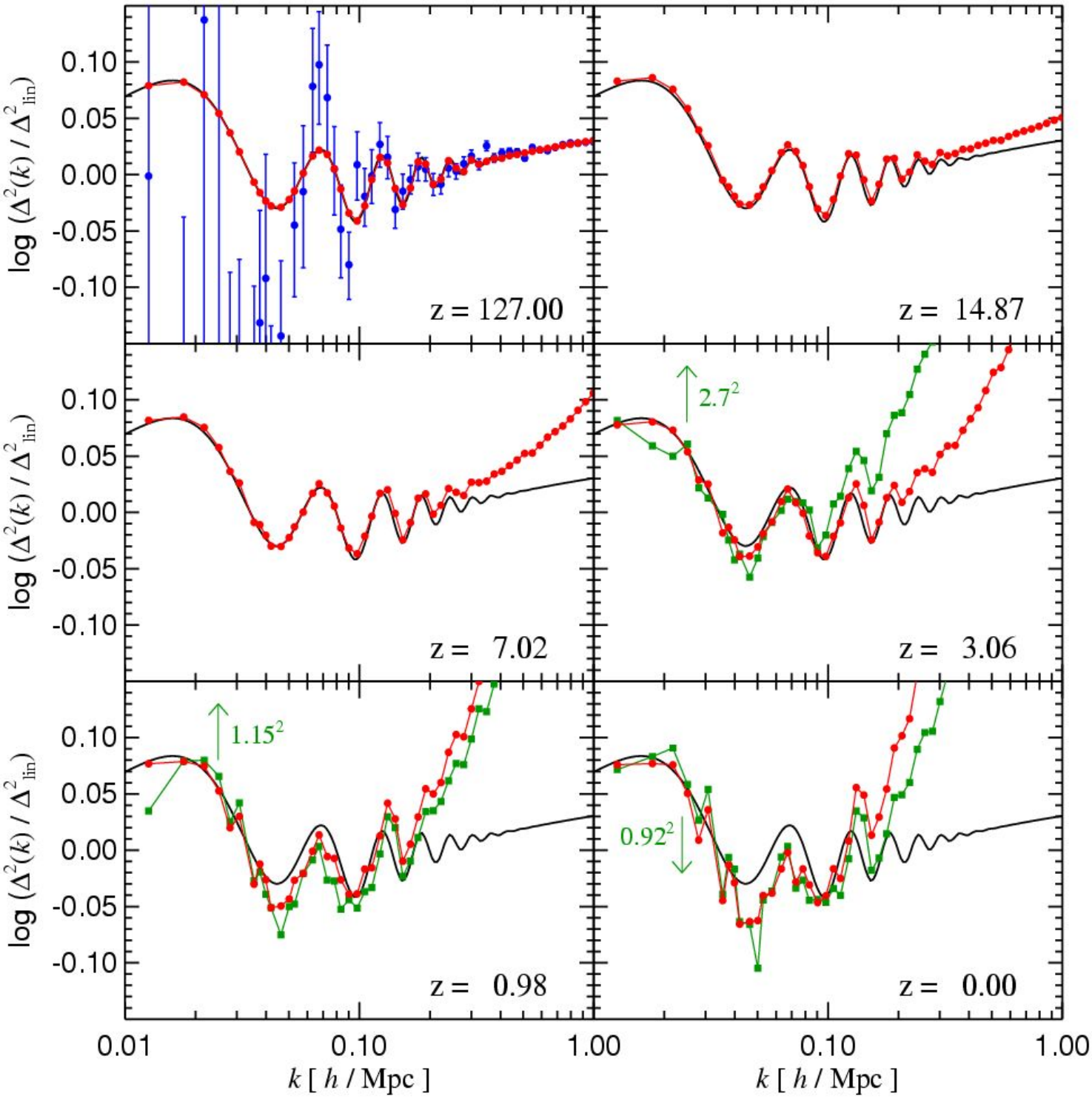






The baryonic wiggles remain visible in the galaxy distribution down to low redshift and may serve as a "standard ruler" to constrain dark energy

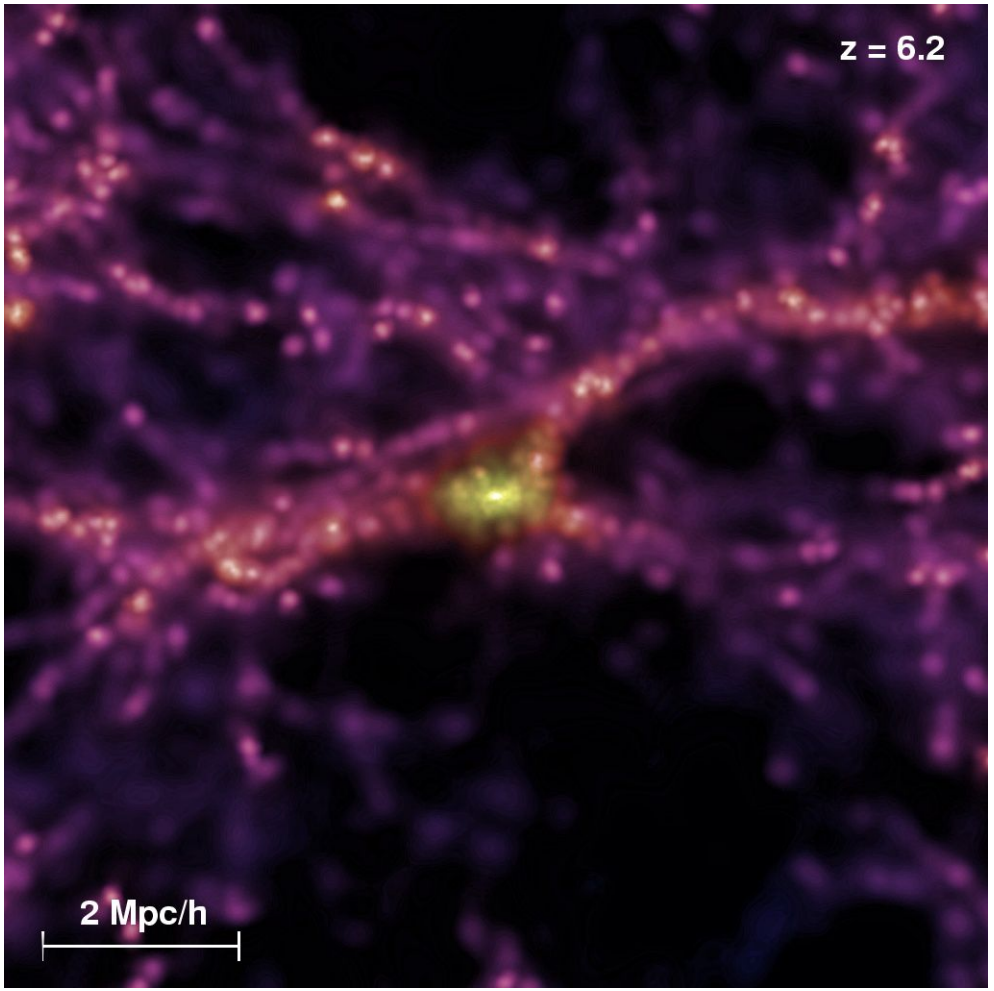
DARK MATTER AND GALAXY POWER SPECTRA IN THE REGION OF THE WIGGLES

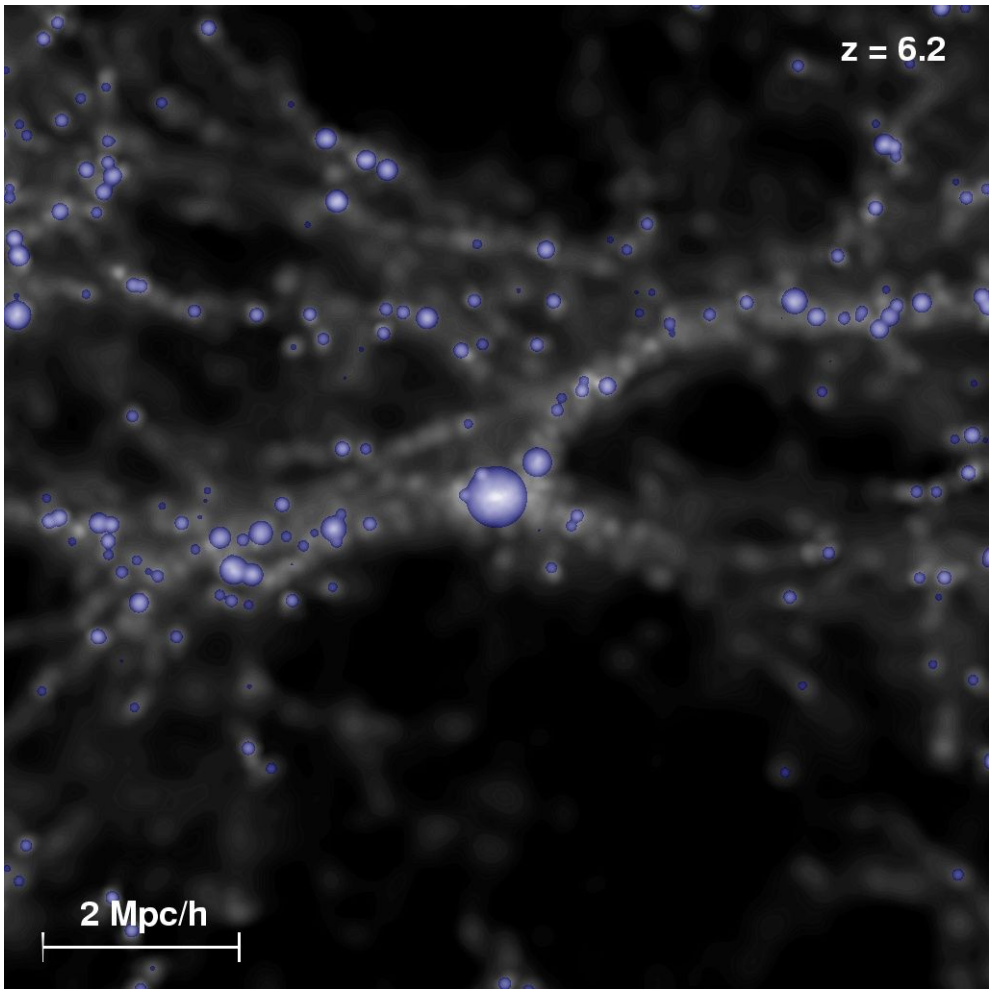


We can identify the halos at z~6.2 as plausible "Sloan" quasar candidates

DARK MATTER AND GALAXY DISTRIBUTION AROUND THE GALAXY WITH THE LARGEST STELLAR MASS AT Z=6.2

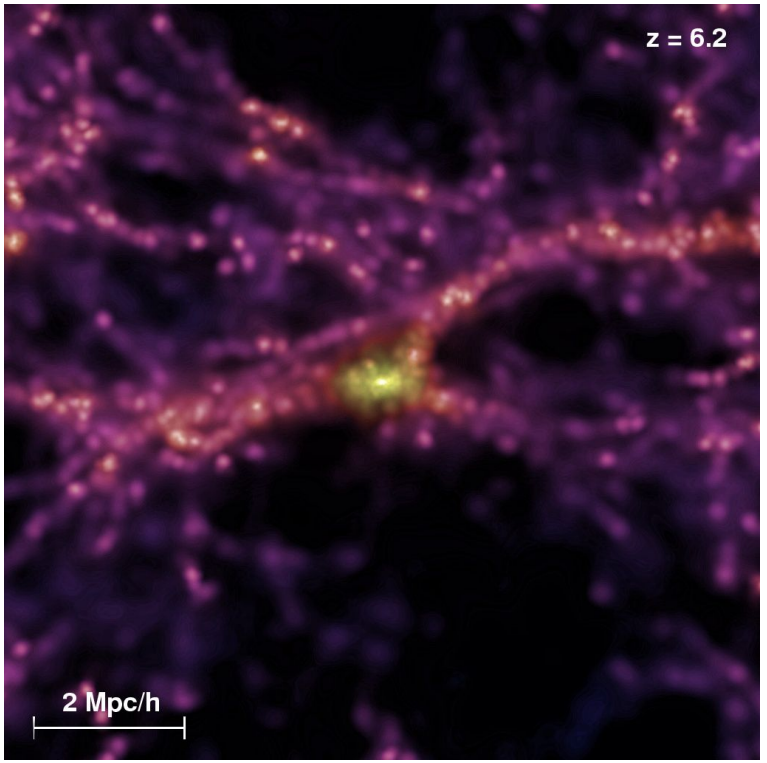
$M_h = 5.3 \times 10^{12}$ $M_* = 8.2 \times 10^{10}$ SFR = 235 $M_\odot$ / yr

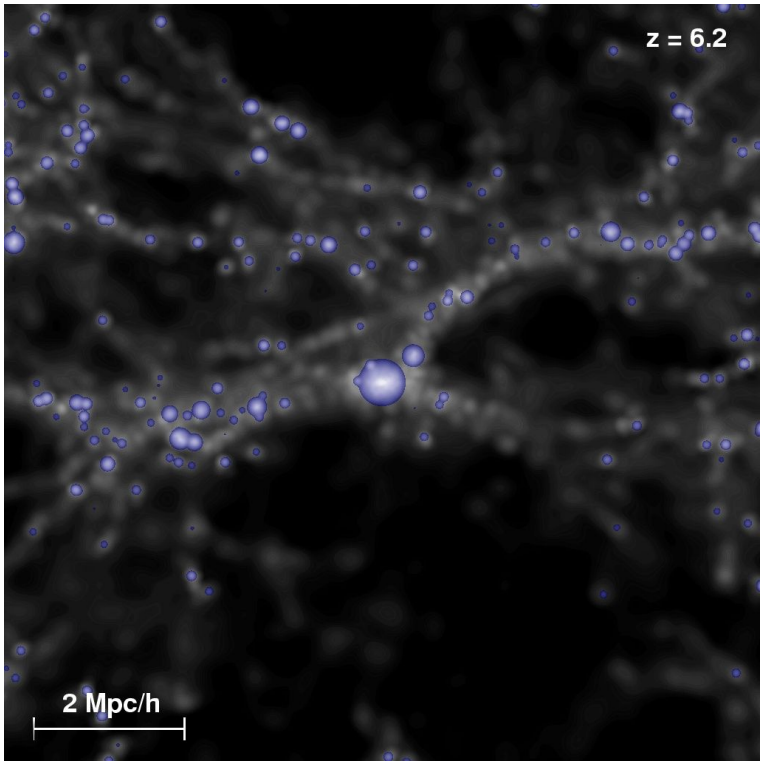


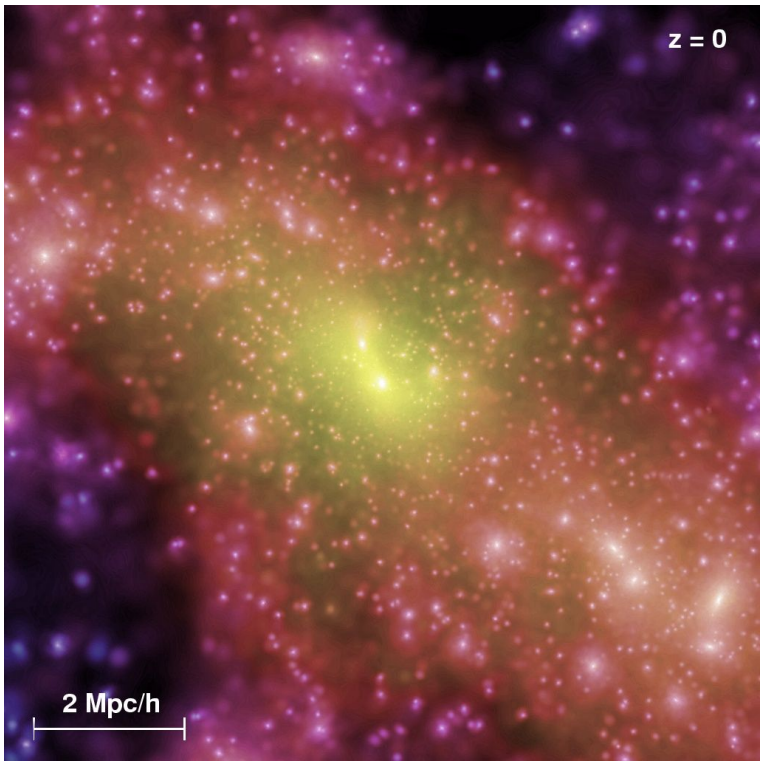


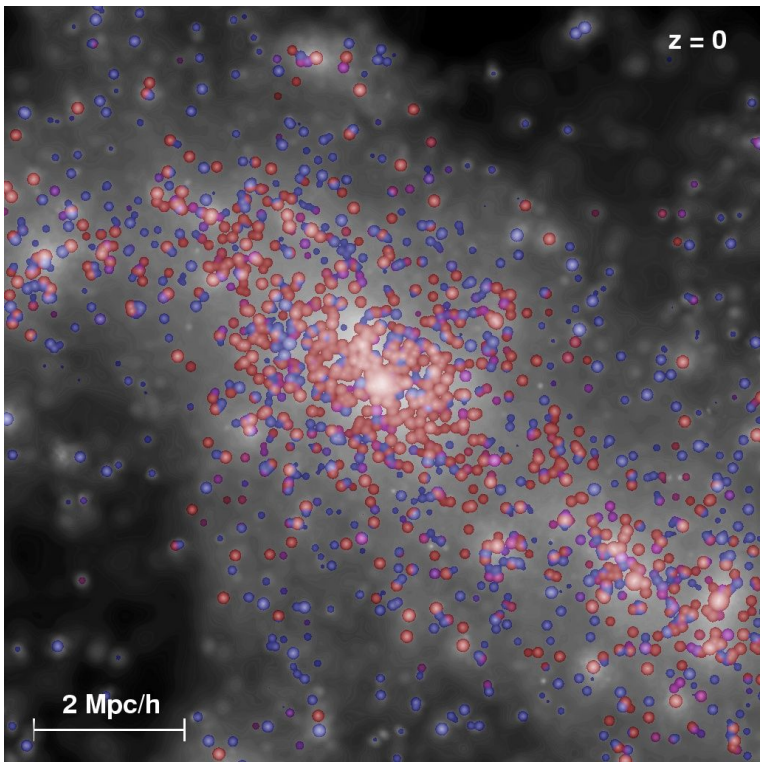
The quasars end up as cD galaxies in rich galaxy clusters today

TRACING GALAXIES OVER COSMIC TIME



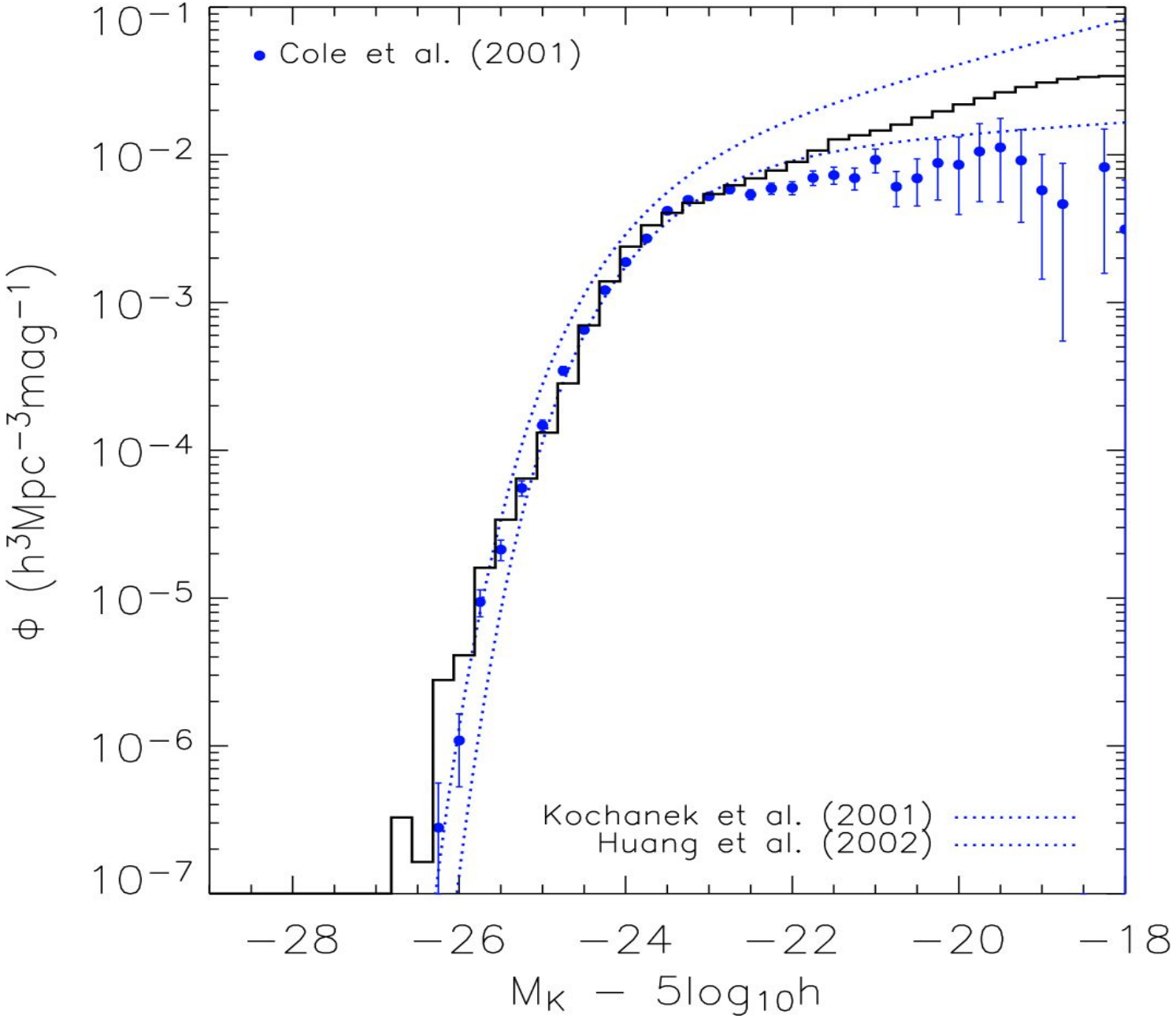






The semi-analytic model fits a multitude of observational data

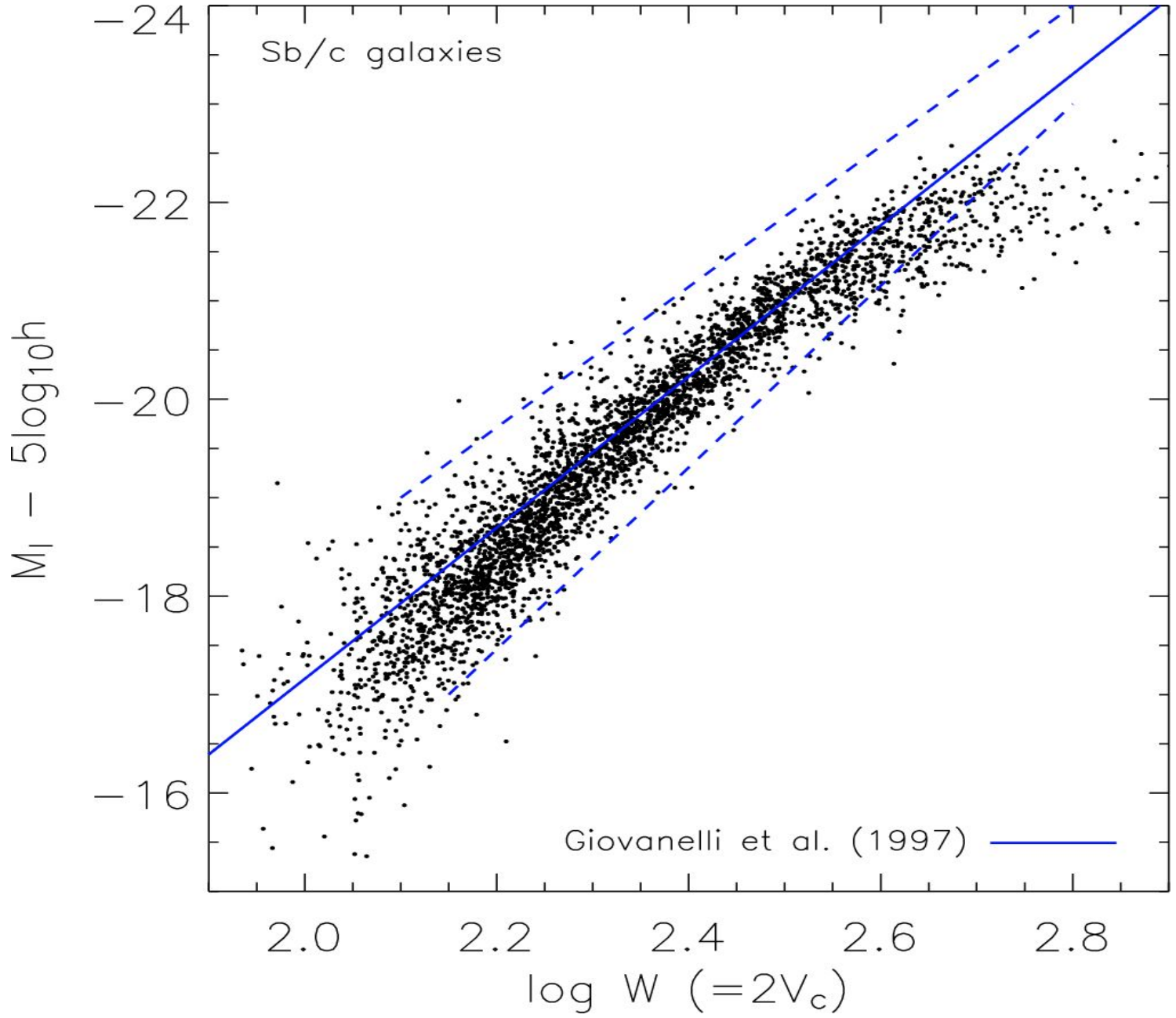
K-BAND LUMINOSITY FUNCTION



Croton et al. (2004)

The semi-analytic model fits a multitude of observational data

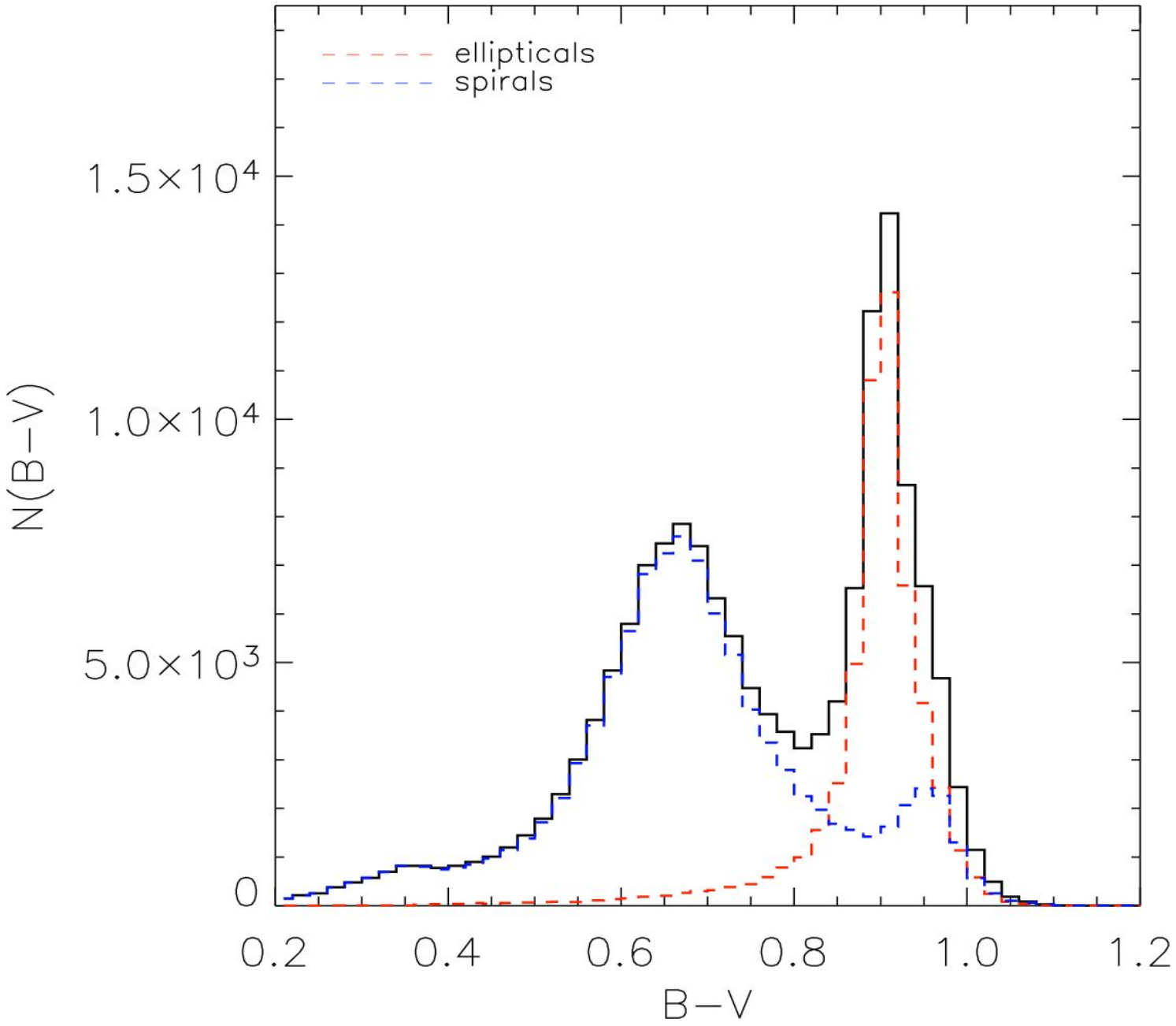
I-BAND TULLY-FISHER



Croton et al. (2004)

The semi-analytic model fits a multitude of observational data

B-V COLOUR DISTRIBUTION



Croton et al. (2004)

The Millennium Run can be used to make accurate predictions for gravitationally induced distortions in the CMB

THE INTEGRATED SACHS-WOLFE AND REES-SCIAMA EFFECTS

$$\frac{\Delta T}{T} = -\frac{2}{c^3}\int \mathrm{d}s \frac{\partial \Phi}{\partial t}$$

The change in the potential can be decomposed into different contributions:

- Decaying potential effect (ISW)
- Non-linear growth of structures (intrinsic Rees-Sciama effect)
- Transverse motion of a structure (proper motion Rees-Sciama effect)

Note: For linear growth in a Ω=1 universe, the potential is frozen in !

For linear growth:

$$\Phi \propto \frac{D(a)}{a}$$

(Rubino-Martin, Hernandez-Monteagudo & Ensslin 2004)

To separate these contributions, I propose the following decomposition:

$$\frac{\partial \Phi}{\partial t} = \frac{\partial}{\partial t}\left(\frac{a}{D(a)}\Phi\right) + \frac{\partial}{\partial t}\left[\left(1 - \frac{a}{D(a)}\right)\Phi\right]$$

Rees-Sciama term | ISW term

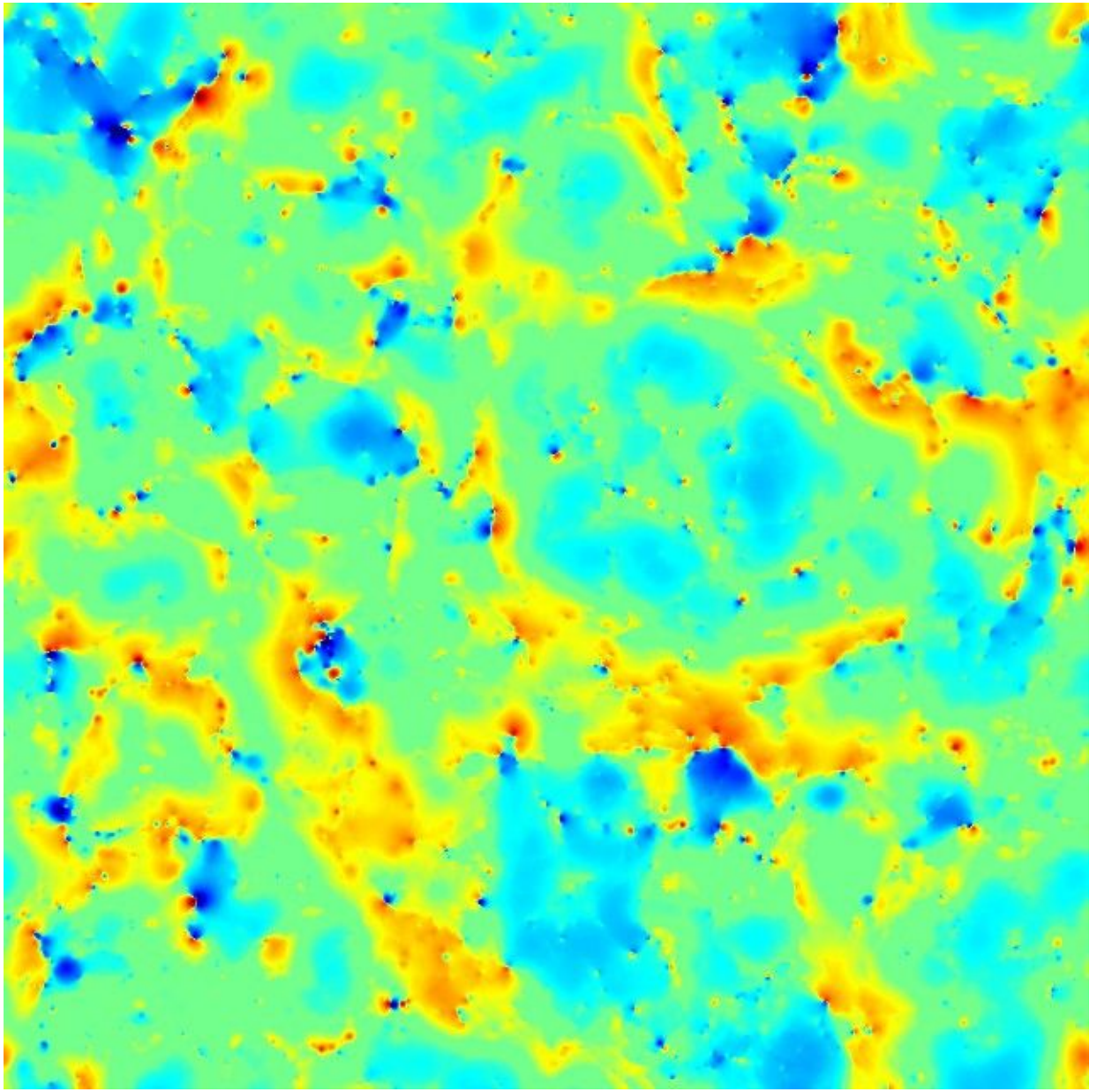
The Rees-Sciama Effect is sensitive to motion of massive systems, and to forming clusters and voids

THE REES-SCIAMA PART OF THE TIME DERIVATIVE OF THE POTENTIAL

The RS constribution

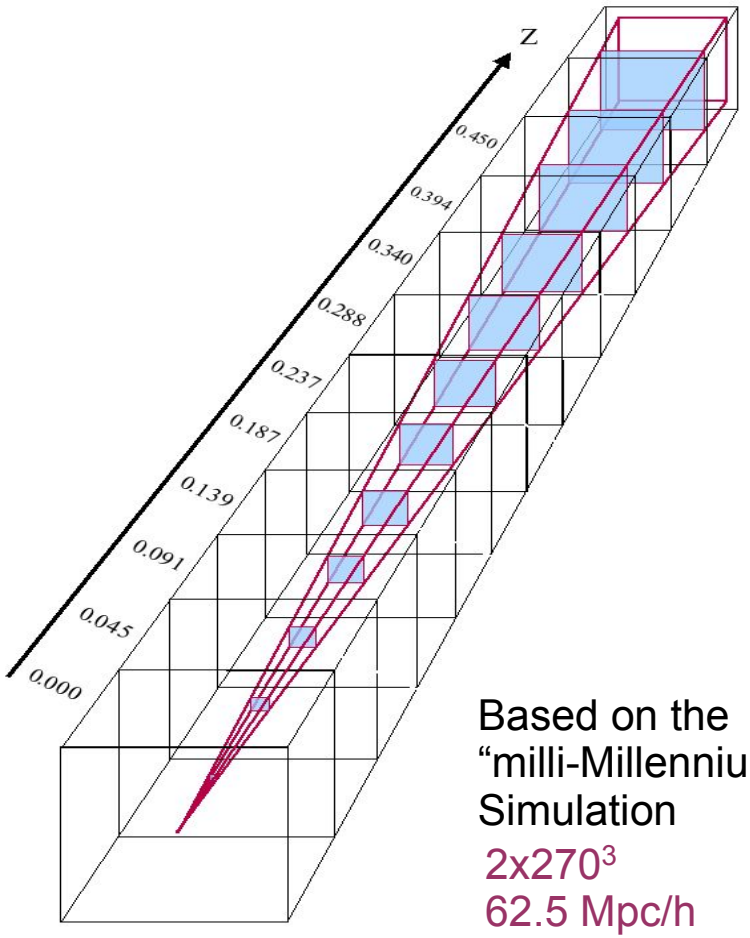
$$\frac{\partial}{\partial t}\left(\frac{a}{D(a)}\,\Phi\right)$$

in a slice through the Millennium Run at z=0



Ray tracing along the backwards light-cone can be used to compute the total temperature anisotropy due to the Rees-Sciama effect

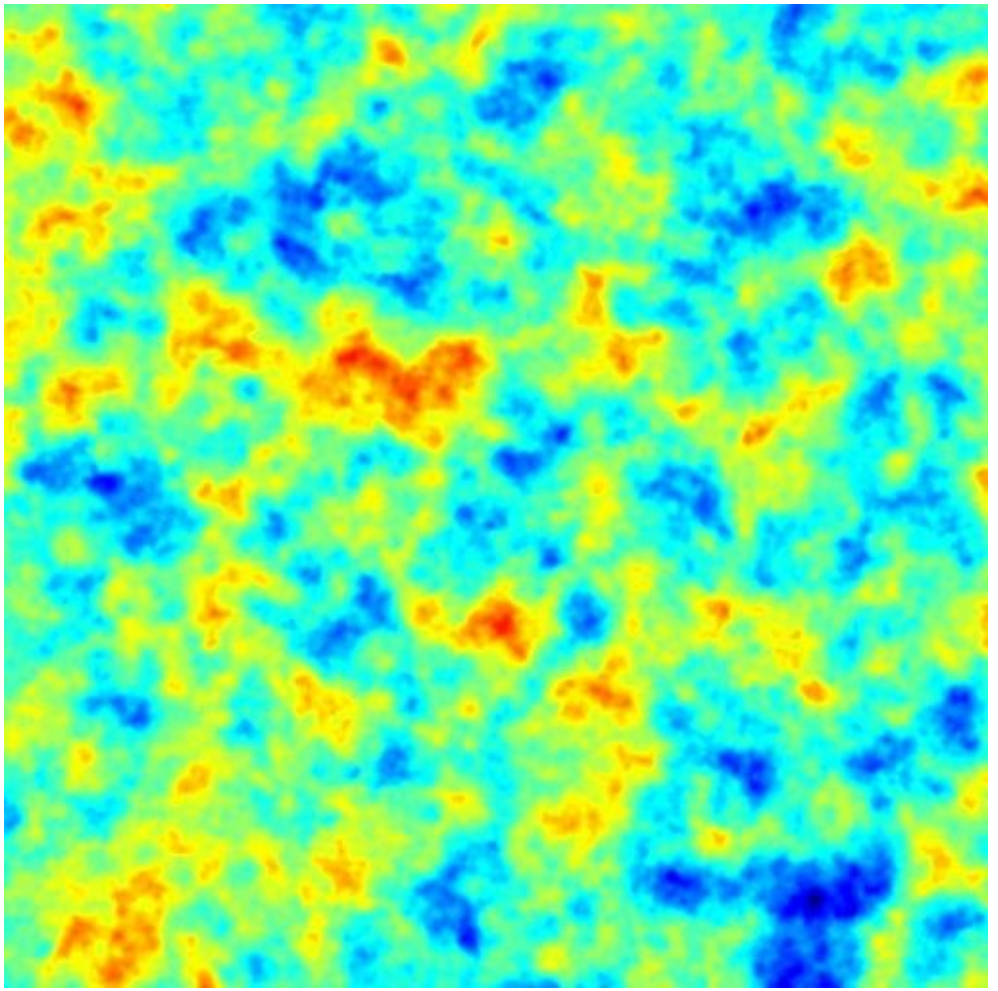
REES-SCIAMA TEMPERTATURE ANISOTROPY MAP



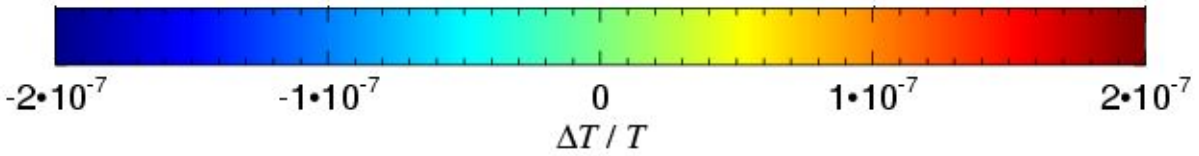
Based on the "milli-Millennium" Simulation

$2x270^3$

62.5 Mpc/h



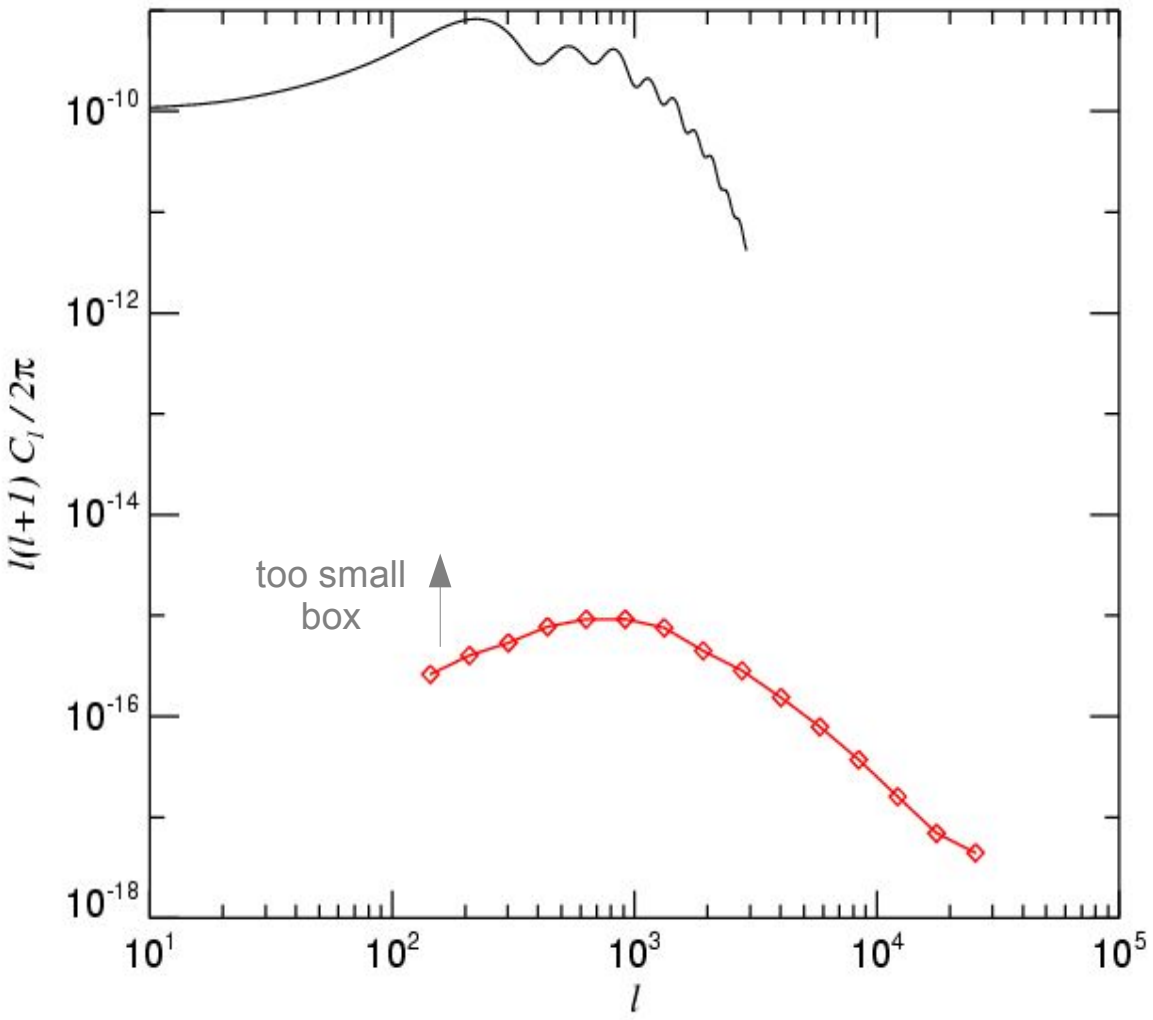
3°



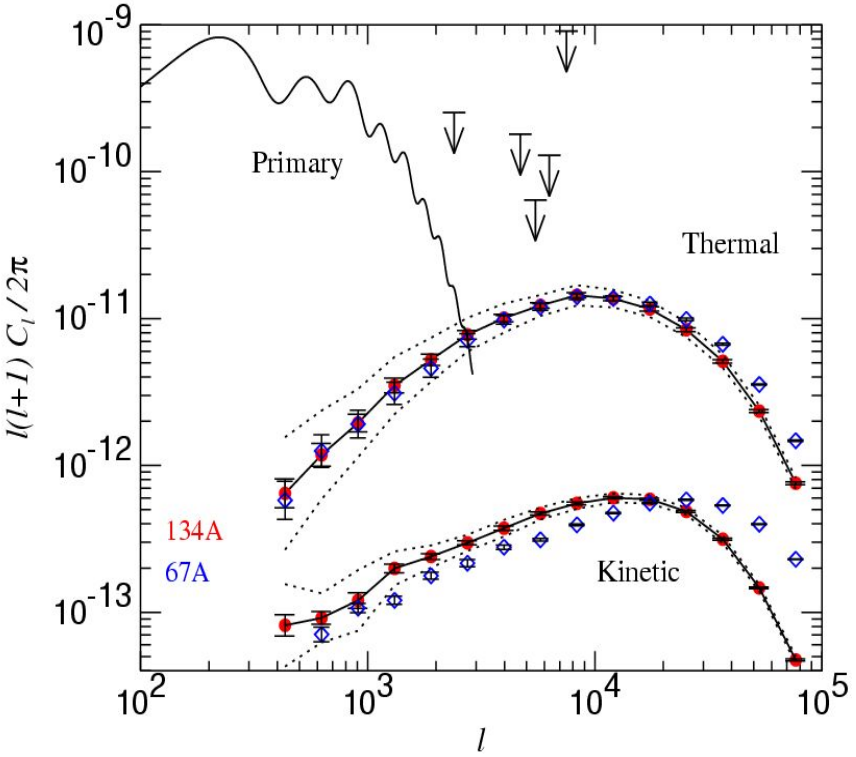
The angular power spectrum of Rees-Sciama anisotropies falls significantly below the primary anisotropies

REES-SCIAMA TEMPERATURE FLUCTUATION SPECTRUM

Rees-Sciama (small box)



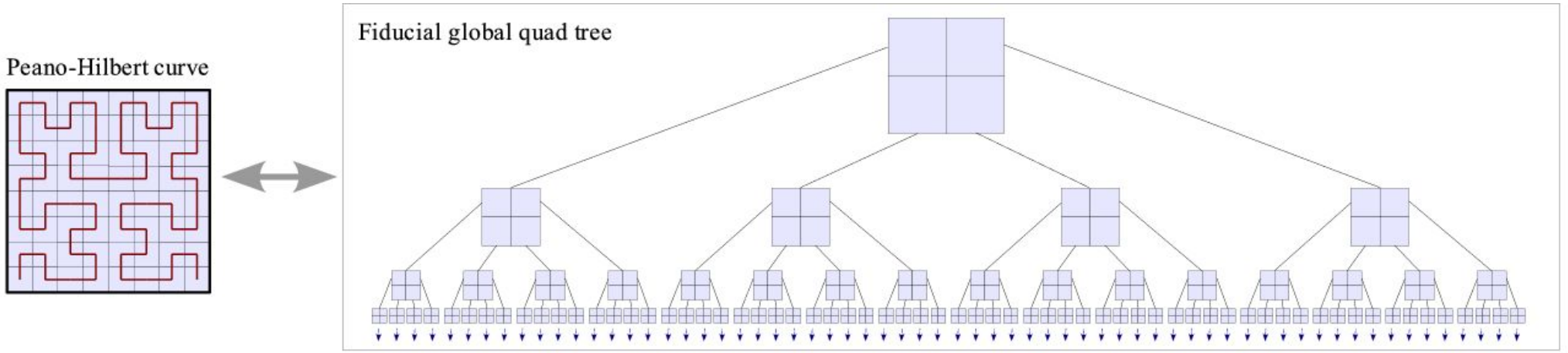
Thermal & Kinetic SZ

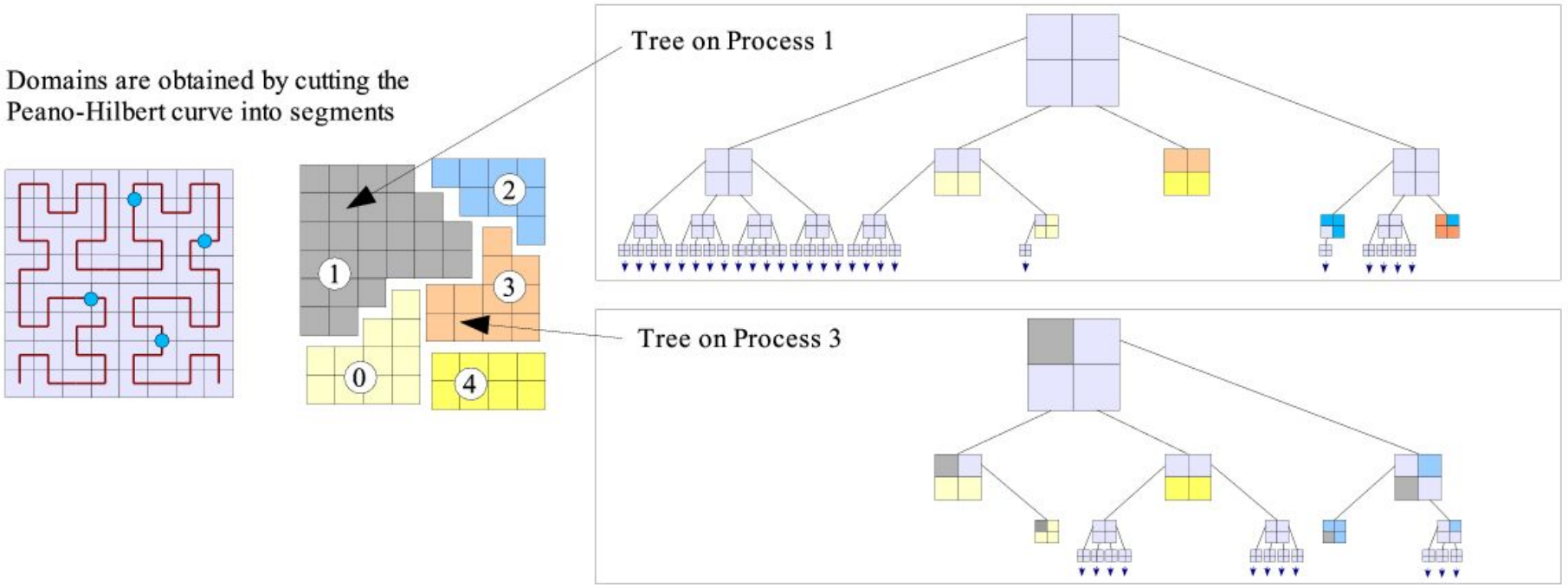


Springel, White & Hernquist (2001)

A space-filling Peano-Hilbert curve is used in GADGET-2 for a novel domain-decomposition concept

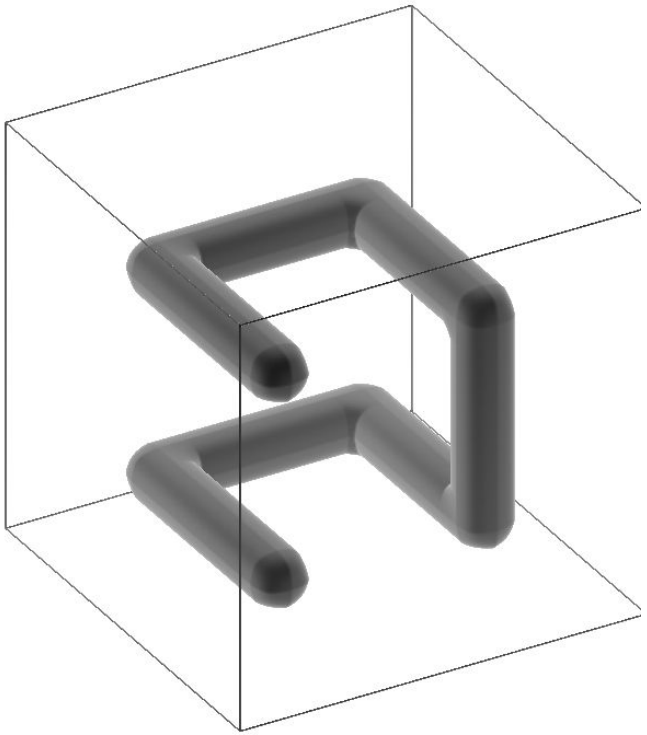
HIERARCHICAL TREE ALGORITHMS

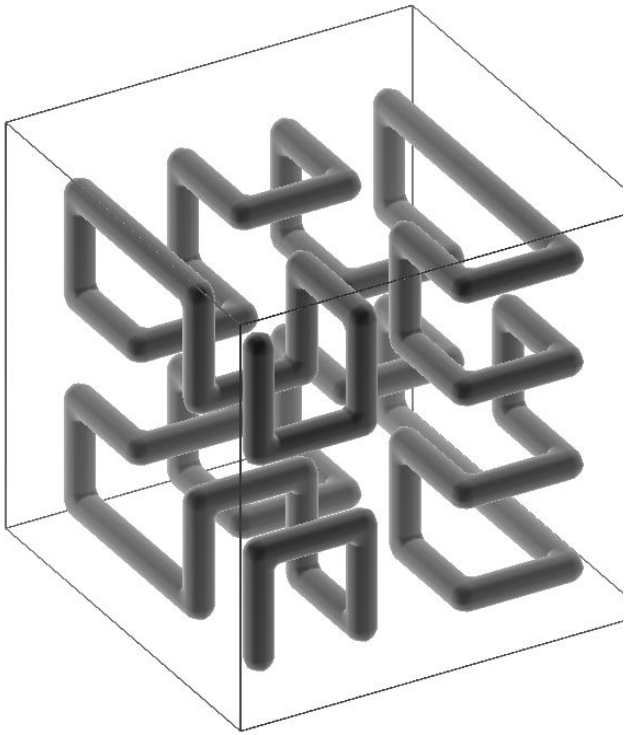


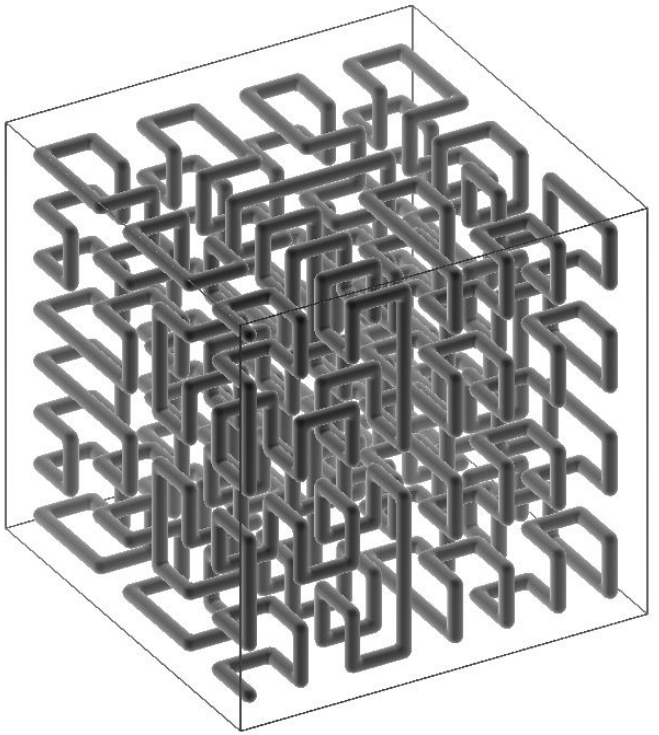


The space-filling Hilbert curve can be readily generalized to 3D

THE PEANO-HILBERT CURVE







The TreePM technique combines the advantages of PM-method and Tree-algorithm

THE TREE-PM FORCE SPLIT

Periodic peculiar potential

$$\nabla^2\phi(\mathbf{x}) = 4\pi G[\rho(\mathbf{x}) - \overline{\rho}] = 4\pi G \sum_{\mathbf{n}} \sum_{i} m_i \left[\tilde{\delta}(\mathbf{x} - \mathbf{x}_i - \mathbf{n}L) - \frac{1}{L^3}\right]$$

Idea: Compute the long-range force with the PM algorithm, and only a **local** short-range force with the tree.

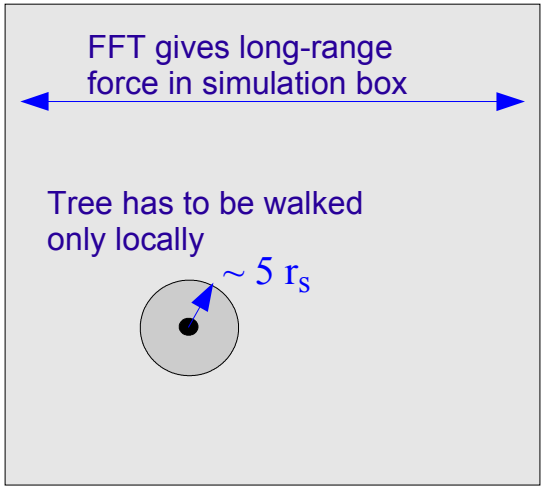
Let's split the potential in Fourier space into a long-range and a short-range part:

$$\phi_{\mathbf{k}}^{\text{long}} = \phi_{\mathbf{k}} \exp(-\mathbf{k}^2 r_s^2)$$

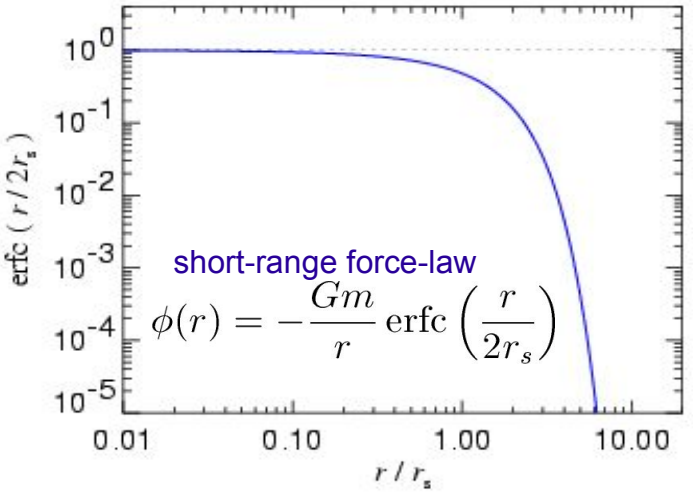
$$\phi_{\mathbf{k}}^{\text{short}} = \phi_{\mathbf{k}} \left[1 - \exp(-\mathbf{k}^2 r_s^2)\right]$$

Solve with PM-method

- CIC mass assignment
- **FFT**
- multiply with kernel
- **FFT backwards**
- Compute force with 4-point finite difference operator
- Interpolate forces to particle positions



Solve in real space with TREE

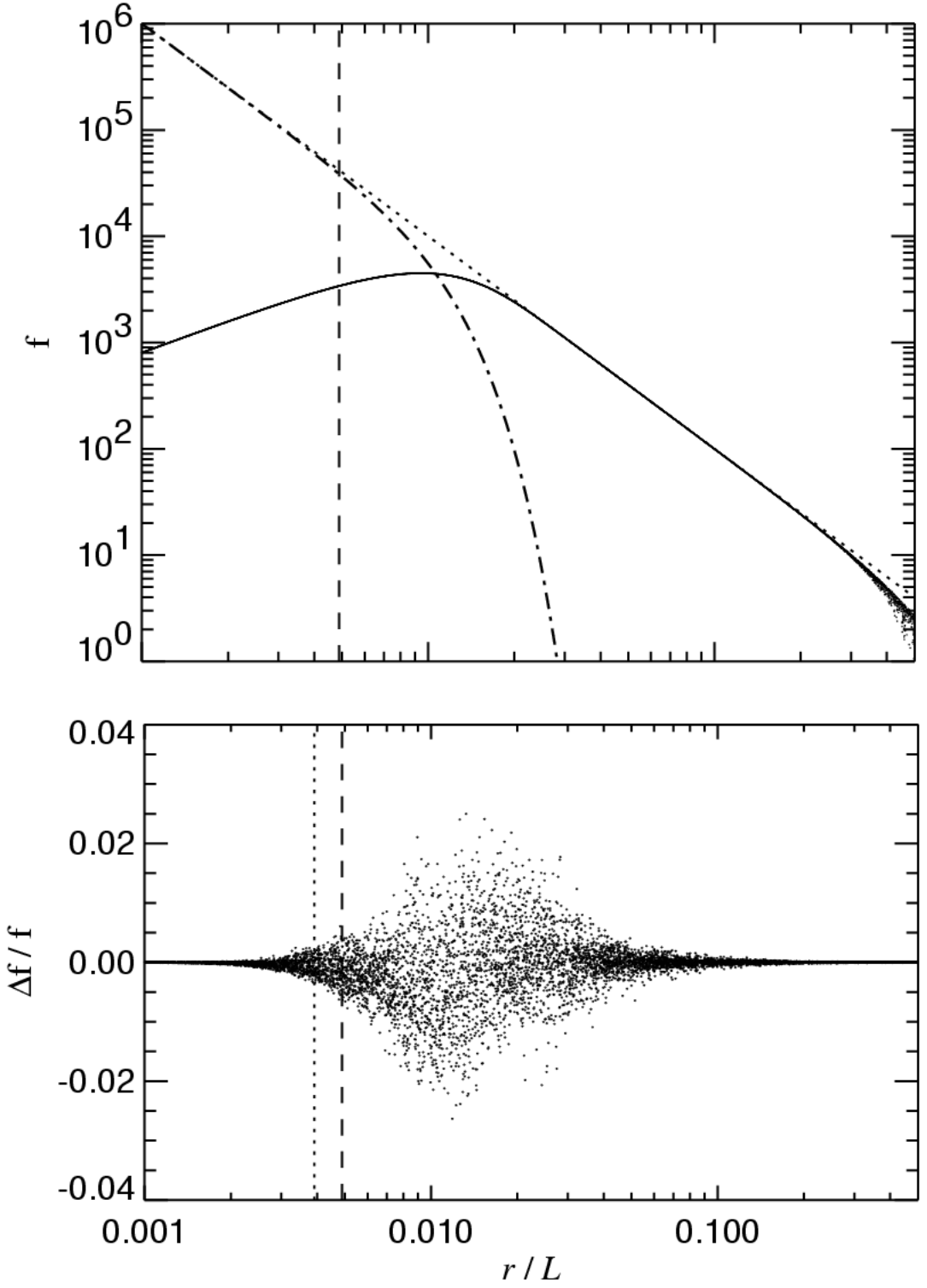


Advantages of this algorithm include:

- Accurate and fast long-range force
- No force anisotropy
- Speed is insensitive to clustering (as for tree algorithm)
- No Ewald correction necessary for periodic boundary conditions

The TreePM technique produces small errors in the matching region between PM and Tree

FORCE DECOMPOSITION AND ERRORS IN THE TREE-PM SPLIT



Symplectic integration schemes can be generated by applying the idea of operating splitting to the Hamiltonian

THE LEAPFROG AS A SYMPLECTIC INTEGRATOR

Separable Hamiltonian

$$H = H_{\mathrm{kin}} + H_{\mathrm{pot}}$$

Drift- and Kick-Operators

$$\mathbf{D}(\Delta t) \equiv \exp\left(\int_t^{t+\Delta t} \mathrm{d}t\, \mathbf{H}_{\mathrm{kin}}\right) = \left\{ \begin{array}{lcl} \mathbf{p}_i & \mapsto & \mathbf{p}_i \\ \mathbf{x}_i & \mapsto & \mathbf{x}_i + \frac{\mathbf{p}_i}{m_i}\Delta t \end{array} \right.$$

$$\mathbf{K}(\Delta t) = \exp\left(\int_t^{t+\Delta t} \mathrm{d}t\, \mathbf{H}_{\mathrm{pot}}\right) = \left\{ \begin{array}{lcl} \mathbf{x}_i & \mapsto & \mathbf{x}_i \\ \mathbf{p}_i & \mapsto & \mathbf{p}_i - \sum_j m_i m_j \frac{\partial \phi(\mathbf{x}_{ij})}{\partial \mathbf{x}_i}\Delta t \end{array} \right.$$

The drift and kick operators are symplectic transformations of phase-space !

The Leapfrog

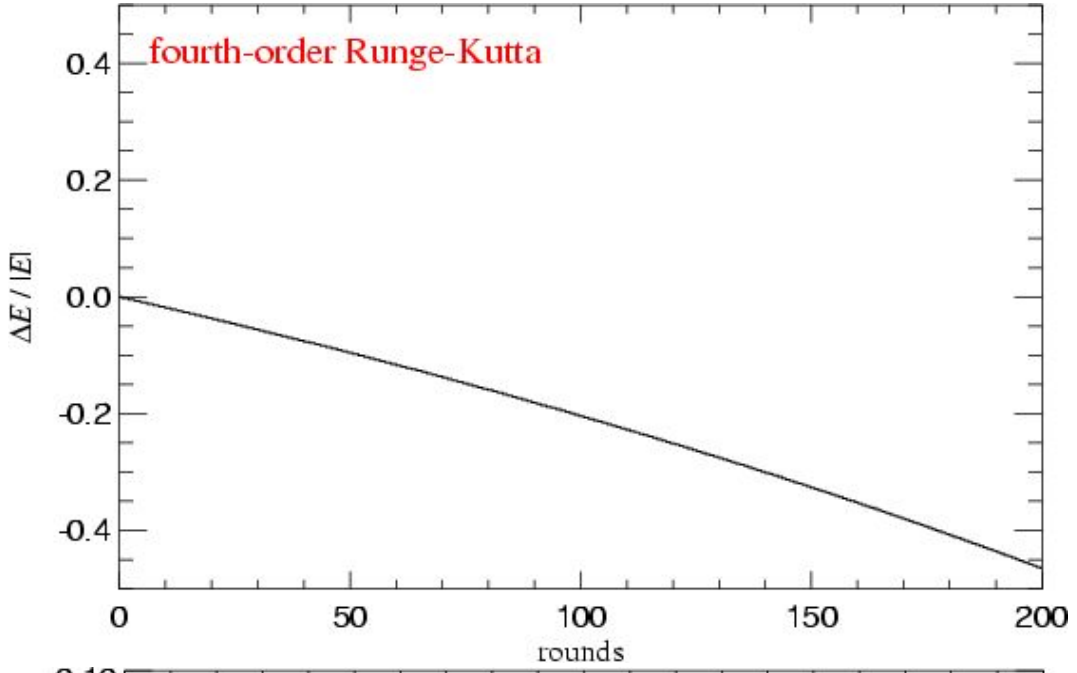
Drift-Kick-Drift: $$\tilde{\mathbf{U}}(\Delta t) = \mathbf{D}\left(\frac{\Delta t}{2}\right)\, \mathbf{K}(\Delta t)\, \mathbf{D}\left(\frac{\Delta t}{2}\right)$$

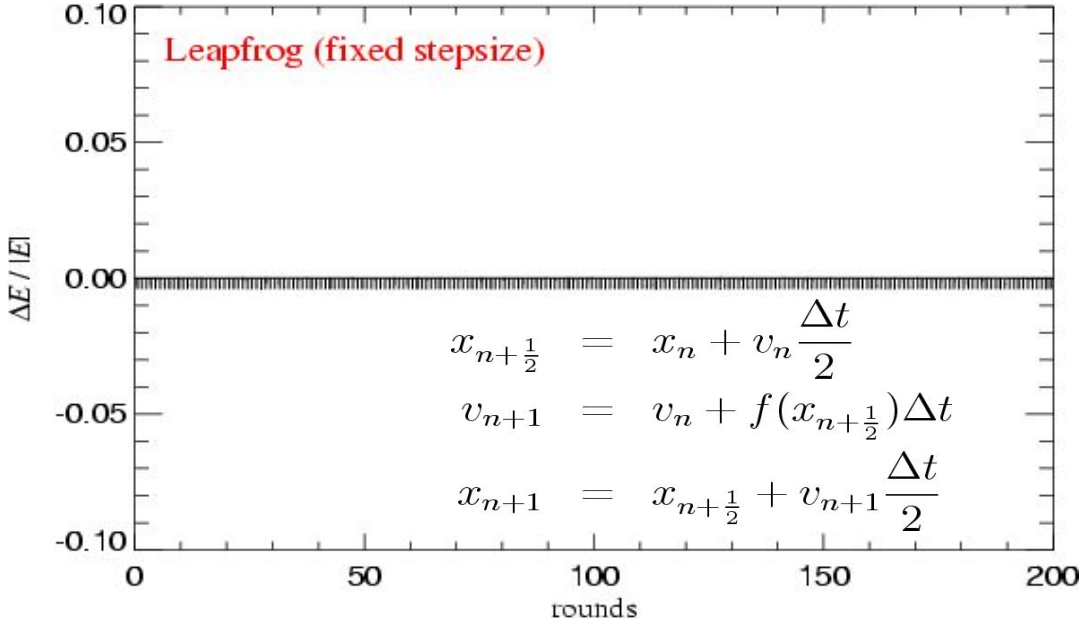
Kick-Drift-Kick: $$\tilde{\mathbf{U}}(\Delta t) = \mathbf{K}\left(\frac{\Delta t}{2}\right)\, \mathbf{D}(\Delta t)\, \mathbf{K}\left(\frac{\Delta t}{2}\right)$$

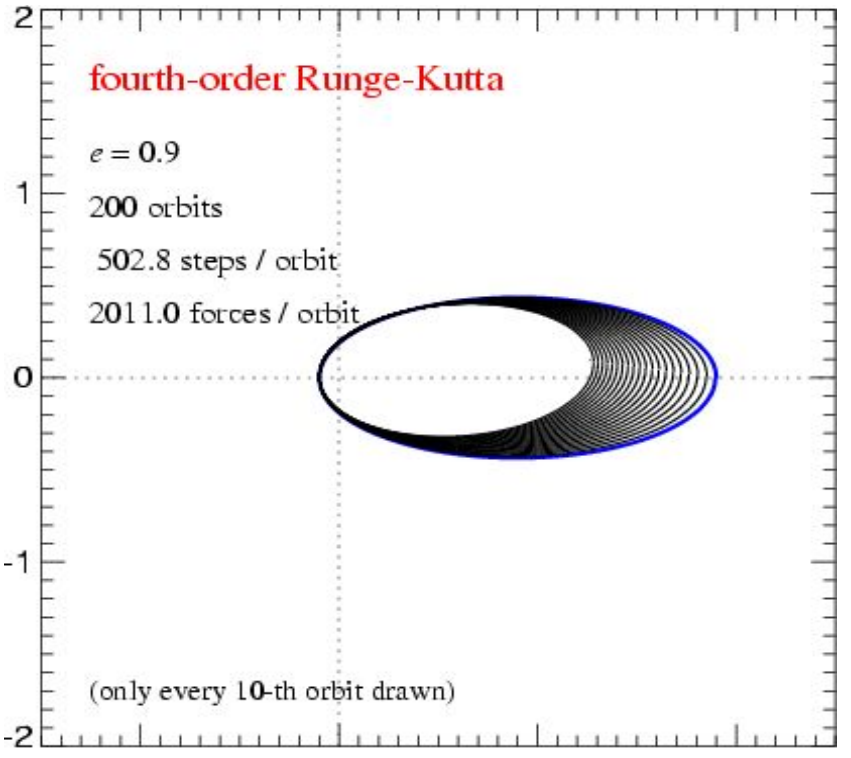
Hamiltonian of the numerical system: $$\tilde{H} = H + H_{\mathrm{err}} \qquad H_{\mathrm{err}} = \frac{\Delta t^2}{12}\left\{\left\{H_{\mathrm{kin}}, H_{\mathrm{pot}}\right\}, H_{\mathrm{kin}} + \frac{1}{2}H_{\mathrm{pot}}\right\} + \mathcal{O}(\Delta t^3)$$

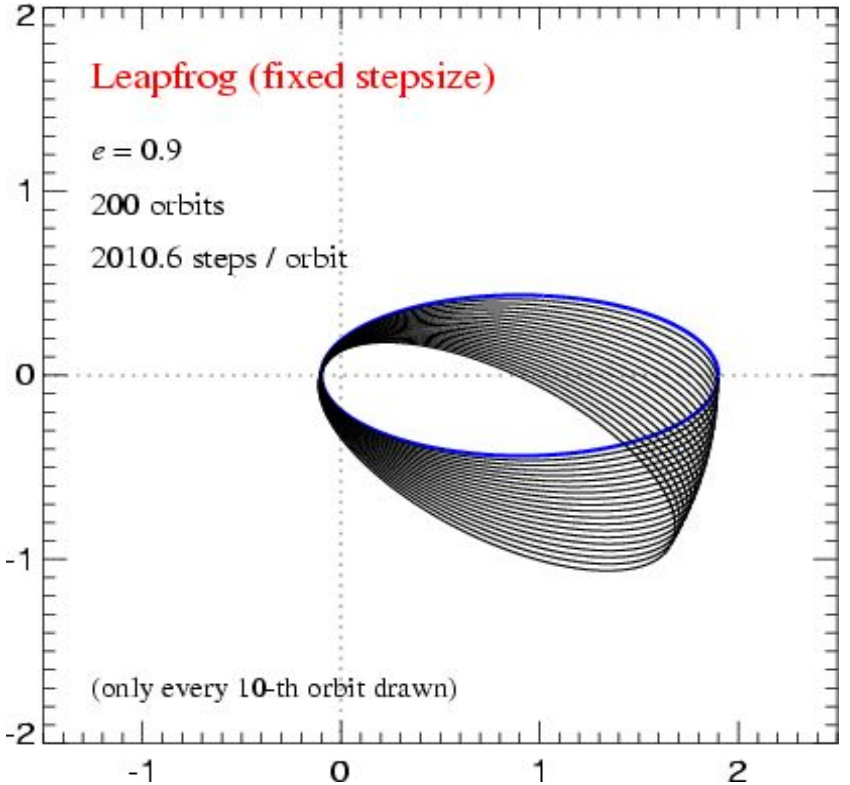
The leapfrog is behaving much better than one might expect...

INTEGRATING THE KEPLER PROBLEM



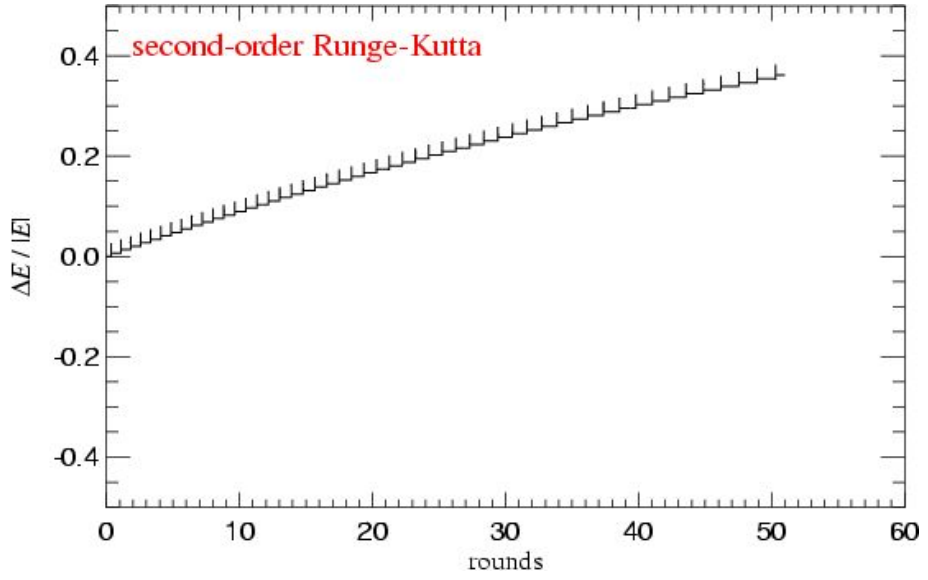


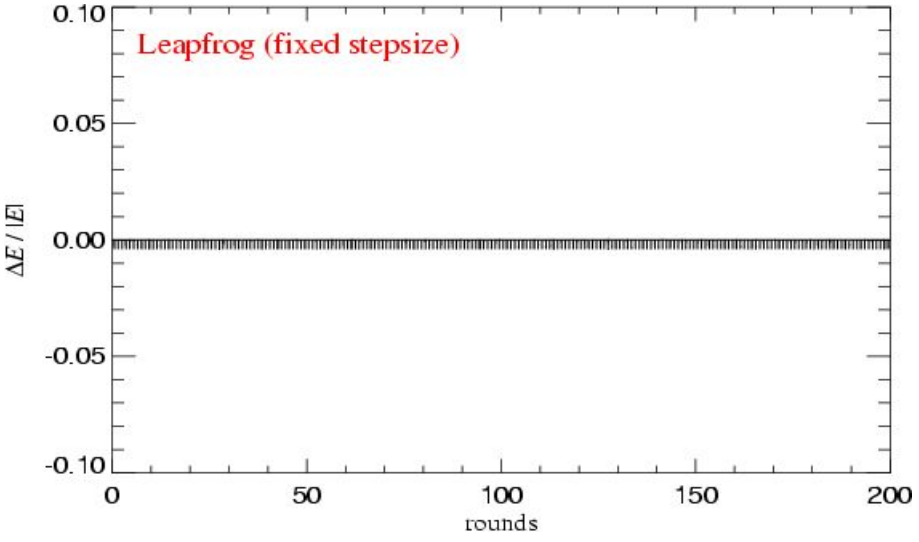


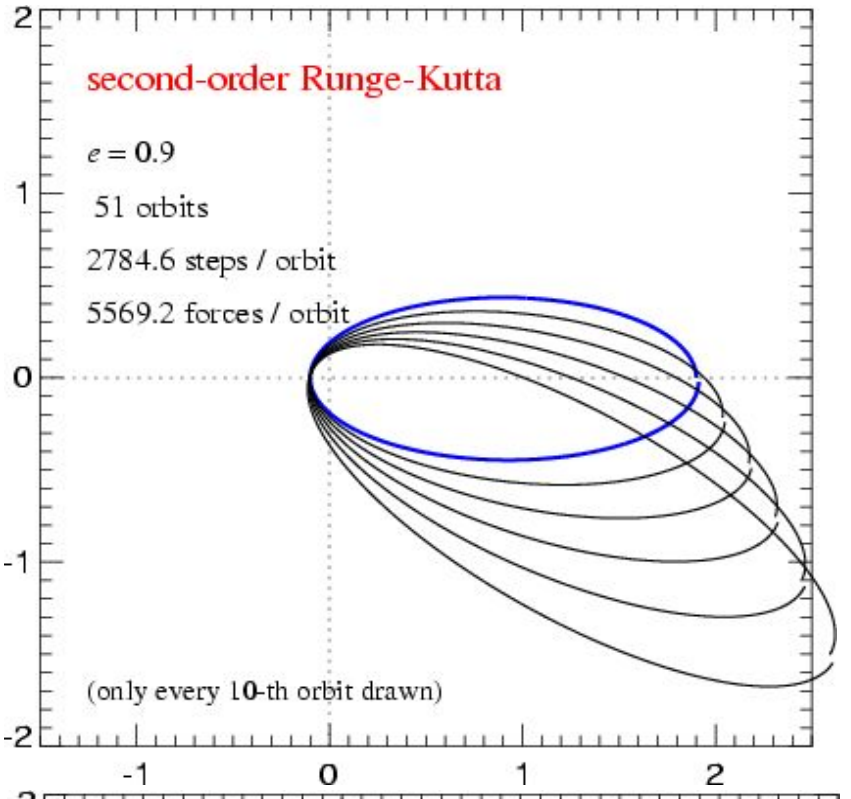


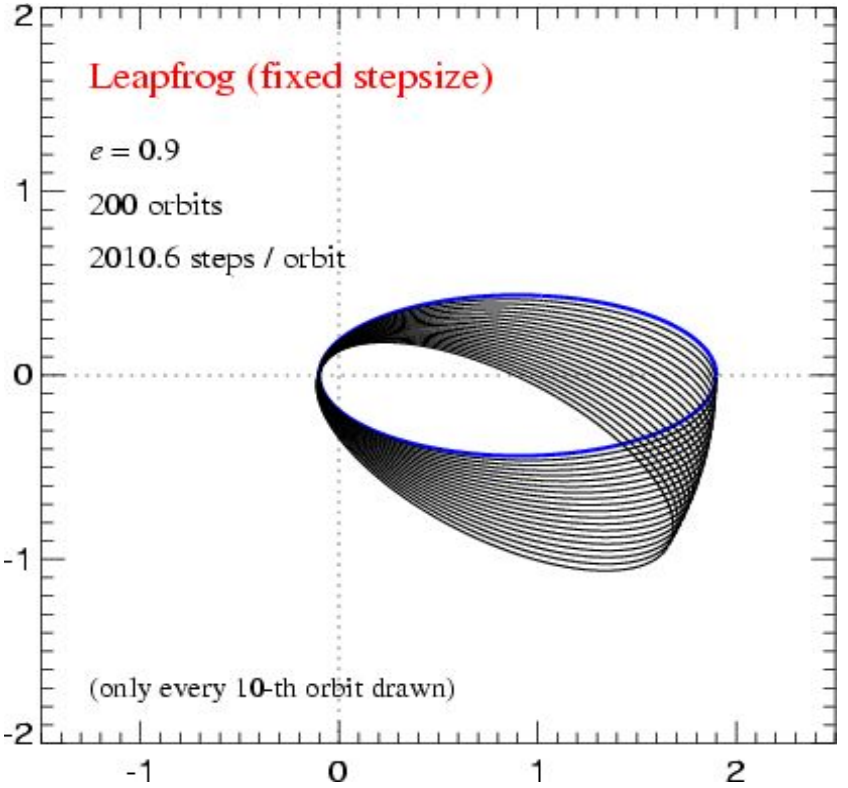
When compared with an integrator of the same order, the leapfrog is highly superior

INTEGRATING THE KEPLER PROBLEM









The force-split can be used to construct a symplectic integrator where long- and short-range forces are treated independently

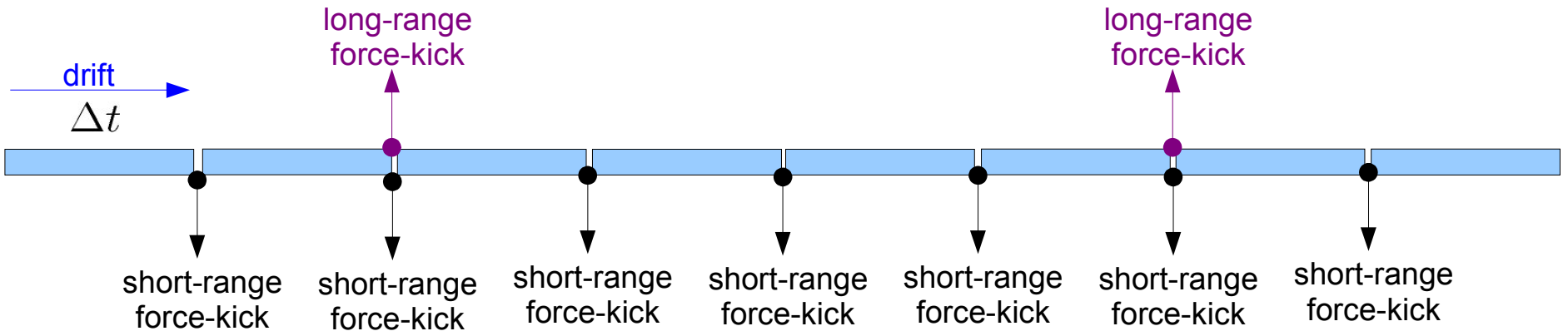
TIME INTEGRATION FOR LONG AND SHORT-RANGE FORCES

Separate the potential into a long-range and a short-range part:

$$H = \sum_i \frac{\mathbf{p}_i^2}{2m_i a(t)^2} + \frac{1}{2}\sum_{ij} \frac{m_i m_j\, \varphi_{\rm sr}(\mathbf{x}_i - \mathbf{x}_j)}{a(t)} + \frac{1}{2}\sum_{ij} \frac{m_i m_j\, \varphi_{\rm lr}(\mathbf{x}_j - \mathbf{x}_j)}{a(t)}$$

The short-range force can then be evolved in a symplectic way on a smaller timestep than the long range force:

$$\tilde{\mathbf{U}}(\Delta t) = \mathbf{K}_{\rm lr}\left(\frac{\Delta t}{2}\right)\left[\mathbf{K}_{\rm sr}\left(\frac{\Delta t}{2m}\right)\mathbf{D}\left(\frac{\Delta t}{m}\right)\mathbf{K}_{\rm sr}\left(\frac{\Delta t}{2m}\right)\right]^m \mathbf{K}_{\rm lr}\left(\frac{\Delta t}{2}\right)$$



Conclusions

- We have implemented new numerical methods which allow us to carry out unprecedently large, high-resolution cosmological N-body simulations.
  We have achieved $N>10^{10}$, with a formal dynamic range of 10^5 in 3D.

- The simulation allows high-precision measurements of the dark matter clustering, like halo mass function, power spectrum, abundance of dark matter substructure, higher-order counts-in-cells, etc.

- An advanced semi-analytic model has been developed. It allows the construction of theoretical mock galaxy catalogues, describing the history of 25 million galaxies. This allows new tests of hierarchical galaxy formation and assessments of the relative importance of different physics for galaxy formation. The galaxy catalogues will form one of the backbones of an emerging *Theoretical Virtual Observatory*.

- The two-point correlation function of galaxies is in very good agreement with observations. The theoretical galaxy model also reproduces the observed trends of clustering strength with magnitude and color.

- Accoustic oscillations are partly washed out by non-linear evolution. While being affected already by non-linear effects, the first and second peak survive as features down to z=0, and should be still be present in the galaxy distribution today.