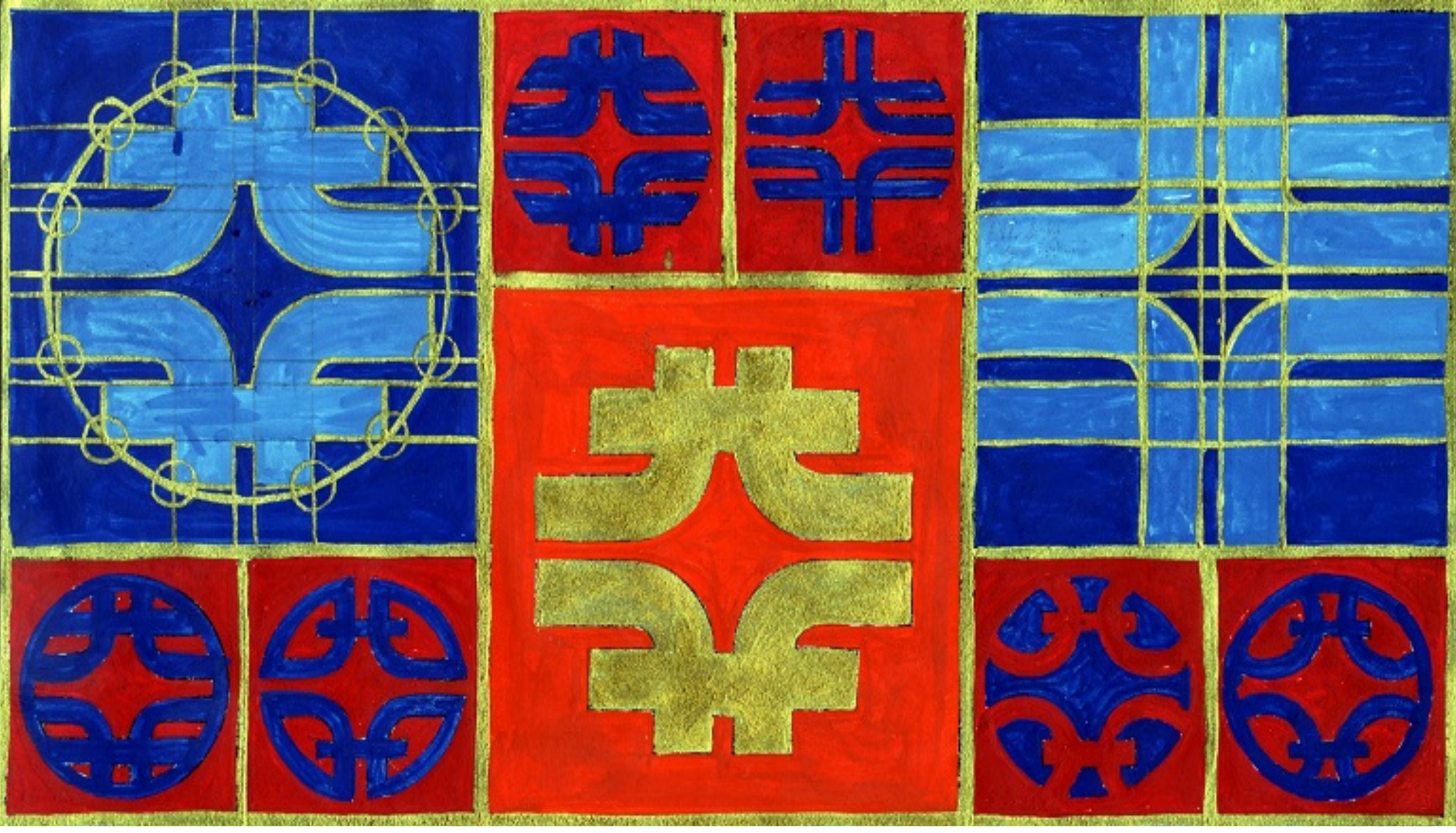


Bayesian Model Averaging

William I. Jay (Fermilab)

JLab Theory Seminary — 21 Sept 2020

Statistics > Methodology

[Submitted on 3 Aug 2020 (v1), last revised 24 Aug 2020 (this version, v2)]

Bayesian model averaging for analysis of lattice field theory results

William I. Jay, Ethan T. Neil

Statistical modeling is a key component in the extraction of physical results from lattice field theory calculations. Although the general models used are often strongly motivated by physics, their precise form is typically ill-determined, and many model variations can be plausibly considered for the same lattice data. Model averaging, which amounts to a probability-weighted average over all model variations, can incorporate systematic errors associated with model choice without being overly conservative. We discuss the framework of model averaging from the perspective of Bayesian statistics, and give useful formulas and approximations for the particular case of least-squares fitting, commonly used in modeling lattice results. In addition, we frame the common problem of data subset selection (e.g. choice of minimum and maximum time separation for fitting a two-point correlation function) as a model selection problem, and study model averaging as a straightforward alternative to manual selection of fit ranges. Numerical examples involving both mock and real lattice data are given.

Comments: 26 pages, 6 figures. v2: updated refs and other minor changes. Submitted to Phys. Rev. D
Subjects: **Methodology (stat.ME)**; High Energy Physics - Lattice (hep-lat)
Report number: FERMILAB-PUB-20-374-T
Cite as: arXiv:2008.01069 **[stat.ME]**
(or arXiv:2008.01069v2 **[stat.ME]** for this version)

Link: 2008.01069

Bayesian Model Averaging

William I. Jay (Fermilab)

JLab Theory Seminary — 21 Sept 2020

Motivation

- Statistical inference appears ubiquitously throughout science: Which model best describes the data?

- Physics often provides well-motivated theoretical models for data (this is *very* special in practical applications of statistics).

- Reliable ***parameter extraction*** is often more important than ***model comparison*** by itself.

  - Frequently the exact microscopic model is known exactly but too cumbersome to use in practice

  - Sometimes an exact model is absent, but phenomenological considerations suggest a class of models

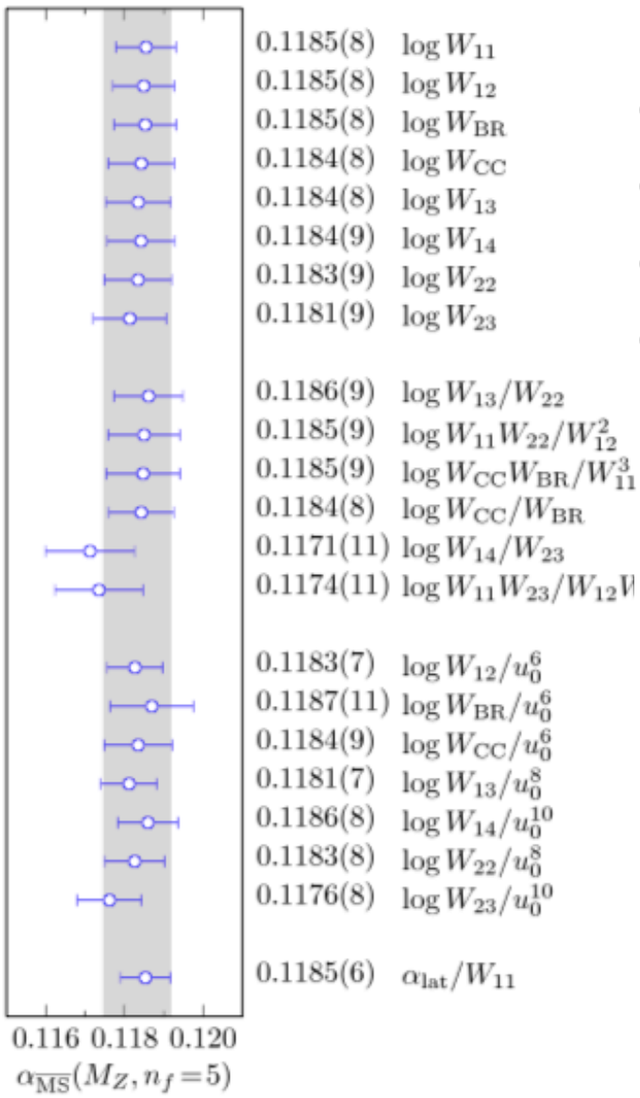
Motivation

- Most of my examples will be from numerical lattice QCD, but the statistical ideas are completely general.

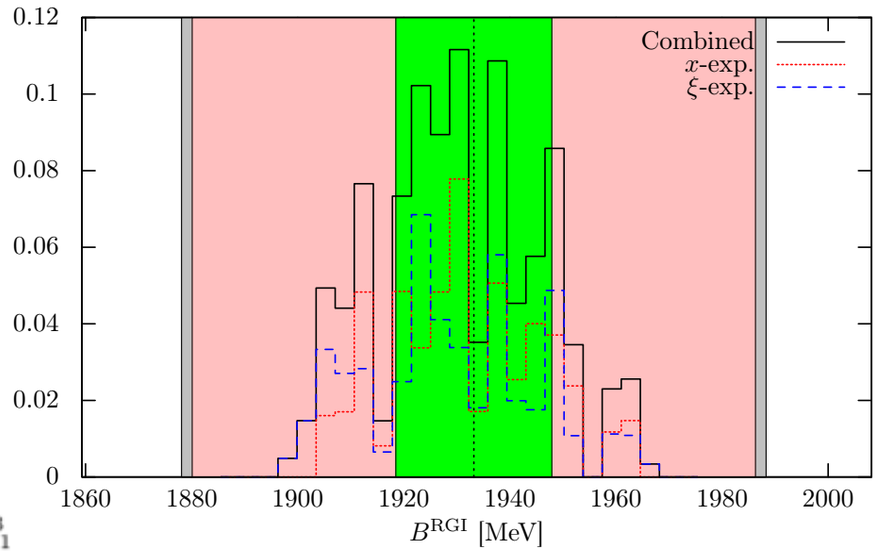
- Examples of model selection in lattice QCD:
  - Is SU(2) or SU(3) χPT more applicable?
  - Does NLO vs NNLO work better?
  - How to handle exact models with infinitely many terms?
- What is the systematic effect of the choice or truncation?
- Is some choice optimal?

Model averaging in the literature

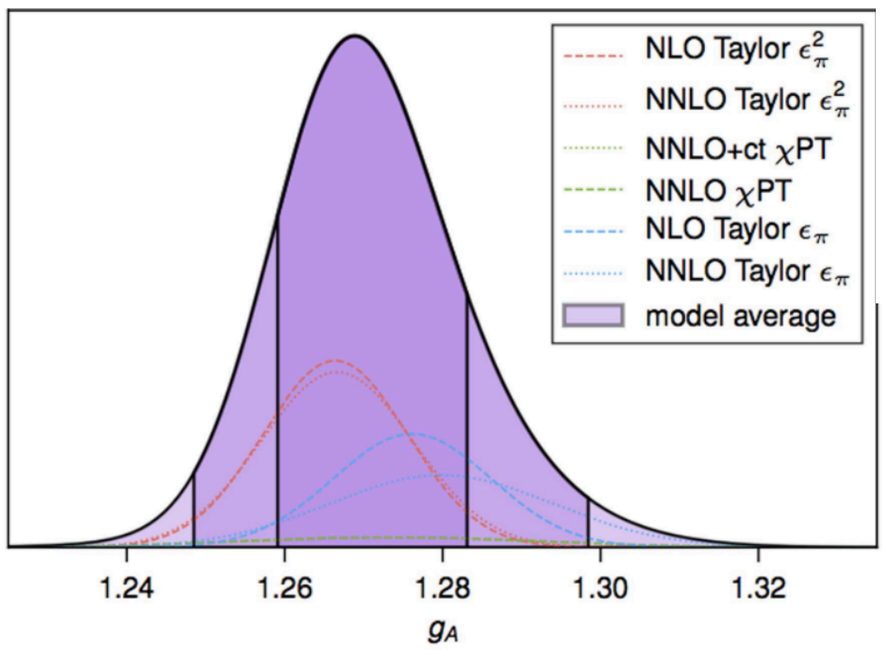
HPQCD: 0807.1687



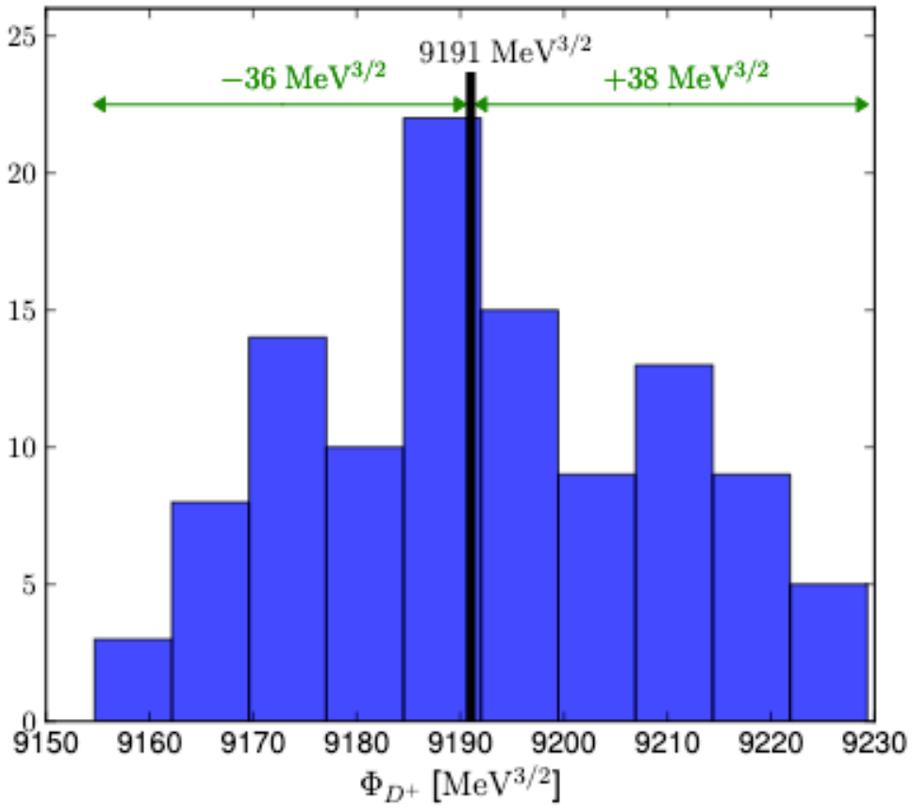
BMW:1310.3626



CalLat: 1805.12130

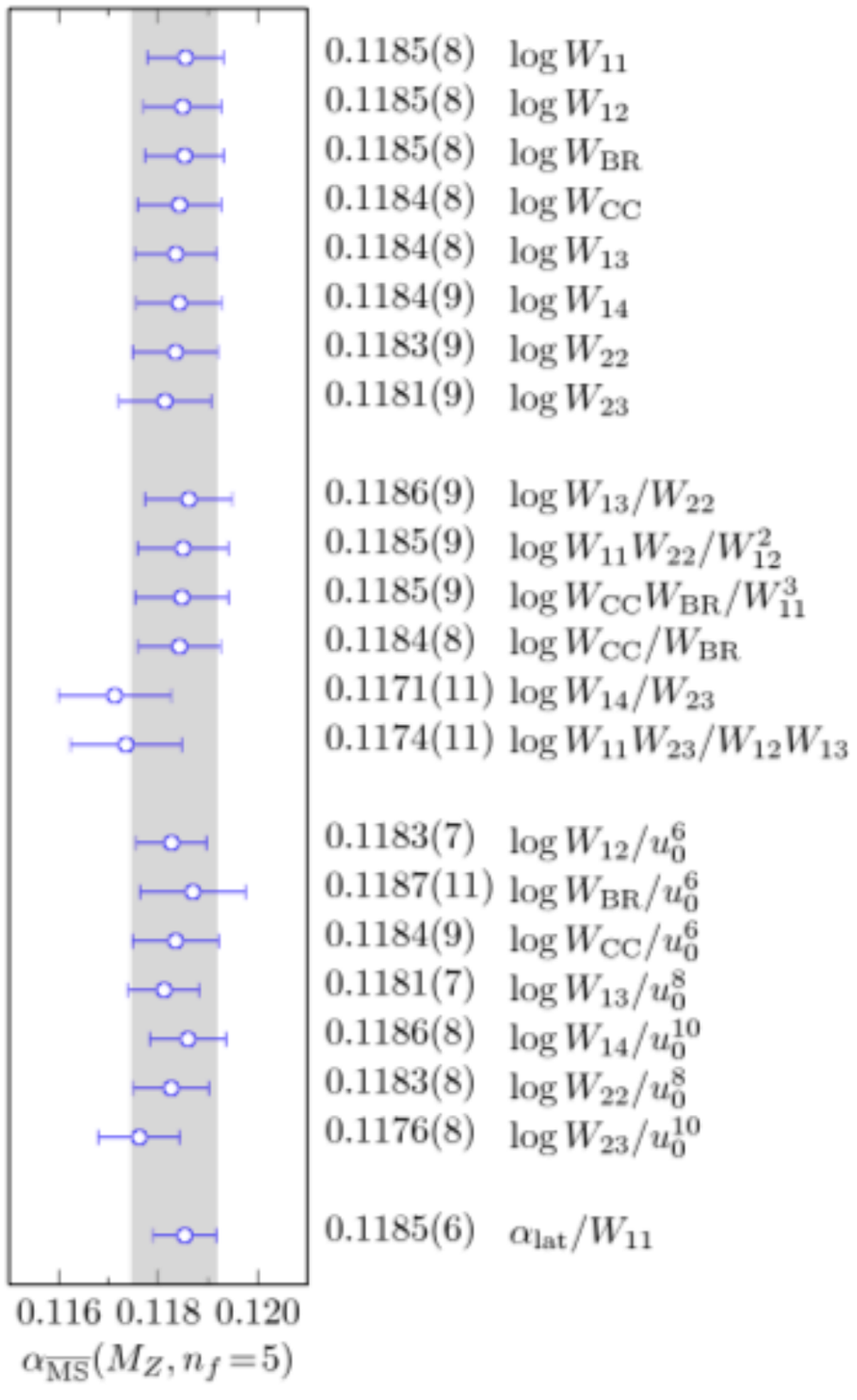


FNAL/MILC:1407.3772



Just a few examples—not an exhaustive list !

Model averaging in the literature



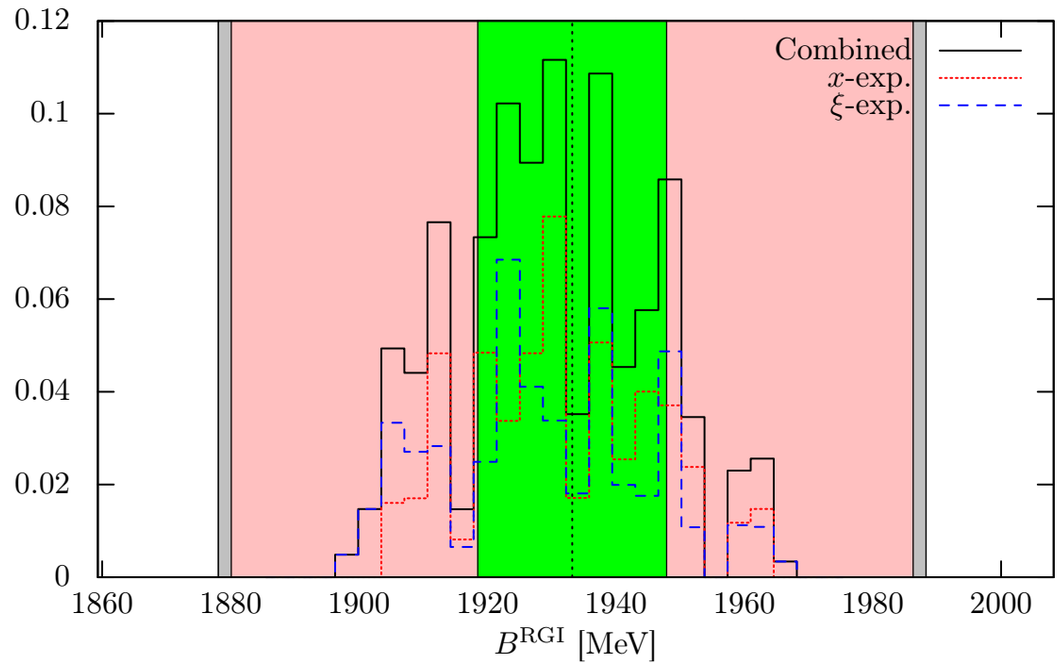
Weighted averages:
Use the variance as weight

HPQCD, arXiv:0807.1687

"We average our 22 different results weighted by their inverse variances, giving more weight to results with smaller variances. The variance for our composite result is the inverse of the average of the inverse variances from the separate determinations."

Model averaging in the literature

Combining distributions:
Total distribution gives mean and error



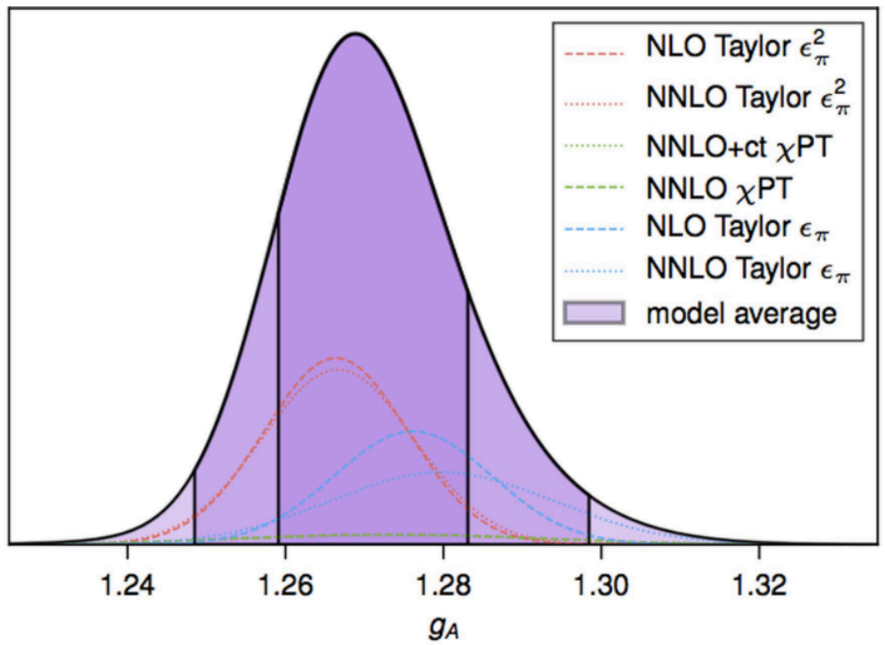
BMW:1310.3626

Figure 12: "Systematic error distributions for the LO LEC… *In the plots, the central, vertical, dotted line is the mean of the total distribution, i.e. our final central value.* The central, vertical green band denotes the systematic error, the larger pink one, the statistical error and the largest gray one, the sum in quadrature of these two errors."

[Emphasis added]

Model averaging in the literature

A Bayesian framework:
Combine results using
Bayesian model weights



(Close to what we advocate)

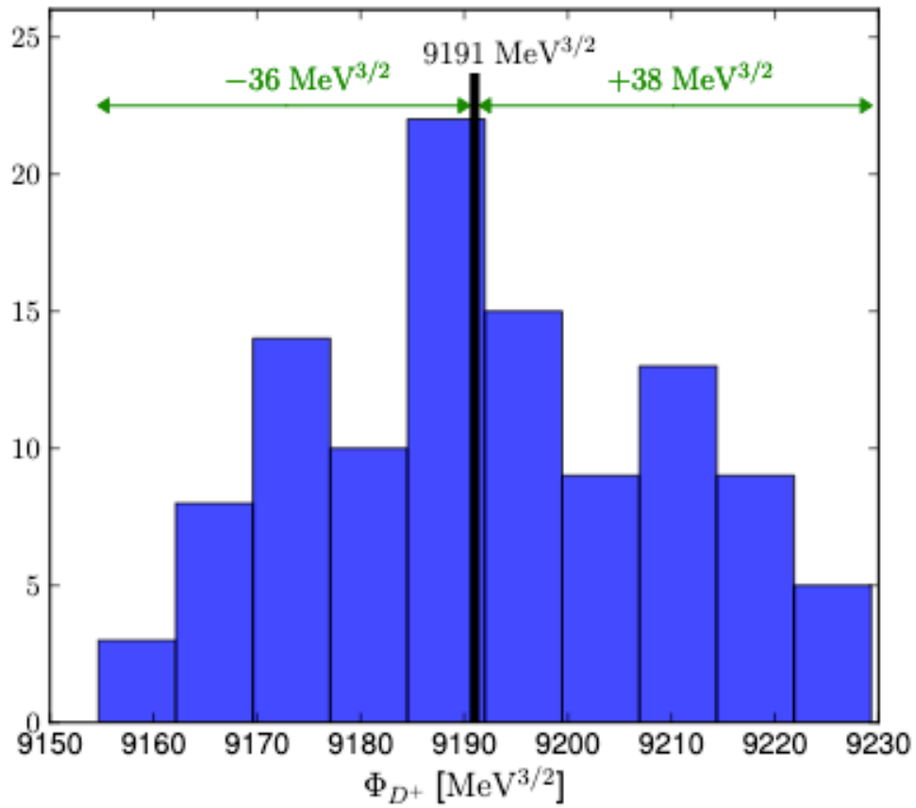
CalLat: 1805.12130

p 20. “The weighted average is performed with **`lsqfit`**.”

(Fantastic fitting code, but default weights are not quite right for model averaging)

Model averaging in the literature

A conservative estimate:
Systematic error from full width of model variations.



FNAL/MILC:1407.3772

FIG. 19: “Histograms of … values obtained from various versions of the continuum/chiral extrapolation and various inputs of quark masses and scale values from the physical- mass analysis. Our central fit gives 9191 MeV$^{3/2}$…those values are marked with vertical black lines. At the top of each histogram, we show the range taken as the systematic error of the self-contained chiral analysis of the current section.

~~Model averaging~~ in the literature

Model selection

Picking fits

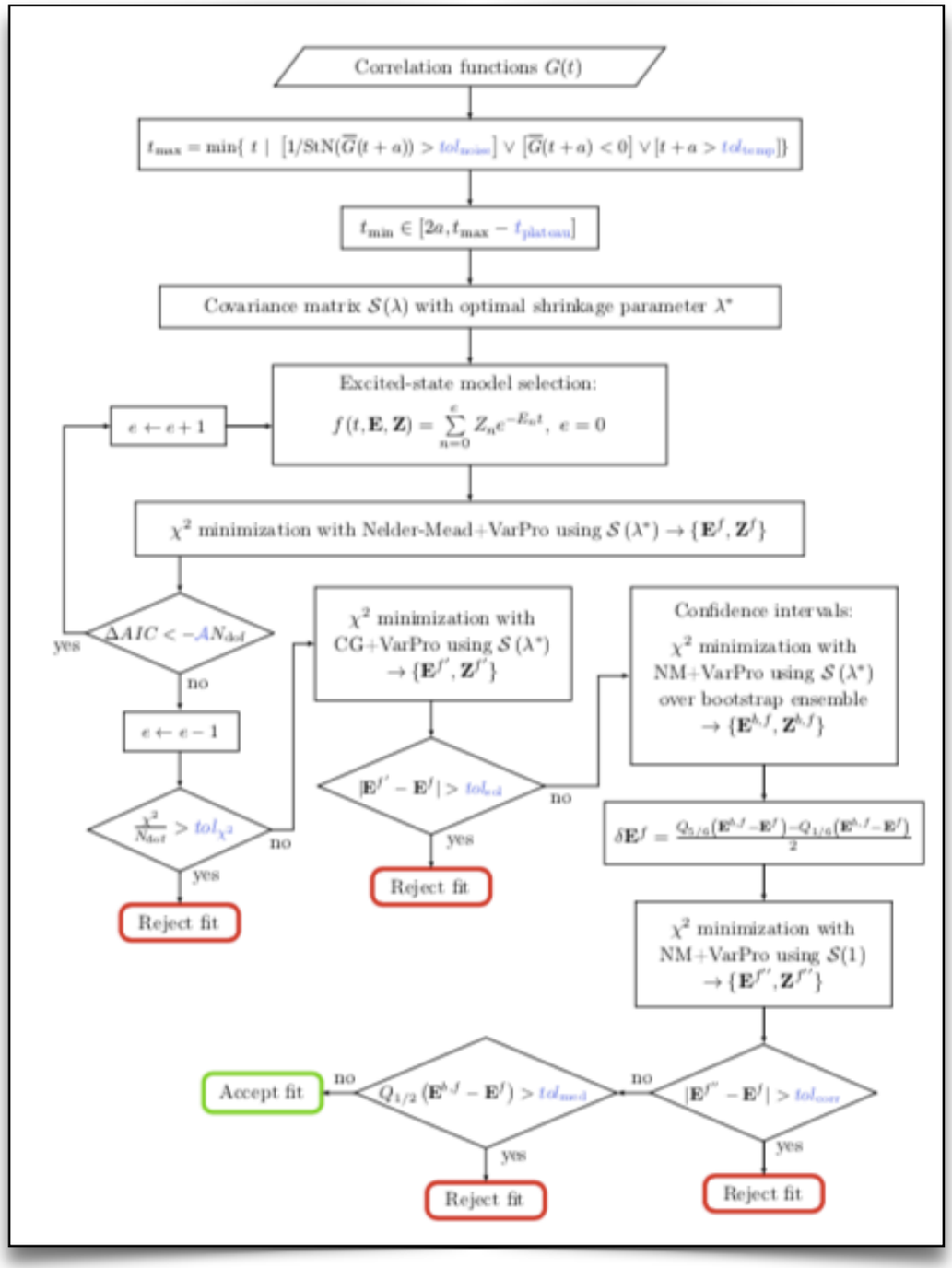
Use the AIC to choose a best fit.

Closely related to what we describe

NPLQCD & QCDSF:2003.12130

“This work employs the Akaike Information Criterion (AIC) with a cutoff chosen to penalize overfitting in which a fit with e excited states is only preferred over a fit with e – 1 excited states if

$$AIC(e) - AIC(e-1) < -A \times N_{DOF}(e)$$



Bayesian model averaging

- Existing approaches to model averaging are often *ad hoc.*
- Bayesian statistics provides a rigorous perspective to this question
- Bayesian model averaging weights models according to their statistical probability. Benefits:
  - Offers better precision than, e.g., just taking the full range of variations
  - Removes dependence on arbitrary decisions by the analyst
  - Helps clarify statistical assumptions related to existing methods

Example:

Lattice QCD hadron spectroscopy

Ex: Lattice QCD hadron spectroscopy

- Hadronic spectrum ⟷ QCD 2pt correlation functions

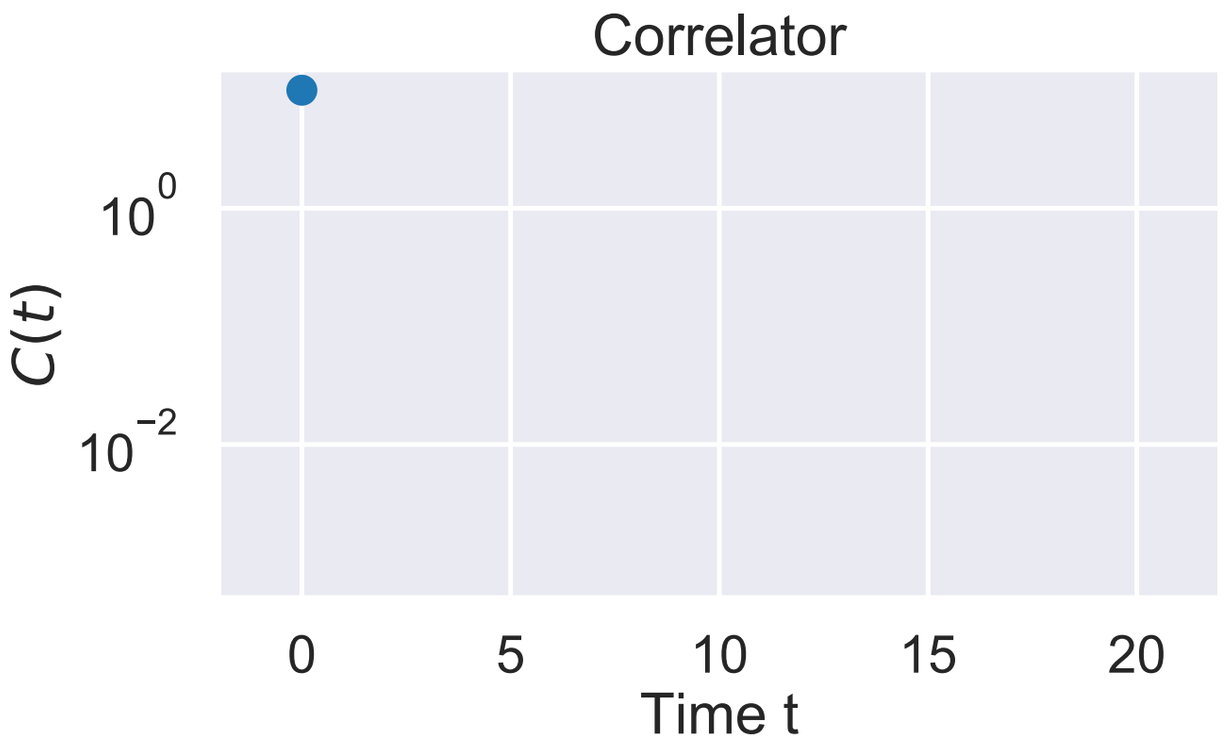
$$\begin{aligned}
\langle O(t)O(0)\rangle &= \langle 0|\, e^{Ht} O(0) e^{-Ht} O(0)\,|0\rangle \\
&= \sum_n e^{-E_n t}\, \langle 0|\, O(0)\,|n\rangle\, \langle n|\, O(0)\,|0\rangle \\
&= \sum_n e^{-E_n t}\, \left|\langle 0|\, O(0)\,|n\rangle\right|^2 \\
&= \sum_n |Z_n|^2 e^{-E_n t}
\end{aligned}$$

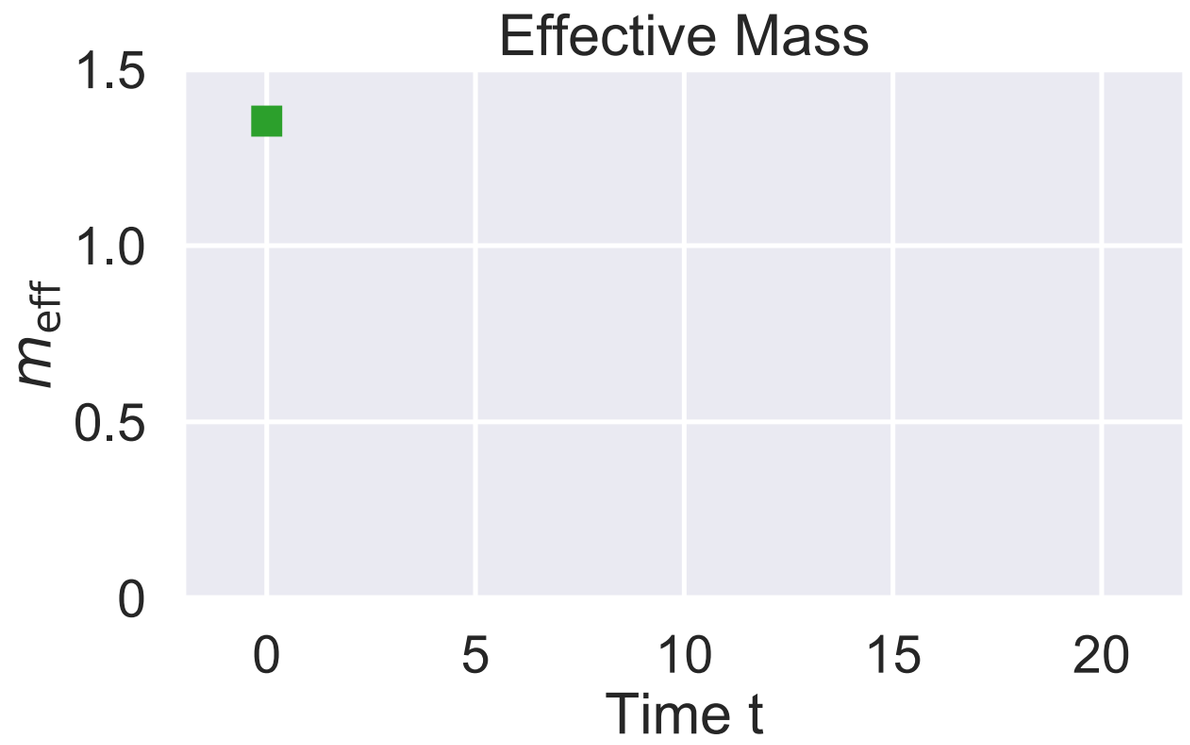
"The operators couple to an infinite tower of states."

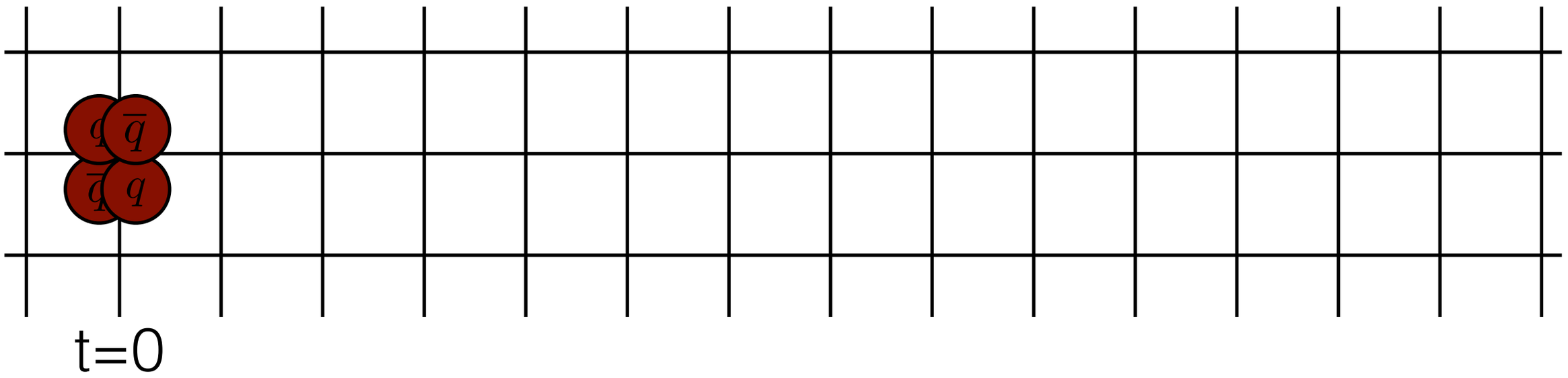
$$m_{\text{eff}} = \log \frac{C(t)}{C(t+1)} \overset{t\to\infty}{\sim} m_0$$

"The ground-state mass asymptotically dominates the Euclidean 2pt function."

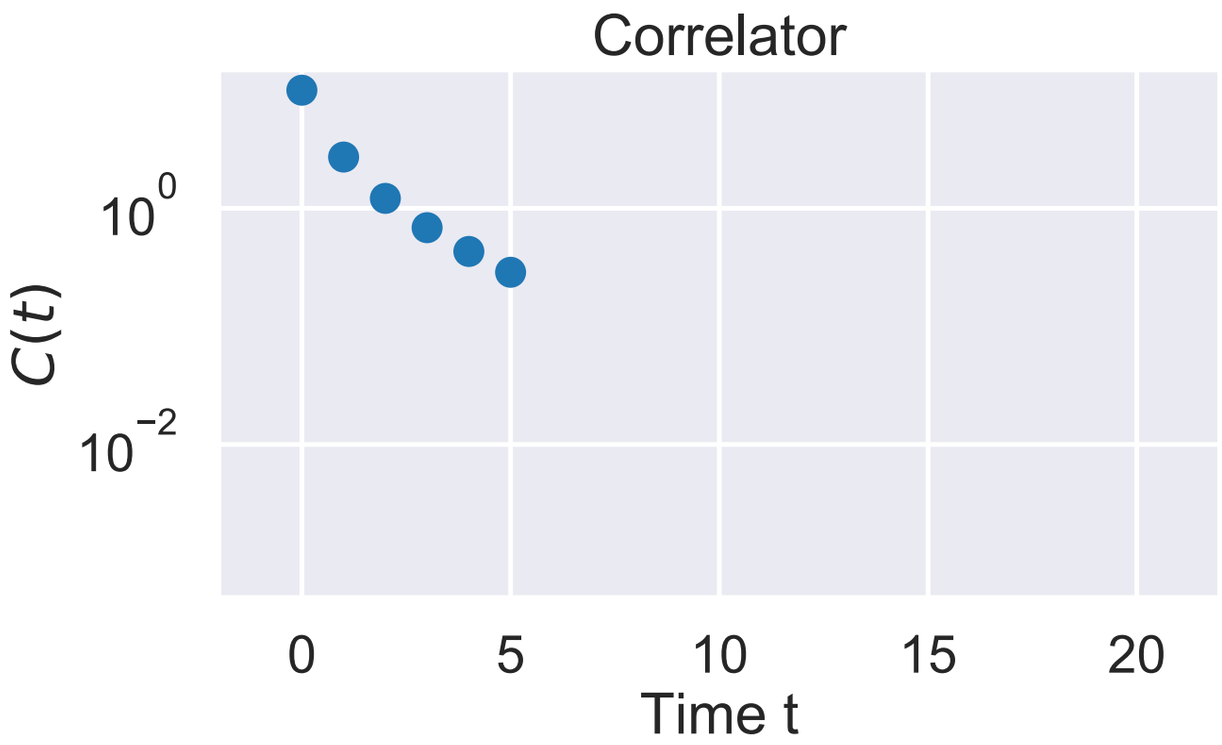
Ex: Lattice QCD hadron spectroscopy

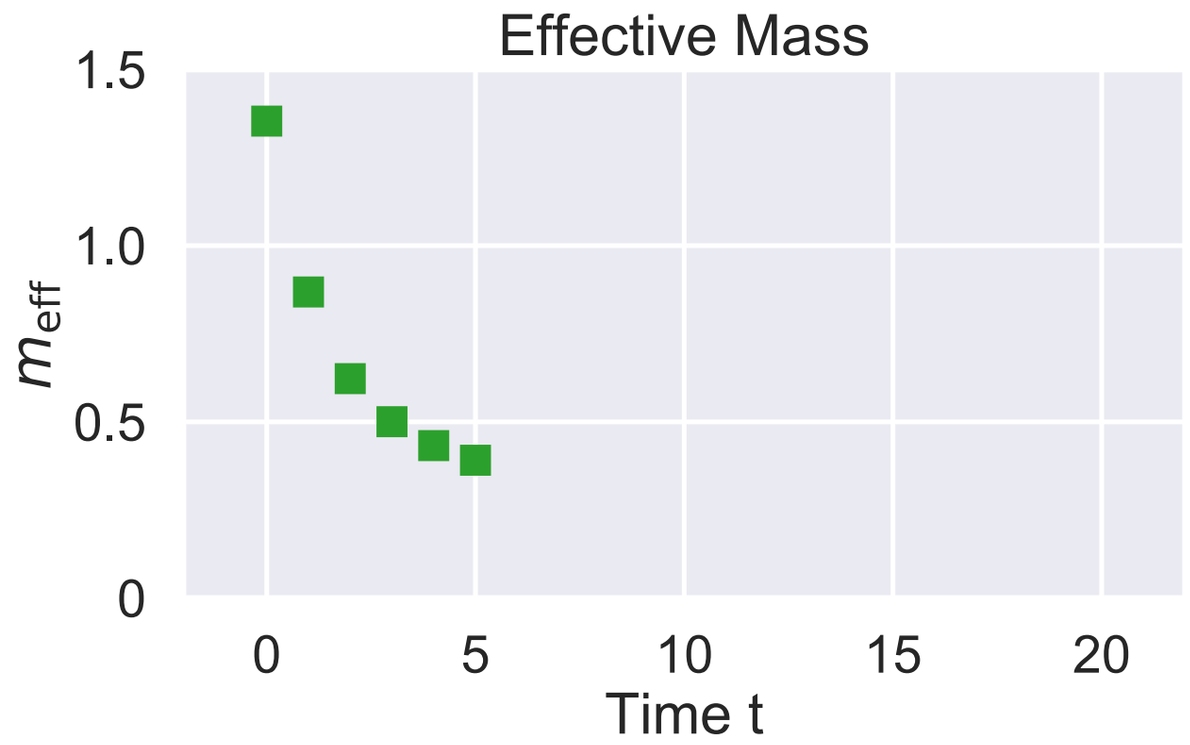


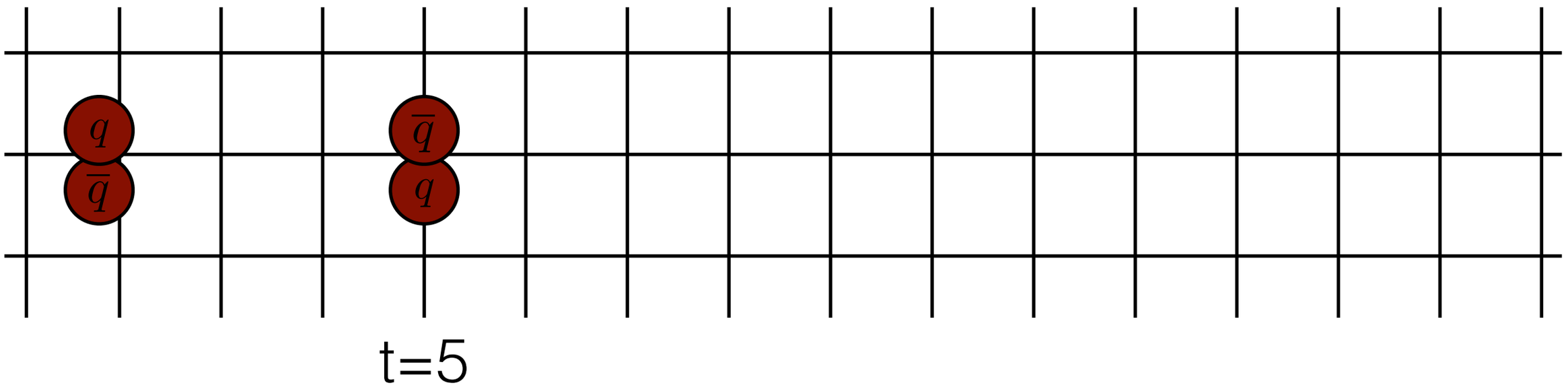




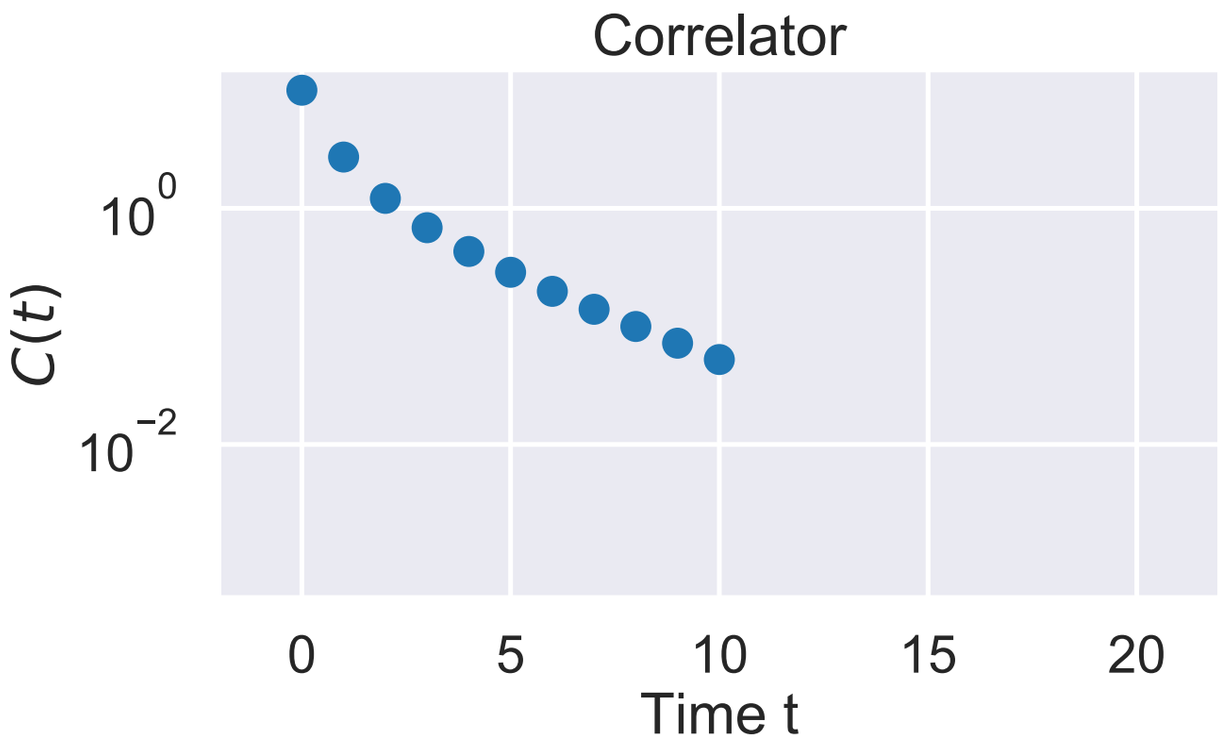
Ex: Lattice QCD hadron spectroscopy

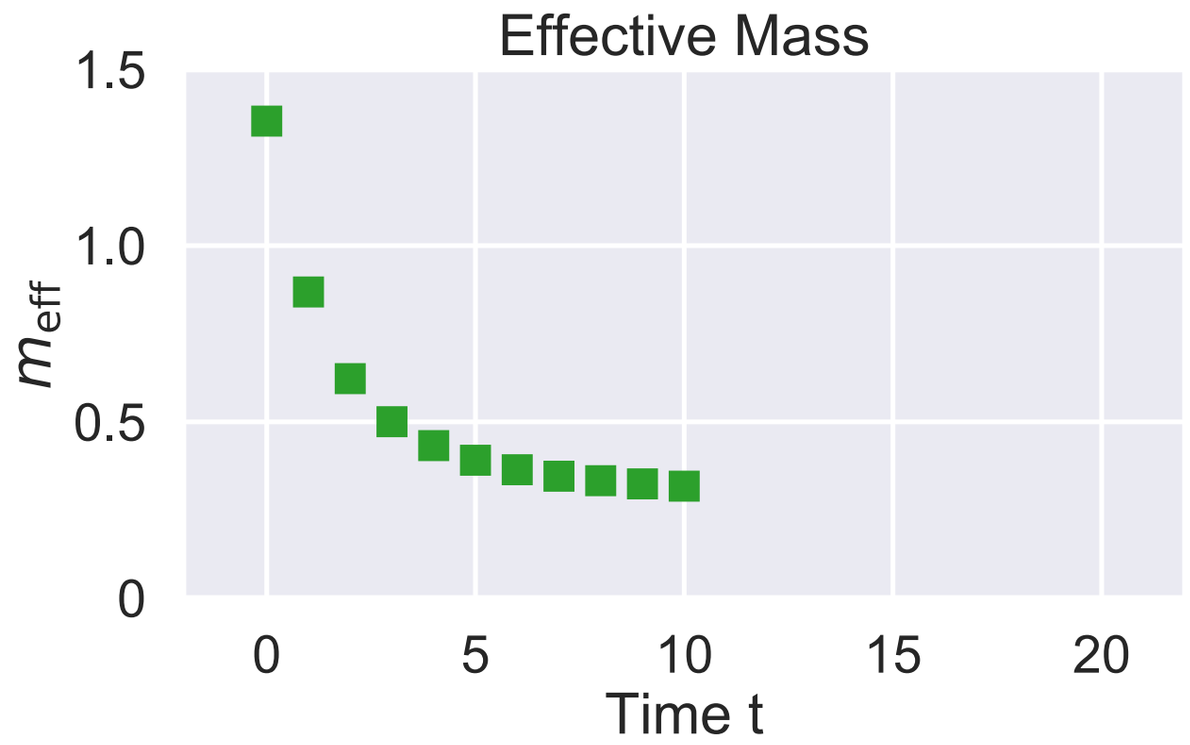


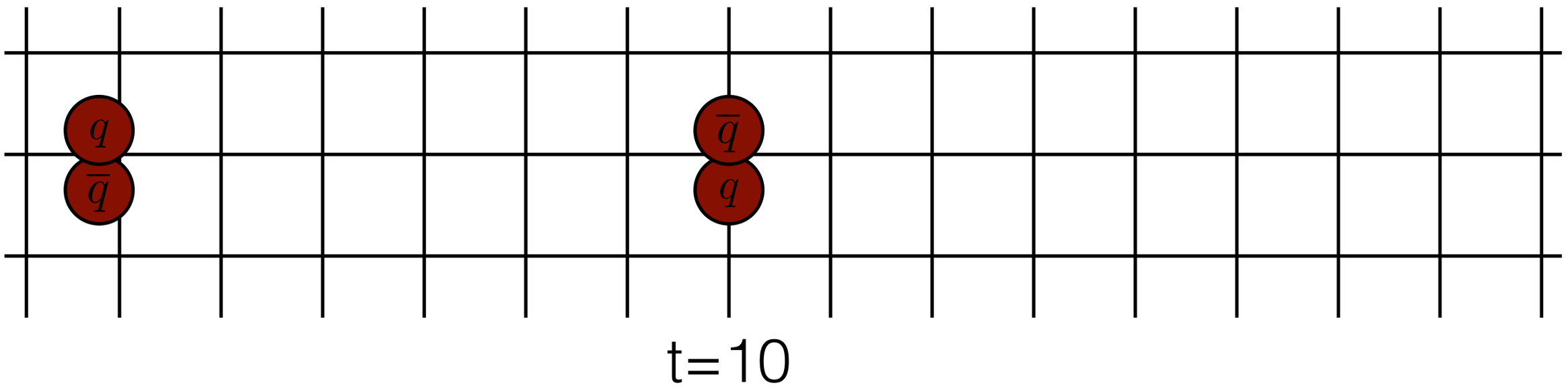




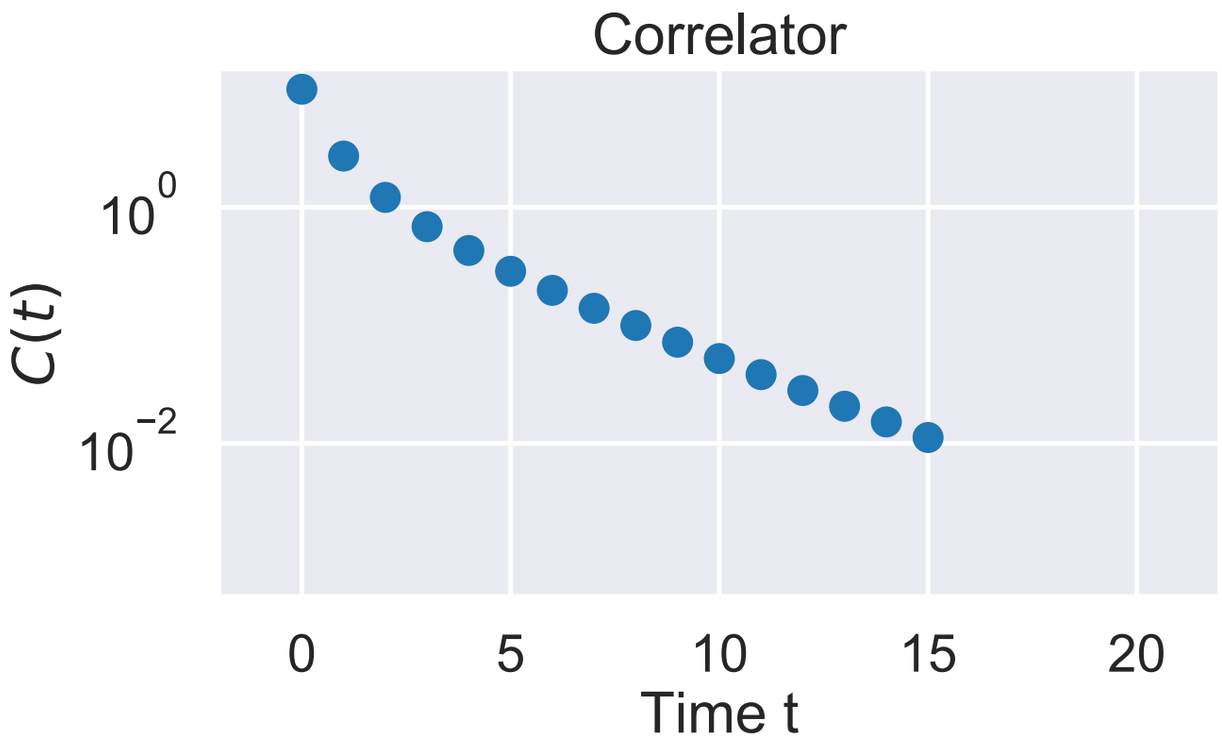
Ex: Lattice QCD hadron spectroscopy

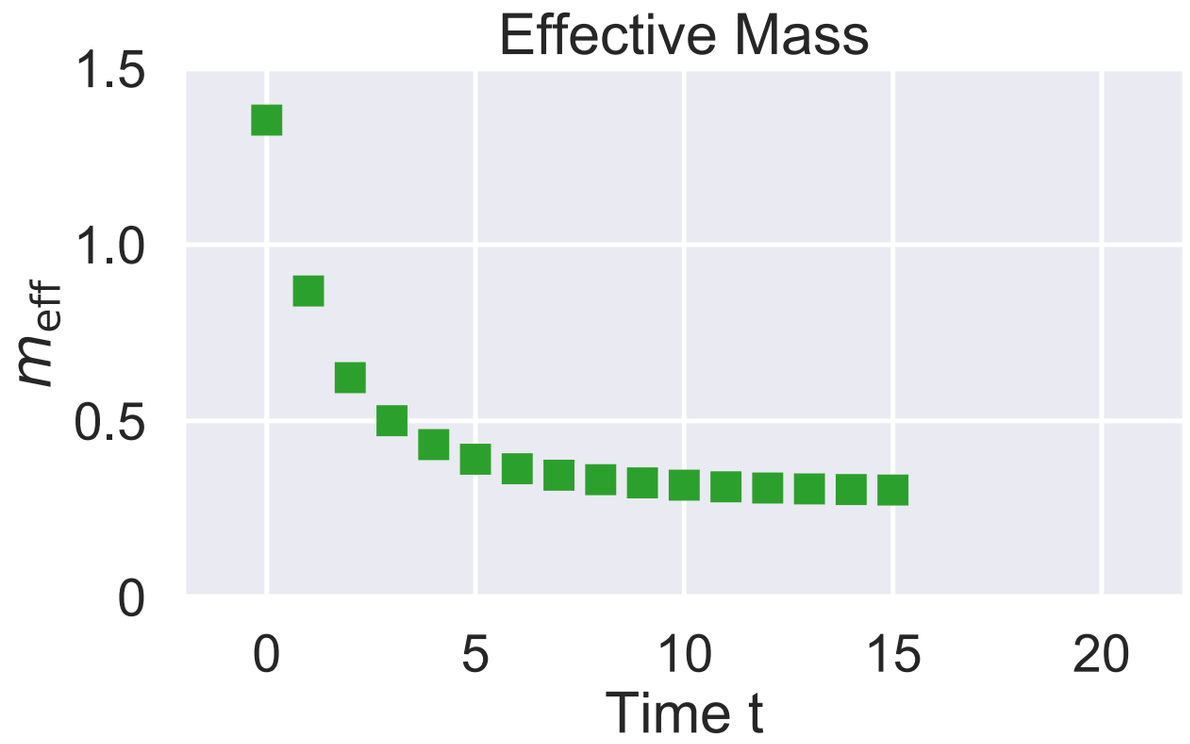


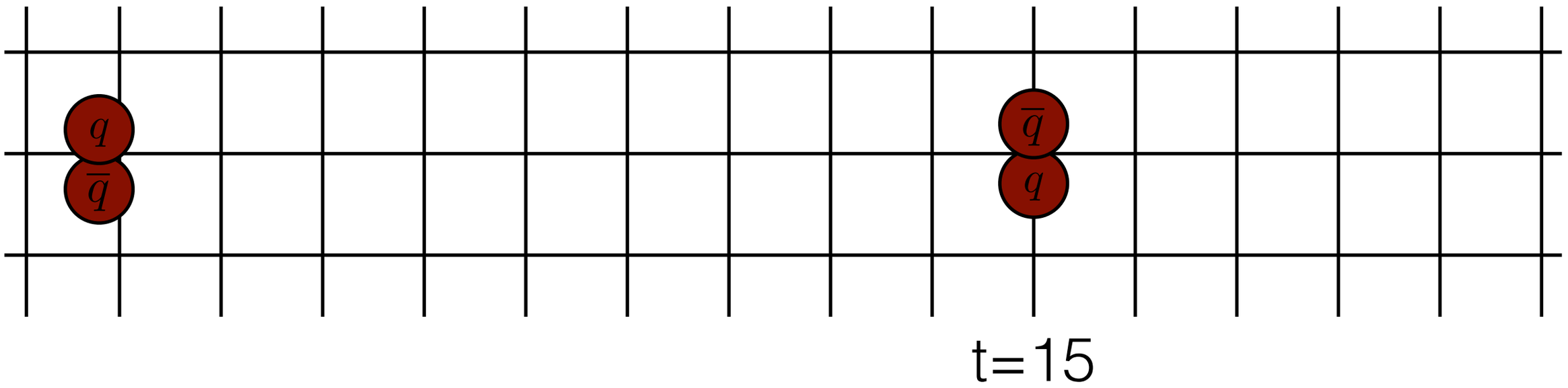




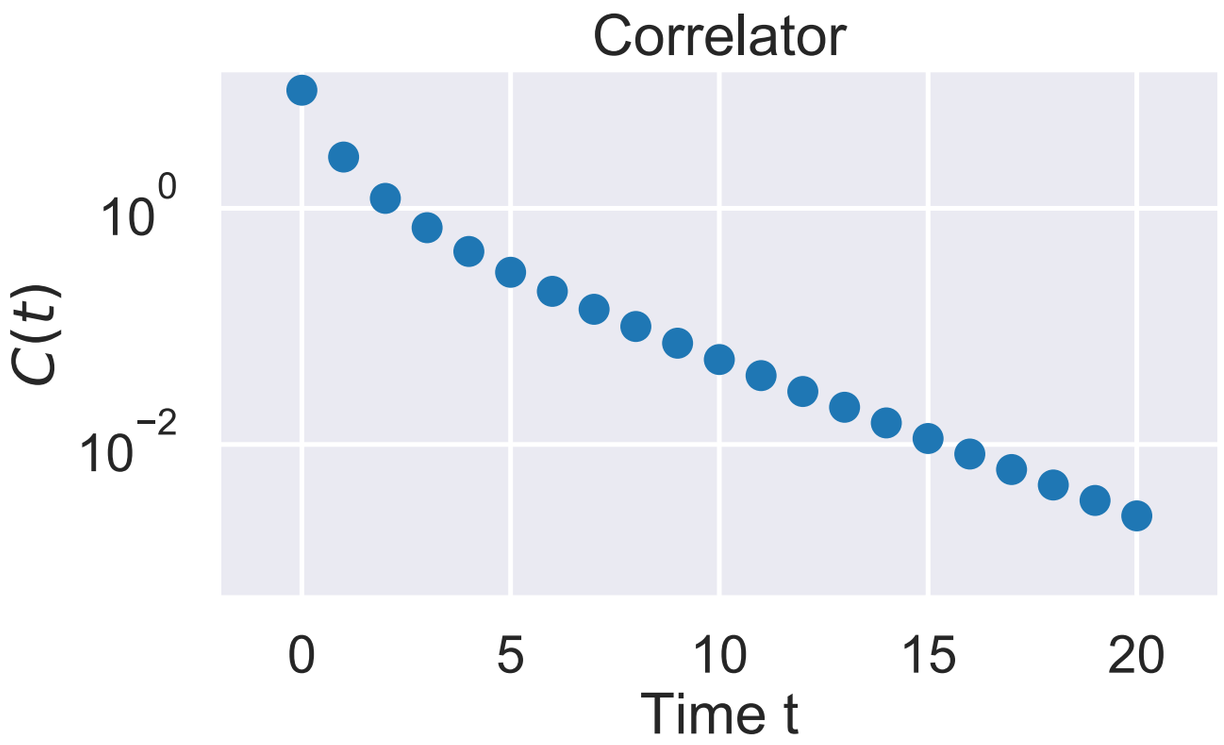
Ex: Lattice QCD hadron spectroscopy

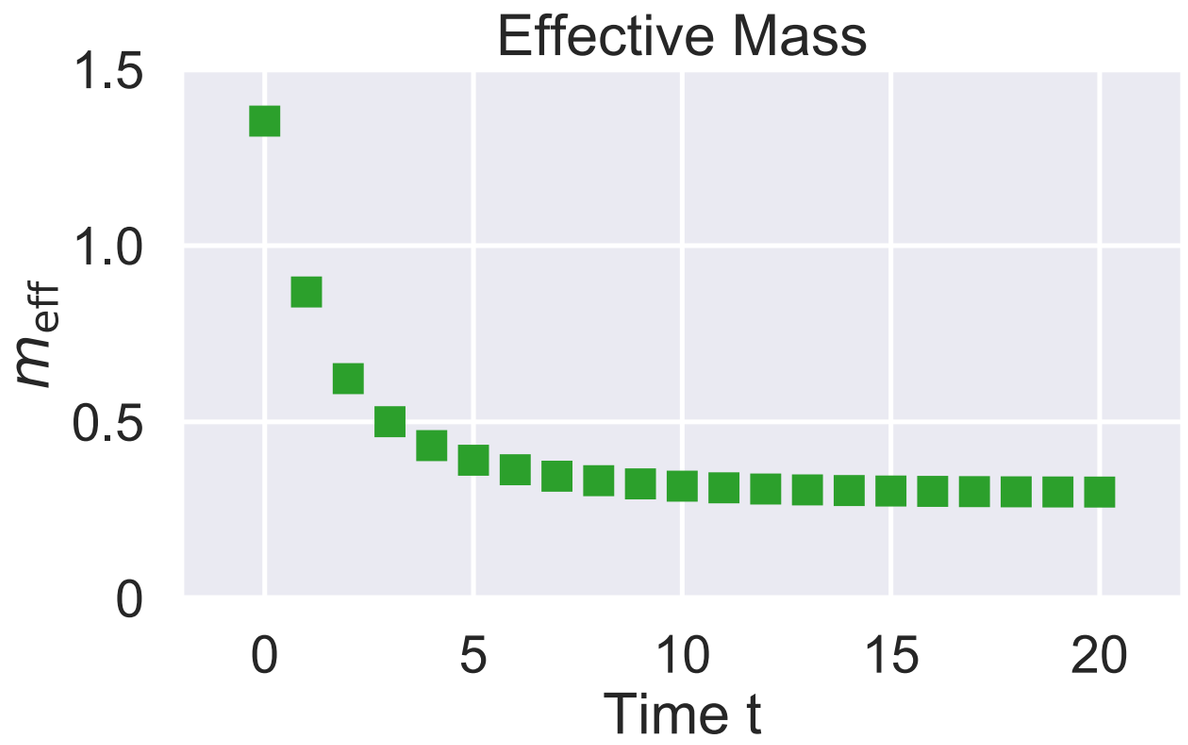


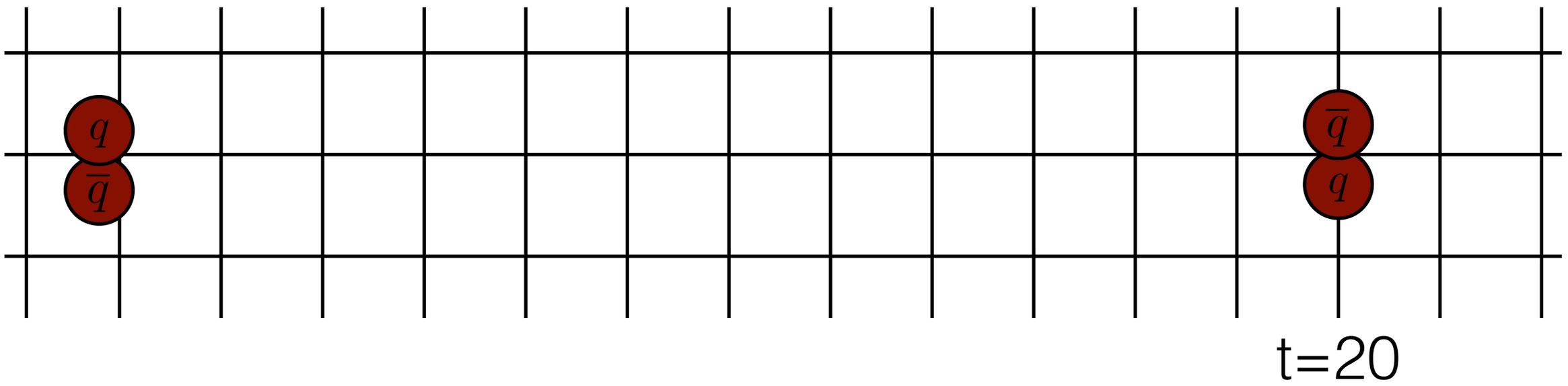




Ex: Lattice QCD hadron spectroscopy







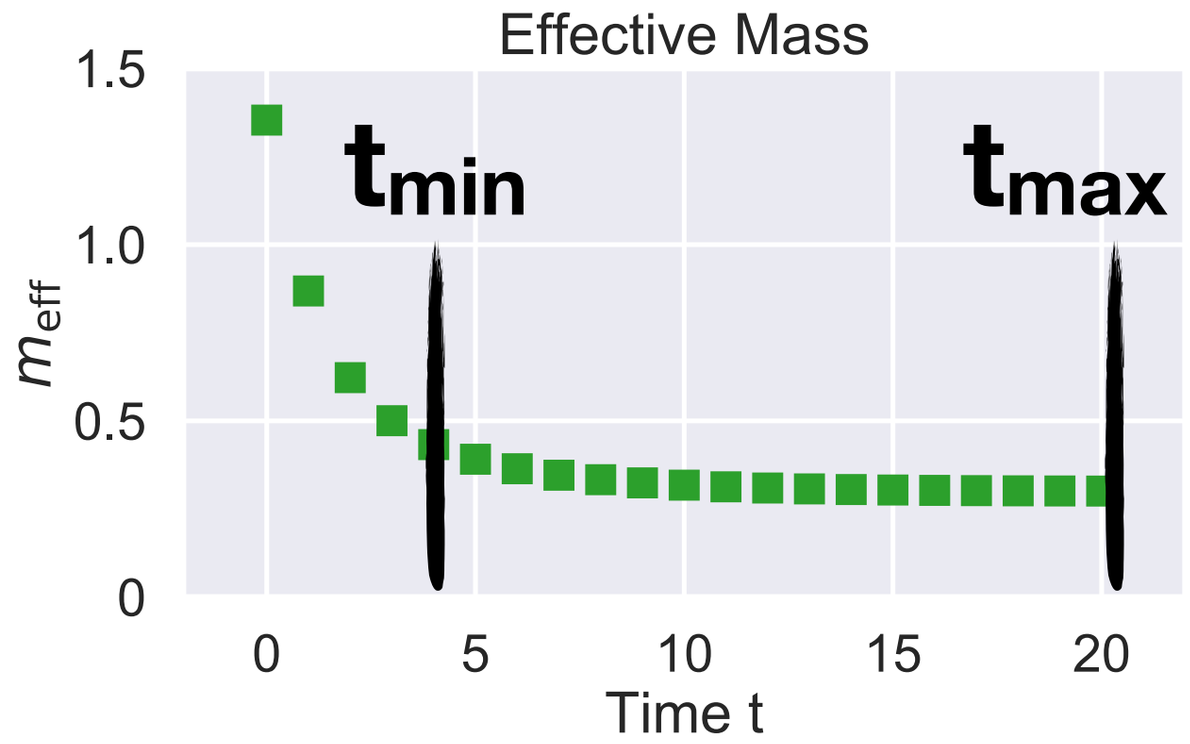
Ex: Lattice QCD hadron spectroscopy

Analysis choices:

1. Fit window (t_{min}, t_{max})?

   - t_{min} too high: throwing away valuable data (signal decays exponentially in time!)

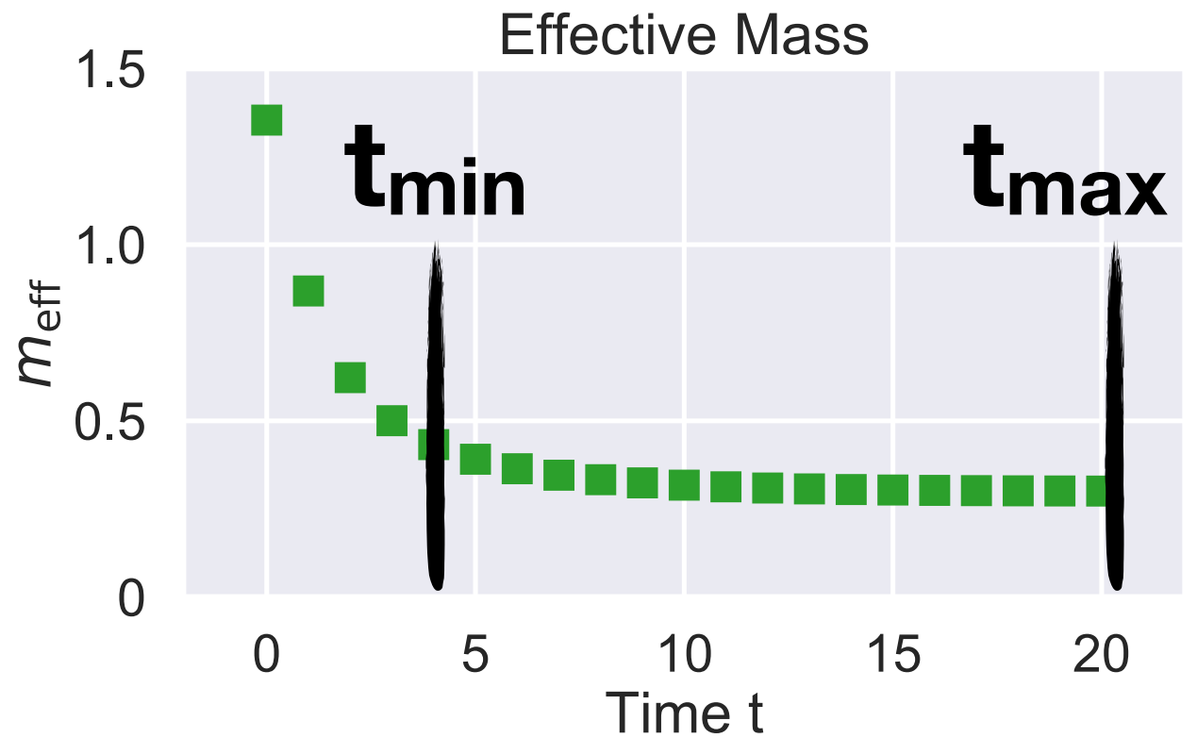
   - t_{min} too low: contamination from excited states

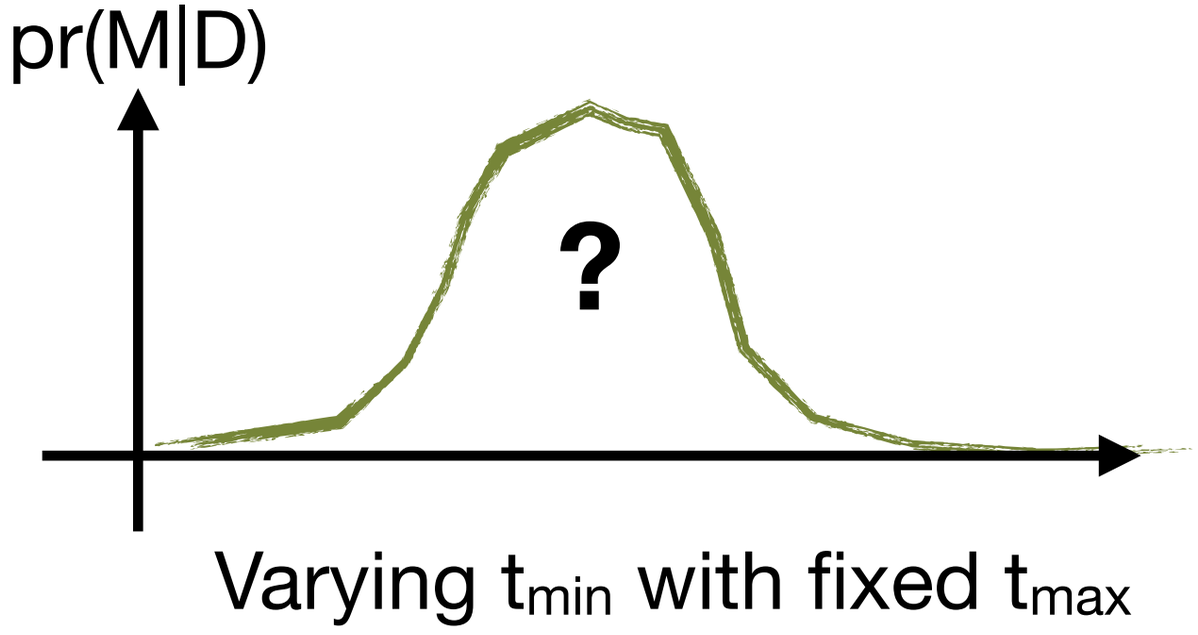


2. How many states to include in the truncated model for C(t)?

Ex: Lattice QCD hadron spectroscopy

- Intuition: “Occam’s razor”
- “Good” models should describe the most data with the fewest parameters
- Further, it should be possible to “marginalize” or “average” over different results to obtain a rigorous bound on the systematic error from “model choice.”
- Might expect:





Formalizing these ideas with Bayesian statistics

Bayesian model averaging

- Consider data **D**, models {**M**}, model parameters **a**

- Group parameters as $\mathbf{a} = \mathbf{a_c} \cup \mathbf{a_m}$ with $\mathbf{a_c}$ as "common parameters" and $\mathbf{a_m}$ as "marginalized parameters"

- We want to average over the space of models

$$\langle f(\mathbf{a}_c) \rangle = \sum_M \langle f(\mathbf{a}_c) \rangle_M \, \mathrm{pr}(M|D)$$

- Key: pr(M|D) is the *model weight* or "*Bayes factor*":

$$\mathrm{pr}(M|D) = \int d\mathbf{a} \, \frac{\mathrm{pr}(D|\mathbf{a}, M)\mathrm{pr}(\mathbf{a}|M)\mathrm{pr}(M)}{\mathrm{pr}(D)}$$

Bayesian model averaging

Model average for single parameter a_0

$$\langle a_0 \rangle = \sum_M \langle a_0 \rangle_M \, \mathrm{pr}(M|D)$$ Central value

$$\sigma^2_{a_0} = \left\langle a_0^2 \right\rangle - \langle a_0 \rangle^2$$ Combined statistical+systematic error

$$= \sum_{i=1}^{N} \left\langle a_0^2 \right\rangle_i \mathrm{pr}(M_i|D) - \left(\sum_{i=1}^{N} \langle a_0 \rangle_i \, \mathrm{pr}(M_i|D) \right)^2$$

$$= \sum_{i=1}^{N} \sigma^2_{a_0,i} \mathrm{pr}(M_i|D) + \sum_{i=1}^{N} \langle a_0 \rangle_i^2 \, \mathrm{pr}(M_i|D) - \left(\sum_{i=1}^{N} \langle a_0 \rangle_i \;\; \mathrm{pr}(M_i|D) \right)^2$$

Bayesian model averaging

Model average for single parameter a_0

Statistical error:
Weighted average of statistical variance

Systematic error:
(a) Special case of uniform weights
→ "variance over space of models"
(b) General case: *not* variance from weighted estimates

$$\sigma_{a_0} = \langle a_0^2 \rangle - \langle a_0 \rangle$$

$$= \sum_{i=1}^{N} \langle a_0^2 \rangle_i \, \mathrm{pr}(M_i|D) - \left(\sum_{i=1}^{N} \langle a_0 \rangle_i \, \mathrm{pr}(M_i|D) \right)^2$$

$$= \sum_{i=1}^{N} \sigma^2_{a_0,i} \mathrm{pr}(M_i|D) + \sum_{i=1}^{N} \langle a_0 \rangle_i^2 \, \mathrm{pr}(M_i|D) - \left(\sum_{i=1}^{N} \langle a_0 \rangle_i \;\; \mathrm{pr}(M_i|D) \right)^2$$

Bayesian model averaging

- Consider an analysis with just two models, M_1 and M_2
- Suppose M1 is strongly favored by the data ($p \ll 1$):
  - $pr(M_1|D) = 1 - p$
  - $pr(M_2|D) = p$

$$\langle a_0 \rangle = \langle a_0 \rangle_1 + (\langle a_0 \rangle_2 - \langle a_0 \rangle_1) p,$$
$$\sigma^2_{a_0} \approx \sigma^2_{a_0,1} + \left[\sigma^2_{a_0,2} - \sigma^2_{a_0,1} + (\langle a_0 \rangle_2 - \langle a_0 \rangle_1)^2\right] p.$$

- Recover results of M_1 as $p \to 0$
- For $p \neq 0$, corrections likely to be small (but depends on sizes!)

Bayesian model averaging

Likelihood function ~Exp[-χ^2]

Prior for parameters in model

Prior for model

$$\mathrm{pr}(M|D) = \int d\mathbf{a}\, \frac{\mathrm{pr}(D|\mathbf{a}, M)\,\mathrm{pr}(\mathbf{a}|M)\,\mathrm{pr}(M)}{\mathrm{pr}(D)}$$

- In principle, the model weight is completely calculable (e.g, using Monte Carlo)
- For large data sets, the integrand becomes tightly localized around the best-fit χ^2 and increasingly Gaussian

Bayesian model averaging

Evaluate integral

→ *Gaussian Approximate Posterior* (GAP)

$$-2\log(\mathrm{pr}(M|D)) \approx GAP_M$$

Likelihood function tightly peaked around best-fit χ^2

Prior covariance matrix

Gaussian fluctuations about the best-fit χ^2

$$GAP_M = -2\log(\mathrm{pr}(M)) + (\chi^{\star}_{aug})^2 + \log\det\tilde{\Sigma} - \log\det\Sigma^{\star}$$

Model prior

Cancels in average when pr(M) uniform

Last two terms: "Information gain" in fit over the prior

Bayesian model averaging

Before using this result to compute pr(M|D), we must first think carefully:

1. Empirical priors and the Jeffreys-Lindley paradox

   (Not specific to our analysis, must confront in *any* real-world application of Bayesian model weights)

2. Bias correction

#1: Empirical priors

- We often use empirical priors and imagine that we can remove them (prior width → ∞) without influencing the best fit too much.

- However, taking prior width to ∞ gives a divergence in the model weight!

$$\log\det\tilde{\Sigma} - \log\det\Sigma^{\star} = 2\sum_i \log\left(\frac{\tilde{\sigma}_i}{\sigma_i^{\star}}\right) \to \infty$$

- *Jeffreys-Lindley paradox:* For models with different numbers of parameters, **the simpler model is pathologically preferred**, regardless of the data.

- Note: divergence cancels when comparing models with common priors.

#1: Empirical priors

- *Formal Solution*: use **cross-validation** to set empirical priors with a partial fit to “training” data. Asymptotically, both covariance matrices then approach the true model. The difference will vanish as 1/N.

- Intuition: If data set is “infinite,” then dividing the data into training and testing samples yields two samples that match the population distribution.

- In practice, we can simply ignore these terms as sub-leading 1/N effects. (Or get serious about using real, fixed priors)

#2: Bias correction

- The sample log likelihood is a biased estimator of the true log likelihood. (Recall: **⟨biased estimator⟩ ≠ asymptotic truth**)

- Why? Roughly, the sample log-likelihood systematically overshoots the true value due to finite-sample-size fluctuations in the data

- The bias correction has been computed in the statistics literature:

$$b = 2\,\mathrm{tr}[J^{-1}I]$$

J = (negative) Hessian matrix

I = Fisher matrix

- The matrix product asymptotically approaches the identity, so the trace counts the *number of parameters* **k**:

$$-2\log \mathrm{pr}(M|D) \approx -2\log \mathrm{pr}(M) + (\chi^{\star}_{\mathrm{aug}})^2 + 2k$$

- This is the well-known **Akaike information criterion**, or AIC.

Tallying progress

- We want to compute model averages. We need pr(M|D):

$$\langle f(\mathbf{a}_c)\rangle = \sum_M \langle f(\mathbf{a}_c)\rangle_M \,\mathrm{pr}(M|D)$$

- So far we've found:

$$-2\log \mathrm{pr}(M|D) \approx -2\log \mathrm{pr}(M) + (\chi^{\star}_{\mathrm{aug}})^2 + 2k$$

- This result applies to different models for fixed data.

- We can extend it to account for cuts on data.

Data subset selection

- Data subset selection can be reinterpreted as a model selection problem with a formal auxiliary model. (Note: we still just fit as usual!)
- Ex: choosing t_{min} for a 2pt correlator fit
  - For $t > t_{min}$: fit to N-state exponential model.
  - For $t \leq t_{min}$: imagine fit to a "perfect model" that interpolates the data.
    - For N_{cut} data points, the "perfect model" can be an order-N_{cut} polynomial.
    - The full model has $k + N_{cut}$ total parameters
- By construction, the perfect model adds nothing to χ^2
- The new parameters do contribute to the model weight via the bias-correction term:

$$-2\log \mathrm{pr}(M|D) \approx -2\log \mathrm{pr}(M) + (\chi^{\star}_{\mathrm{aug}})^2 + 2k + 2N_{\mathrm{cut}}$$

Extra piece

Data subset selection

- Data subset selection can be reinterpreted as a model selection problem with a formal auxiliary model. (Note: we still just fit as usual!)
- Ex: choosing t_{min} for a 2pt correlator fit

Key point:
This is a formal argument.
In practice, fit to the usual model.
Interpret the result with the modified weight.

al.

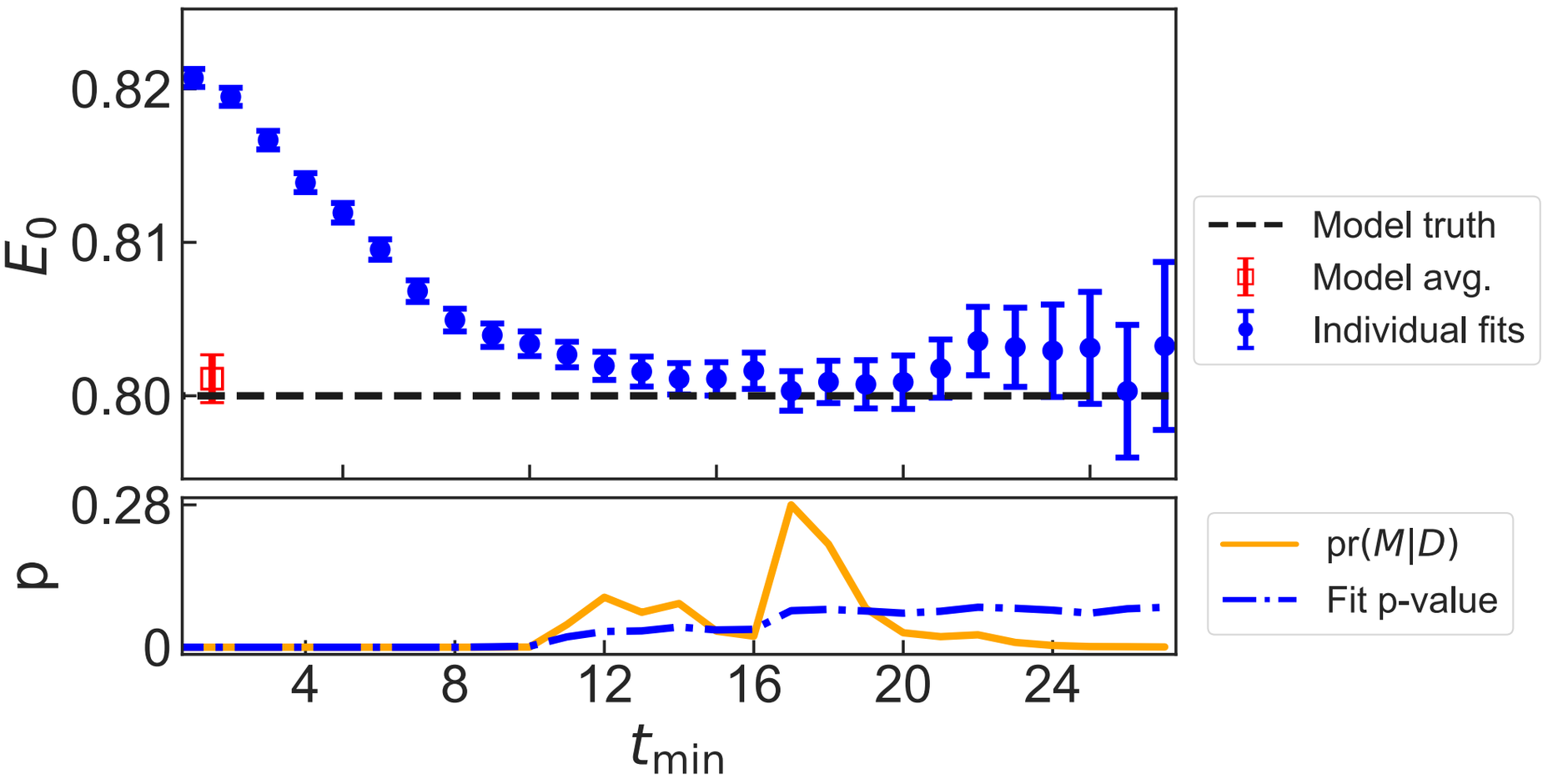
- By construction, the perfect model adds nothing to χ^2
- The new parameters do contribute to the model weight via the bias-correction term:

$$-2\log \mathrm{pr}(M|D) \approx -2\log \mathrm{pr}(M) + (\chi^{\star}_{\mathrm{aug}})^2 + 2k + 2N_{\mathrm{cut}}$$

Extra piece

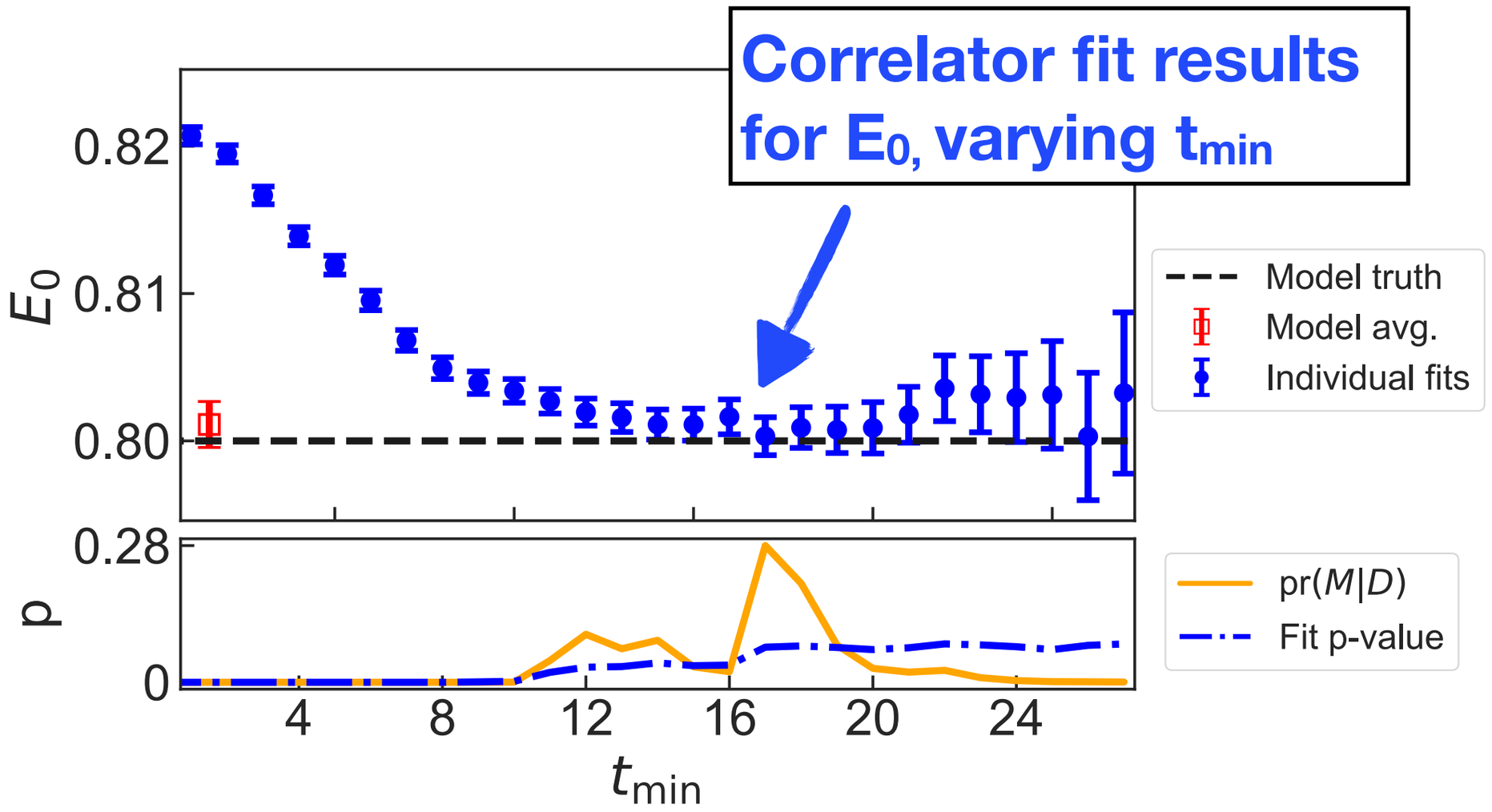
Ex. 1: t_{min} averaging (toy data)

Two-exponential “mock correlator”model truth, fit to single exponential



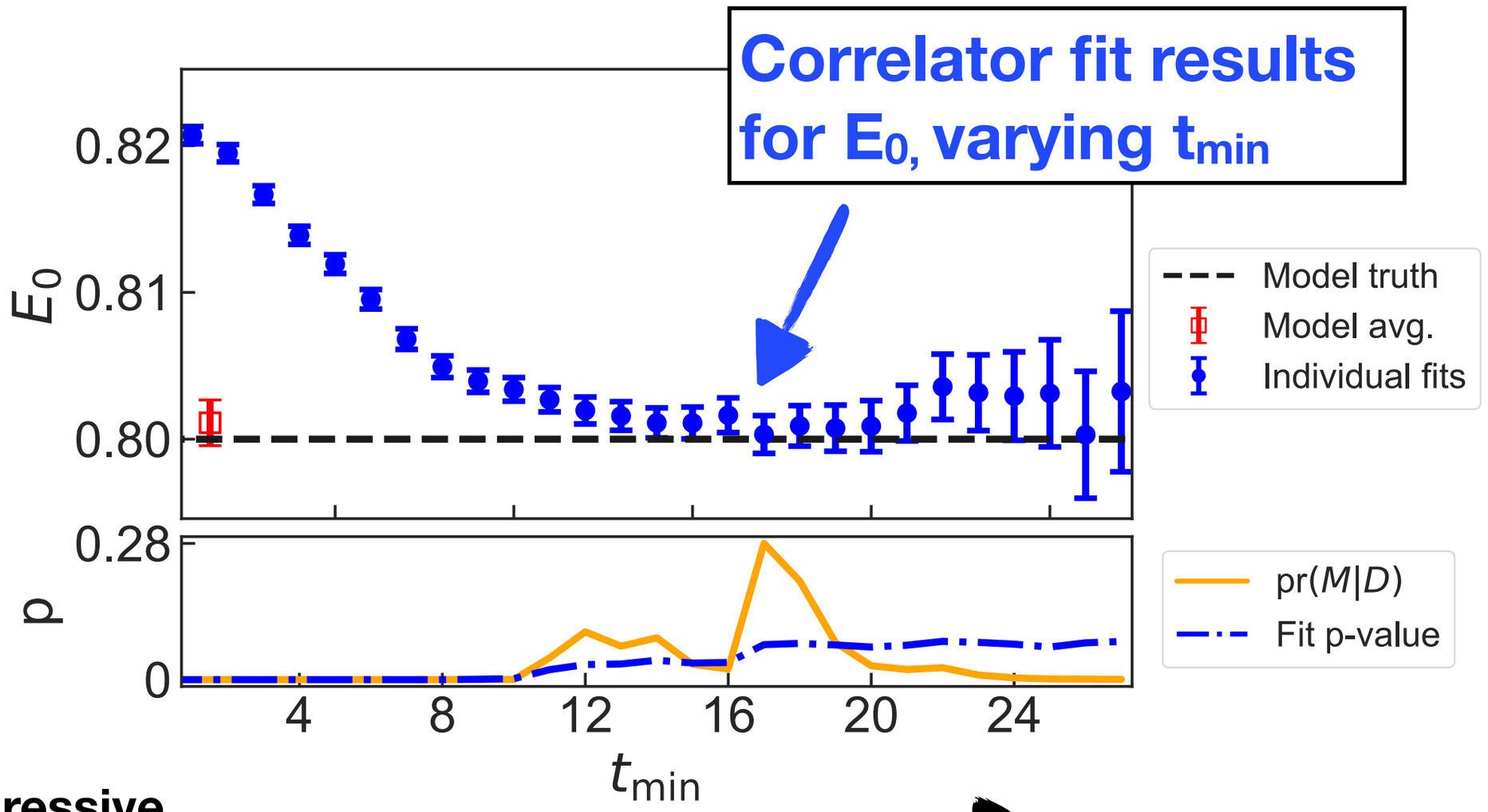
Ex. 1: t_{min} averaging (toy data)

Two-exponential "mock correlator"model truth, fit to single exponential



Ex. 1: t_{min} averaging (toy data)

Two-exponential "mock correlator"model truth, fit to single exponential



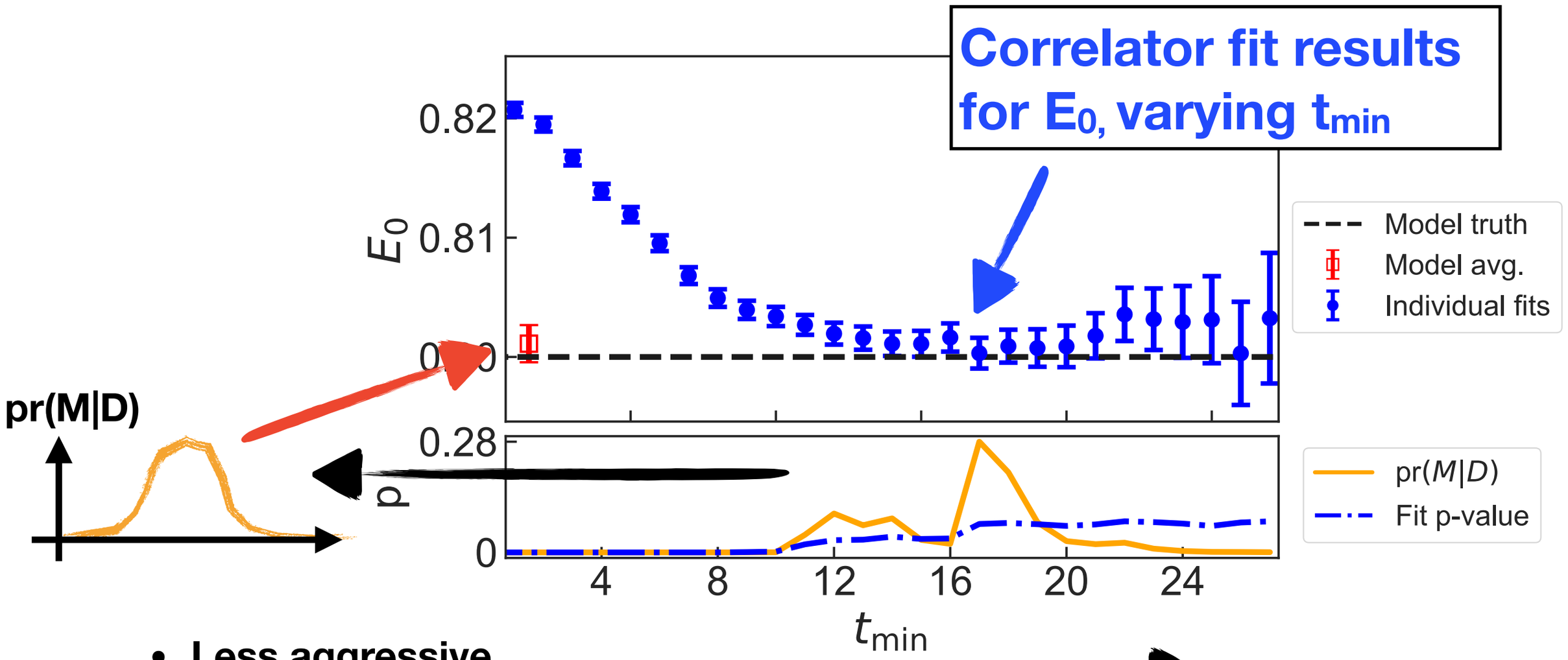
- **Less aggressive**
- **More total data**
- **More excited-state contamination**

Cut time t_{min}

- **More aggressive**
- **Less total data**

Ex. 1: t_{min} averaging (toy data)

Two-exponential "mock correlator"model truth, fit to single exponential



- Less aggressive
- Less total data
- More excited-state contamination

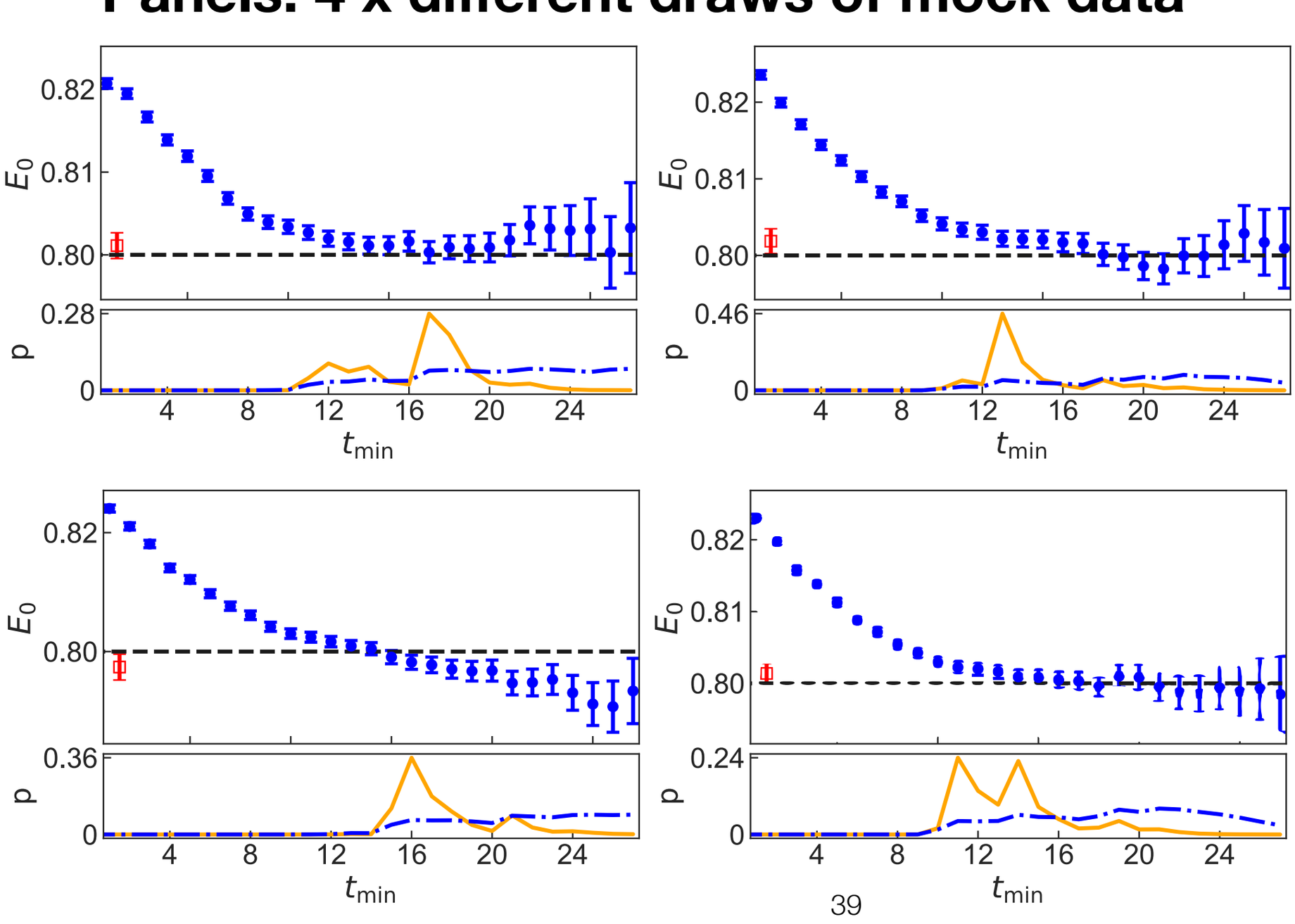
Cut time t_{min}

- More aggressive
- Less total data

Ex. 1: t_{min} averaging (toy data)

Two-exponential “mock correlator”model truth, fit to single exponential

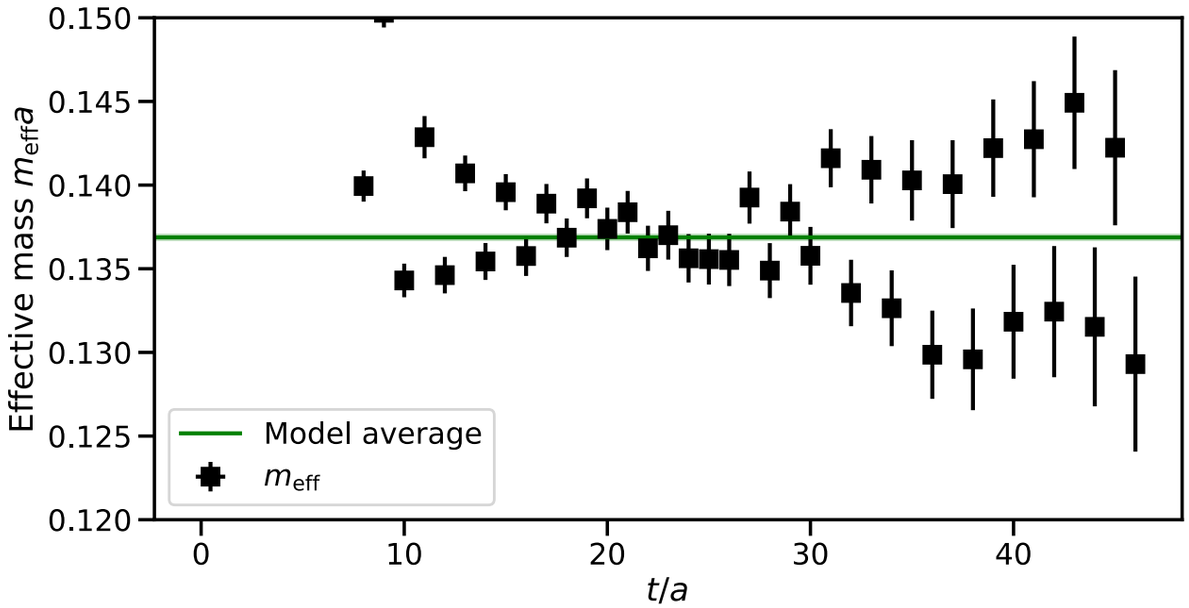
Panels: 4 x different draws of mock data

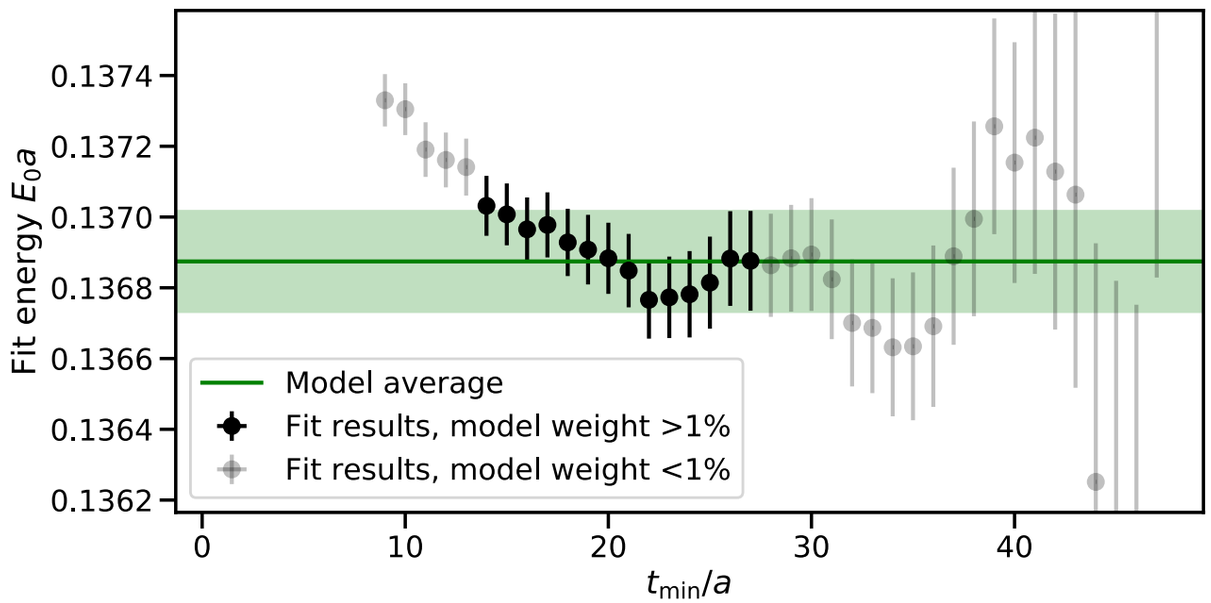


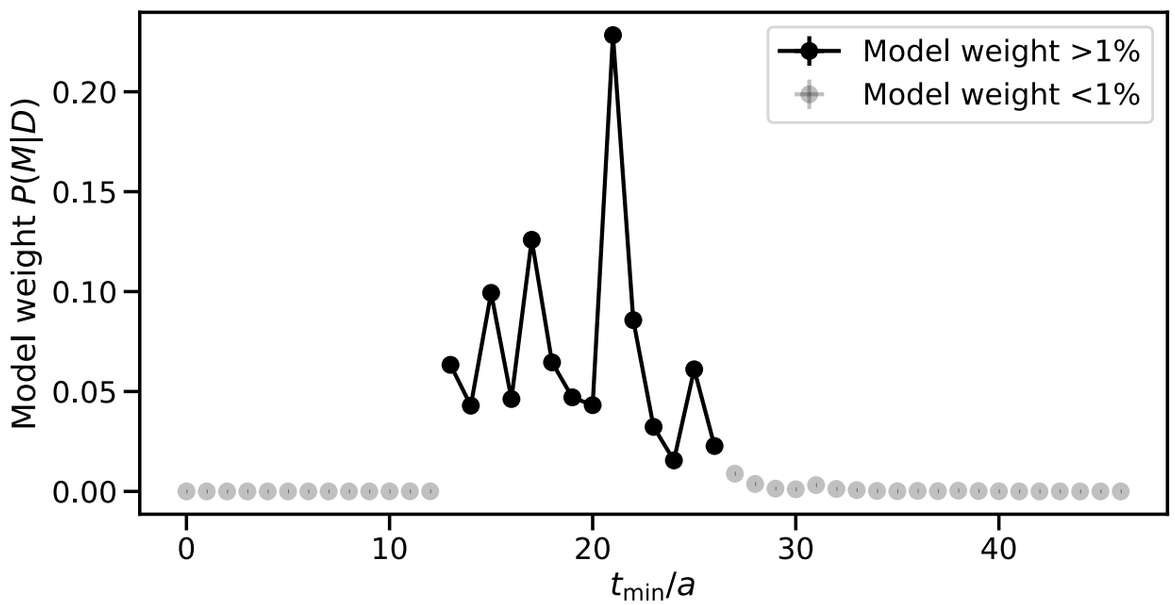
Model average agrees well with true ground-state

Ex. 2: t_{min} averaging (lattice data)

Real lattice data, staggered fermions, fit to (1+1) states







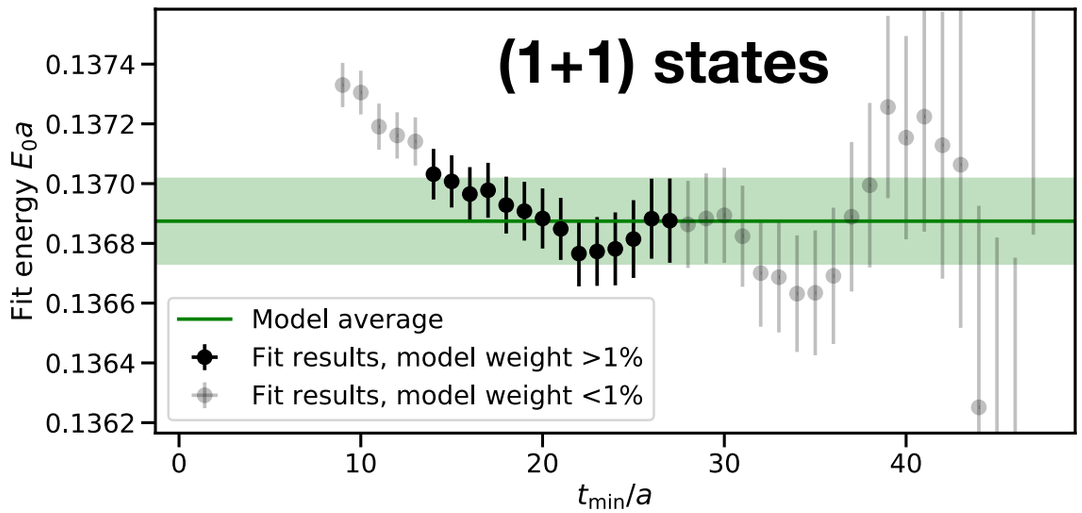
- **Model average** agrees closely (<1σ) with published best-fit result from **arXiv:1809.02827**

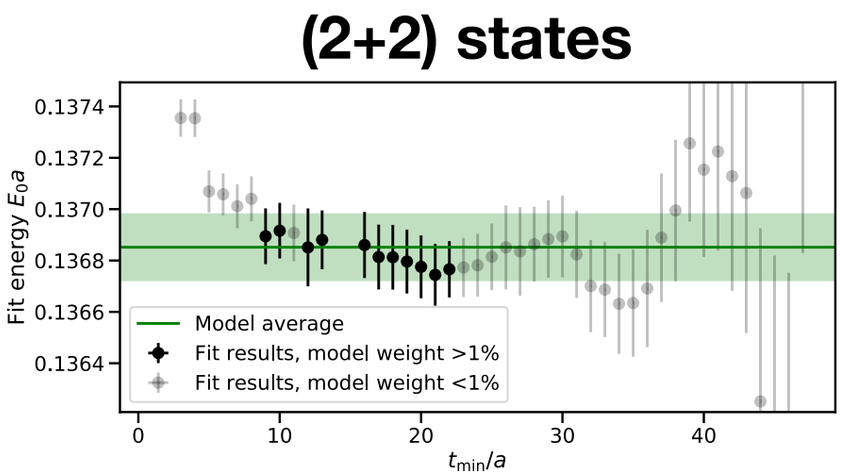
Ex. 2: t_{min} averaging (lattice data)

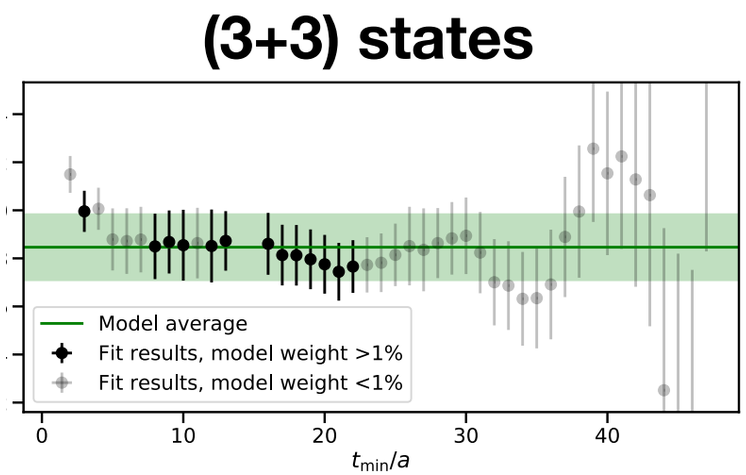
Real lattice data, staggered fermions, fit to (1+1) states

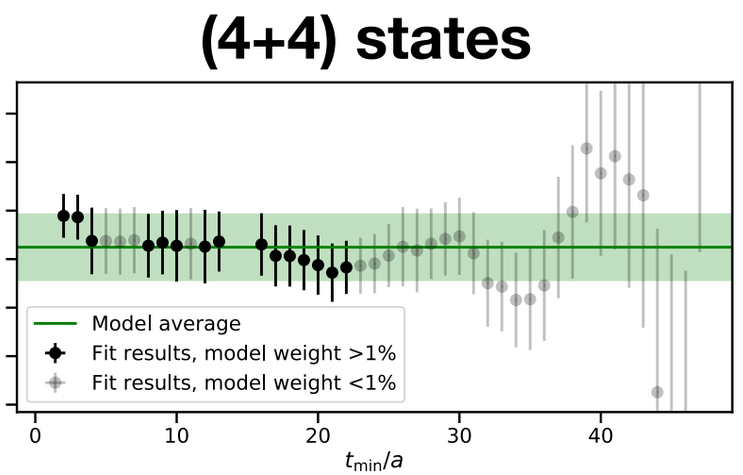
- Repeat: (2+2), (3+3), and (4+4) states.
- Better fits for smaller t_{min}
- **Model average** is quantitatively unchanged
- **Model average** agrees closely (<1σ) with published best-fit result from **arXiv:1809.02827**

[Repeated from previous slide]



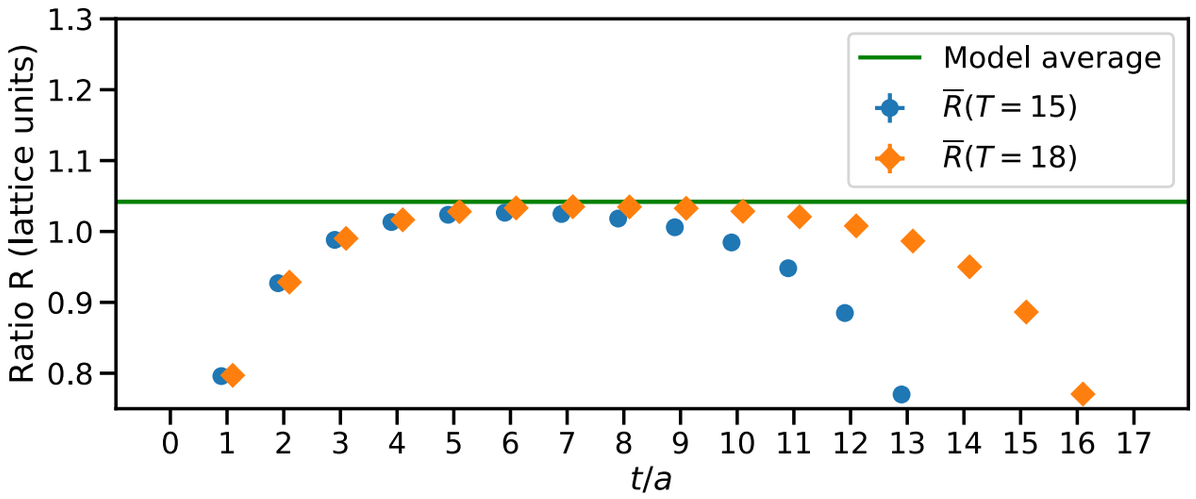


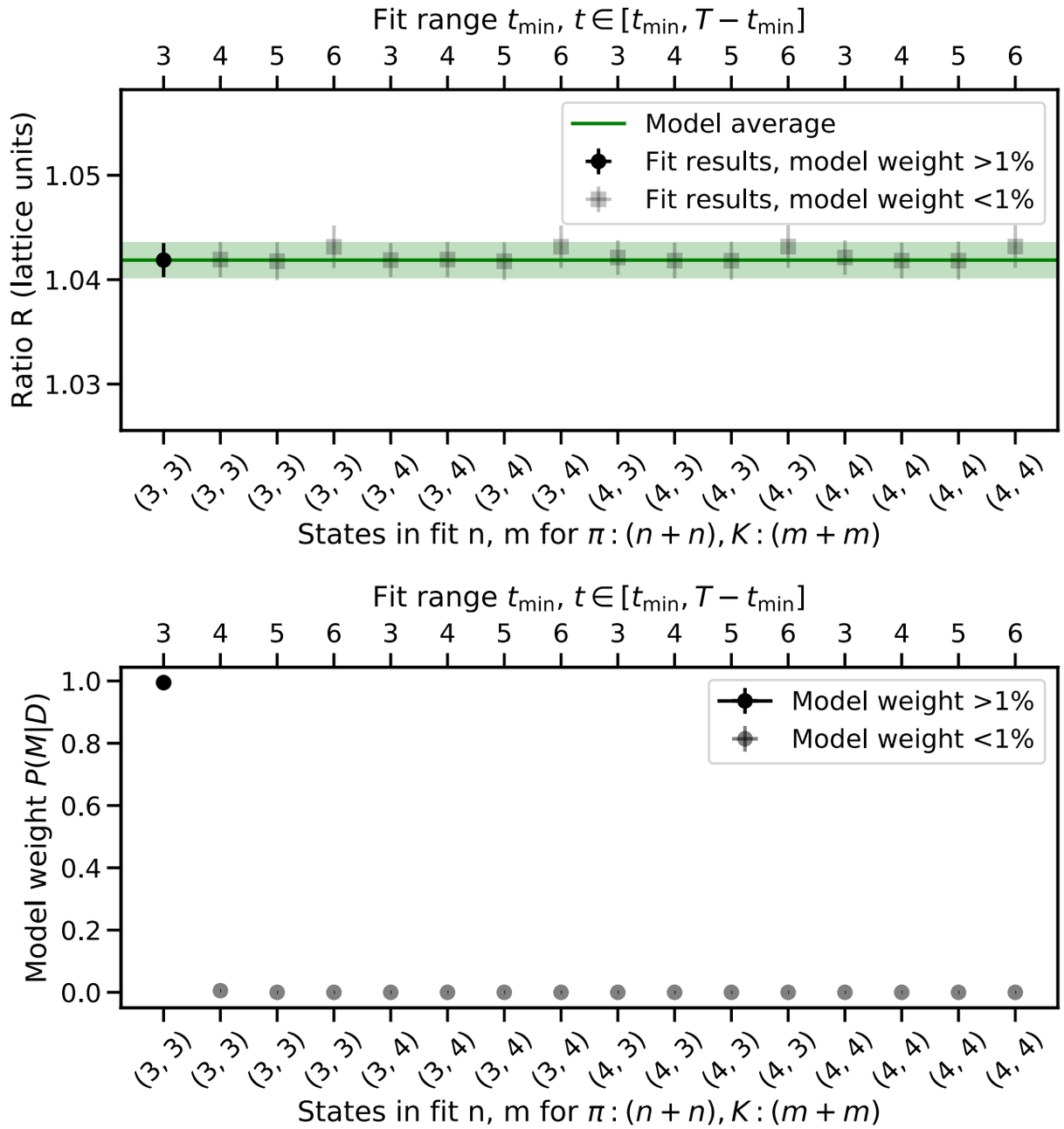




Ex. 3: Matrix elements (lattice data)

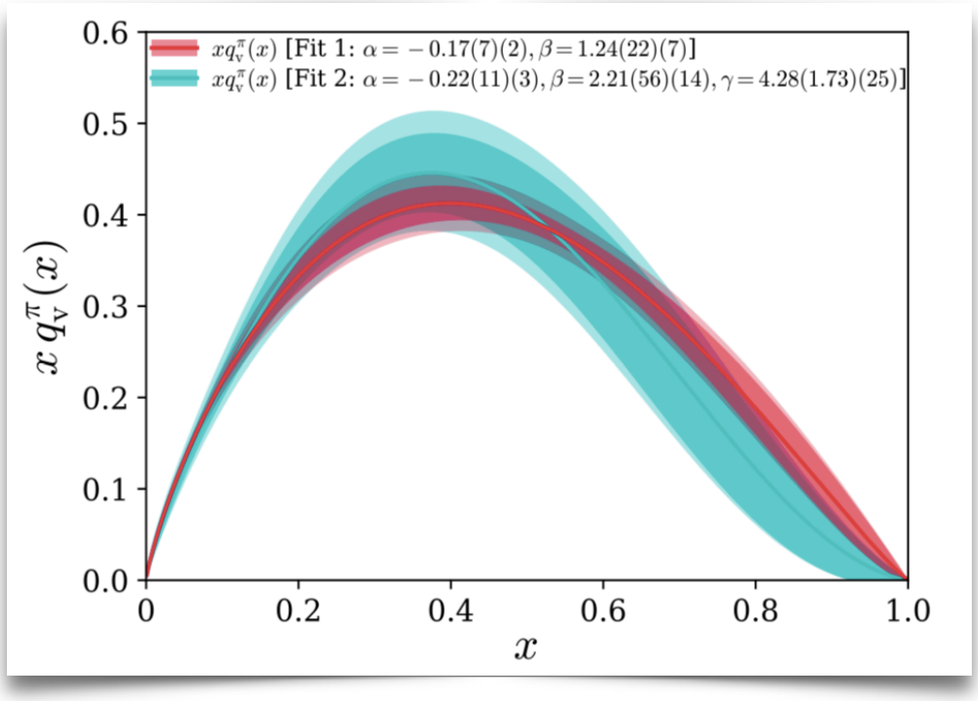
Real lattice data, staggered fermions





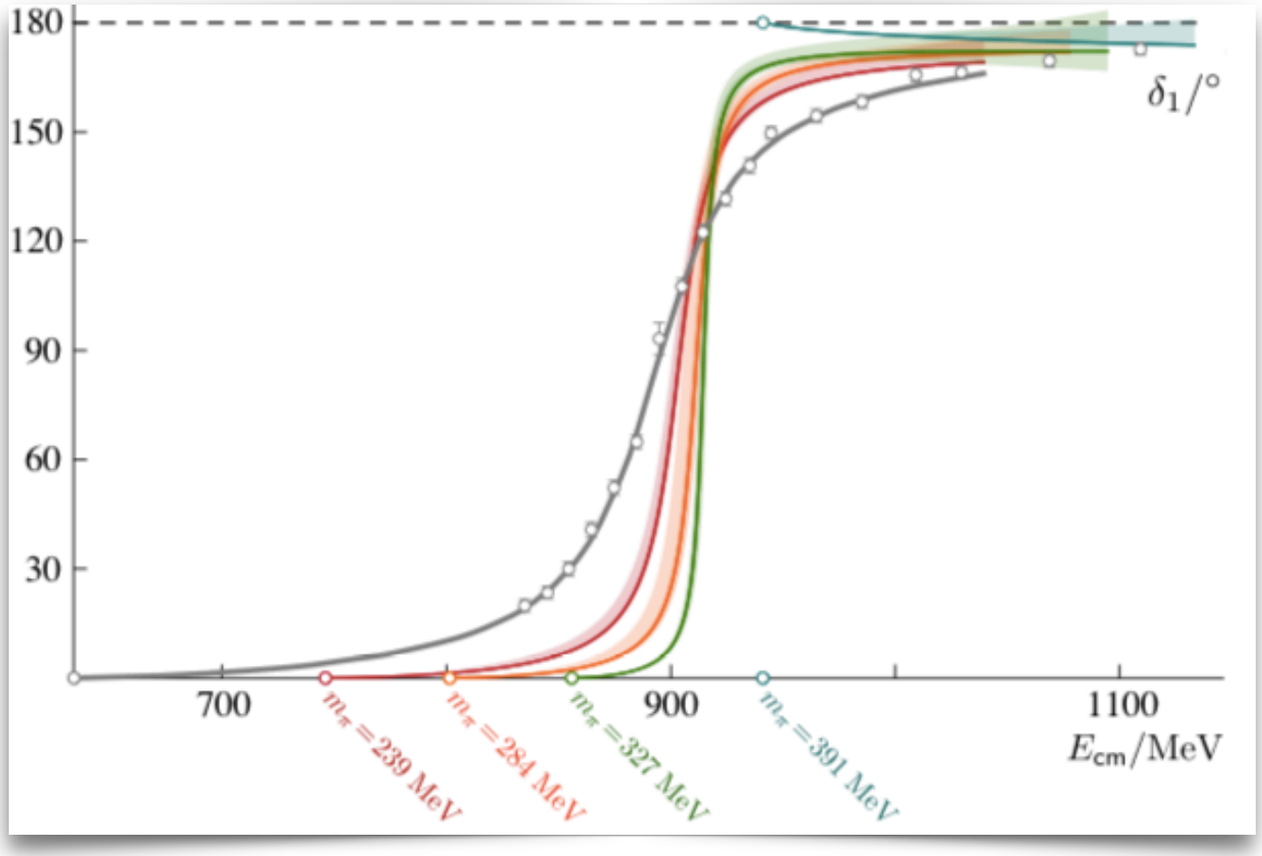
- Model weights are tightly peaked around a single fit
- *Occam's razor*: fits using fewer parameters to describe more data are preferred.
- Model averaging → Model selection
- The result agrees closely (<1σ) with published best-fit result from **arXiv: 1809.02827**

Possible applications at JLab?



"[The pion valence quark distribution] is achieved by numerically evaluating the convolution of the NLO kernel Eq. (8) and …phenomenologically motivated functional forms of the PDF."

R.S. Sufian et al.
PRD 102 (2020) 5, 054508
arXiv:2001.04960



"To avoid bias, we consider a wide selection of scattering amplitude parameterizations that fall into four familiar categories…

D.J. Wilson et al. [hadspec]
PRL 123 (2019) 4, 042002
arXiv:1904.03188

Conclusions

- Bayesian model averaging is a statistically rigorous way to handle uncertainty in model specification without being overly conservative

- Examples are well-matched to standard analysis problems in practical lattice problems:
  - ❖ What are acceptable t_{min} / t_{max} values in a correlator fit?
  - ❖ How many excited states to keep?
  - ❖ What mass range to use in a χPT fit?
  - ❖ How many terms should appear in an EFT fit?
- Implementation is easy: just need χ^2 and the number of parameters in the model

- We've seen good performance in tests with both mock and real data

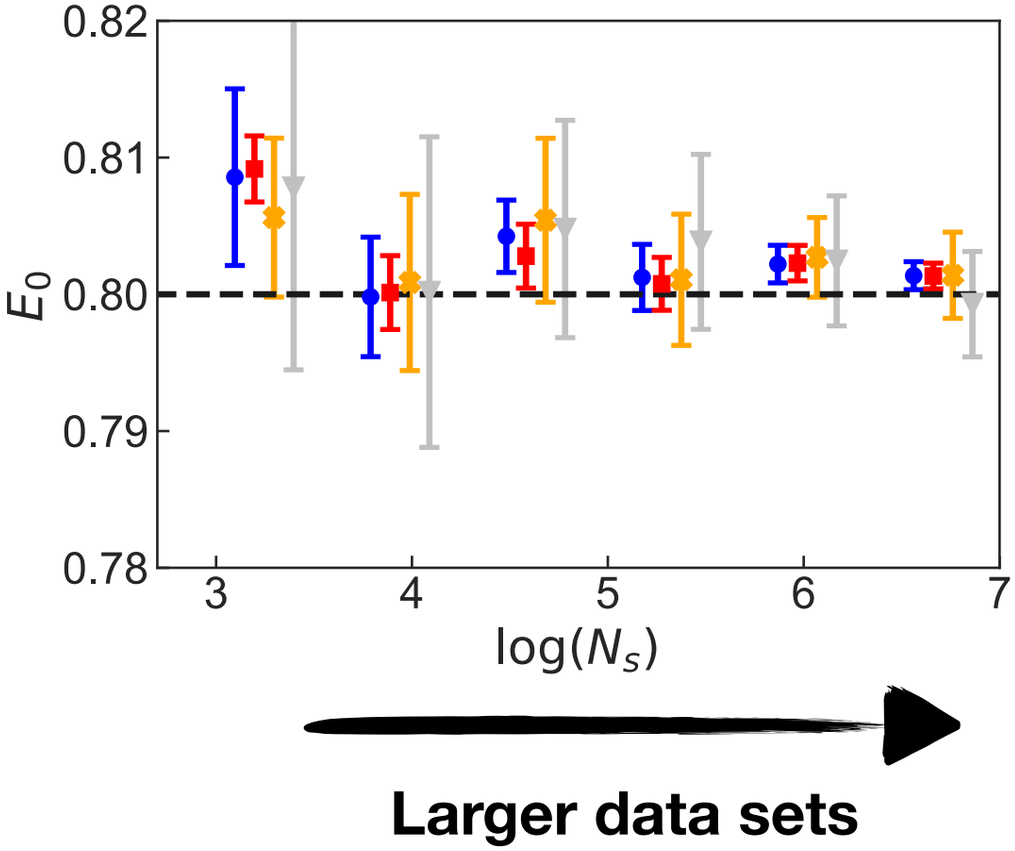
Backup slides

Ex. 1: t_{min} averaging

Two-exponential "mock correlator"model truth, fit to single exponential

Scaling results

$$-2\log \mathrm{pr}(M|D) \approx (\chi^{\star}_{\mathrm{aug}})^2 + 2k + 2N_{\mathrm{cut}}$$



Larger data sets

Model truth
Model avg. (AIC)
Fixed t_{min} = 14
Full-width systematic
Model avg. (naive)

"Choose best-fit t_{min}=14"

Conservative systematic (p-value > 0.1 only)

$$-2\log \mathrm{pr}(M|D) \approx (\chi^{\star}_{\mathrm{aug}})^2$$

Model Subset selection

- Consider the data sample y_i on the i^{th} gauge configuration. y_i is generally a vector: C(t), $t \in [1,\ldots N_t]$

- Partition the data samples $y_i = (y_i^{cut}, y_i^{keep})$

- Define a joint model:

Sample mean

$$y_i - g_M(\mathbf{a}, P) = \begin{cases} y_i - \bar{y}^{\mathrm{cut}}, & y_i \in y_i^{\mathrm{cut}} \\ y_i - f_M(\mathbf{a}), & y_i \in y_i^{\mathrm{keep}} \end{cases}$$

Model Subset selection

$$-2\log\ \mathrm{pr}(D|\mathbf{a},M) = \sum_{i=1}^{N}\chi_i^2(P)$$

$$= \sum_{i=1}^{N}(y_i - g_M(\mathbf{a},P))^T\Sigma^{-1}(y_i - g_M(\mathbf{a},P))$$

$$= \sum_{i=1}^{N}(y_i^{\mathrm{keep}} - f_M(\mathbf{a}))^T\Sigma_P^{-1}(y_i^{\mathrm{keep}} - f_M(\mathbf{a})) + (\mathrm{const})$$

All other terms involving the cut data contain the difference [$\bar{y}^{cut}$ - $g_M(\mathbf{a}, P)$] at least once. Therefore, they vanish identically by construction.