

Dark Quest II

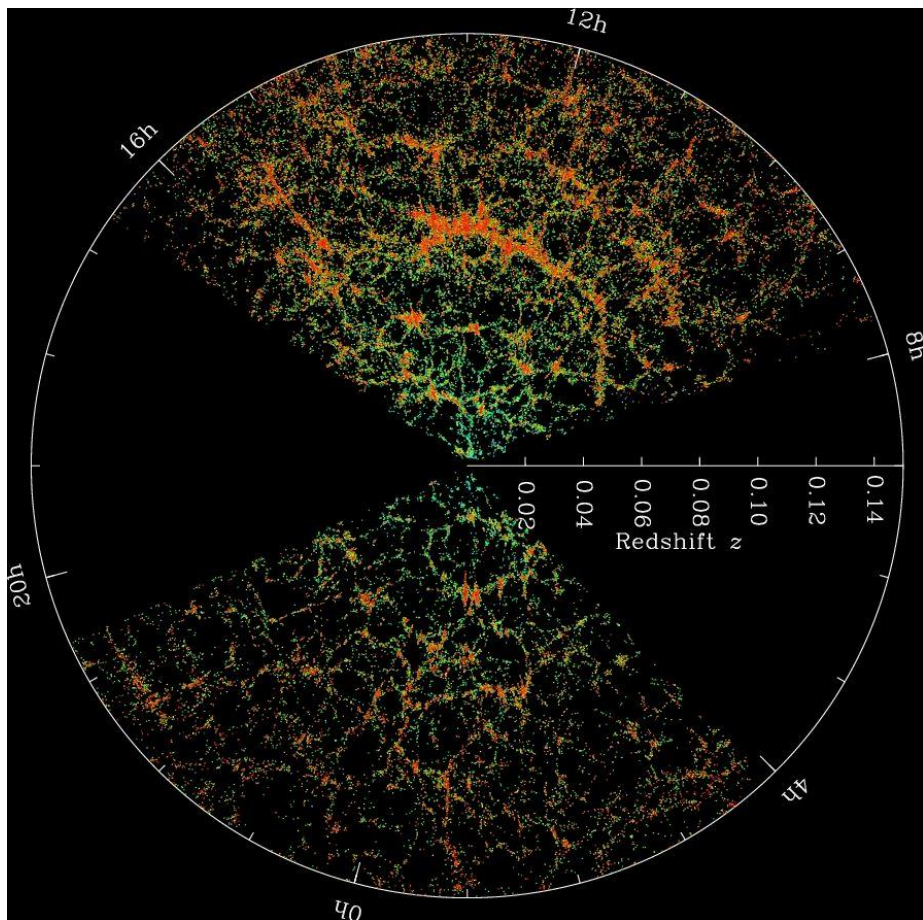
エミュレータ開発の進捗

田中 賢(京大基研)

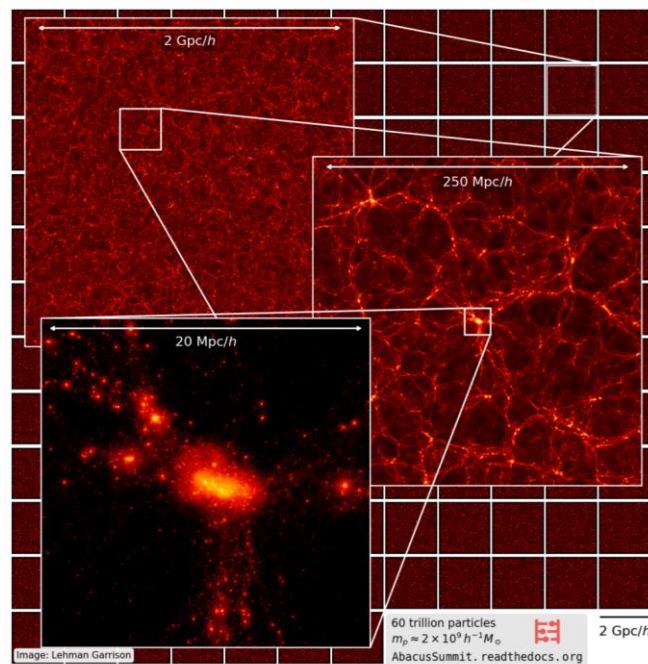
西道 啓博 (京大基研), 小林 洋祐 (アリゾナ大学)

第10回観測的宇宙論ワークショップ
2021/11/17 - 19

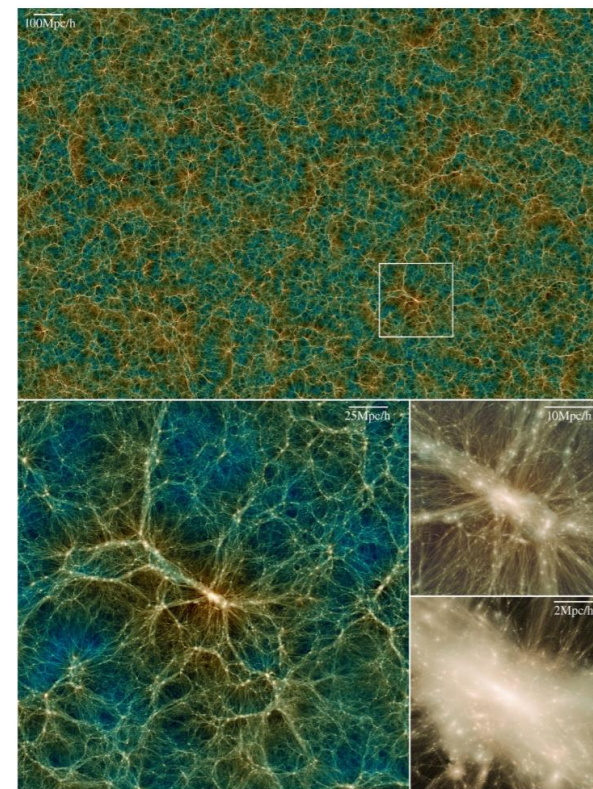
宇宙論的N-bodyシミュレーション



©SDSS



Maksimova+ `21
Abacus Simulation , $\sim 8640^3$ particles



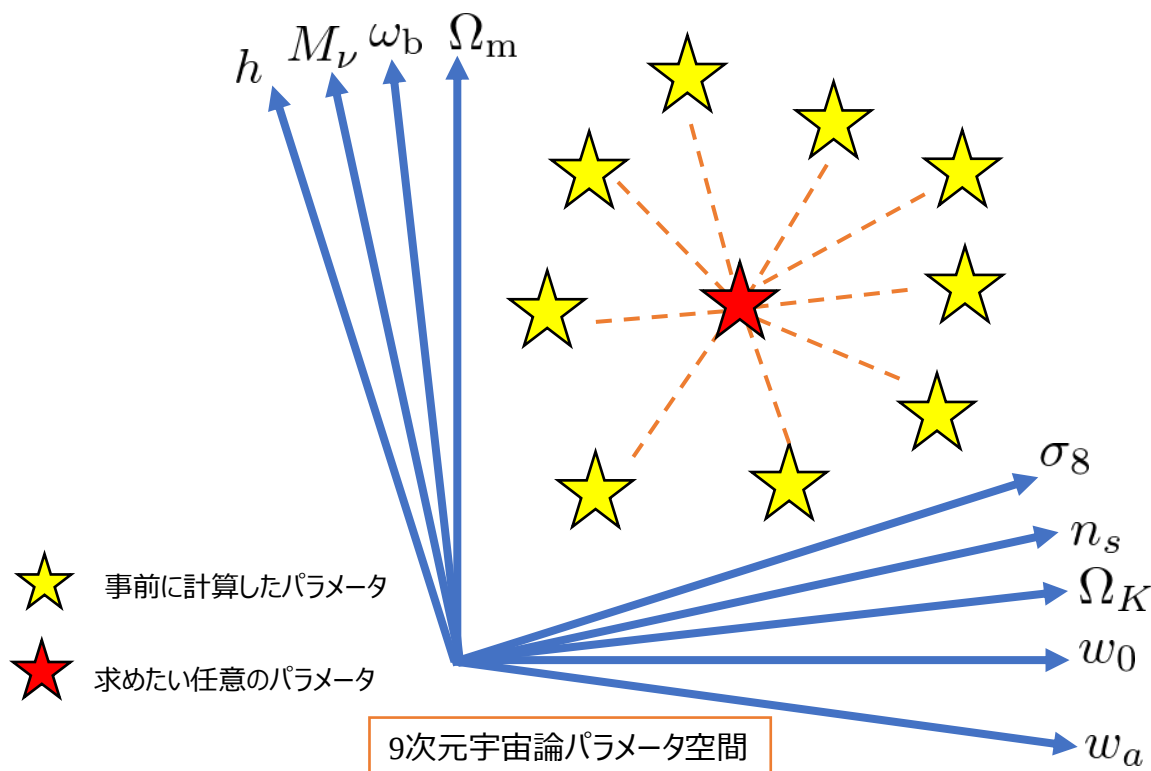
Ishiyama+ `21
Uchuu Simulation , $\sim 12800^3$ particles

富岳をすべて使えば(メモリのには)40000³粒子以上の計算も不可能ではない時代になってきている

- 大規模構造シミュレーションにおいてN-bodyシミュレーションは強力なツール
- 特定の宇宙論パラメータに対して1兆粒子以上の大きなN体シミュレーションは数多く出てきている ->石山さんトーク

Emulator

- 観測から得られる宇宙論パラメータは多次元かつ幅広く分布しており、それを埋め尽くシミュレーションを行うことは計算手法や計算資源が発達した現在でも非常に困難
- Dark Quest project I (DQ I, Nisimichi+ '19) では**エミュレータ技術**を使い、高い精度で任意の6次元宇宙論パラメータでのハローや銀河の統計量を求めることに成功している



ルールに則った方法で決定したパラメータ空間★をN体シミュレーションで計算

その結果を元にパワースペクトルやハロー質量関数などの統計量を求めそれを学習データとして機械学習を用いてエミュレータを構築

エミュレータによって任意のパラメータ空間での統計量を推定

N体シミュレーションをするパラメータの数や粒子数の規模は使える計算資源との兼ね合い

宇宙大規模構造 Emulator

	MiraTitan Femulator (Heitmann+`16, Lawrence+`17,Bocquet+`20)	EuclidEmulator (Knabenhans+`20)	Dark quest emulator I. (Nishimichi+ `19)	Dark quest emulator II. (本講演)
宇宙論モデル	8D $w_0w_a\text{CDM} + M_\nu$	8D $w_0w_a\text{CDM} + M_\nu$	6D flat $w\text{CDM}$	9D $w_0w_a\text{CDM} + M_\nu + \Omega_K$
シミュレーションコード	HACC	PKDGRAV3	GADGET-2 base	FDPS base
ターゲット	Matter power spectra Halo Mass Function	Matter power spectra	銀河やハローの統計量 RSD	Matter power spectra Halo Mass Function
(最高)解像度	2.1 Gpc/h, 3200^3	1 Gpc/h, 3000^3	1 Gpc/h, 2048^3	1,2, 4 Gpc/h , 1024^3, 2048^3, 3000^3
実験計画	多層ラテン超方格	多層ラテン超方格	多層ラテン超方格	Sequential design
回帰モデル	Gaussian process	polynomial chaos expansion (PCE)	Weighted PCA + Gaussian process	ニューラルネット(FNN) super resolution emulation
データセット数	111 params	254 params	101 params	2100(low), ~100(mid), ~ 20(high)
精度	2%程度	1%程度	2% (halo-matter cross-corr., relevant for galaxy- galaxy weak lensing) 4%(halo auto-corr., relevant for galaxy clustering corr.)	powerやHMFで 1%以下が目標

FDPS based code

- ✓ 粒子数が十分多い場合にはGadget2(Springel, 2005) コードの10倍以上の高速化が可能
- ✓ 強スケーリングで 90% の高並列性を実現
- ✓ Fast Multipole Method (FMM) を使った公開コード PKDGRAV3 (Potter+, 2017) に比べても総計算時間では有利

- “Linear Response” method をPM部分に適用することにより様々な効果を追加

- Massive neutrino
- Massless neutrino
- Photon
- Metric PT
- DE clustering

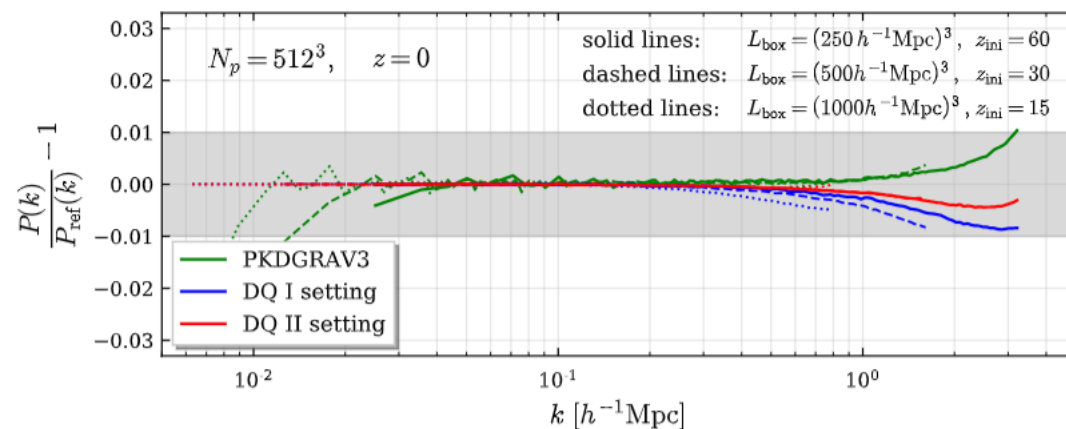
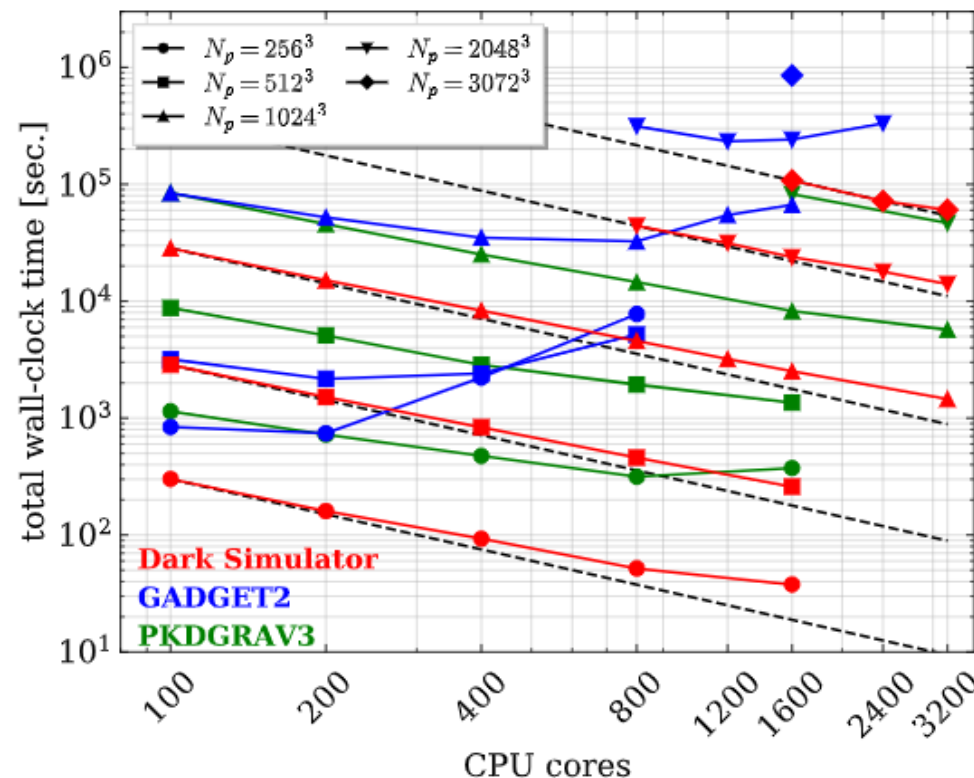
$$\frac{d^2 T_{cb}(k)}{d(\ln a)^2} + \left[1 + \frac{d \ln \mathcal{H}}{d \ln a} \right] \frac{dT_{cb}(k)}{d \ln a} = \frac{3}{2} \left[\Omega_{cb}(a) T_{cb}(k) + \mathcal{S}_{\text{other}}(k) \right]$$

Expansion history

External sources

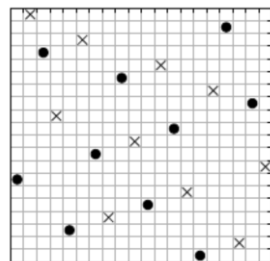
Transfer function for CDM+baryon fluid

Extract from CLASS



Experimental Design in DQ1

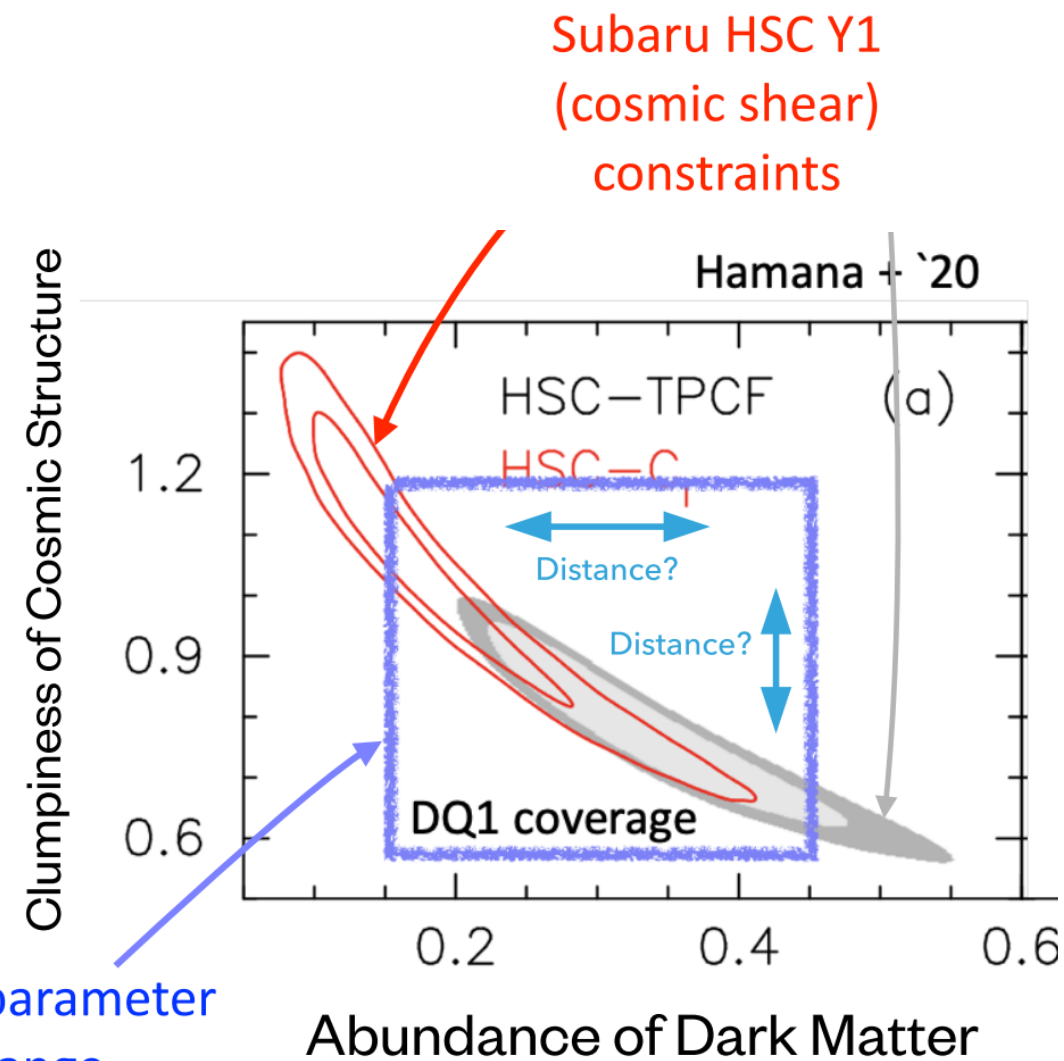
- ✓ HSCの観測から示唆されるパラメータ空間はDQ1のときよりも膨大
- ✓ 例えば $\Omega_m - \sigma_8$ パラメータ空間は“バナナ” 状の形をしている
 - 格子パターンのパラメータを決めるのは困難
- ✓ 学習データがどれくらいあれば十分なのかはやってみないとわからない
- ✓ 先行研究などでよく使われている**Latin-Hypercube design**では最初にパラメータの範囲とパラメータ数を決めるのであとからパラメータを新たに追加することが難しい



- 各列と各行ではサンプルが重複しないようにする
- データ点の最小距離が最大になるように選ぶ (Maximin)

DQ2ではニューラルネットワークでの学習を念頭にパラメータの位置や数を柔軟に変更できる
Sequential designを採用した

DQ1 parameter range



Experimental Design

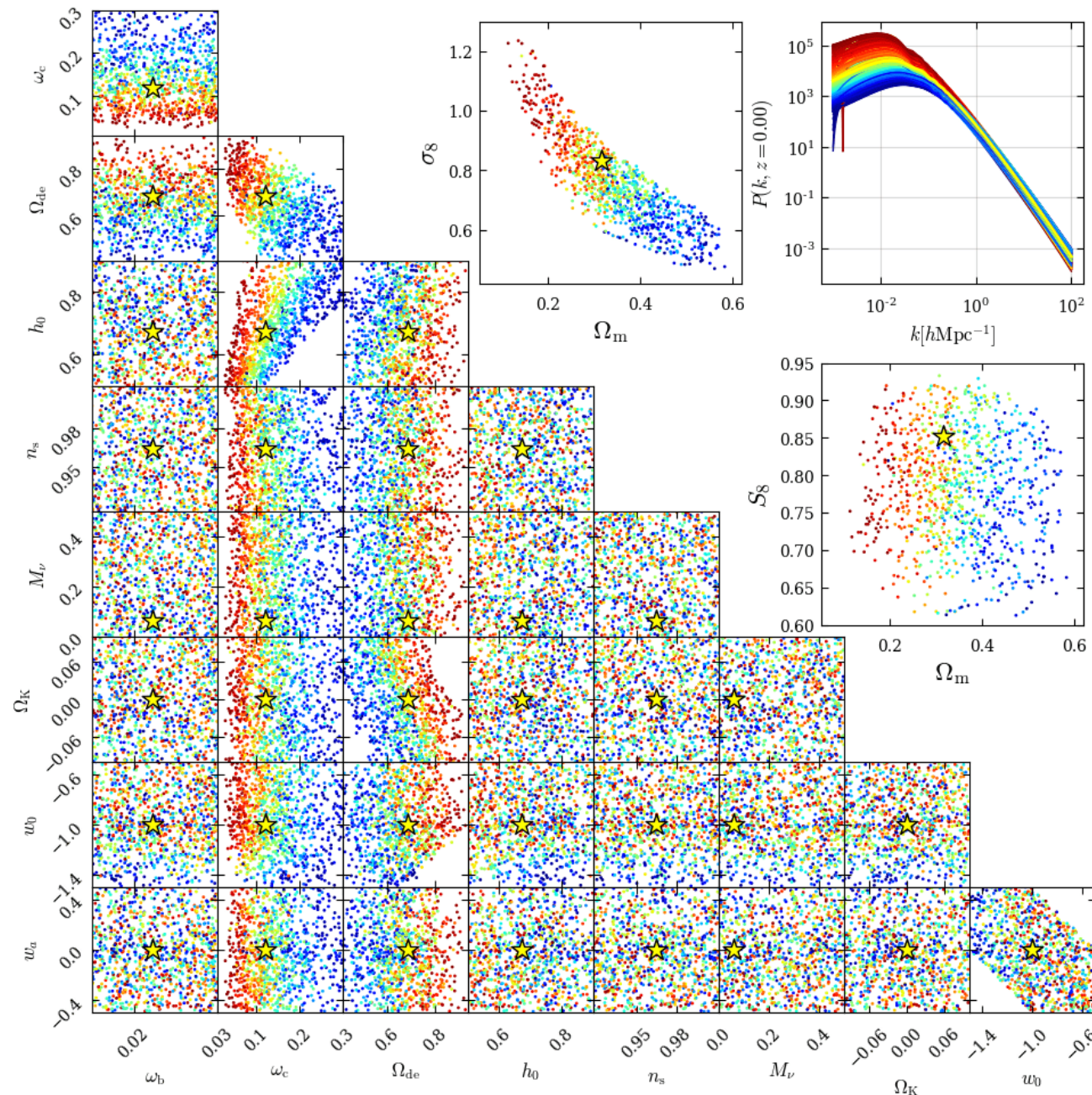
➤ Sequential design

- 最初に $\Omega_m - \sigma_8$ のバナナ状パラメータ空間の中心付近に重みをつけてランダムに振る
- $\omega_c, \omega_b, M_\nu$ などの Ω_m の従属パラメータは整合性が取れるようにランダムに振る
- CLASS(Lesgourgues '11) で Transfer function, linear power spectrum を計算し、fiducial パラメータでの linear power と **距離** が離れすぎているものを採用する

distance criteria of linear power

$$d = \frac{1}{N_k} \sum_i^{N_k} \frac{|P_{\text{lin, fid}} - P_{\text{lin}, i}|}{\min(P_{\text{lin, fid}}, P_{\text{lin}, i})} < 5$$

$$\begin{aligned} 0.0562 < \Omega_m < 0.6, & \quad 0.9163 < n_s < 1.0127, \\ 0.54 < \sigma_8 < 1.4, & \quad 0.0 < M_\nu < 0.5, \\ 0.015 < \omega_b < 0.03, & \quad -0.1 < \Omega_K < 0.1, \\ 0.01 < \omega_c < 0.3, & \quad 0.5 < h_0 < 0.9, \\ 0.4125 < \Omega_{\text{de}} < 0.9385, & \quad -1.5 < w_0 < -0.5, \\ & \quad -0.5 < w_a < 0.5, \end{aligned}$$



星印はfiducial

Learning Dataset

□ Low resolution models

- CDM粒子数 $1,024^3$
- 1辺のシミュレーション領域 1Gpc/h
- 初期密度ゆらぎの乱数固定 1,000 モデル
- 乱数ランダム 1,000 モデル(宇宙論パラメータは固定モデルと同じセット)
- fiducial parameterで乱数ランダム 100 モデル

(CfCA XC50で半年弱でデータ構築完了)

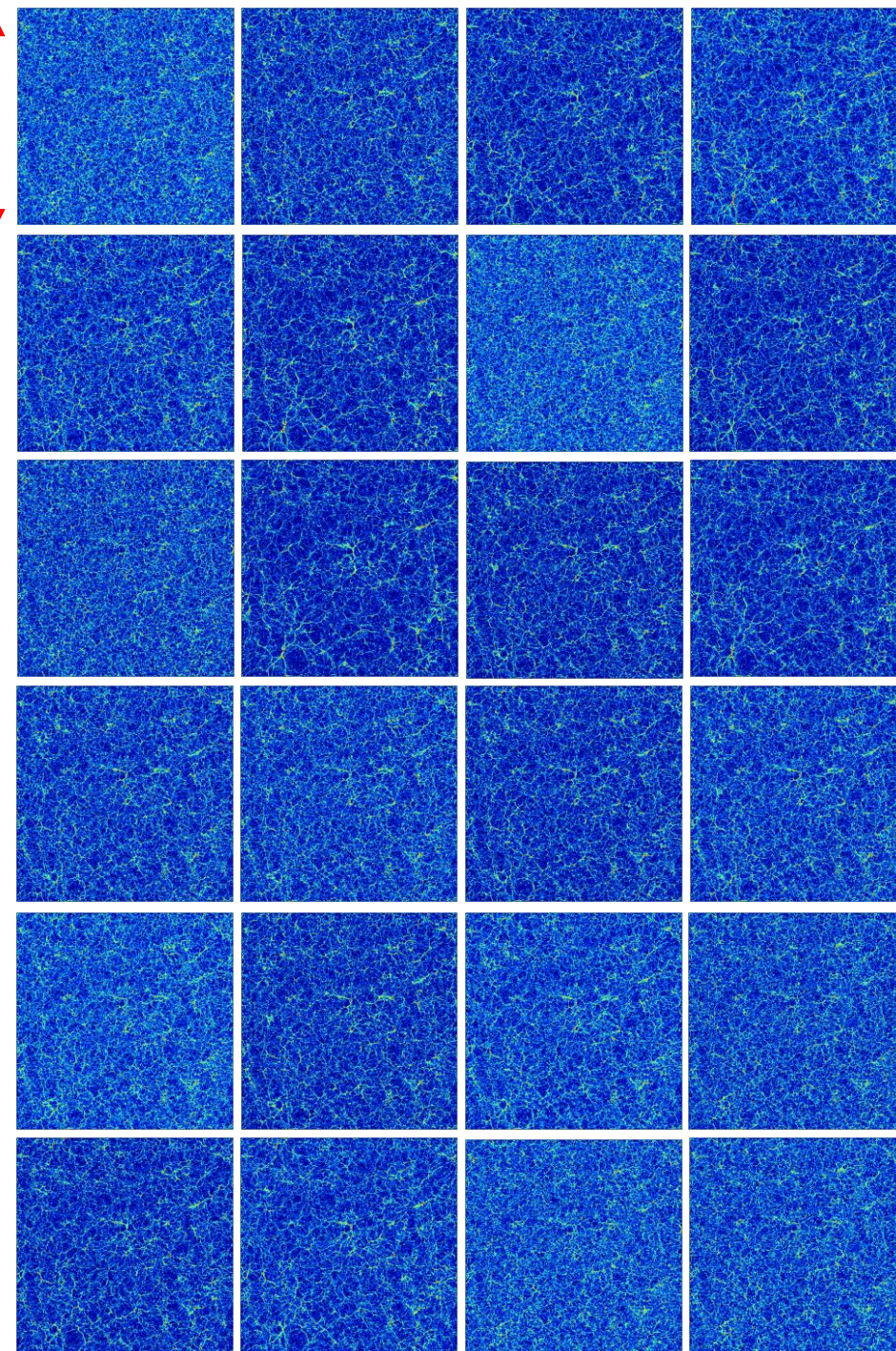
初期密度ゆらぎの乱数を振ることによってcosmic varianceの影響を機械学習に反映できるかを調査する

- 各モデル 11 スナップショット
- 全学習データ数 ~ 23,000
- 生のN-bodyデータサイズ ~ 1.1 PB
- 加工後のデータサイズ ~ 150 TB

(power spectrum, halo mass function, 3D density fields)

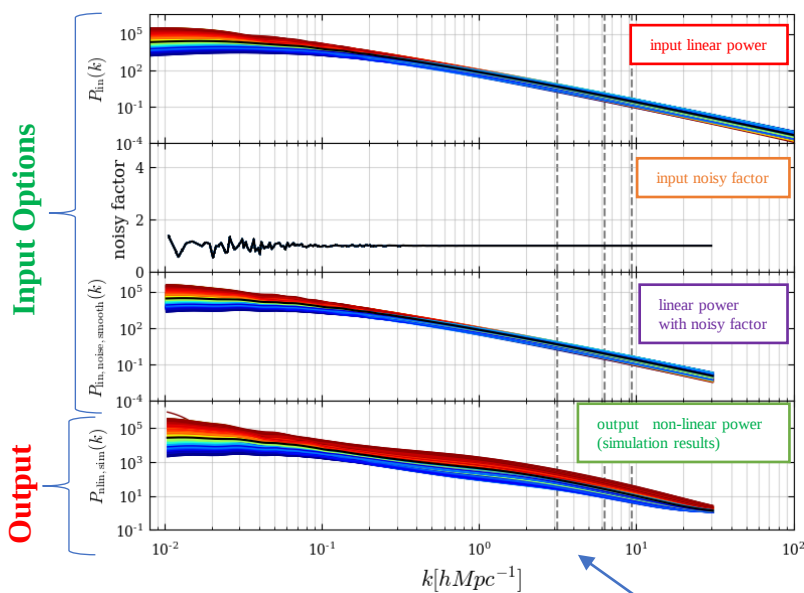
Cosmic varianceに
対応させるために初
期密度ゆらぎを考慮

1Gpc/h



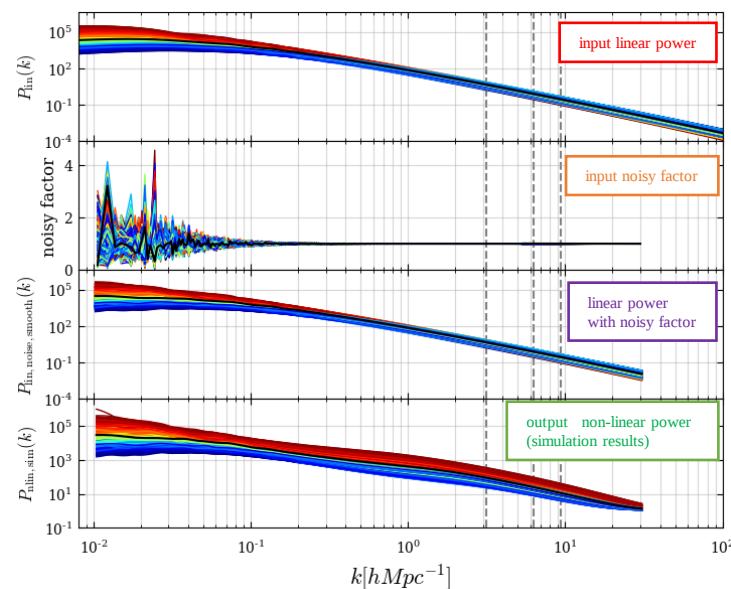
Input Parameters and output power spectrum

乱数固定 1,000 モデル

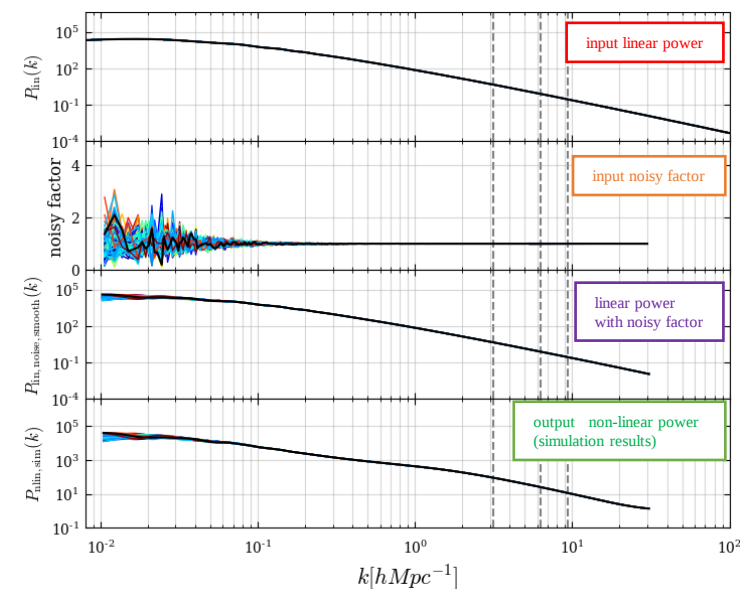


点線は1Gpc/hに対する粒子数 $1024^3, 2048^3, 3072^3$ でのナイキスト周波数。
粒子の相関を使いlarge-kまでパワーを引いてるが基本的に信用できるのはここまで

乱数ランダム 1,000 モデル



fiducial 乱数ランダム 100 モデル

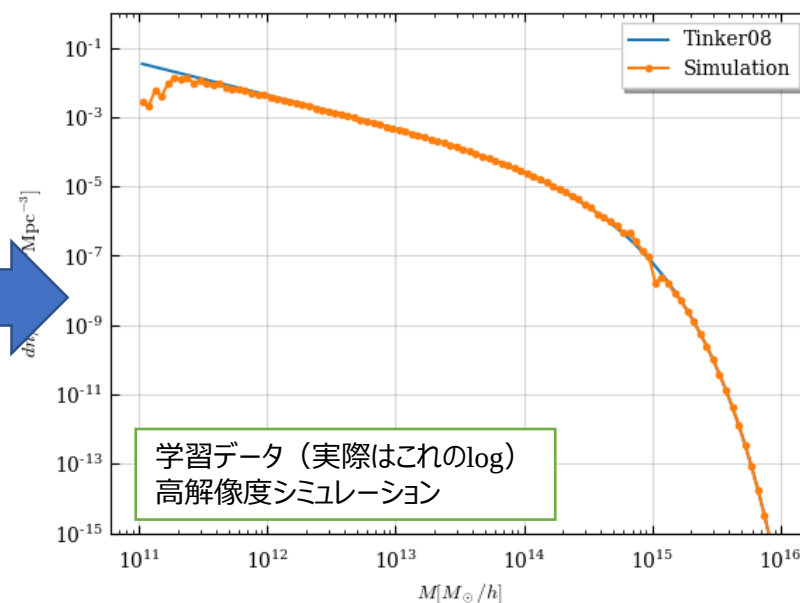
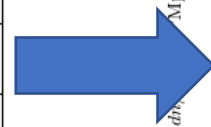
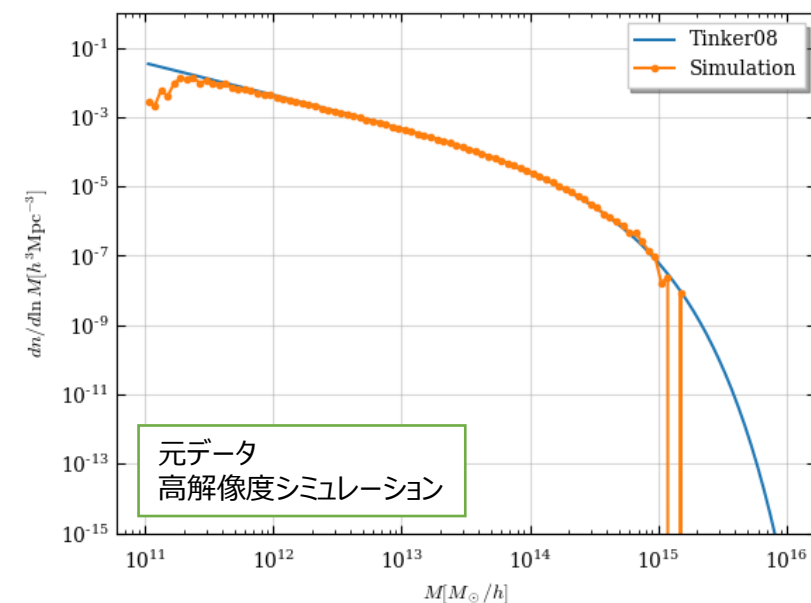
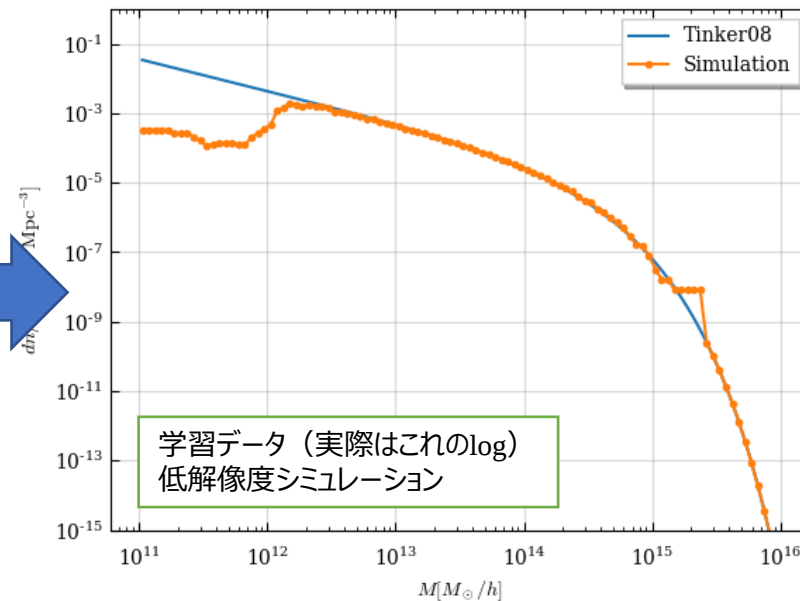
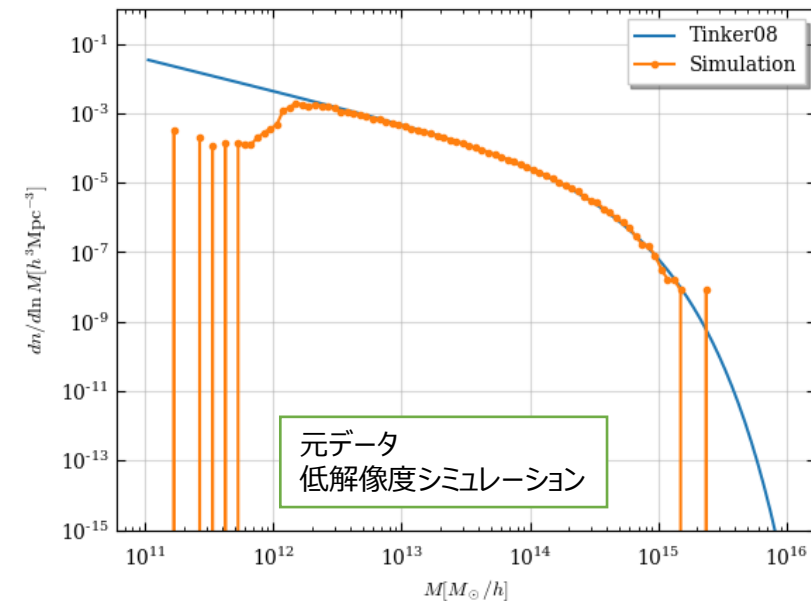


同じ宇宙論パラメータでも初期密度ゆらぎパターンを変えるとsmall-kで差が現れる

- ✓ 上記は $z=0$ の図。これ以外にも各モデルつき計11個の赤方偏移[3,0]のインプット、アウトプットの組を用意
- ✓ 9次元宇宙論パラメータだけでなくlinear powerや乱数の情報を組み合わせて学習を行う

9D params , 9D params + linear power, 9D params + linear power with noisy factor ...

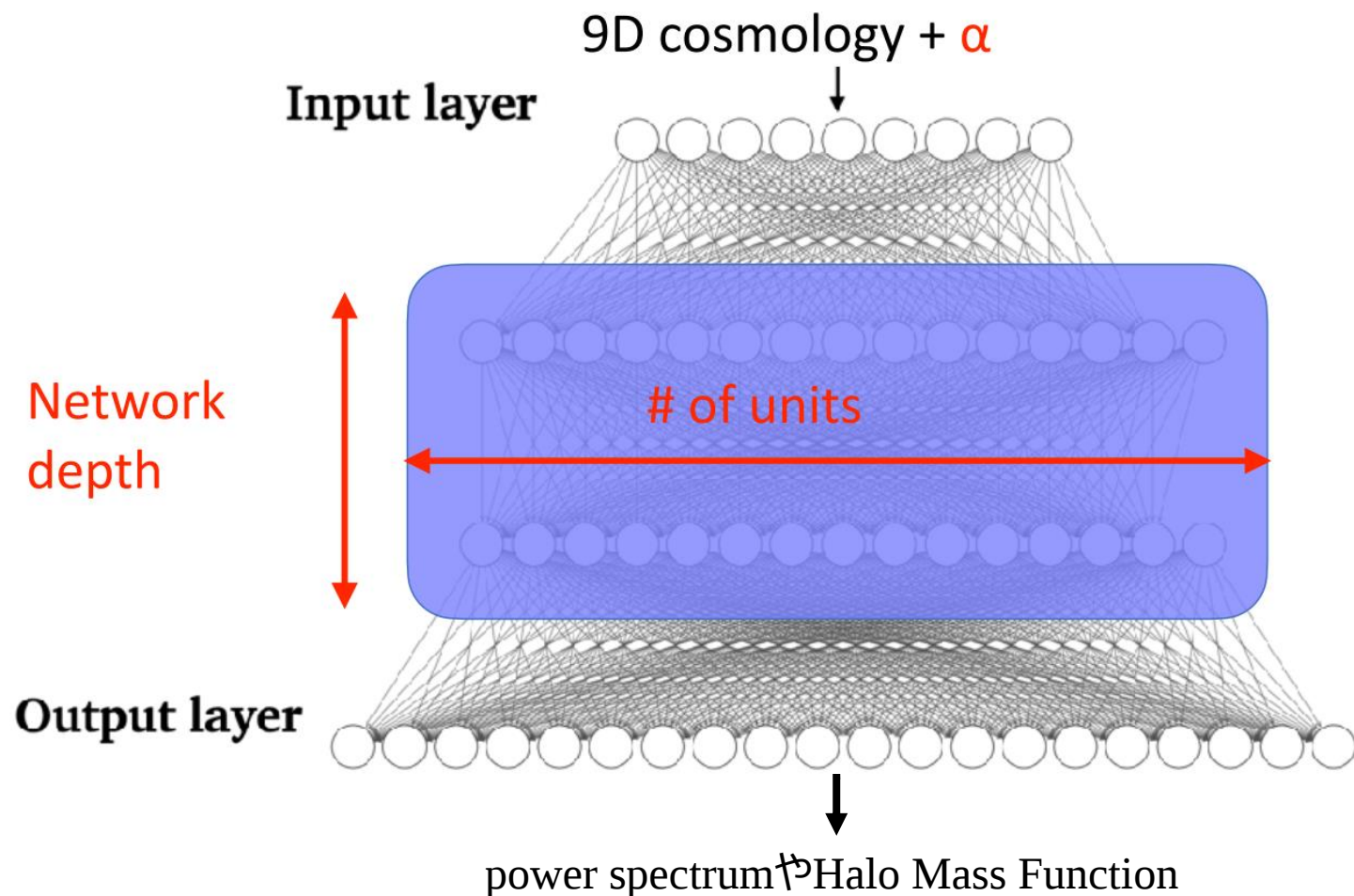
Output Halo Mass function



- 低解像度計算やhigh-zなどハロー質量関数が0になる部分ができるためlogを取れなくなり都合が悪い
- シミュレーションデータ点がある場所はそのを使用
- 大質量側のデータがない部分はTinker08のハロー質量関数でデータ点補間
- 小質量側では既存のデータ点から補間

Regression strateg

- PyTorch Libraryを使った
Feed-Forward Neural Network
- Hyperparameter
 - ✓ 隠れ層のサイズ
 - ✓ ネットワークノード数
 - ✓ 活性化関数
 - ✓ 損失関数
 - ✓ 最適化関数
 - ✓ スケジューラー
 - ✓ エポック数
 - ✓ ミニバッチサイズ ...
- Tuning
 - ✓ Automatic Mixed Precision
 - ✓ Cross-Validation
 - ✓ Batch Normalization
 - ✓ Dropout ...



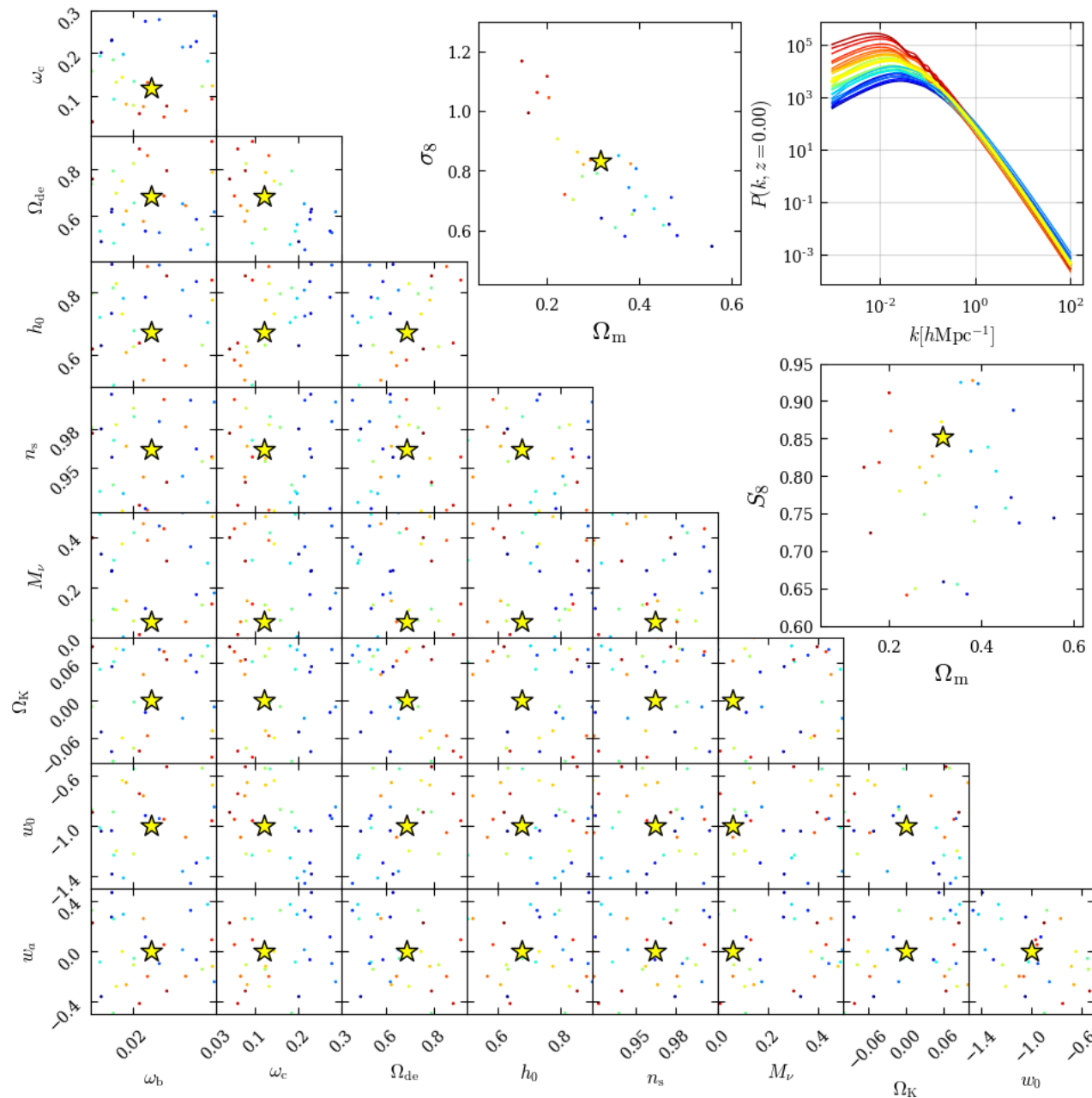
- | | |
|--|--------------------------|
| ➤ [隠れ層、ニューロン]: [2, 200], [4, 400], [6, 600], [8, 800] | ➤ min-batch size : 128 |
| ➤ エポック数 : 20,000 or 100,000 | ➤ Active function : ReLU |
| ➤ cross-validation : 100分割 <ul style="list-style-type: none">✓ 全データに対して99%を学習データ、
1%を検証データとして選択し、100回入れ替えながら学習 | ➤ Loss function : RMSE |
| ➤ repeated cross-validation : 10回 <ul style="list-style-type: none">✓ cross-validationの選択パターンを10回ランダムに変更して繰り返す | ➤ Optimizer : NVLAMB |

Testset

- 学習、検証にはには使わないで学習結果の評価をするためのデータセット
- テストセットの構築もSequential designにを使うと確率的にfiducialパラメータ付近に多くなるのでLatin-Hypercube designで構築
 - Latin-Hypercube designでのバナナ状のサンプル決定は厄介で最初に300個程度の全領域のパラメータを求め、バナナ状内に距離指標など満たすサンプル数がちょうど30個になるまでLatin-Hypercube designの初期値を変えて繰り返し計算し直す
- 初期密度ゆらぎの乱数固定 30 テスト
- 乱数ランダム 300 テスト
 - 30個の同じ宇宙論パラメータ
×10 realizations セット

testsetの距離指標一覧

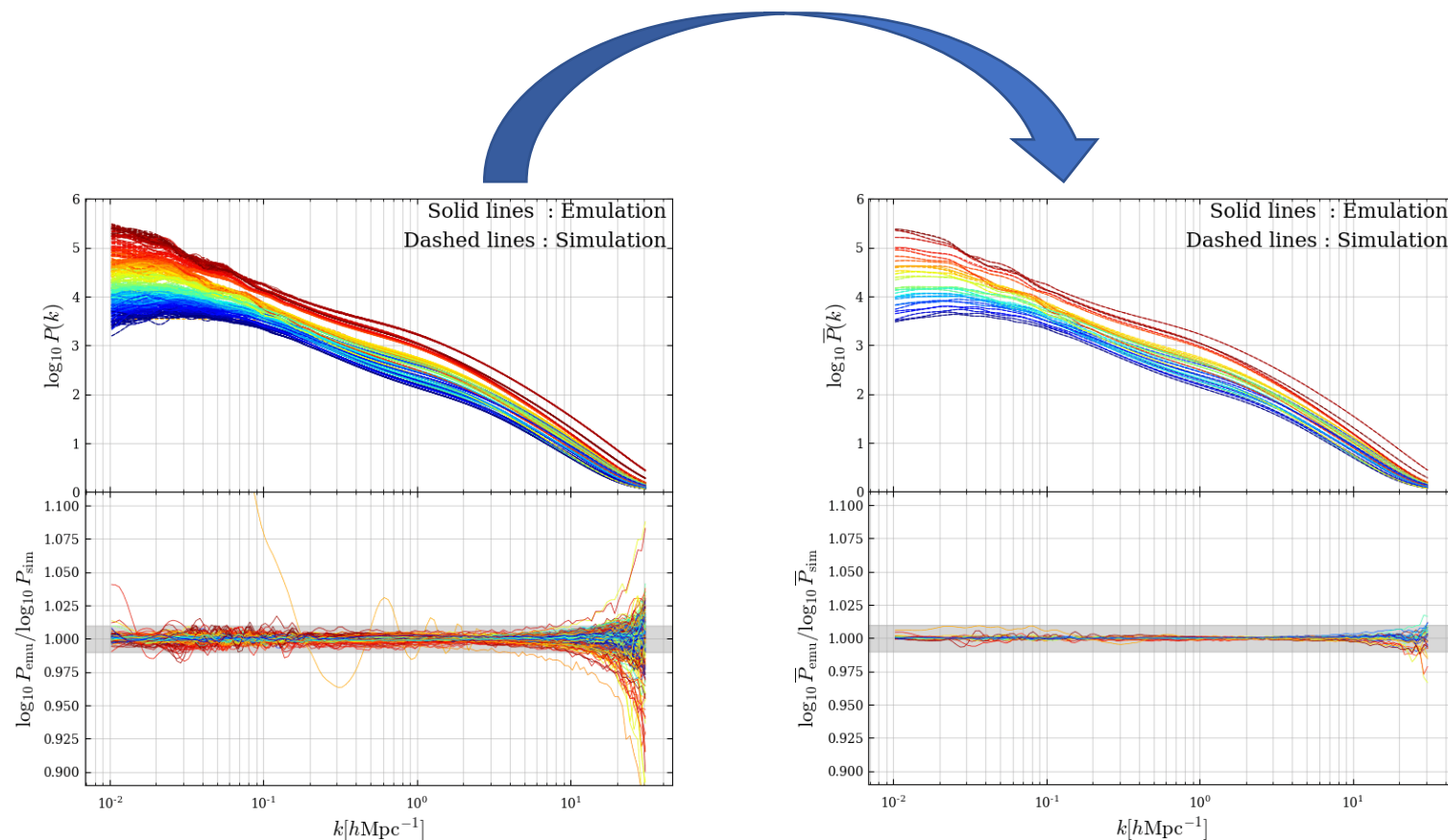
0, 0.064, 0.207, 0.252, 0.273, 0.291, 0.426, 0.466, 0.509, 0.515, 0.625, 0.631, 0.639, 0.645, 0.773, 0.781, 0.796, 0.836, 1.329, 1.383, 1.393, 1.460, 1.476, 1.603, 1.717, 2.204, 2.325, 2.497, 2.739, 3.338, 4.608



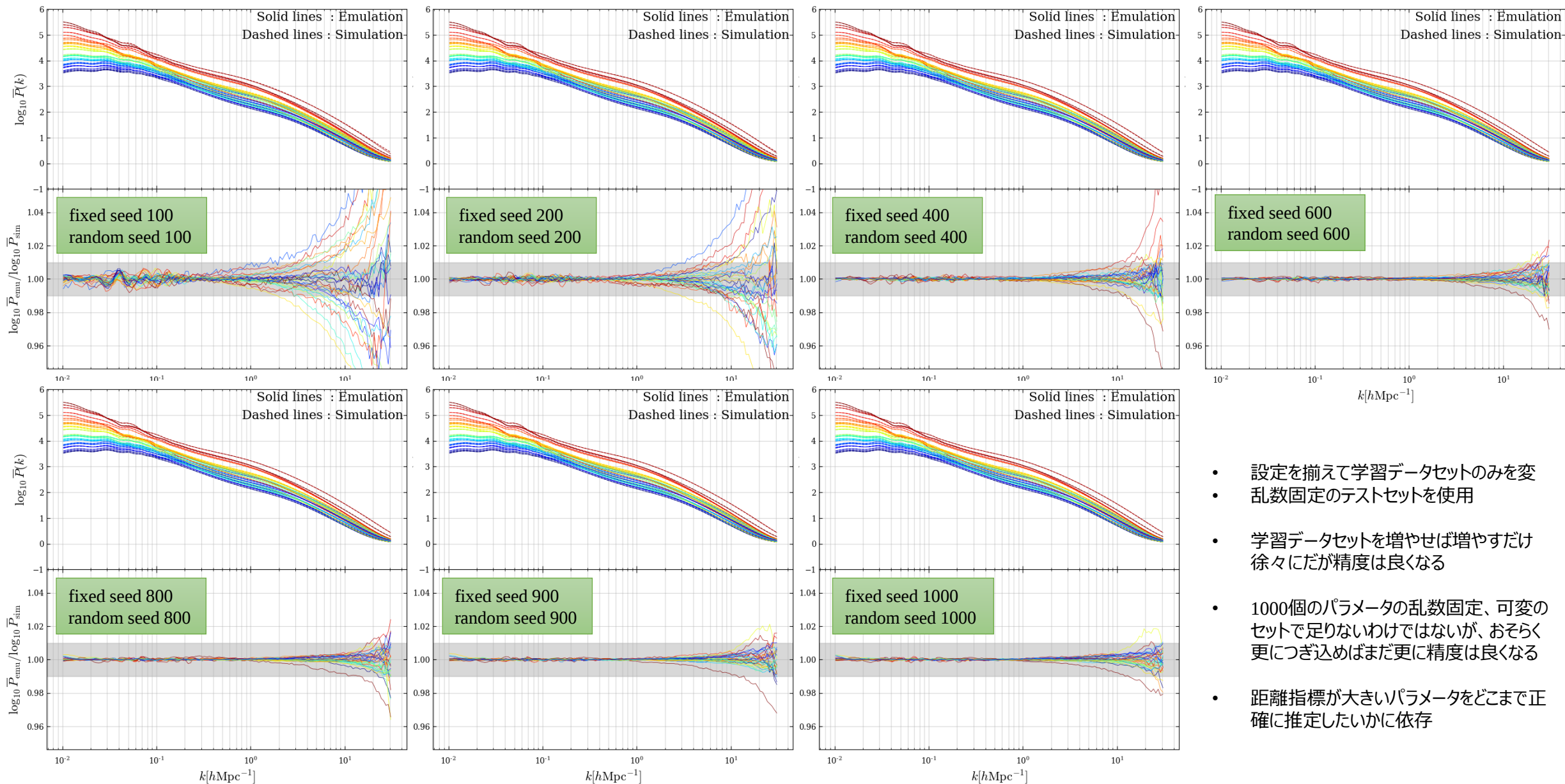
星印はfiducial

Testsetとの比較

- 学習は $\log P(k)$ の形で行っているので比較も $P(k)$ ではなく $\log P(k)$ で行う
- 初期密度乱数を振ったものは一つの宇宙論パラメータに対して10本の線が引ける
 - ✓ 個々の結果に余り意味がないので平均化した値で比較
- 上パネルで $\log P(k)$ が0と交差する付近で相対誤差が定義できないのでエラーが大きくなることもある
 - ✓ 以降の結果の large- k 側の大きなエラーは基本的にはこれは関係ない範囲
 - ✓ 高解像度合わせた計算の結果はこれが影響する箇所もある



最低限必要な学習データセット数の確認

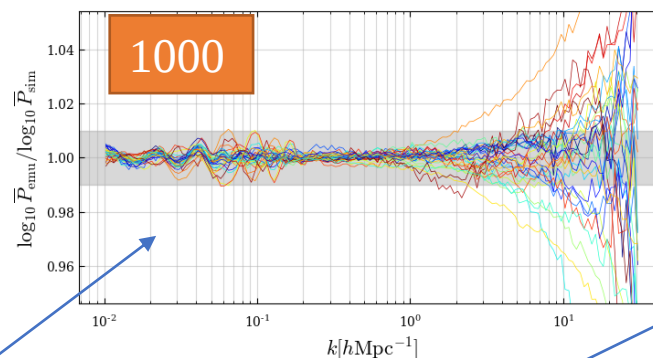
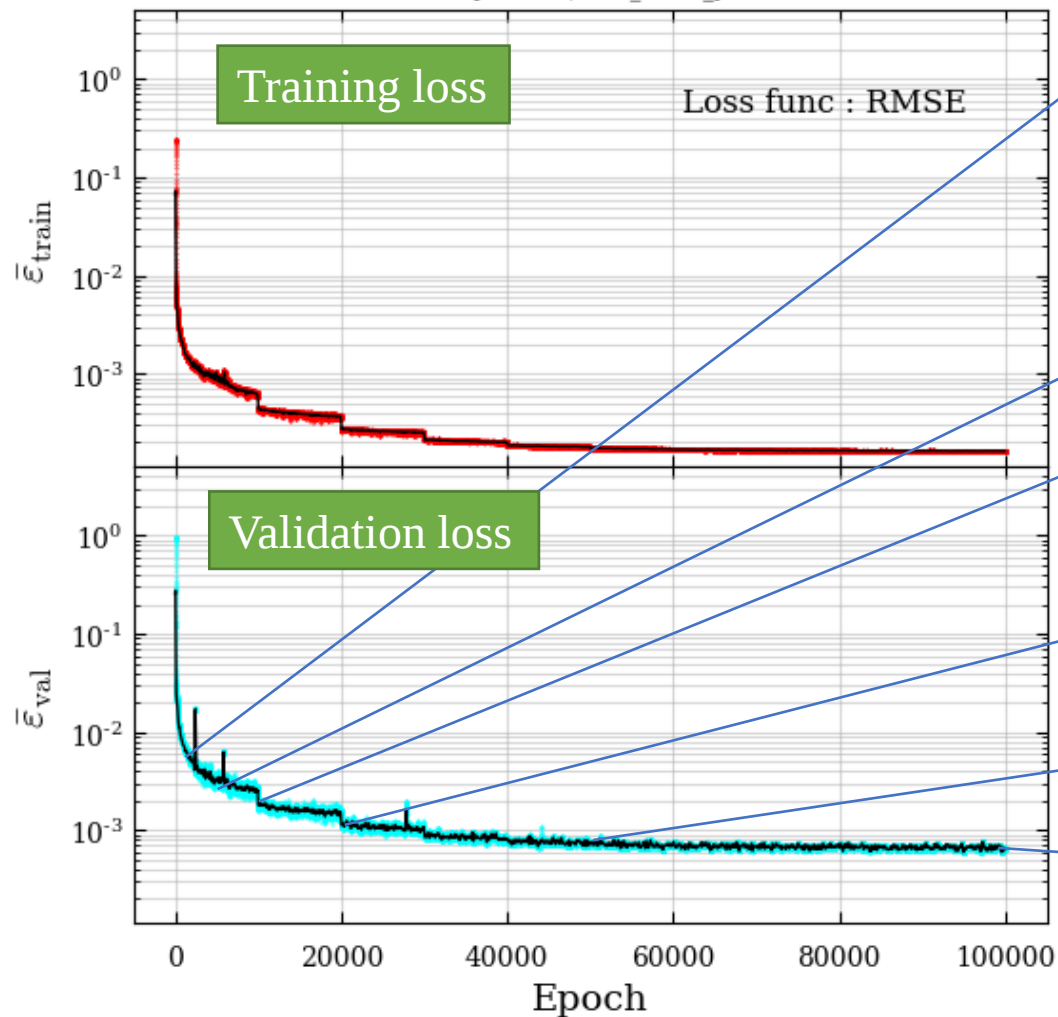


- 設定を揃えて学習データセットのみを変
- 乱数固定のテストセットを使用
- 学習データセットを増やせば増やすだけ徐々にだが精度は良くなる
- 1000個のパラメータの乱数固定、可変のセットで足りないわけではないが、おそらく更につぎ込めばまだ更に精度は良くなる
- 距離指標が大きいパラメータをどこまで正確に推定したいかに依存

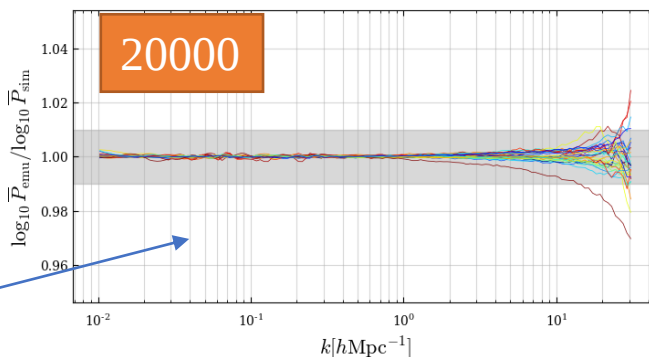
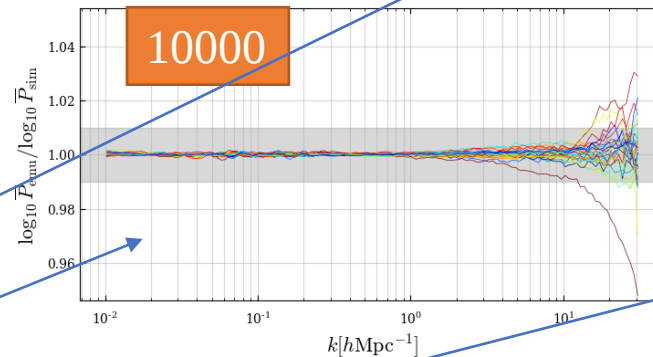
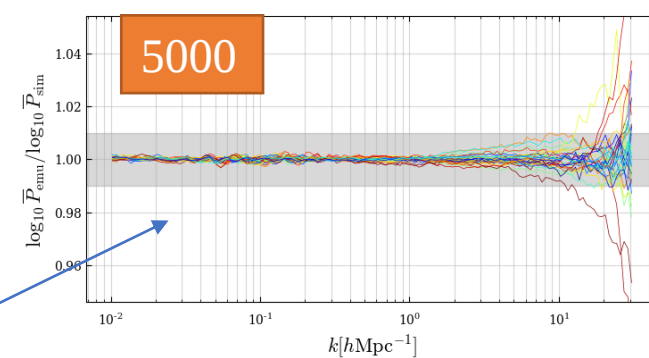
必要なエポック数

- 設定を揃えてエポック数のみを変更
- 乱数固定のテストセットを使用

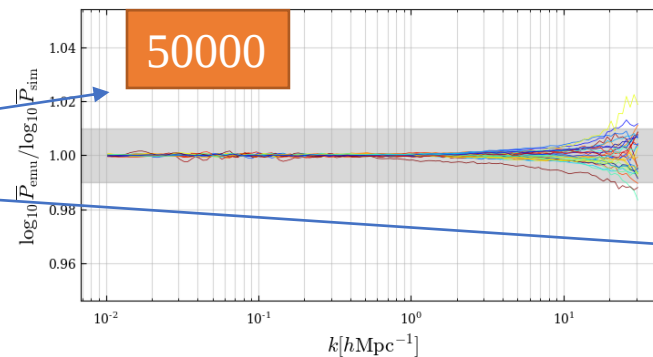
$N_{\text{param}} = 2102$, $N_{\text{batch}} = 128$, $N_{\text{hidden}} = 4$, $N_{\text{net}} = 400$,
 $N_{\text{epoch}} = 100000$, $N_{\text{CV}} = 100$, $N_{\text{RepCV}} = 10$,
TYPE=param9d, GRF_linear_power



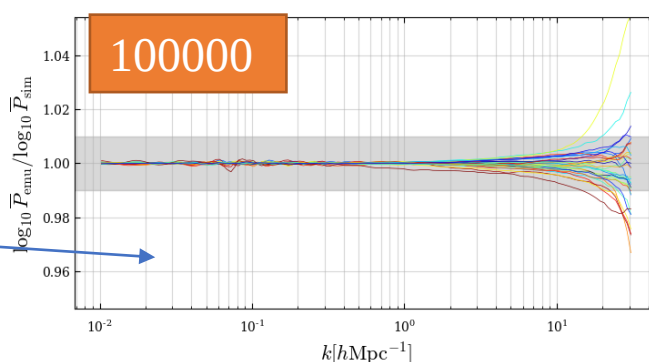
5000回くらい回せばそれなりに回帰可能



この場合は50000回くらいが最適

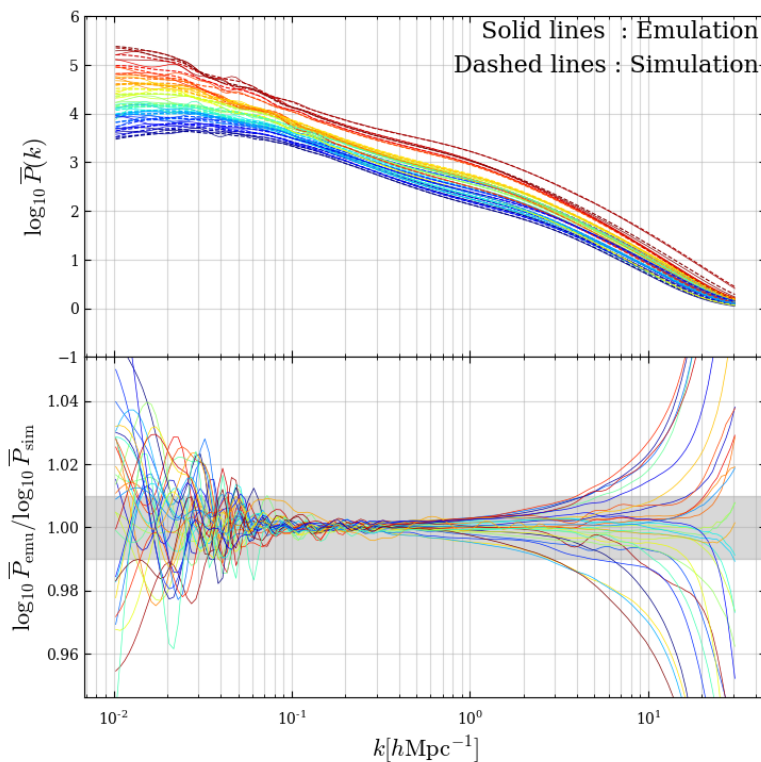


Lossは50000とあまり変わらないがエラーが増す

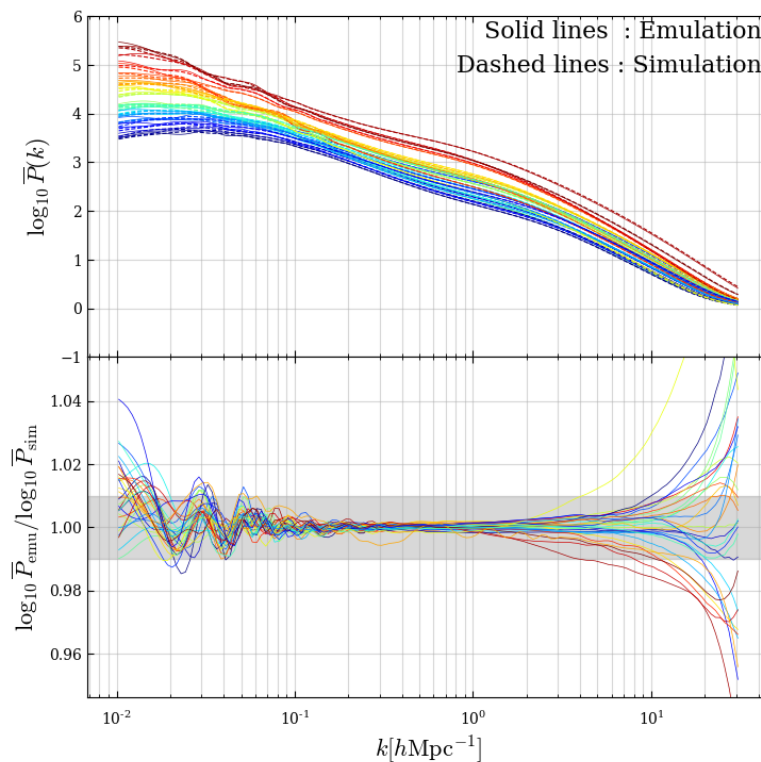


最適なInput parameterの確認

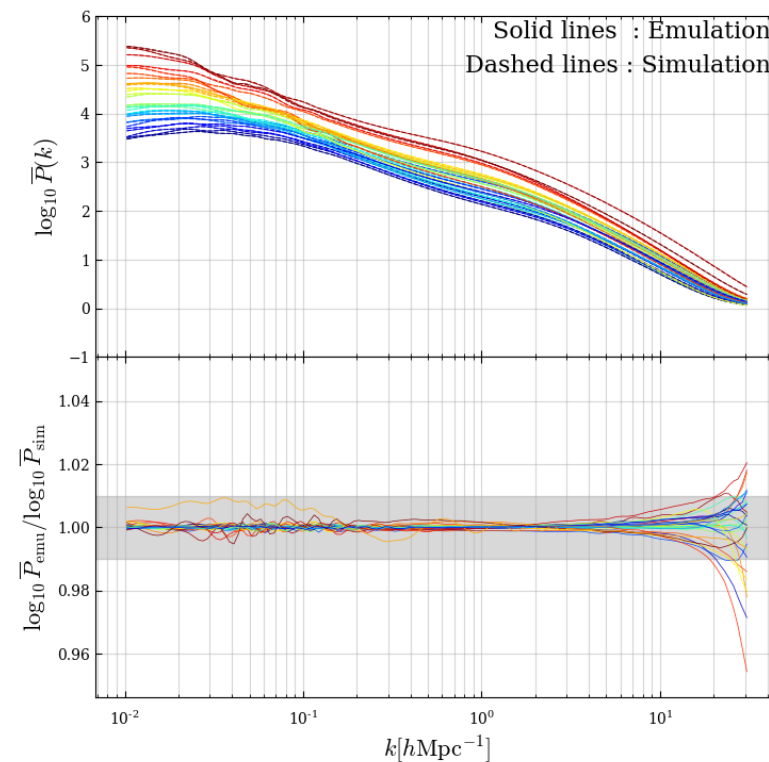
9D parameter のみ



9D parameter + lin P(k)



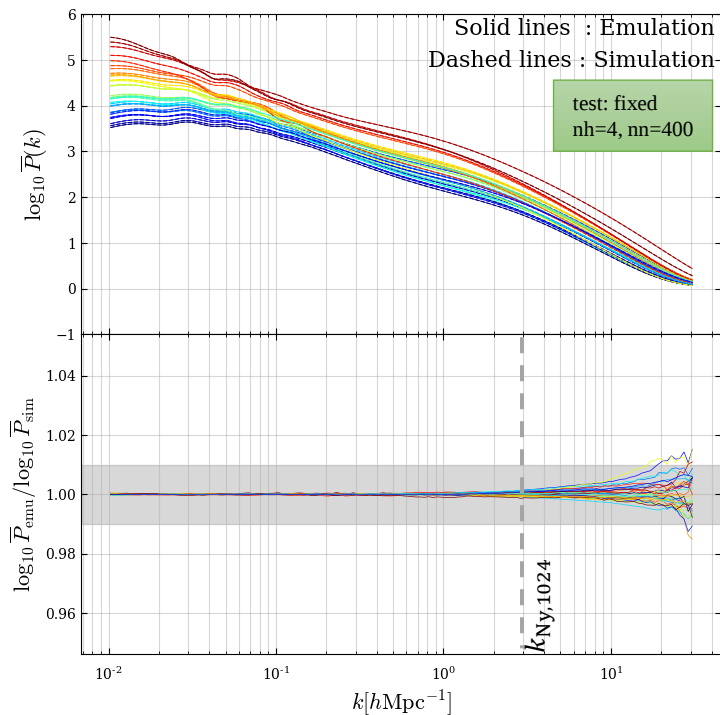
9D parameter + lin P(k) w/ 初期ゆらぎ



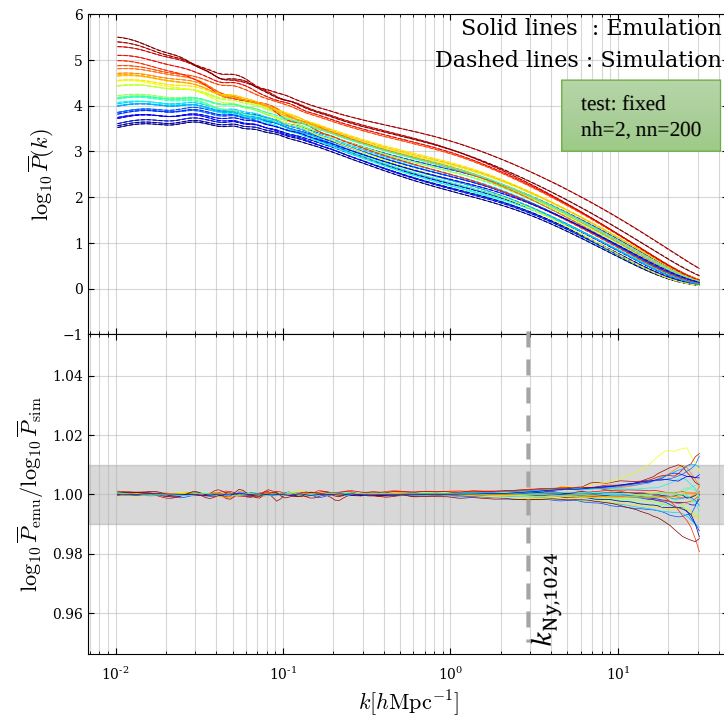
- ✓ ランダムテストセットに対する10realizationのエラーの平均
- ✓ 9Dパラメータのみやlin P(k)を使うだけでは初期ゆらぎ密度のランダムを捉えることが不可能
- ✓ 初期密度ゆらぎのランダム具合を考慮した情報を入れると大スケールのばらつきなどが解消

Power spectrum : 乱数固定データとの比較

fixed seedによる学習



fixed + random seed による学習

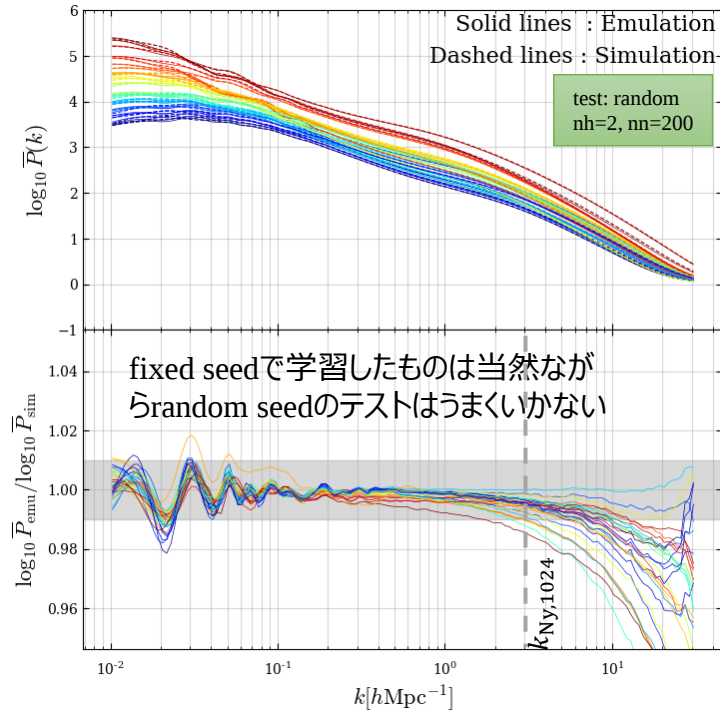


nh : 隠れ層の数
nn : ニューロンの数
学習条件によって最適なハイパーパラメータも
変わるので試した範囲で最も良いものを採用

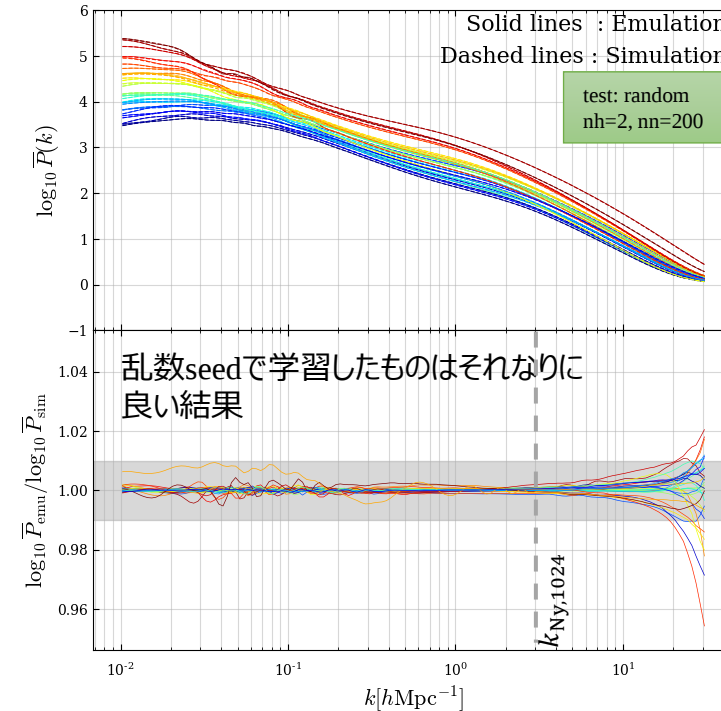
- ✓ fixedテストセットに対してfixed seedのみで学習したものは k_{Ny} 以下では0.1%以下レベルで再現可能
- ✓ fixedテストセットに対してfixed + random seedで学習したものは若干振動が目立つが基本的に1%以下のエラー
- ✓ 小スケール(k_{ny} の3-4倍)に行くにつれてエラーが大きくなる傾向にある

Power spectrum : 乱数を振ったテストセットとの比較

fixed seedによる学習



fixed + random seed による学習

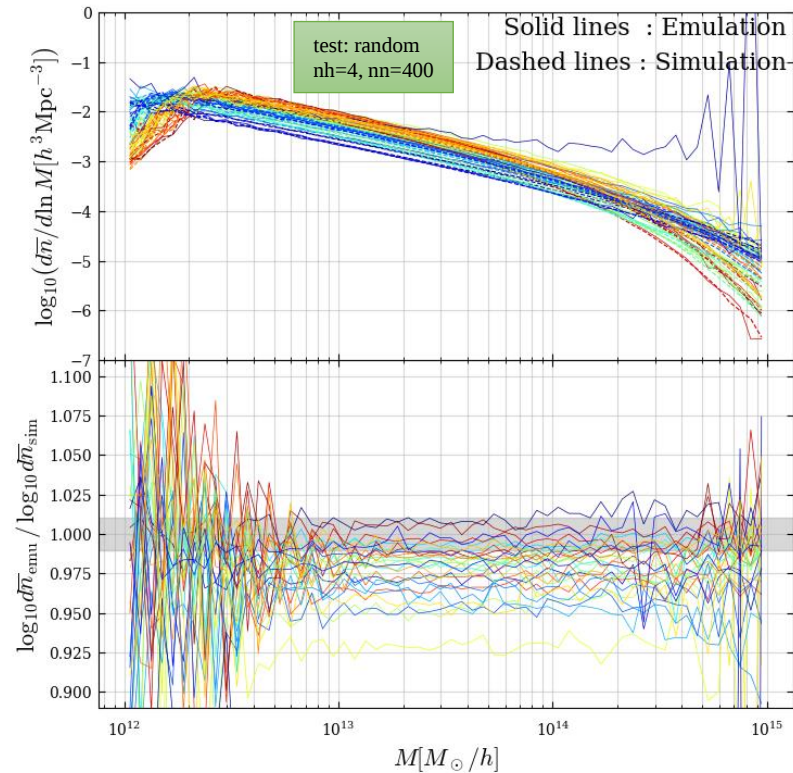


それぞれ10realizationある30個
のパラメータセットの平均をプロット

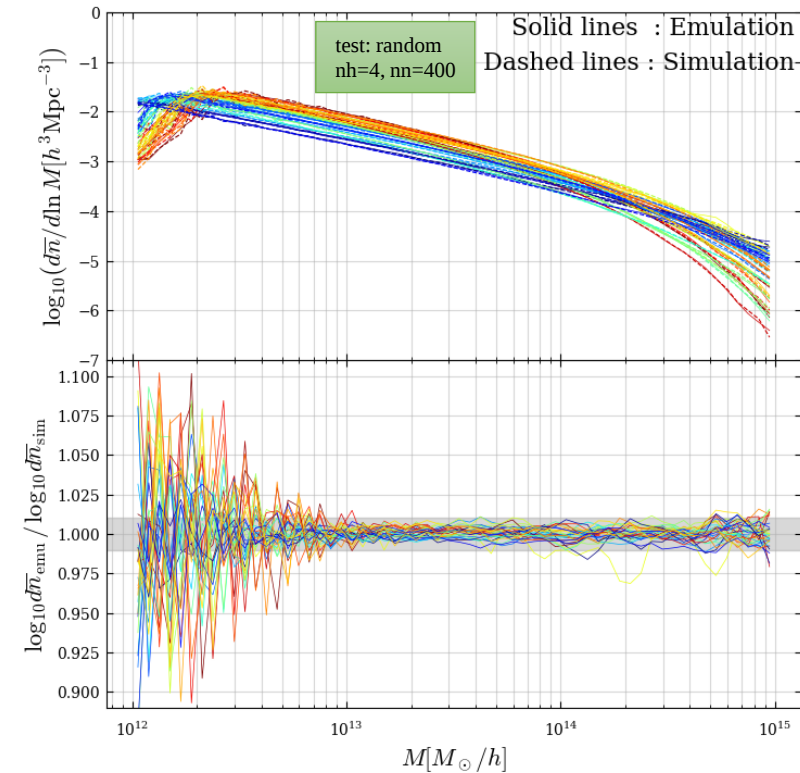
- fixed seedのみで学習したものはcosmic varianceを考慮したパワースペクトルは再現できない
- 一方でfixed + random seedで学習したものはrealizationの平均を取れば右端を除いて1%以下の精度
 - ✓ 個々のrealizationを見ただけでは比較できないので注意

Halo Mass Function : 乱数ランダム 30 × 10 realization 平均

fixed seedによる学習



fixed + random seed による学習

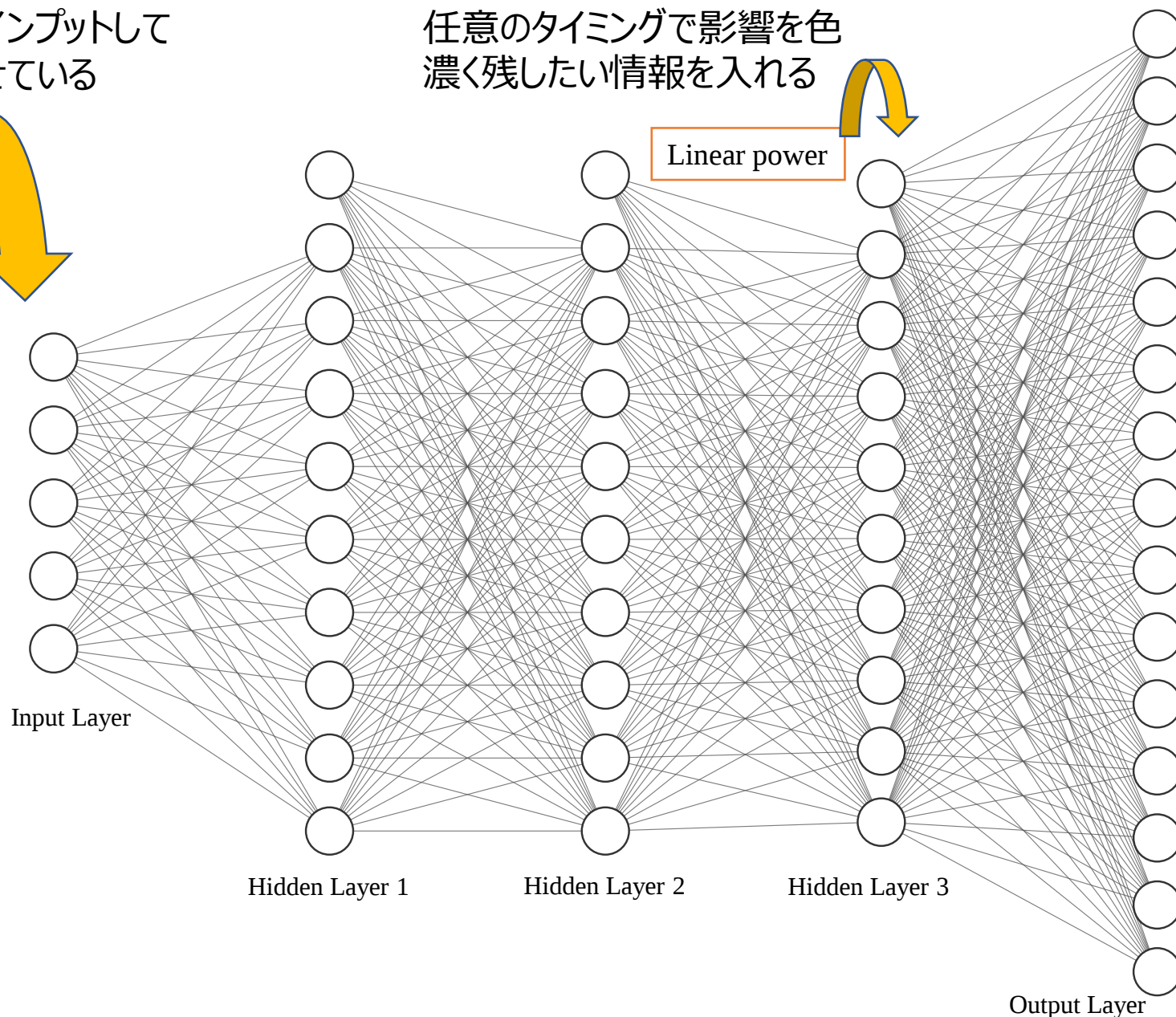
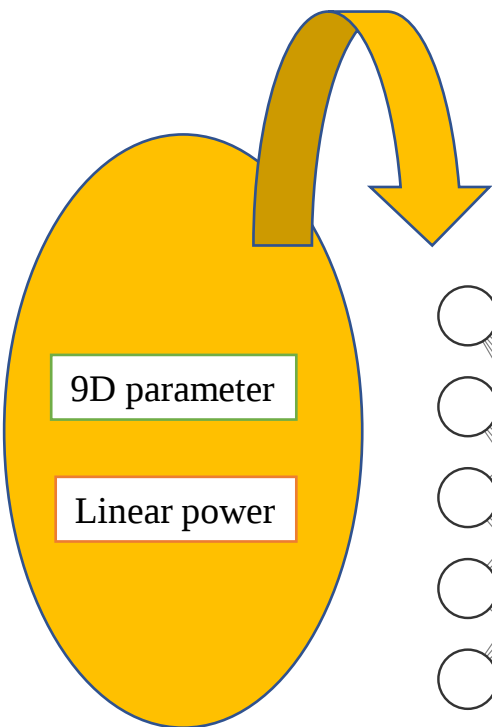


- 同様に10 realization の平均同士で比較
- fixed seedの学習結果ではrandomテストセットのばらつきを捉えることがこちらでもできない
- fixed + random seedの学習結果ではrandomテストセットに対して $8 \times 10^{12} < M < 1 \times 10^{15}$ では概ね1%以下のエラーを達成

入力パラメータを渡すタイミングの調査

今は全てまとめてインプットして
全結合で学習させている

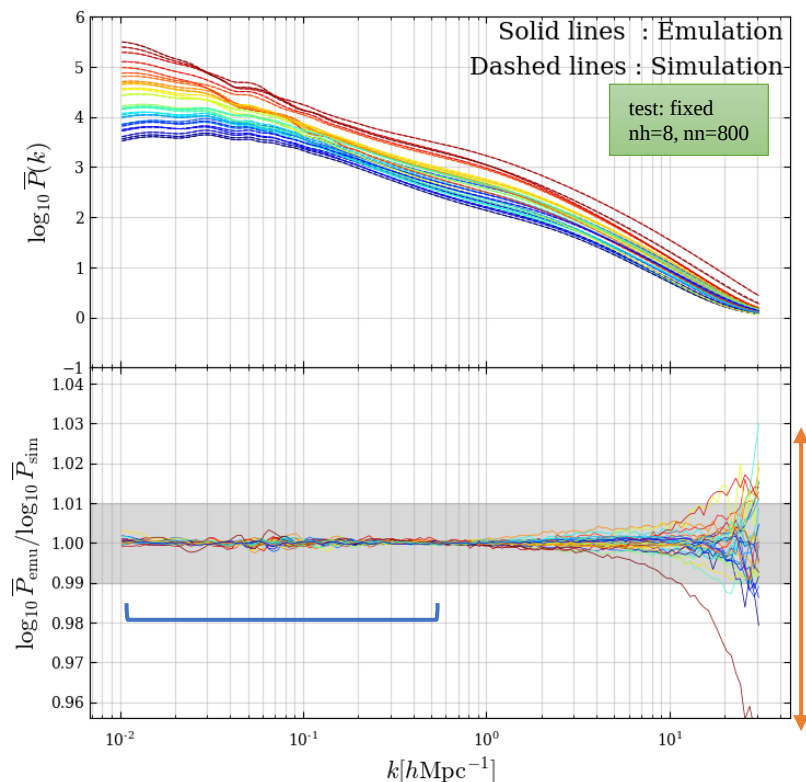
任意のタイミングで影響を色
濃く残したい情報を入れる



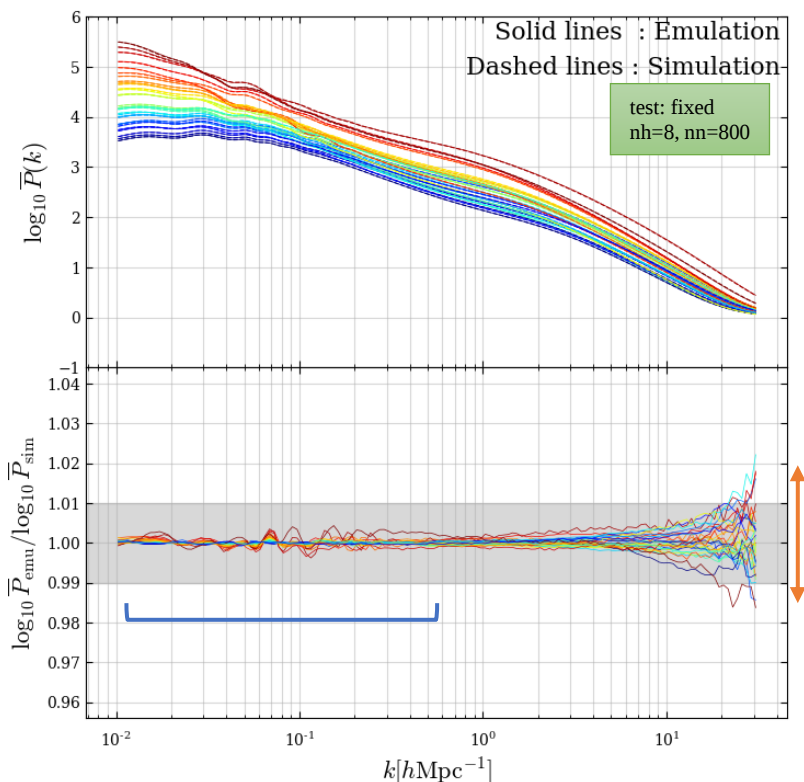
- 全ての情報を最初に一括で入れるのではなく適当な層の学習で入れるという選択肢もある
- 非線形変換で薄れていく線形パワーの情報をほぼそのまま使うことが可能になるので(良いか悪いかは不明だが)何らかの結果を期待できる

入力パラメータを渡すタイミングの調査

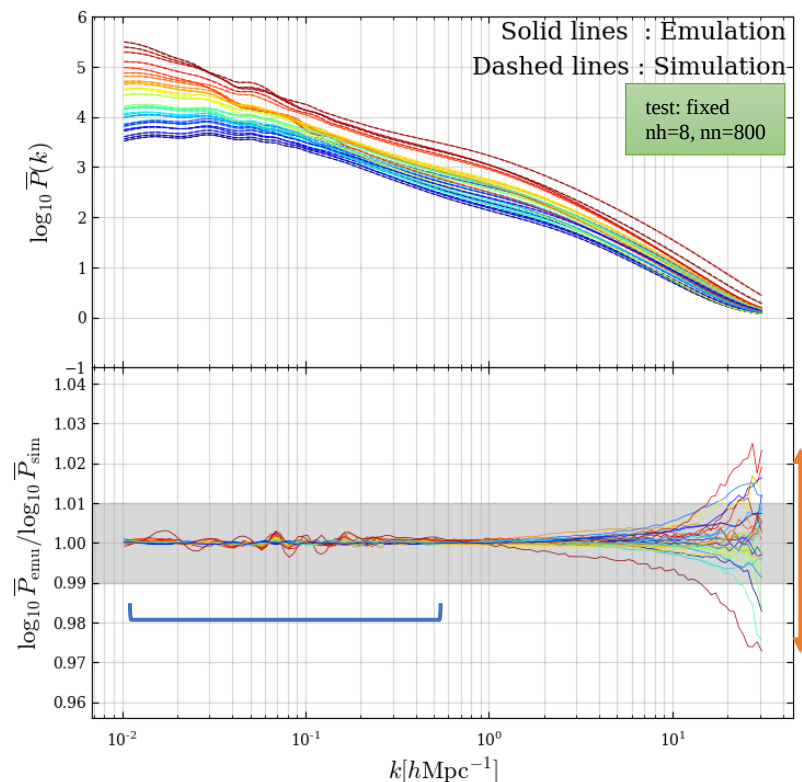
全パラメータを最初から使用



P_{lin} のみ最後の層で投入



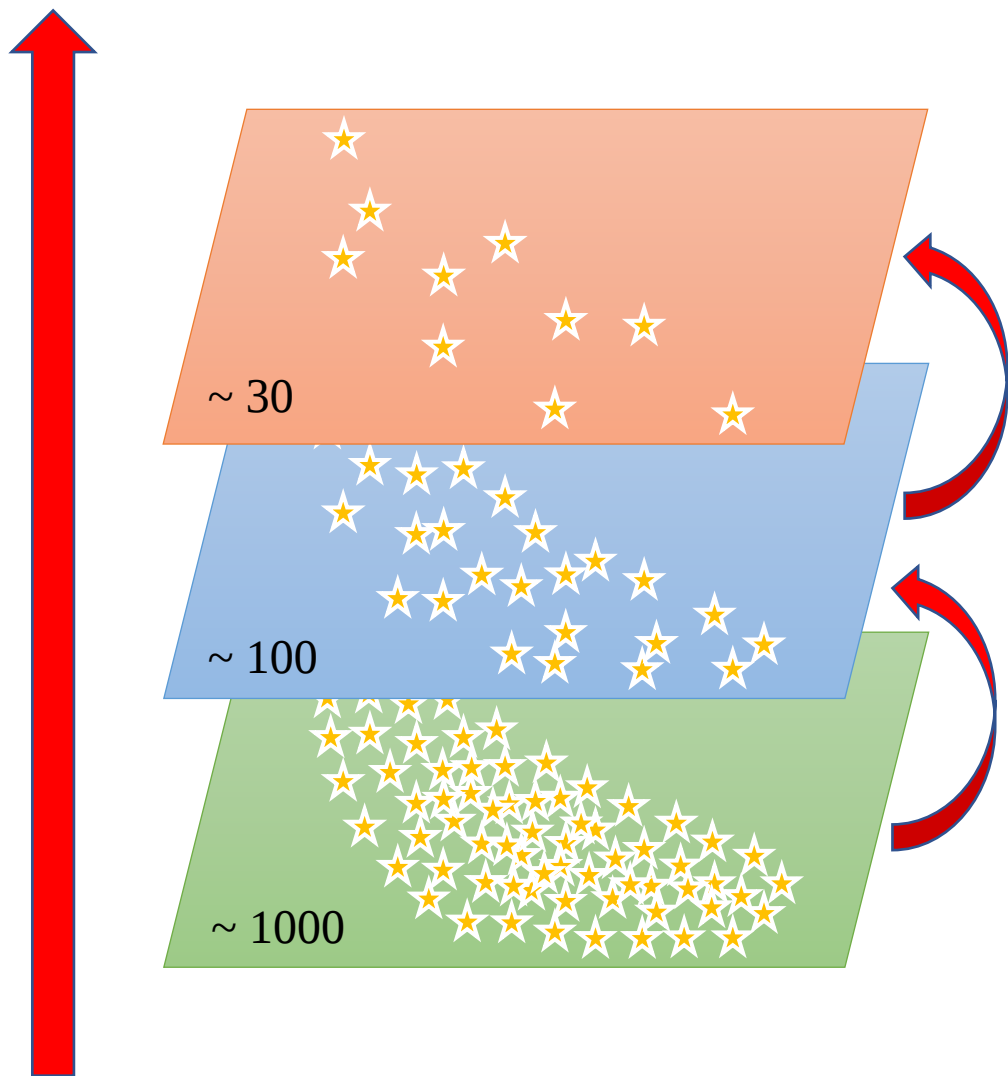
全パラメータを最初から使用しつつ
 P_{lin} を最後の層で再び投入



- 最初から全部のパラメータを一斉に渡す場合より、 P_{lin} を最後に投入したほうが小スケールのエラーを抑えることができる
- 逆に若干BAOの振動付近での推定が甘くなる(サブパーセントレベルは保てているが)
- 試しに両方を取り入れたら若干ながら2つの良いところが反映されてそうではあるがまだ改良はできそう

Super-resolution emulate

高解像度。サンプル少



低解像度。サンプル多

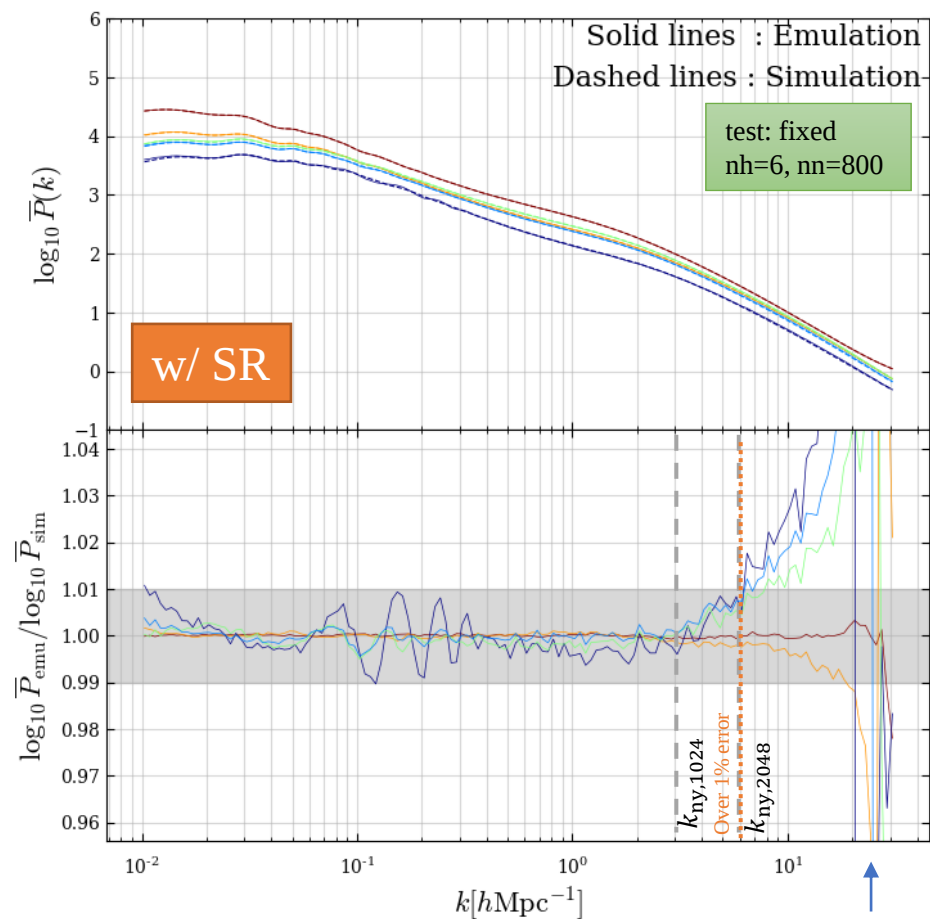
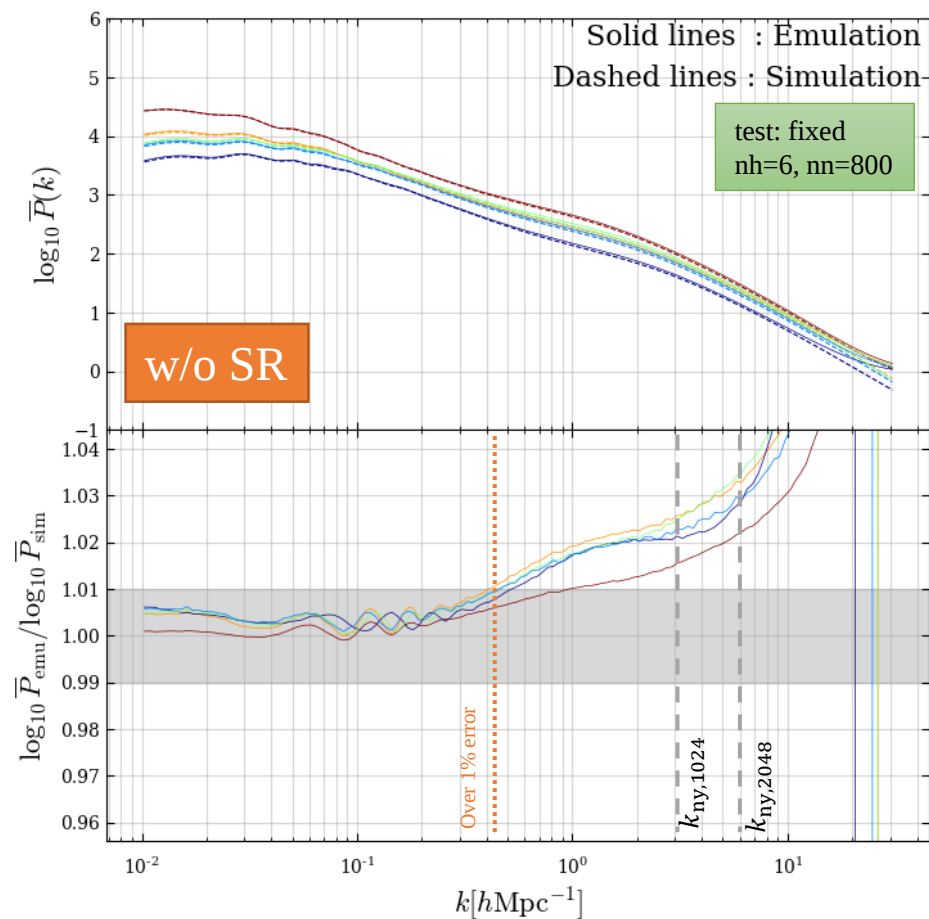
- 最初の目標であるパラメータ空間を埋め尽くす大規模な low-resolution モデルを構築
- 同様の方法を1000モデル以上の高解像度シミュレーションで構築するのは非効率かつ非現実的
- 低解像度と同じパラメータ空間で中解像度、高解像度のシミュレーションをいくつか行い、解像度の違いを学習
 - どれくらいの数が必要か
 - ✓ 初期密度ゆらぎ固定と乱数のバランス
 - パラメータの選び方の自由度
 - ✓ fiducial付近を中心的に埋めるか
 - ✓ 全領域を埋めるようにバラけさせるか



最終的には低解像度でしか埋めていないようなパラメータ空間周りでも高解像度シミュレーション相当の結果を得られるようにする

Power Spectrum for SR emulation

低解像度1000個のfixed
中解像度7個の fixed seedで学習

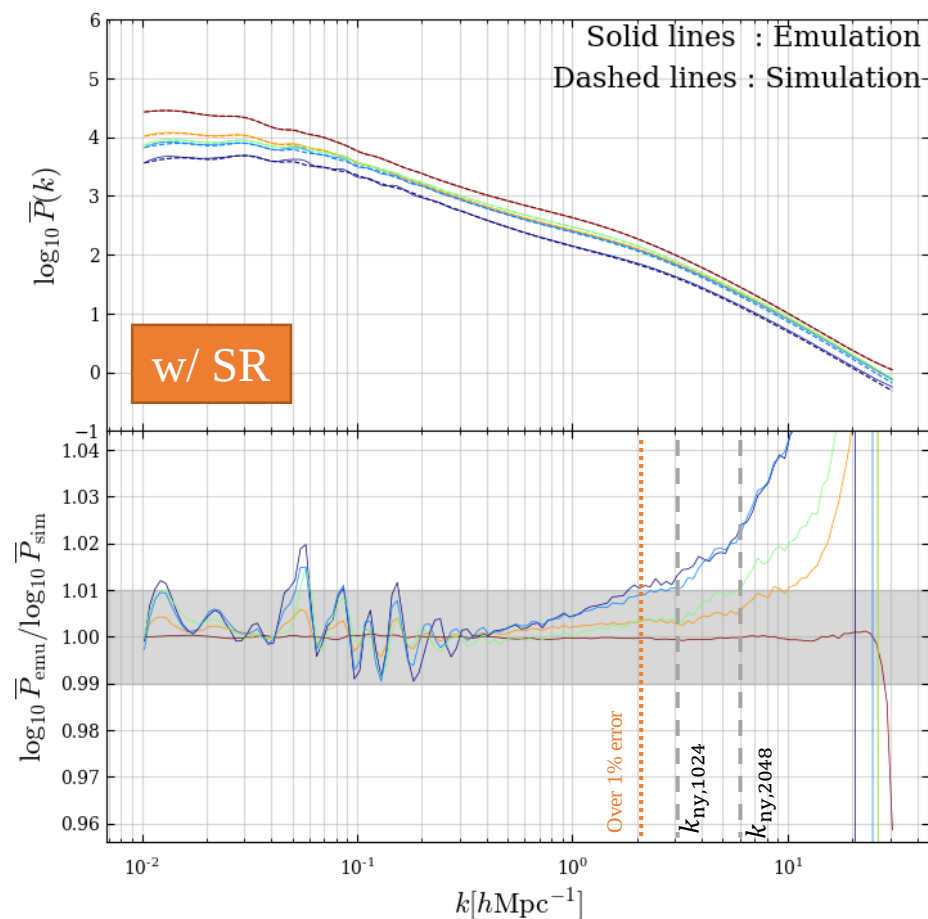
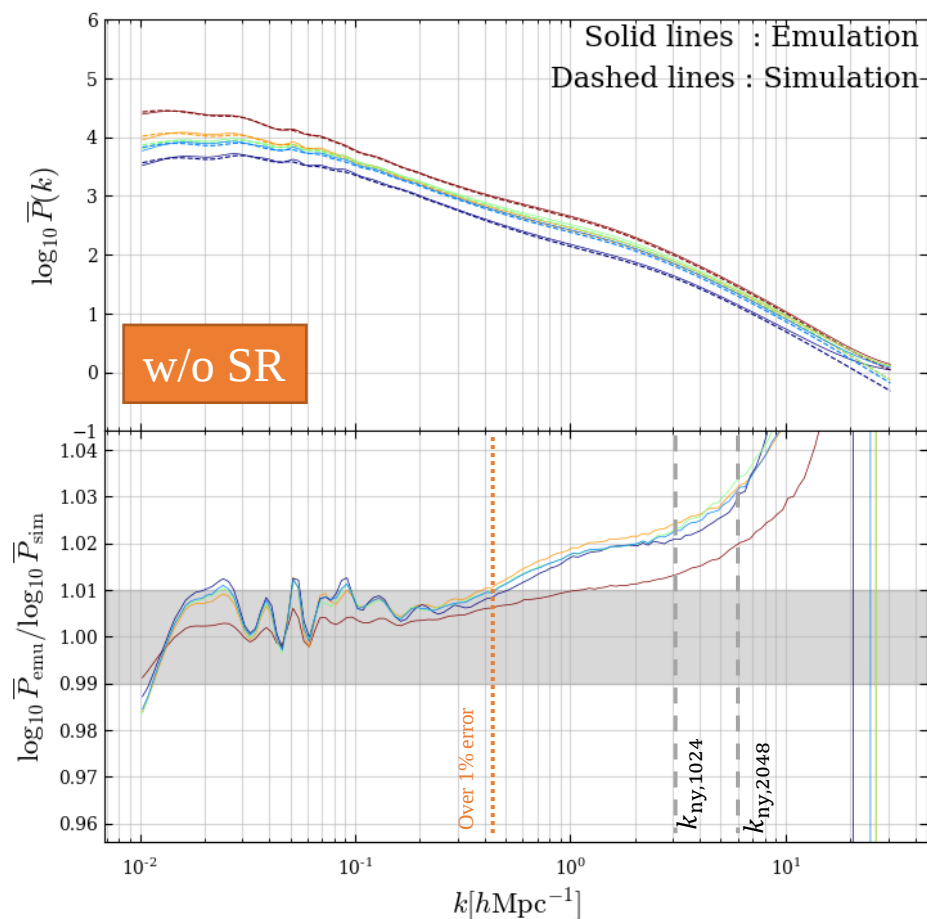


↑ $\log P(k)$ が0付近の場所では相対誤差が定義できないことも影響

- ✓ 2048³ 粒子、1Gpc/hの中解像度データを追加
- ✓ randomのテストセットに対してはまだエラーが大きいので省略

Power Spectrum for SR emulation

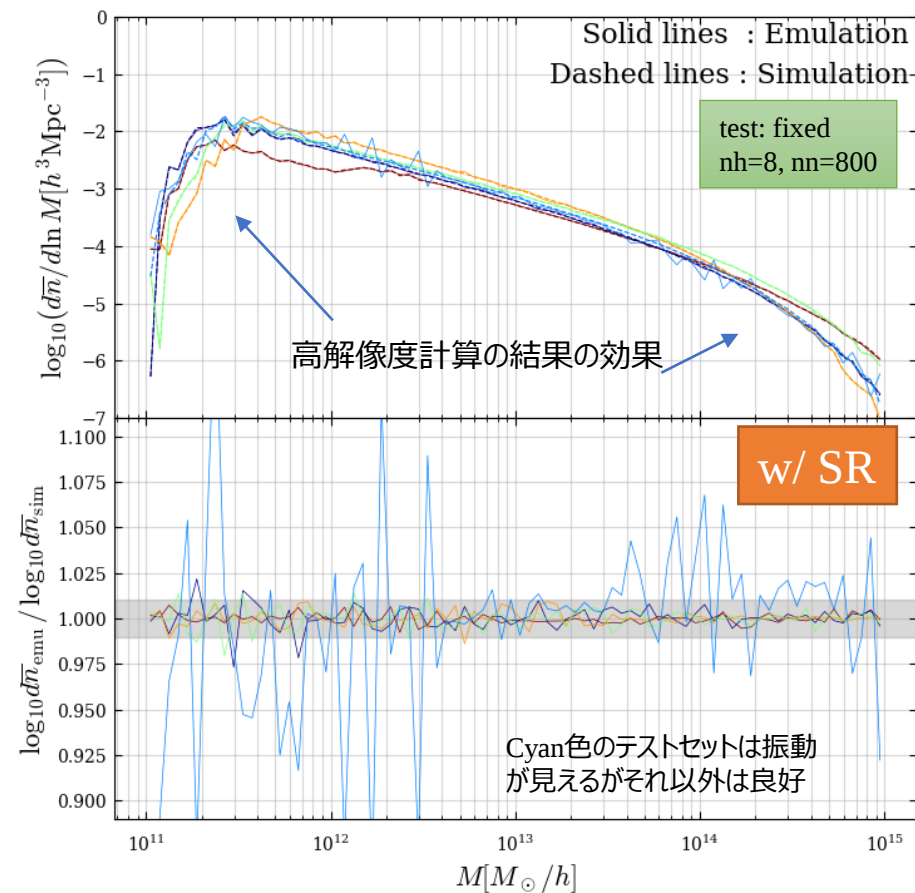
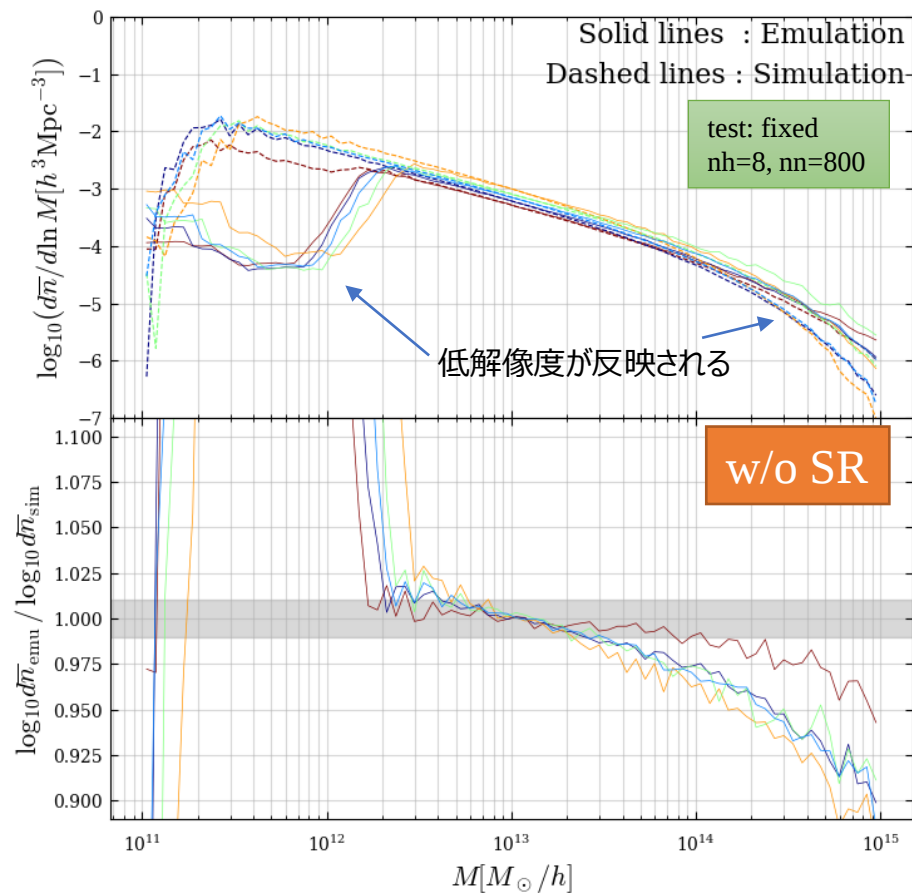
低解像度2000個のfixed and random seed
中解像度7個の fixed seed, 30個の random seedで学習



- ✓ 中解像度 random seed の学習データの質がよくないのでノイズとして振る舞いエラーが大きくなってしまった
 - ✓ 3個の宇宙論パラメータの10 realization の30個を追加した形

Halo Mass Function for SR emulation

低解像度2000個のfixed and random seed
中解像度7個のfixed seed, 30個のrandom seedで学習



- ✓ Power spectrumに比べてSuper-Resolutionの恩恵が得られた
- ✓ randomのテストセットに対してはまだエラーが大きいので省略

Summary & Future

□ **Dark Quest II エミュレータ構築**に向けて大規模低解像度データセットを構築した

- 1000パラメータの学習データセットを構築することによりサブパーセントレベルでパワースペクトルやハロー質量関数の推定可能が可能になった
 - ✓ パワースペクトルの小スケールのエラーがまだ課題
 - ✓ ハロー質量は適用範囲に注意をすれば十分適用可能
- 初期の密度ゆらぎのランダムさを考慮することによりcosmic varianceに対応させた学習
 - ✓ 同じ宇宙論パラメータに対して10 realizationの平均を取るという作業が必要
- 中高解像度シミュレーションの情報を追加することにより**Super-resolution**がそこそこ成功
 - ✓ ランダムseedテストセットに対する学習がまだ課題

□ 今後

- 中、高解像度シミュレーションの学習データの数と質の向上
- 3次元密度場レベルの情報を活用