Machine Learning Does It and Does It Better:

Unearthing Primordial Dark-Matter Velocities from the Matter Power Spectrum

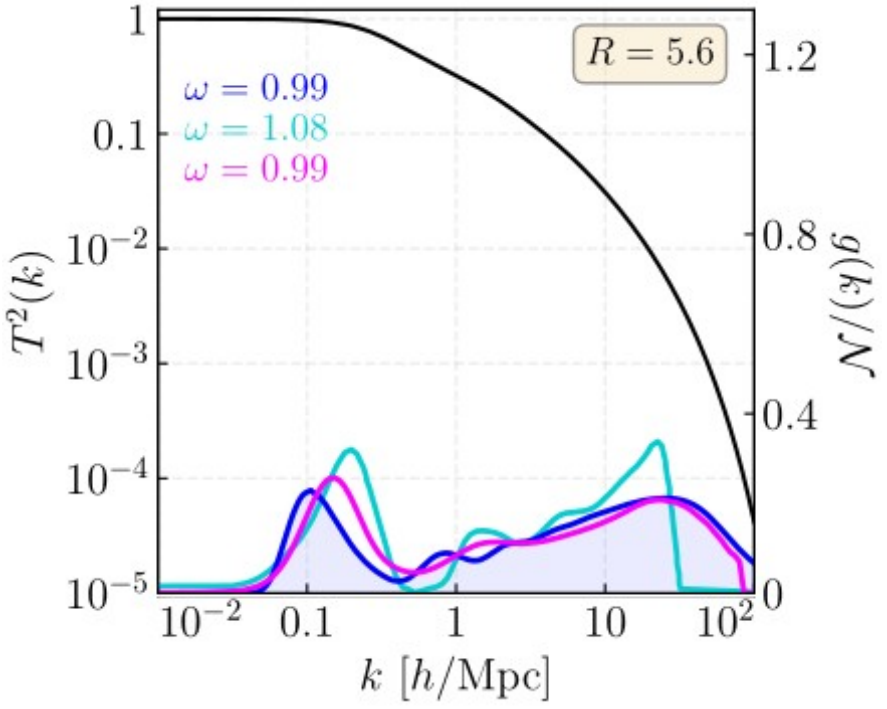
Brooks Thomas

LAFAYETTE COLLEGE

Based on work done in collaboration with:

- **Keith Dienes, Jessica Howard, Fei Huang, Yuanzhen Li [arXiv:2605.xxxxx]**

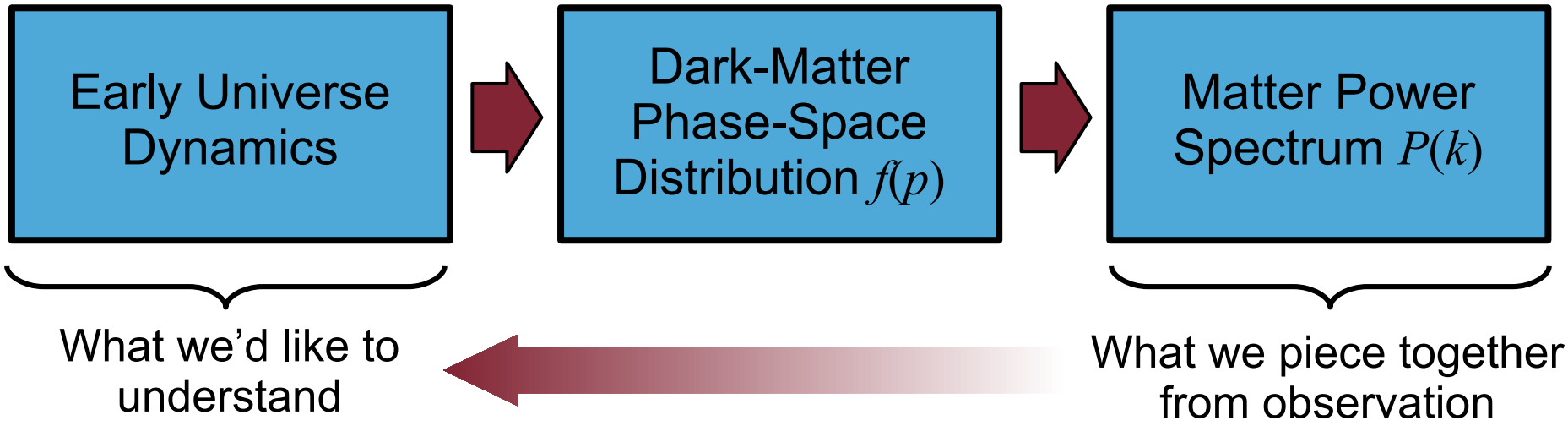
This research is supported in part by NSF



PHENO 2026, May 12th, 2026

The Basic Question

- The early-universe dynamics which produces the dark matter gives rise to a particular dark-matter phase-space distribution $f(p)$. This, in turn, affects the shape of the matter power spectrum $P(k)$.



Q To what extent can we work backwards and ***reconstruct*** the properties of $f(p)$ – and the dynamics that gave rise to it – from information encoded in $P(k)$?

- While the maps from the underlying physics to $f(p)$, and from $f(p)$ to $P(k)$ are clearly not invertible, it is nevertheless possible to "work backwards" and obtain substantial information about the dark sector from information contained in the matter power spectrum.

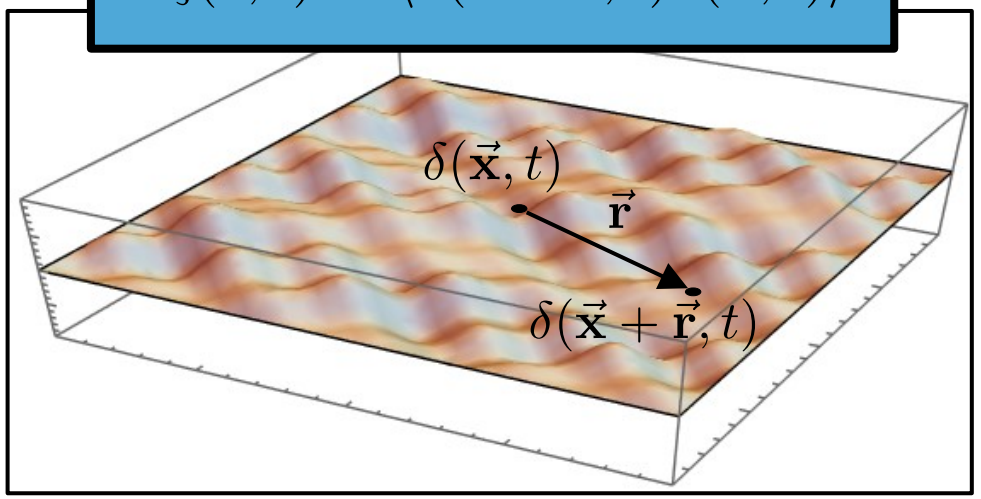
Matter Power Spectrum

- The **matter power spectrum** encapsulates information about the spatial distribution of matter in our universe.

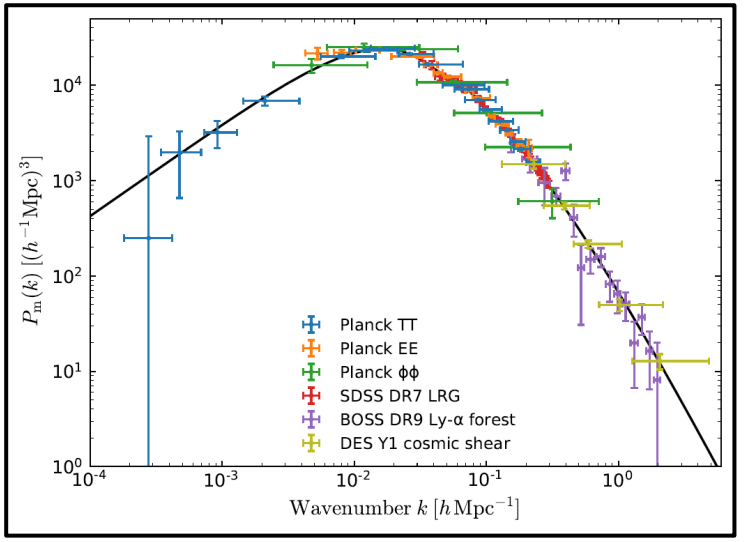
$$P(k,t) \equiv 4\pi \int dr\, r^2 \frac{\sin(kr)}{kr} \xi(r,t)$$

Two-point correlation function for the fractional matter overdensity

$$\xi(\vec{\mathbf{r}},t) = \langle \delta(\vec{\mathbf{x}}+\vec{\mathbf{r}},t)\delta(\vec{\mathbf{x}},t)\rangle$$



Linear Matter Power Spectrum



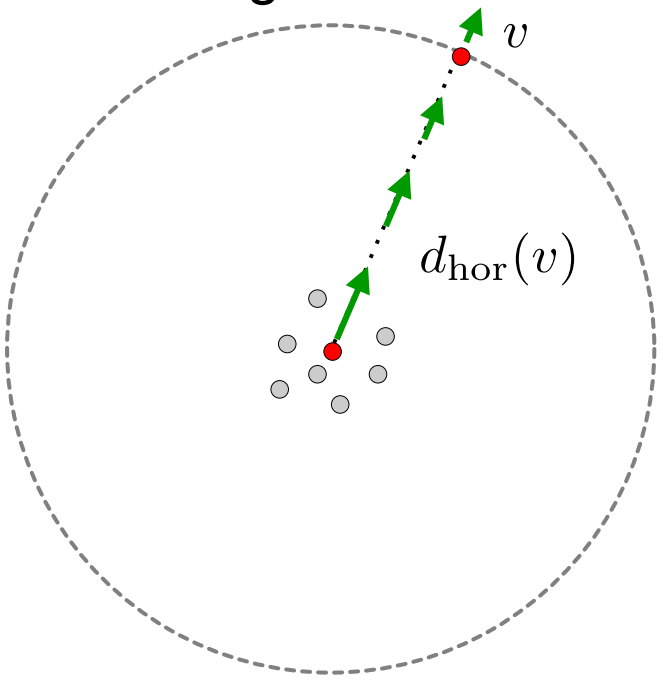
- Not directly observable, but can be probed using CMB data, Lyman-α-forest data, 21-cm forest, etc., up to $k \sim \mathcal{O}(100)\ \mathrm{Mpc}^{-1}$.
- **Robust predictions** for $P(k)$ can be reliably obtained – at least within the linear regime – over a broad range of redshifts through use of numerical tools like CLASS.

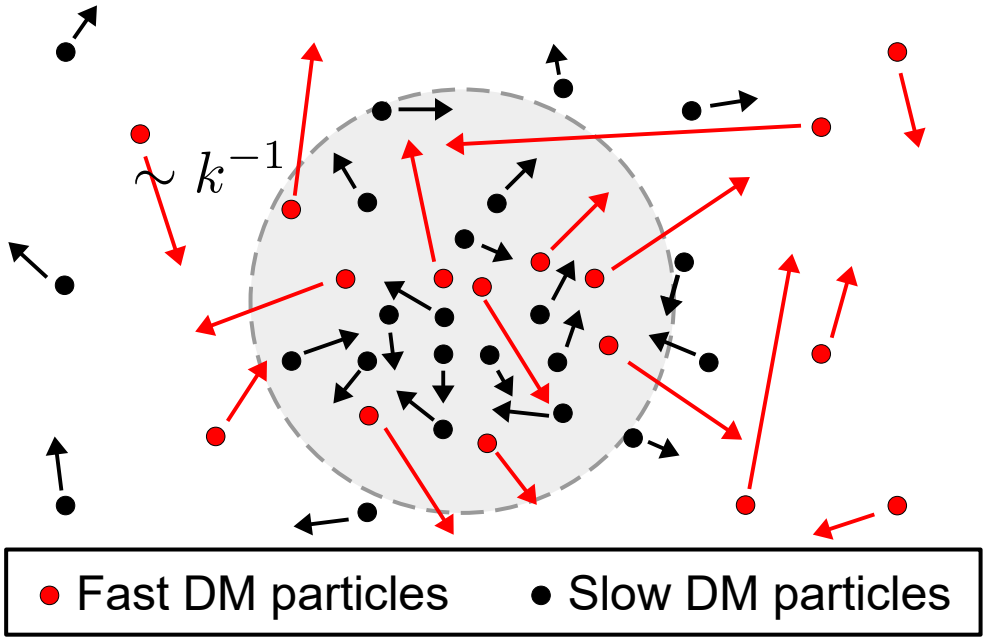
Free Streaming and Dark-Matter Speeds

- The properties of the dark matter can impact $P(k)$ in a variety of ways.
- Deviations from the shape that $P(k)$ takes relative to its shape $P_{\rm CDM}(k)$ in the case of ultra-cold dark-matter with only gravitational interactions are parametrized by the **transfer function** $T(k)$.

$$T^2(k) = \frac{P(k)}{P_{\rm CDM}(k)}$$

- One of the most important of these is **free-streaming** due to **particle horizons** when the dark matter is not ultra-cold.
- Our focus going forward will be on the impact $f(p)$ has on $P(k)$ through free-streaming.





Describing the Phase-Space Distribution

- We're going to find it useful to describe the DM phase-space distribution in a slightly atypical way.

$$x(t) = x(t')\frac{a(t)}{a(t')}$$

$$p(t) = p(t')\frac{a(t')}{a(t)}$$

$$\Rightarrow \quad \frac{d\log(p)}{dt} = -H(t)$$

- Physical number density:

$$n(t) = \frac{g_{\rm int}}{2\pi^2}\int dp\, p^2 f(p,t)$$

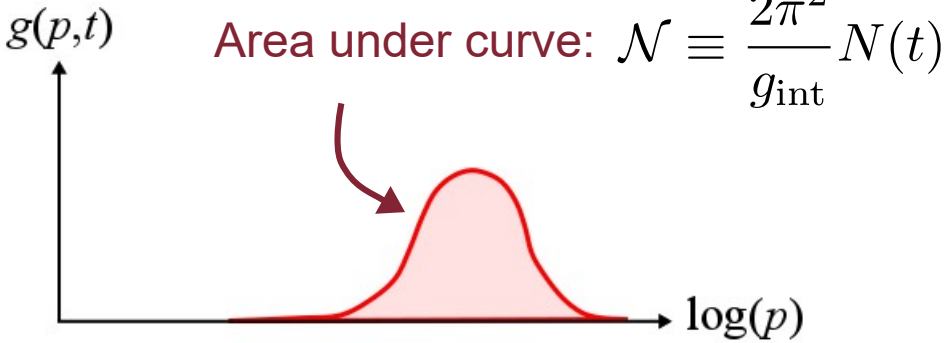
Number of internal degrees of freedom

- Comoving number density:

$$N(t) = n(t)a^3(t) = \frac{g_{\rm int}}{2\pi^2}\int d\log p\, p^3 a^3 f(p,t)$$

- This motivates us define the comoving "log-space" DM phase-space distribution $g(p,t)$:

$$g(p,t) = a^3(t)p^3 f(p,t)$$



Area under curve: $\mathcal{N} \equiv \frac{2\pi^2}{g_{\rm int}}N(t)$

The Usual Approach

- The free-streaming horizon for a particle of mass m and present-day momentum p in an expanding universe is

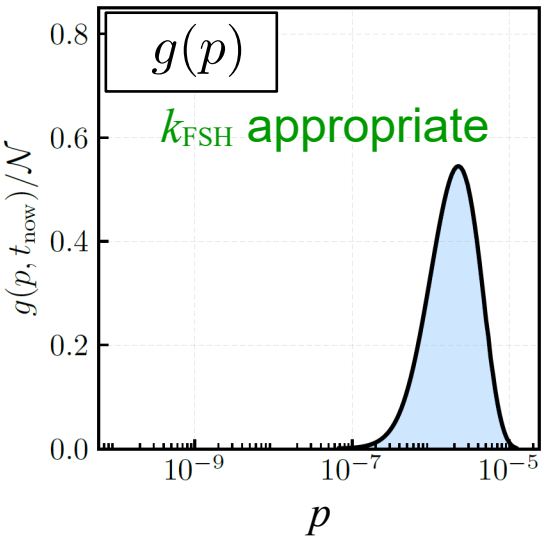
$$k_{\rm hor}(p) \equiv \xi \left[\int_{t_{\rm prod}}^{t} v(p,t) \frac{dt}{a(t)} \right]^{-1} = \xi \left[\int_{a_{\rm prod}}^{1} \frac{da}{Ha^2} \frac{p}{\sqrt{p^2 + m^2 a^2}} \right]^{-1}$$

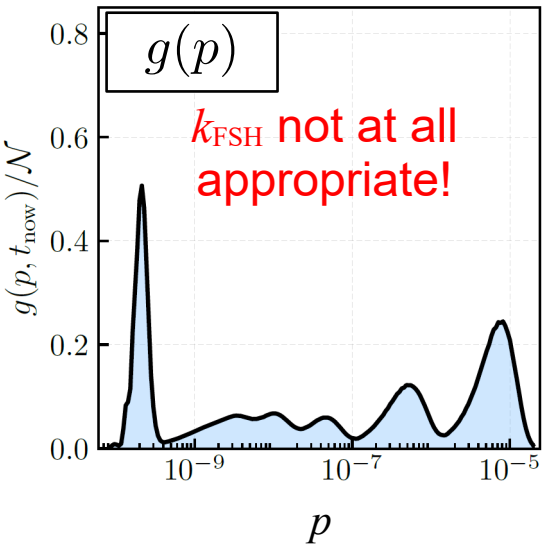
$\mathcal{O}(1)$ constant

- The usual approach (e.g., for warm DM) is to define a single "free-streaming-horizon" scale $k_{\rm FSH}$ using the ***average*** DM velocity $\langle v(t) \rangle$:

$$k_{\rm FSH} \sim \left[\int_{t_{\rm prod}}^{t} \langle v(t) \rangle \frac{dt}{a(t)} \right]^{-1}$$

- This works reasonably well when $g(p)$ is **unimodal** and **narrow**, but not when $g(p)$ is multimodal and/or broad.





An Alternative Approach

- By contrast, we shall consider a somewhat unorthodox procedure in which we regard $k_{\rm hor}(p)$ as a **functional map** between p and k.

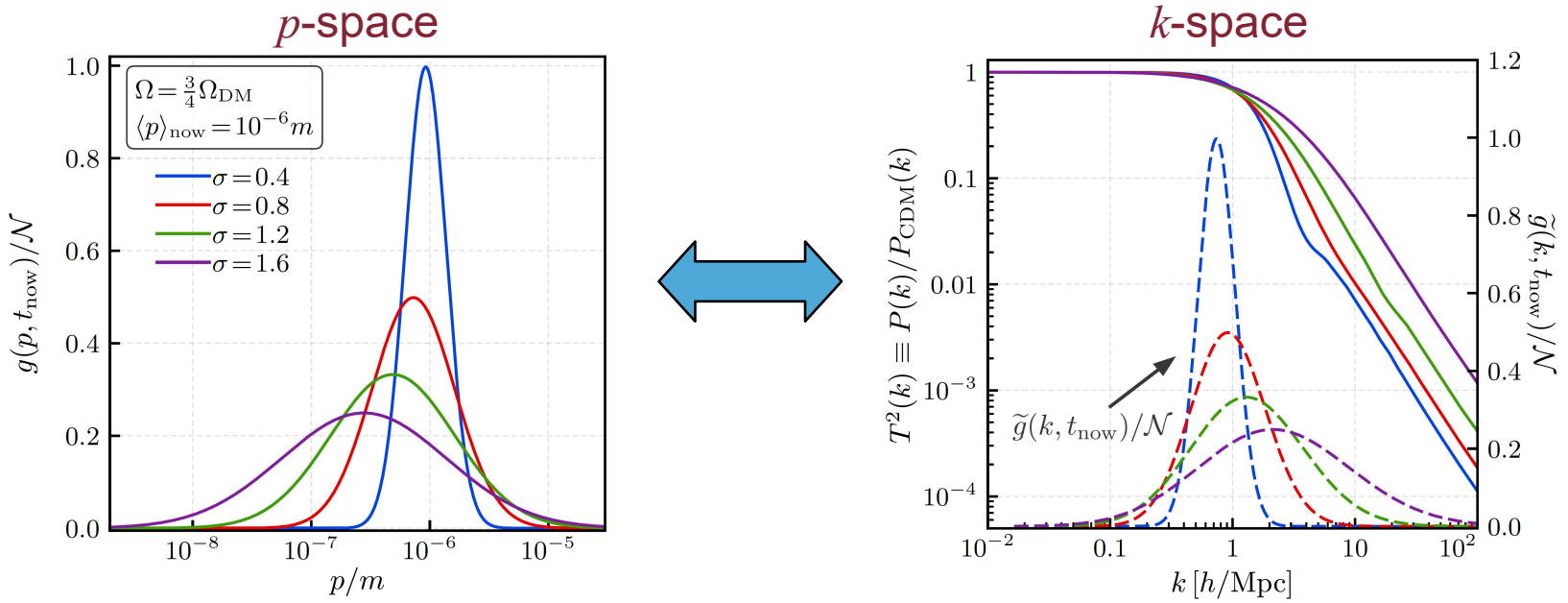
Dienes, Huang, Kost, Su, BT [arXiv:2001.02193]

- We can use this map to define a **phase-space distribution in k-space** which correspond to $g(p)$ in momentum space.

Inverse of $k_{\rm hor}(p)$ — Jacobian

$$\widetilde{g}(k) \equiv g\left(k_{\rm hor}^{-1}(k)\right) \left|\frac{d\log p}{d\log k}\right|$$

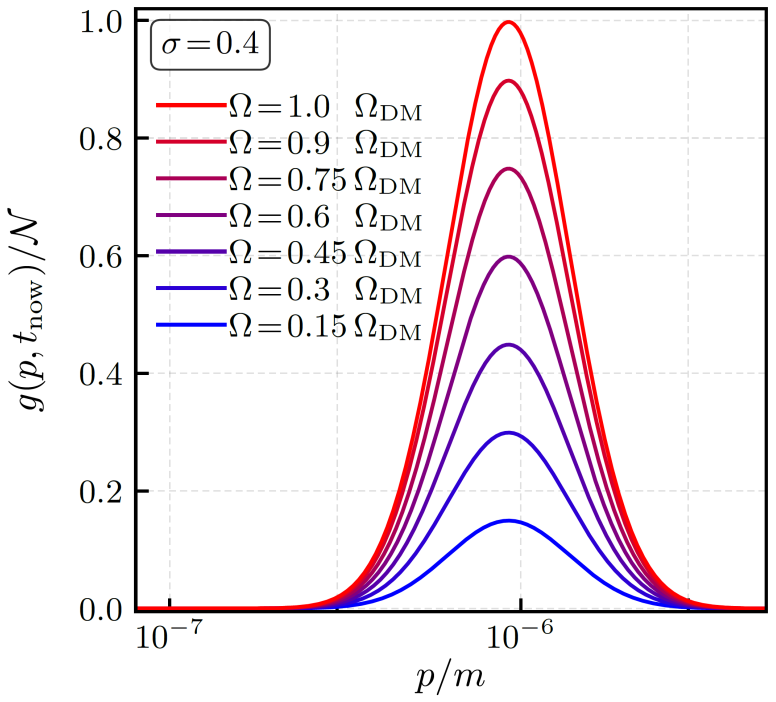
- This means that, in a sense, we can plot the dark-matter phase-space distribution and the matter power spectum **on the same axis** and explore correlations between them.

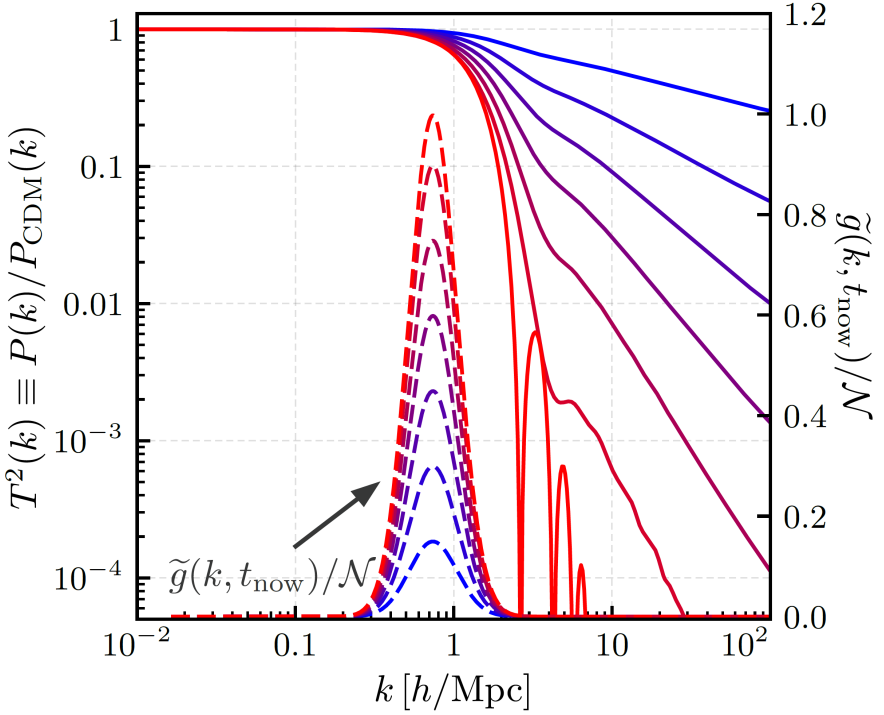


Relating $g(p)$ to $T^2(k)$

- Let's first consider the case of a simple $g(p,t_{\rm now})$ which consists of a single log-normal peak with average momentum $\langle p \rangle$ and width σ.
- We'll begin simply by fixing $\langle p \rangle$ and σ and **varying the normalization** of the peak, assuming that the rest of $\Omega_{\rm DM}$ is made up by cold DM.
- Increasing the abundance Ω associated with the peak, we find...

Correlations between $\tilde{g}(k,t_{\rm now})$ and $T^2(k)$ are **local**: features in $\tilde{g}(k,t_{\rm now})$ are correlated with features in $T^2(k)$ that appear at/around the same k.

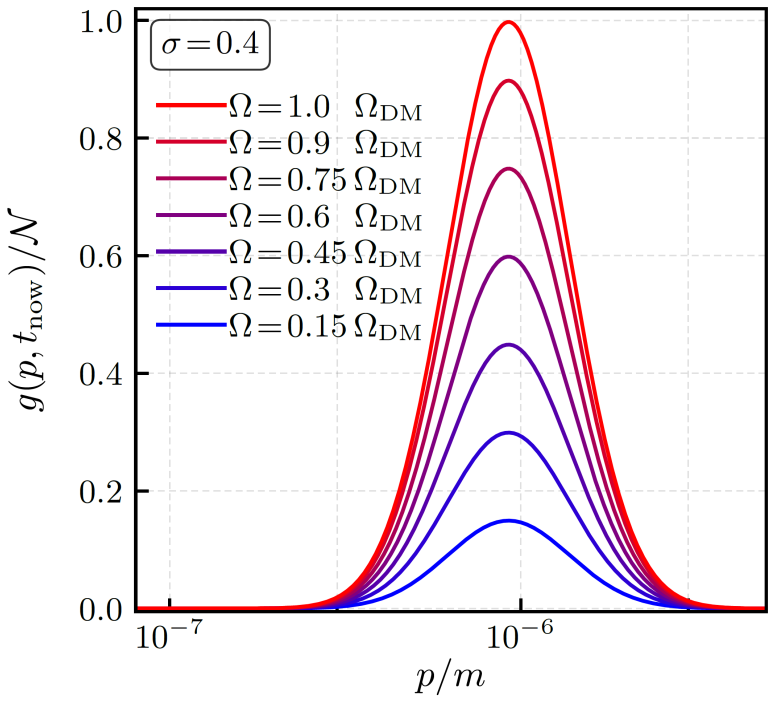


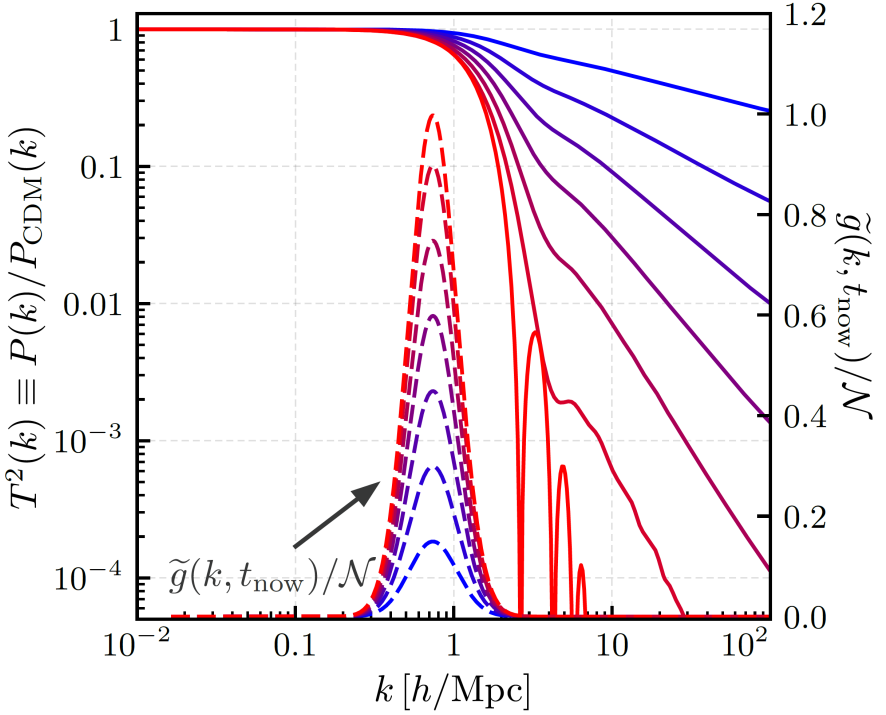


Relating $g(p)$ to $T^2(k)$

- Let's first consider the case of a simple $g(p,t_{\rm now})$ which consists of a single log-normal peak with average momentum $\langle p \rangle$ and width σ.
- We'll begin simply by fixing $\langle p \rangle$ and σ and **varying the normalization** of the peak, assuming that the rest of $\Omega_{\rm DM}$ is made up by cold DM.
- Increasing the abundance Ω associated with the peak, we find...

The abundance associated with a peak in $\tilde{g}(k,t_{\rm now})$ is correlated with the **change in the slope** of $T^2(k)$.





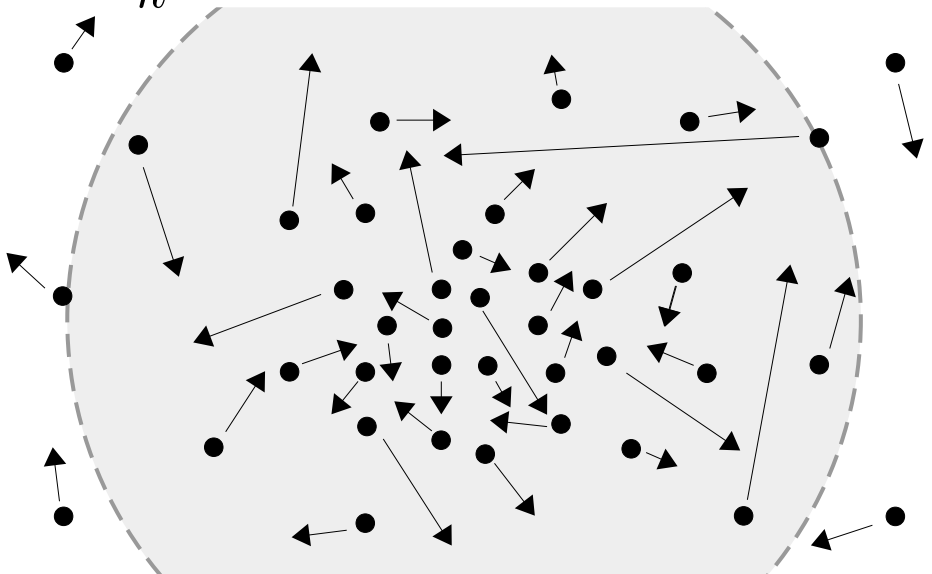
The Hot-Fraction Function

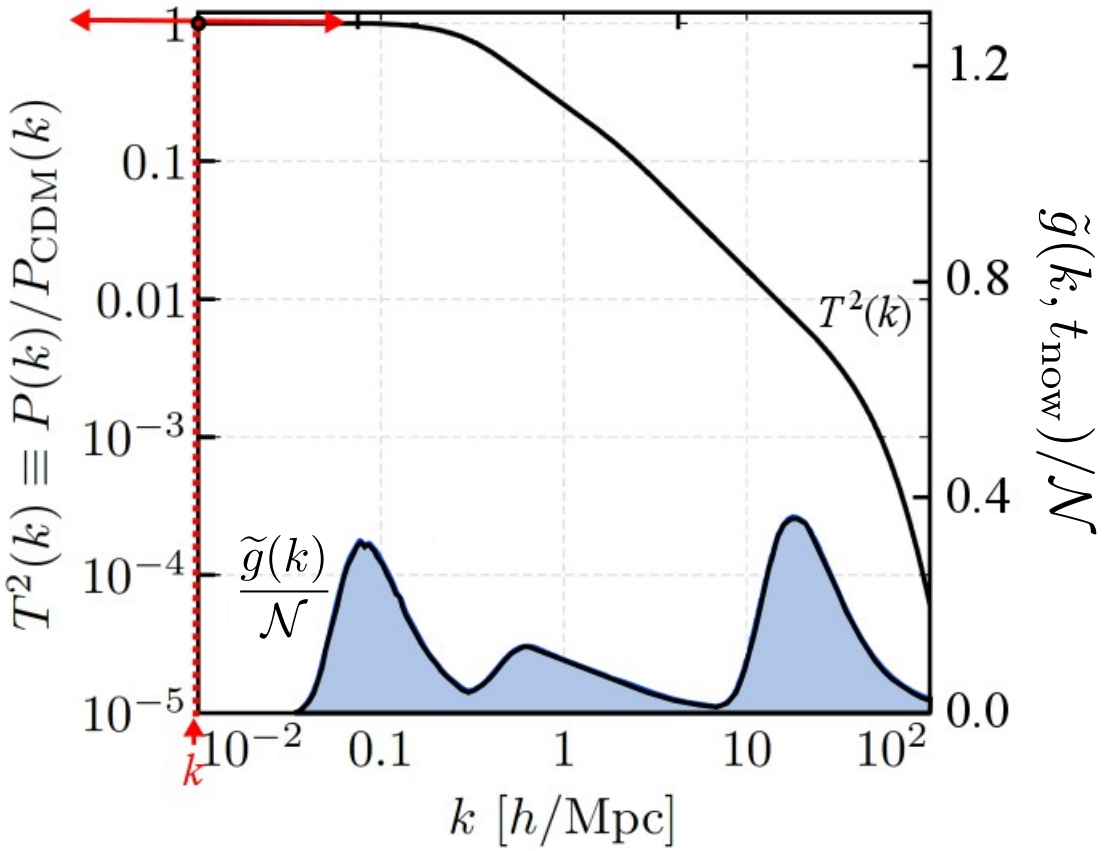
- The slope of the transfer function at a given value of k seems to correlate with the the total **number density of particles which can free-stream** at that value of k – particles with momenta $p > k_{\rm hor}^{-1}(k)$.

- Motivated by these empirical findings, let us define the "**hot-fraction function**" $F(k)$ as follows:

$$F(k) = \frac{\int_{-\infty}^{\log k} \widetilde{g}(k) d\log k'}{\int_{-\infty}^{\infty} \widetilde{g}(k) d\log k'}$$

Fraction of DM number density which free-streams at k





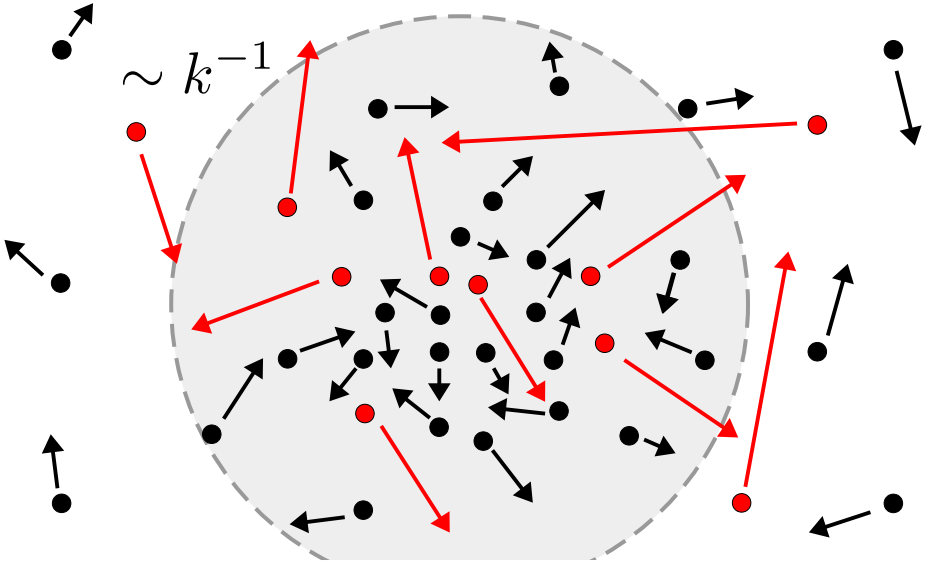
The Hot-Fraction Function

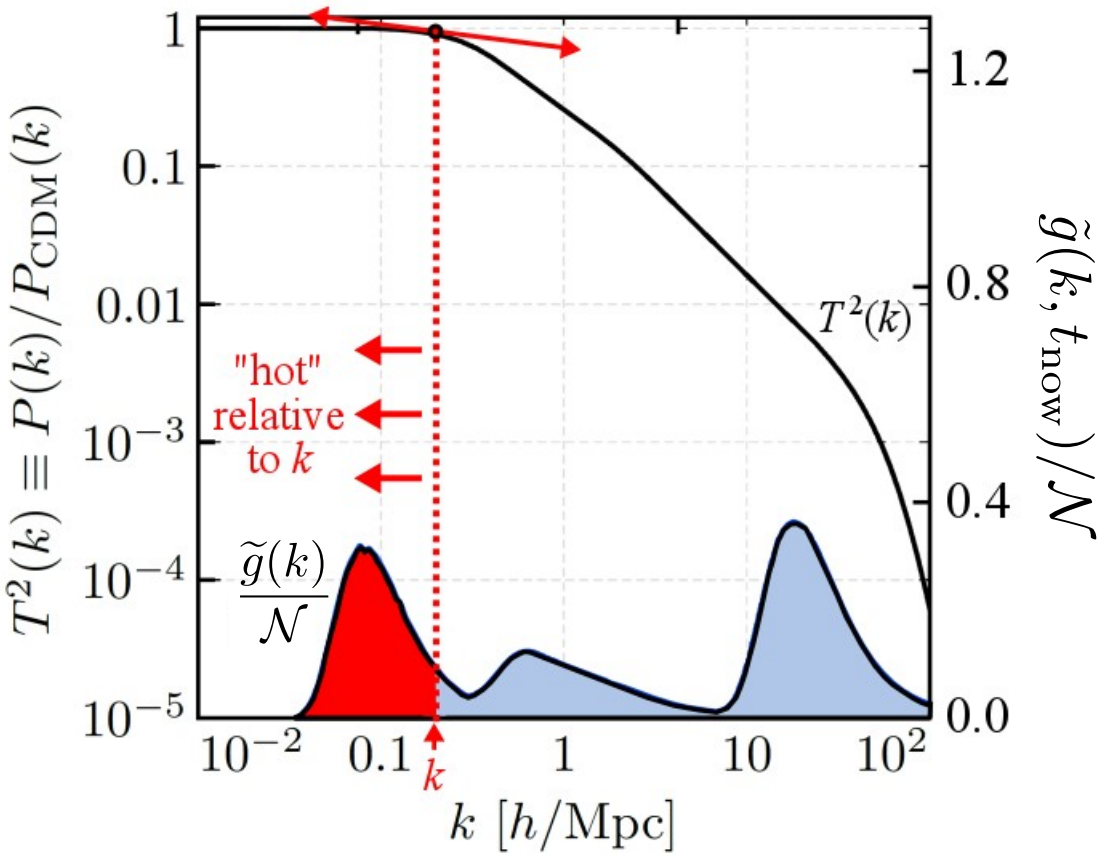
- The slope of the transfer function at a given value of k seems to correlate with the the total ***number density of particles which can free-stream*** at that value of k – particles with momenta $p > k_{\mathrm{FSH}}^{-1}(k)$.

- Motivated by these empirical findings, let us define the "***hot-fraction function***" $F(k)$ as follows:

$$F(k) = \frac{\int_{-\infty}^{\log k} \widetilde{g}(k)\, d\log k'}{\int_{-\infty}^{\infty} \widetilde{g}(k)\, d\log k'}$$

Fraction of DM number density which free-streams at k





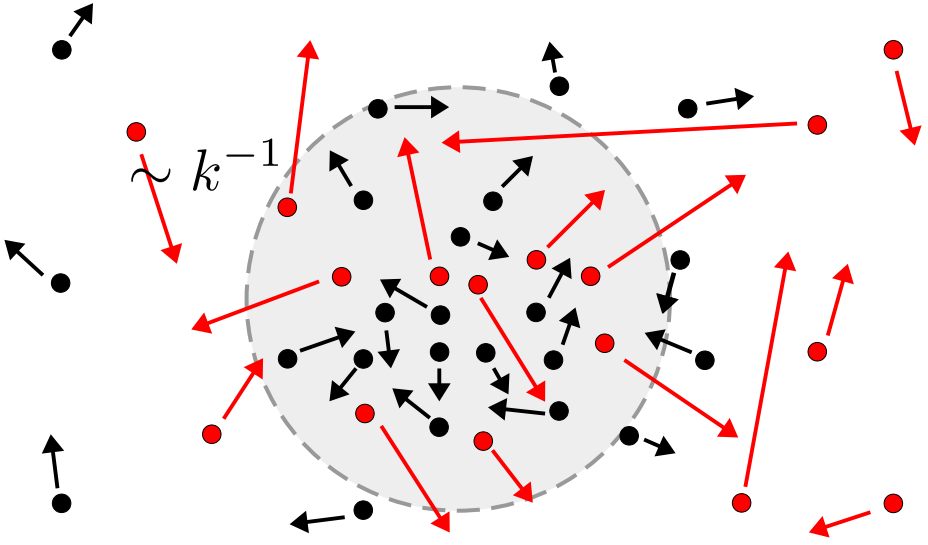
The Hot-Fraction Function

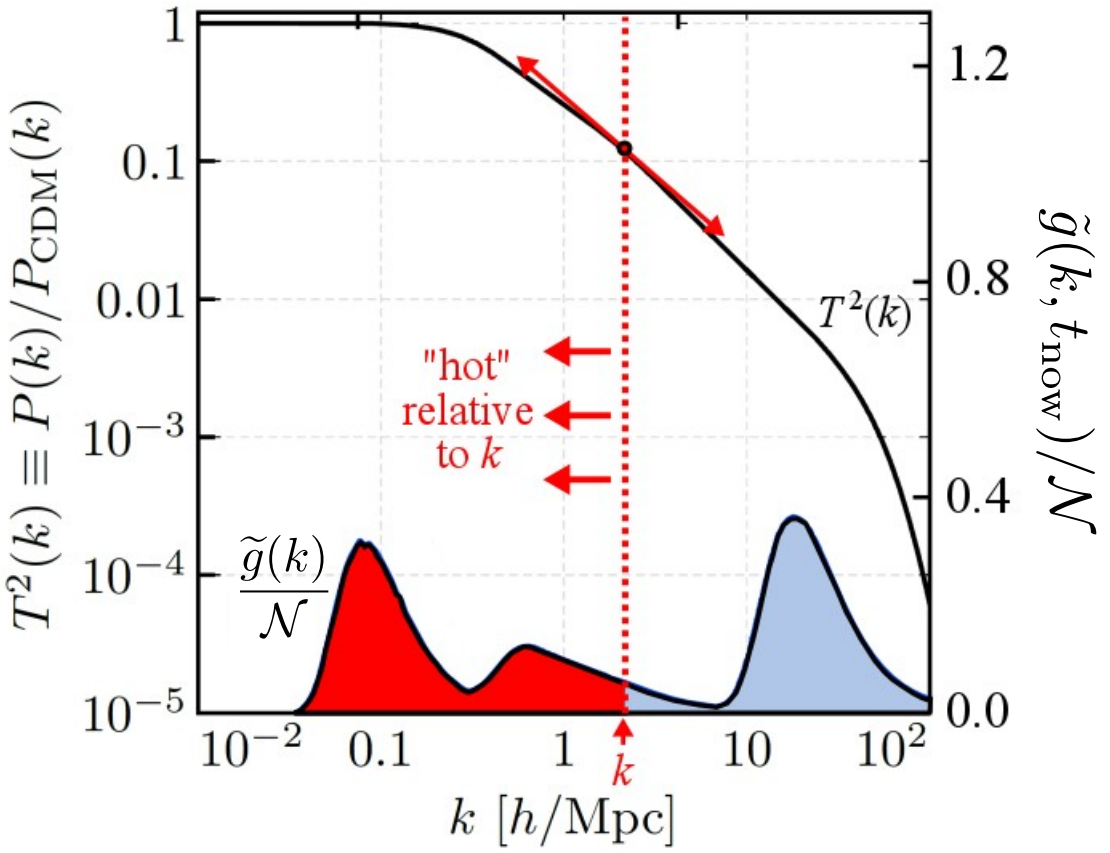
- The slope of the transfer function at a given value of k seems to correlate with the the total ***number density of particles which can free-stream*** at that value of k – particles with momenta $p > k_{\mathrm{FSH}}^{-1}(k)$.

- Motivated by these empirical findings, let us define the "***hot-fraction function***" $F(k)$ as follows:

$$F(k) = \frac{\int_{-\infty}^{\log k} \tilde{g}(k)\, d\log k'}{\int_{-\infty}^{\infty} \tilde{g}(k)\, d\log k'}$$

Fraction of DM number density which free-streams at k





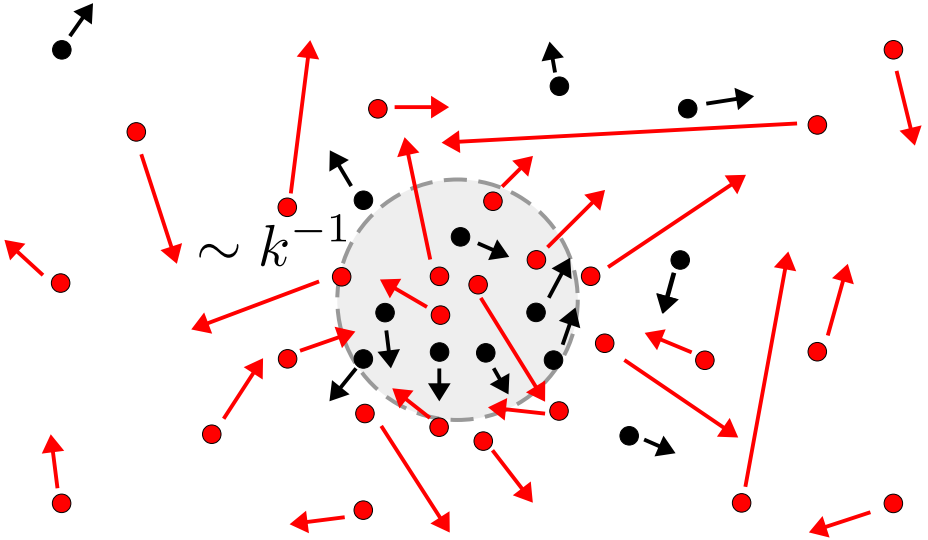
The Hot-Fraction Function

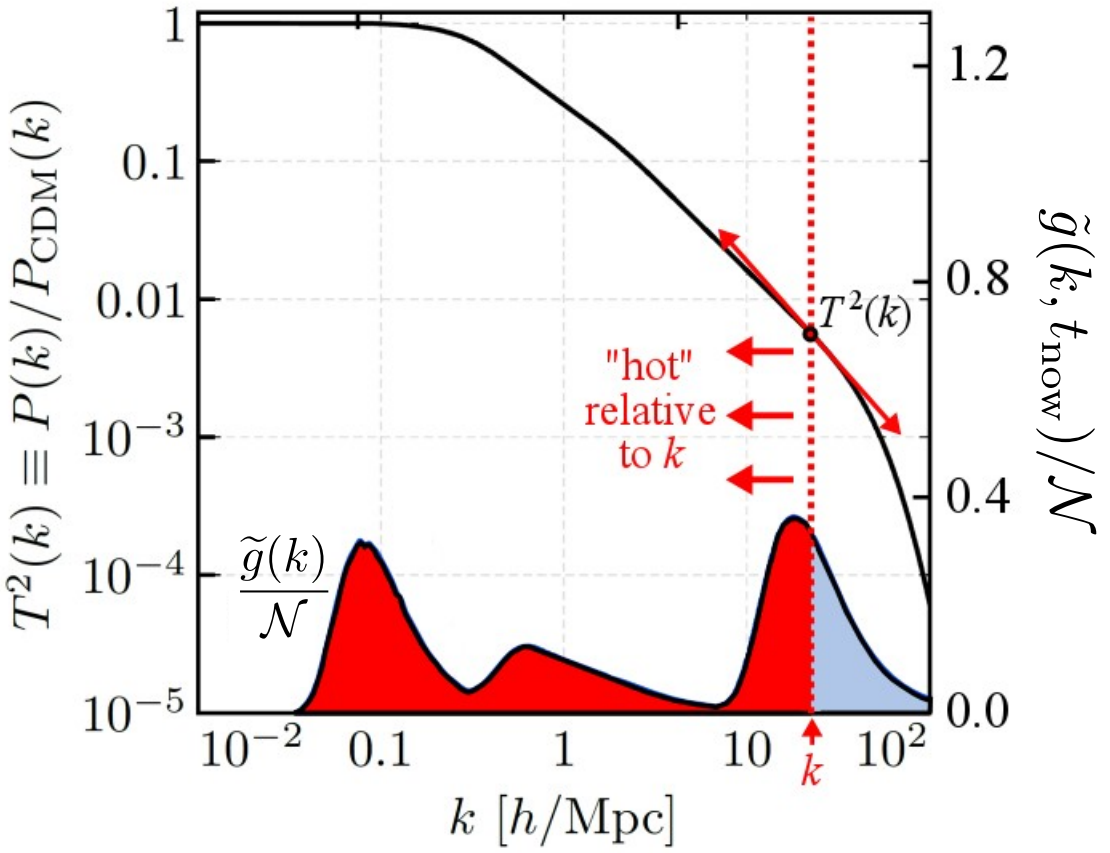
- The slope of the transfer function at a given value of k seems to correlate with the the total ***number density of particles which can free-stream*** at that value of k – particles with momenta $p > k_{\rm FSH}^{-1}(k)$.

- Motivated by these empirical findings, let us define the "***hot-fraction function***" $F(k)$ as follows:

$$F(k) = \frac{\int_{-\infty}^{\log k} \widetilde{g}(k) d\log k'}{\int_{-\infty}^{\infty} \widetilde{g}(k) d\log k'}$$

Fraction of DM number density which free-streams at k





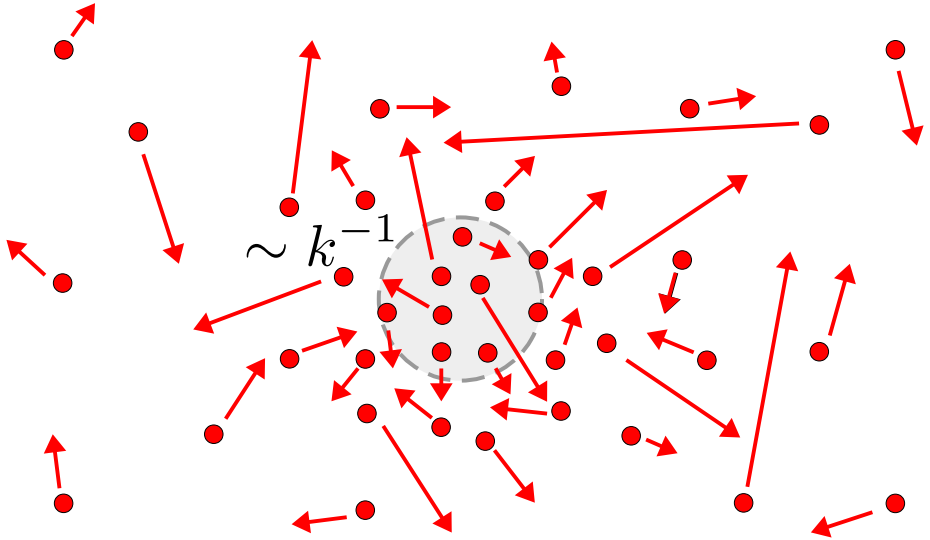
The Hot-Fraction Function

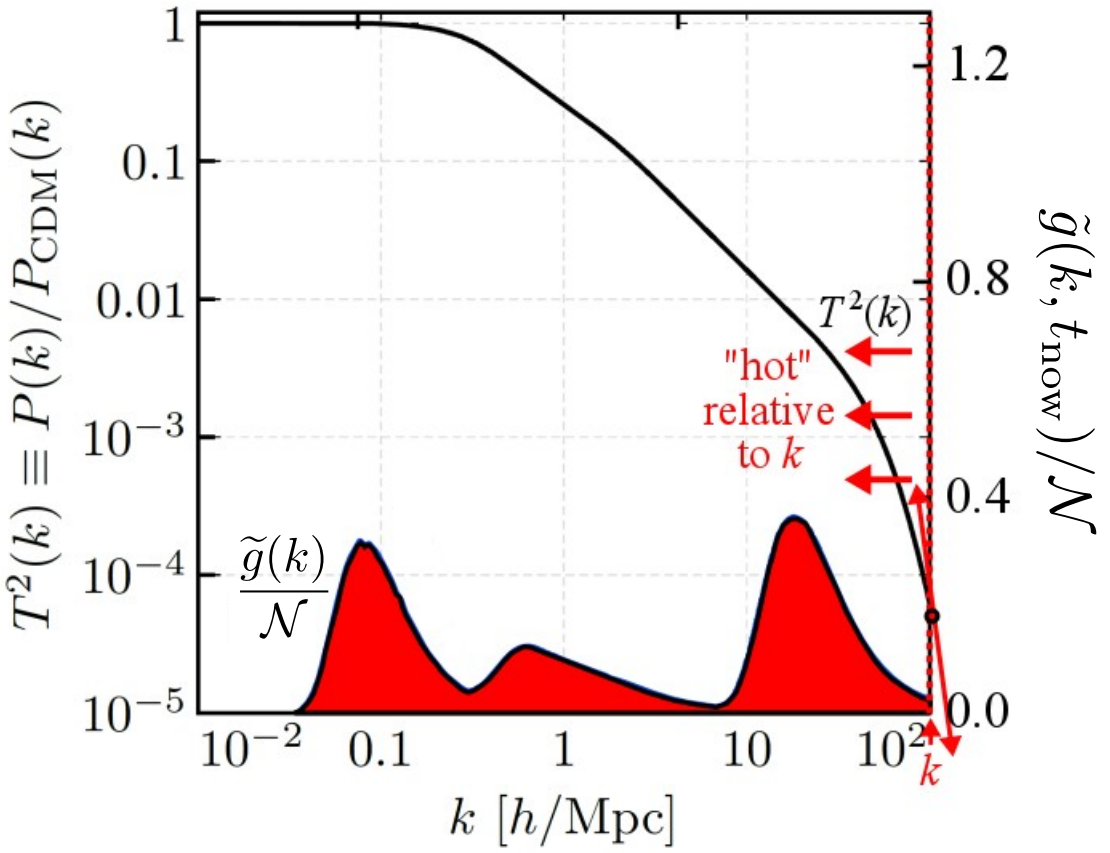
- The slope of the transfer function at a given value of k seems to correlate with the the total ***number density of particles which can free-stream*** at that value of k – particles with momenta $p > k_{\rm FSH}^{-1}(k)$.

- Motivated by these empirical findings, let us define the "***hot-fraction function***" $F(k)$ as follows:

$$F(k) = \frac{\int_{-\infty}^{\log k} \widetilde{g}(k)\, d\log k'}{\int_{-\infty}^{\infty} \widetilde{g}(k)\, d\log k'}$$

Fraction of DM number density which free-streams at k





A Reconstruction Conjecture

- Our conjecture, then, is that there exists some ***invertible functional relationship*** between $F(k)$ and $T^2(k)$.
- Empirically, from numerical investigations (using CLASS) of the relationship between these two quantities, we find that

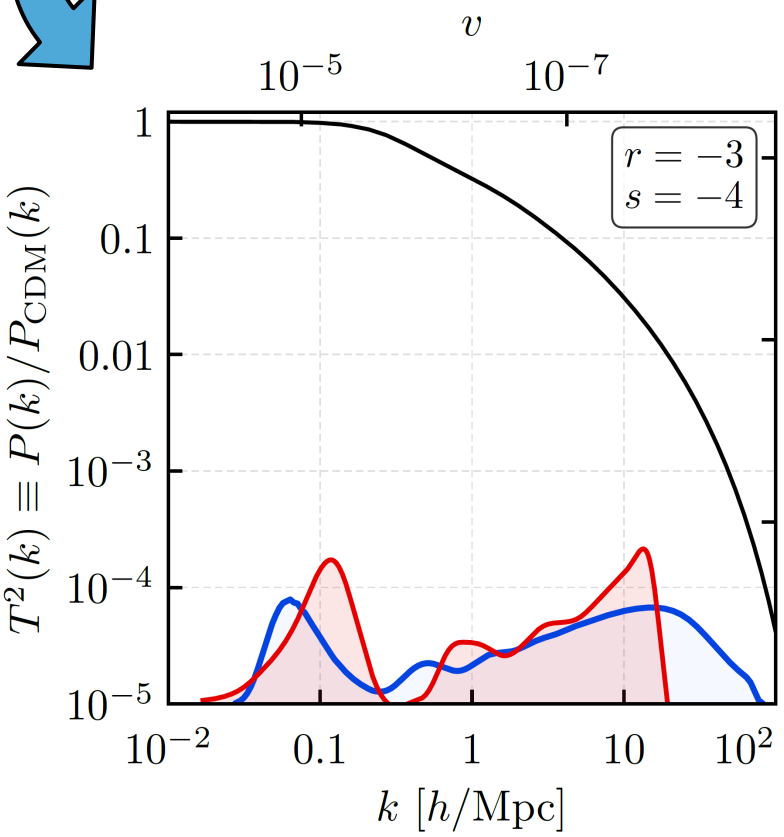
$$\left|\frac{d\log T^2(k)}{d\log k}\right| \approx F^2(k) + \frac{3}{2}F(k)$$

- Taking the derivative of both sides, we obtain an approximate analytic expression for reconstructing $\tilde{g}(k)$ from $T^2(k)$.

$$\frac{\tilde{g}(k)}{\mathcal{N}} \approx \frac{1}{2}\left(\frac{9}{16} + \left|\frac{\log T^2(k)}{\log k}\right|\right)^{-1/2} \left|\frac{d^2\log T^2(k)}{(d\log k)^2}\right|$$

How Well Does This Work in Practice?

- We can assess how well this reconstruction conjecture works in practice by applying it to the $g(p)$ distributions that emerge from concrete example models.
- One class of models which can give rise to highly trivial, multi-modal $g(p)$ distributions is that involving **multi-step decay chains** within an extended dark sector.



Upshot: It Works Surprisingly Well!

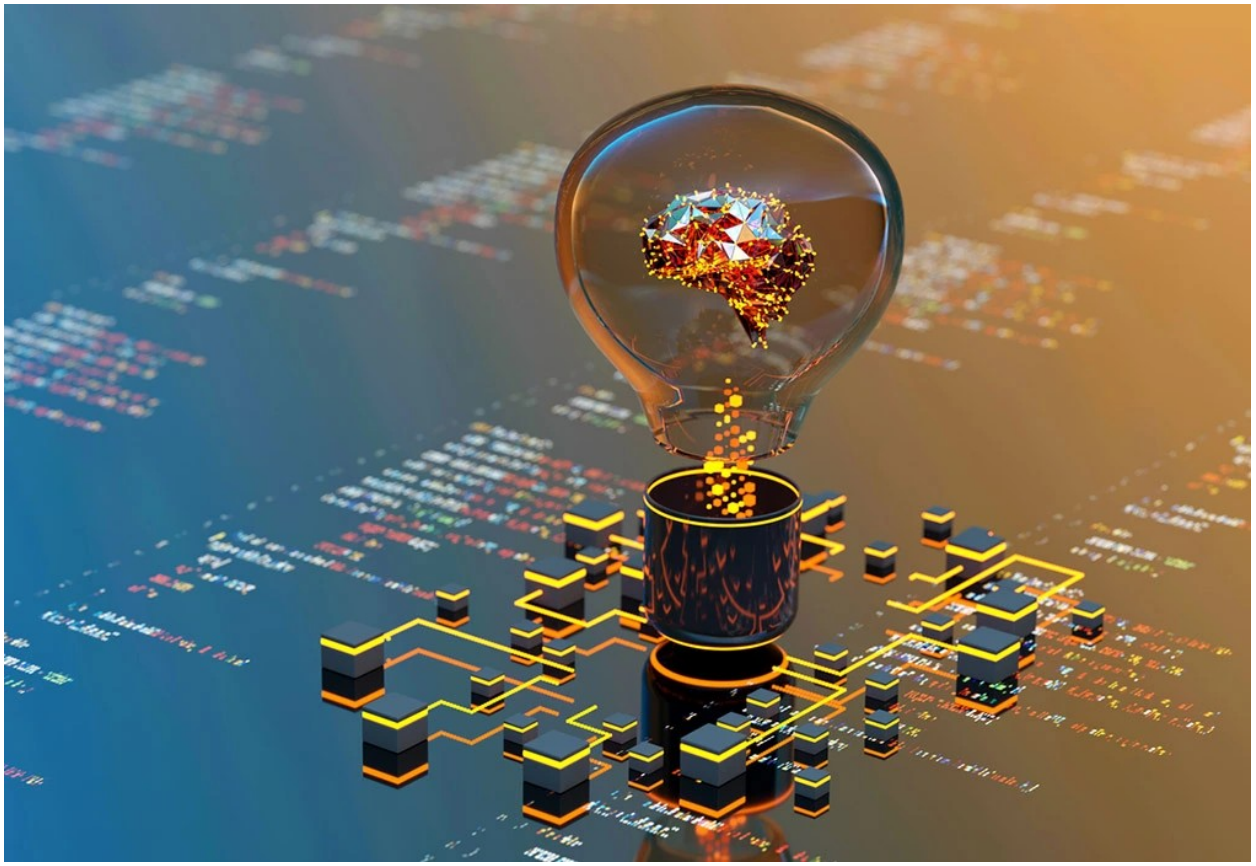
- We find that our reconstruction conjecture reproduces the broad-brush features of the DM velocity distribution quite robustly in all cases.
- However, the accuracy of the reconstruction could be improved… and fortunately there's a straightforward way of accomplishing that.

Blue: actual $\widetilde{g}(k)$ distribution for the model in question

Red: reconstructed $\widetilde{g}(k)$ distribution using our procedure

Machine Learning

- Uncovering relationships between sets of distributions is a task that machine learning is particularly well suited for.



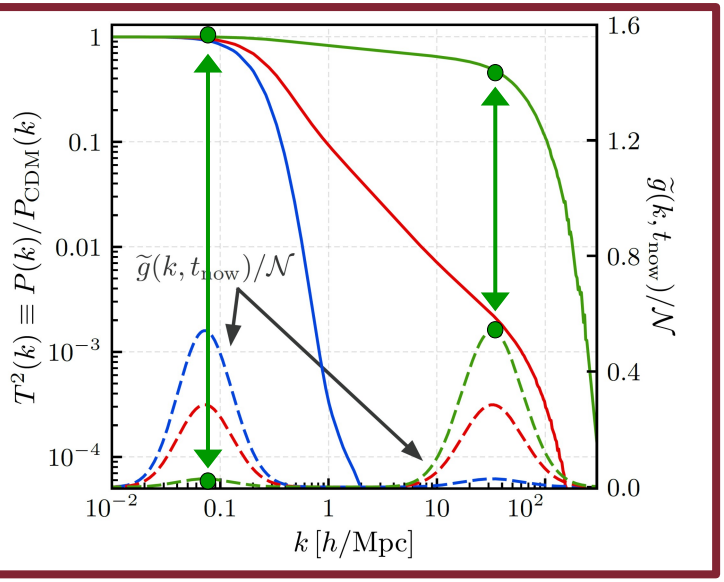
- The intuition that we've developed in deriving our empirical reconstruction conjecture will guide us in terms of network architecture, how we train the network, etc.

A Little Human Intuition

- In constructing and training our machine, we'll exploit the physical intuition we've developed – and in particular, **two guiding principles**:

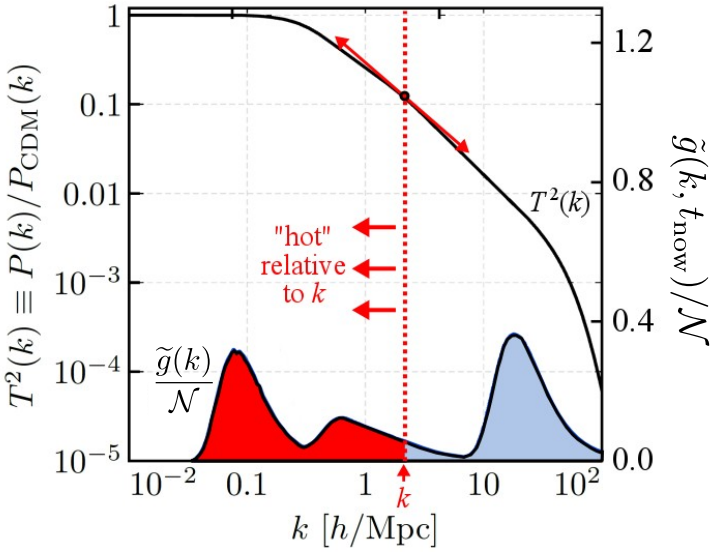
1 k-Locality

The reconstructed value of $g(k)$ at a given k should only depend on the properties of $T^2(k)$ at/around that same value of k.



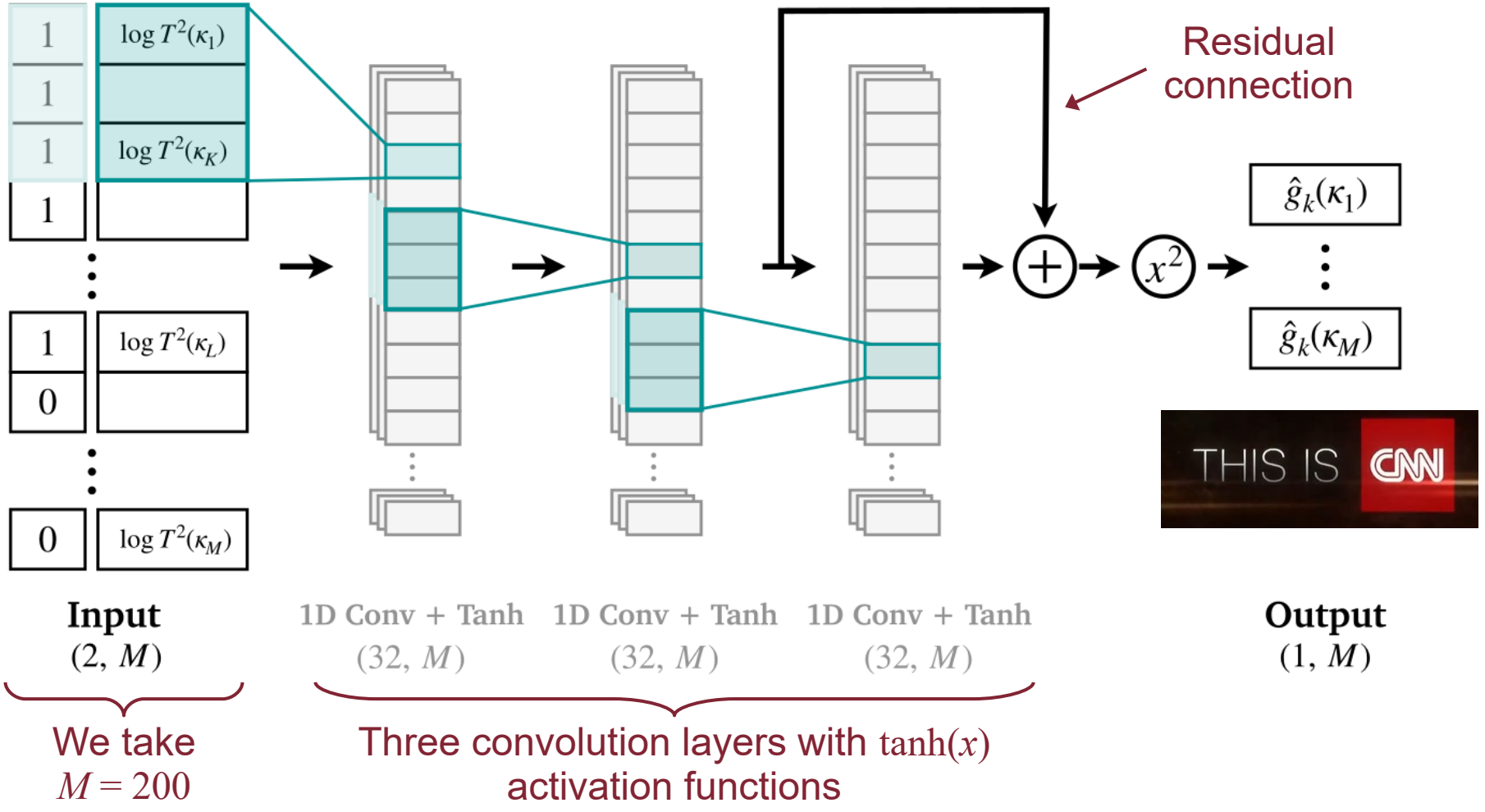
2 Horizon Thresholds

The value of $g(k)$ at a particular value of k should have no impact whatsoever on the shape of $T^2(k)$ anywhere below that value of k.



This in CNN: Reconstruction with ML

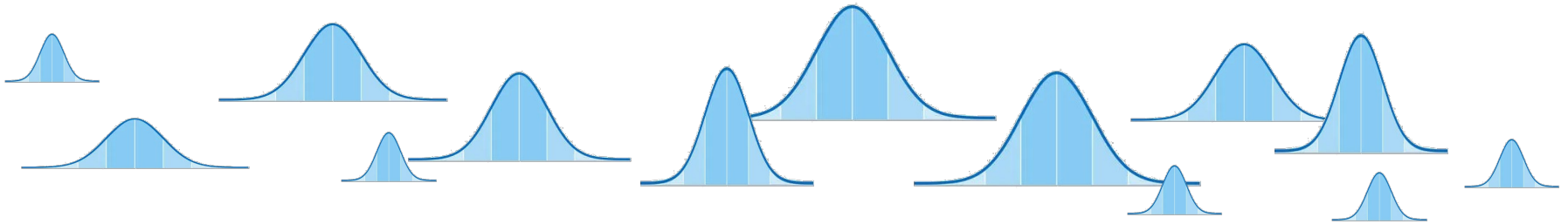
- We find that a **convolutional neural network** (CNN) provides the best performance out of a wide range of possible network architectures.



- Indeed, a CNN incorporates an inductive bias that **emphasizes local features** of $T^2(k)$ while still allowing for non-local connectivity.

Training Data and Truncation

- Our core training data consists of $g(k)$ created by combining $N = 30$ or $N = 80$ individual Gaussian distributions with randomly distributed means, widths, and amplitudes.



- We also include:
  - $g(k)$ distributions created by combining smaller numbers ($N = 1$, $N = 2$, and $N = 3$) of such Gaussians.
  - $g(k)$ distributions associated with DM freeze-in and freeze-out.
- While training, we implement a **truncation procedure** wherein we randomly select some k value $k_{\rm L}$, truncate $\log T^2(k)$ for $\log k > \log k_L$, and forward-fill the truncated values.
- Training on these truncated distributions weakly encourages the CNN to learn something akin to our **horizon-threshold principle**.

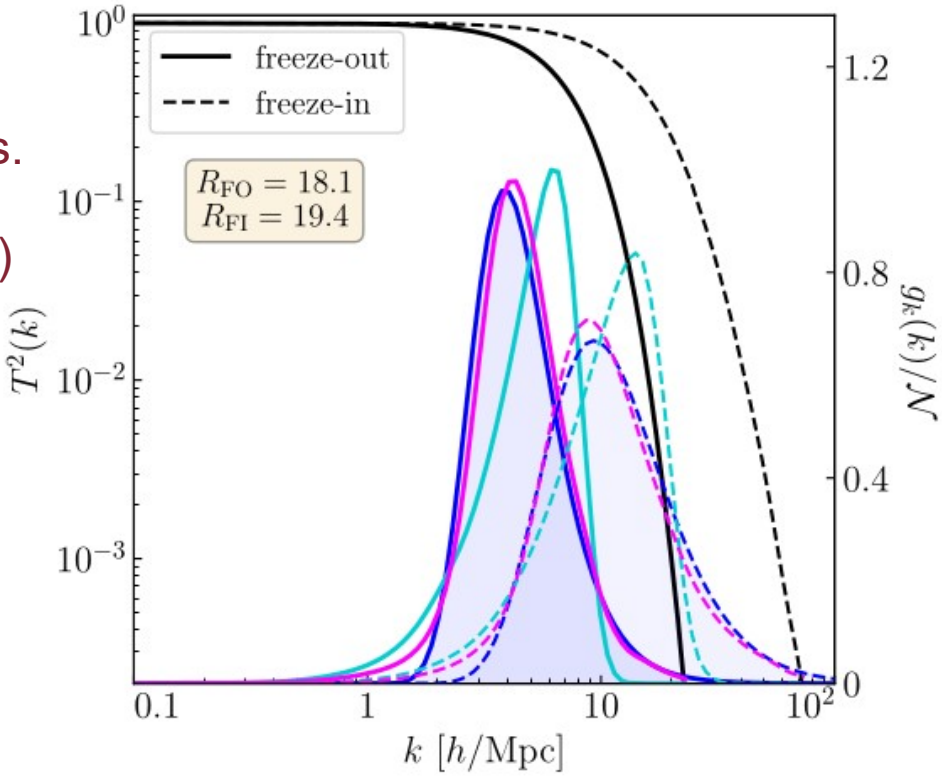
Machine Learning Does it Better

- We begin by examining the trained network's performance for transfer functions obtained from **simple, unimodal** $g(k)$ distributions.
- In particular, we apply them to $g(k)$ distributions obtained from DM freeze-in and freeze-out.
- We can assess the accuracy of a reconstructed phase-space distribution $\hat{g}(k)$ via the mean-squared error (MSE) function

No. of pts. sampled ($M = 200$)

$$L_{\rm MSE} = \frac{1}{M}\sum_{i}^{M}\left|\hat{g}(\kappa_i) - g(\kappa_i)\right|^2.$$

- The ratio $R \equiv L_{\rm MSE}^{\rm (ML)}/L_{\rm MSE}^{\rm (emp)}$ of MSE functions provides a measure of comparing the accuracy of empirical and ML reconstructions.
- In all cases, the CNN **outperforms** our empirical reconstruction conjecture by a significant margin!



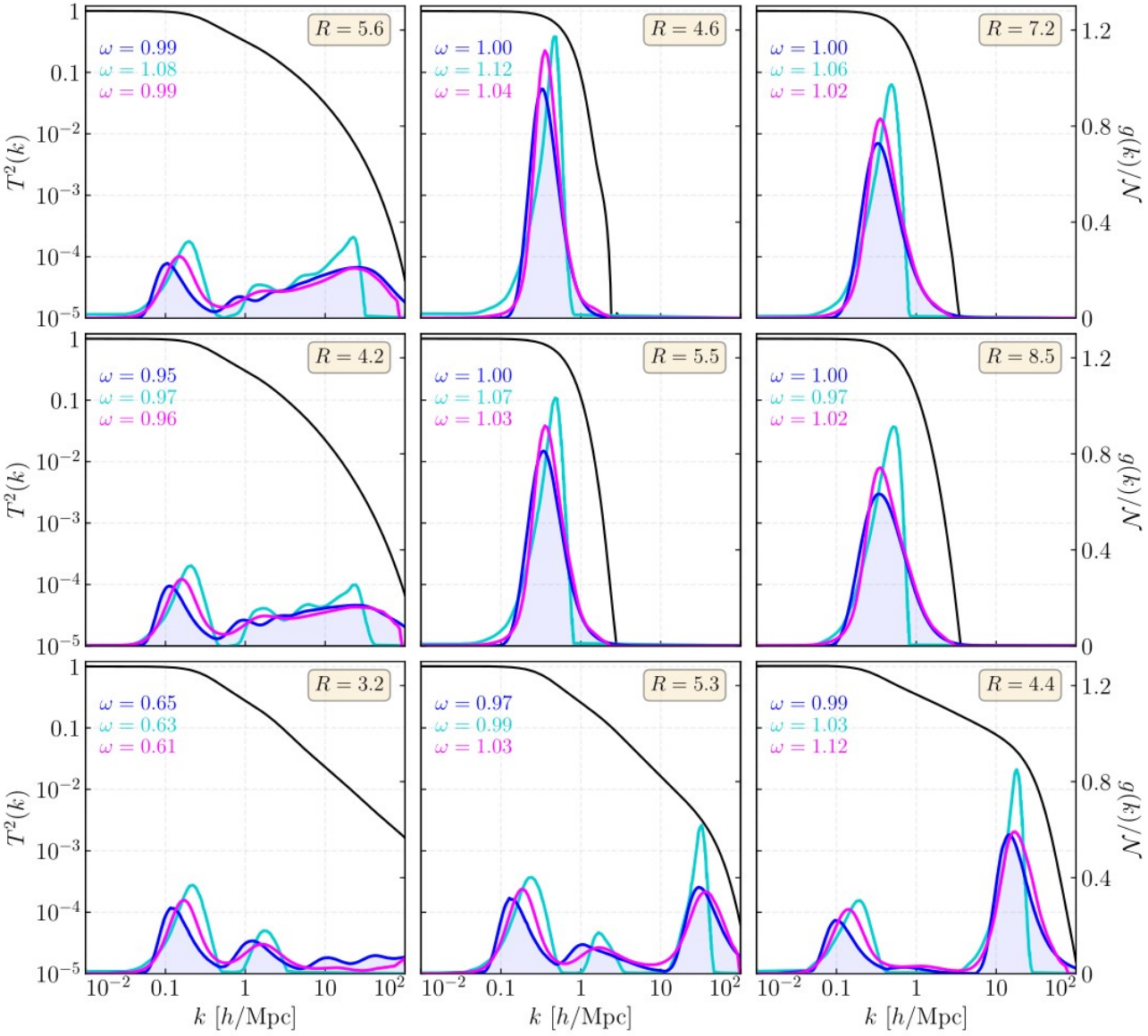
"Truth" $g(k)$ distribution

Analytic reconstruction

ML reconstruction

Machine Learning Does it Better

- We also examine the trained network's performance for transfer functions obtained from **complicated, multi-modal** $g(k)$ distributions.
- We use the CNN to reconstruct g(k) distributions obtained from cascade-decay models.
- Once again, the CNN **outperforms** our empirical reconstruction conjecture by a significant margin.



"Truth" $g(k)$ distribution

Analytic reconstruction

ML reconstruction

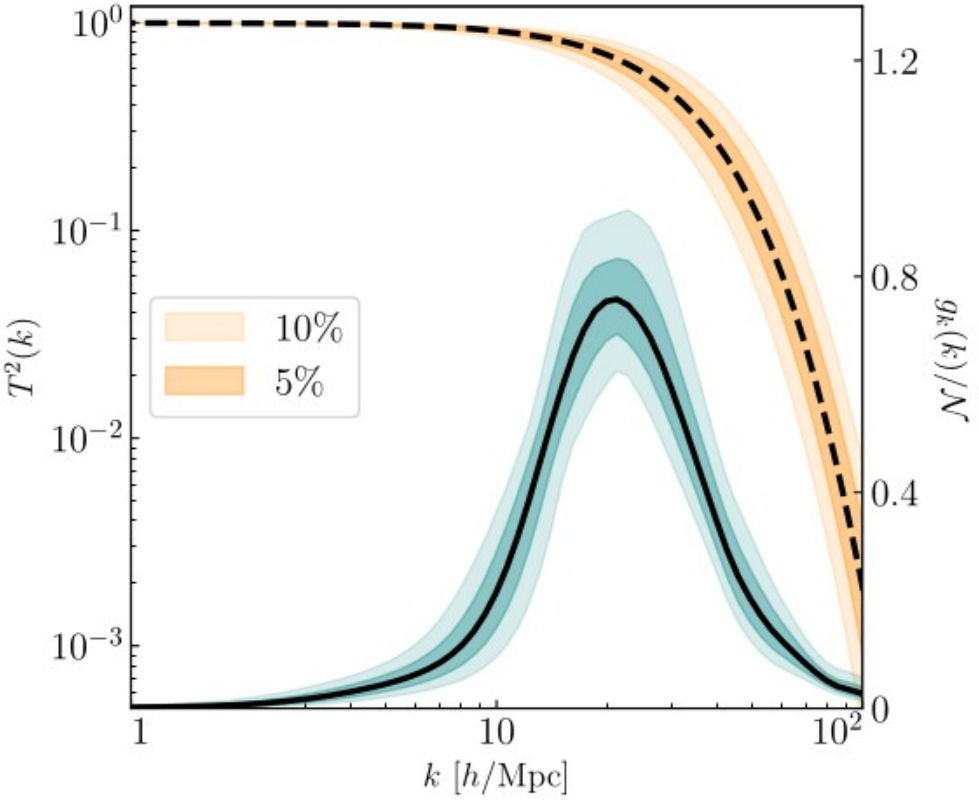
Stability of the Solutions

- We can test the stability of the solutions by perturbing $T^2(k)$ in a controlled way and examining the effect on the corresponding reconstructed $g(k)$ distribution.

Unimodal

$$T^2(k) = \left[1 + (\alpha k)^{\beta}\right]^{2\gamma}$$

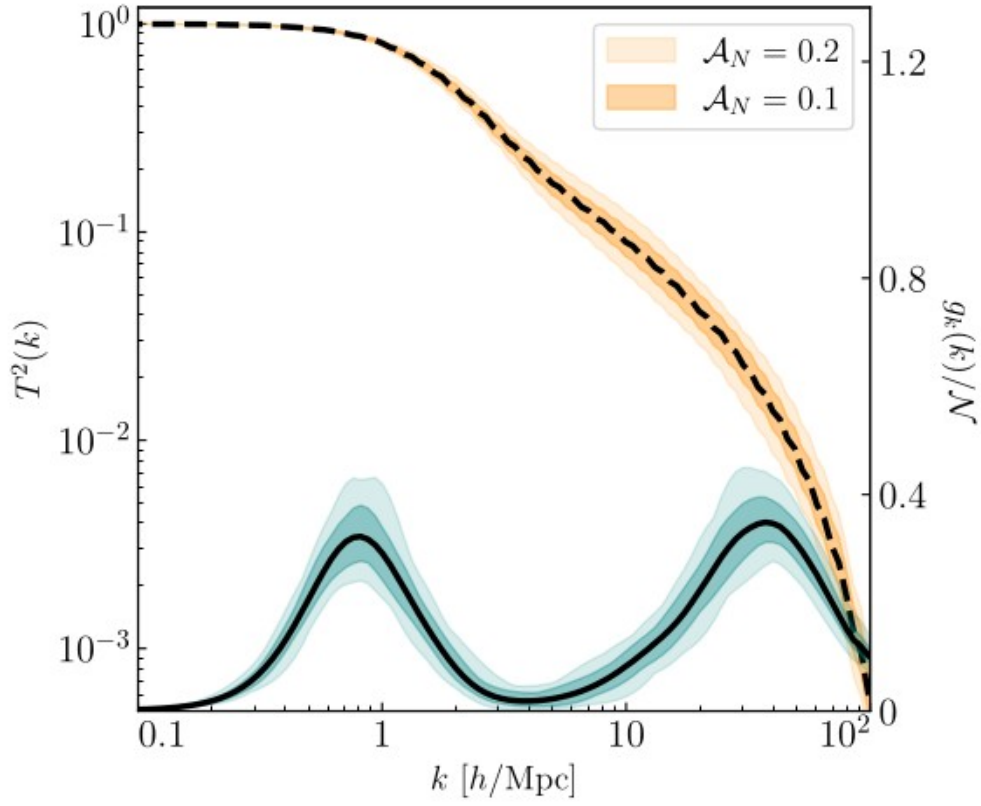
Vary $\{\alpha,\beta,\gamma\}$ within a specified range.



Bimodal

Apply Gaussian noise with amplitude $\mathcal{A}_N$ and smooth with a Gaussian filter.

Vary $\mathcal{A}_N$ within a specified range.



Summary

- Non-trivial dynamics in the early universe can lead to a complicated – and even multi-modal – DM velocity distribution, which in turn affects the shape of the matter power spectrum.
- In an effort to work backwards, we have studied the relationship between $P(k)$ and the DM velocity distribution and found that the ***slope of the transfer function*** at a given k is related to the ***fraction of the DM number density which can free-stream*** on the scale k.
- Motivated by these results, we have formulated a conjecture for ***reconstructing the primordial dark-matter velocity distribution*** from the shape of the matter power spectrum.
- Moreover, we have seen that **machine learning** can significantly improve the accuracy of such reconstruction efforts.