2 Quality criteria, averaging and error estimation

The essential characteristics of our approach to the problem of rating and averaging lattice quantities have been outlined in our first publication [1]. Our aim is to help the reader assess the reliability of a particular lattice result without necessarily studying the original article in depth. This is a delicate issue, since the ratings may make things appear simpler than they are. Nevertheless, it safeguards against the common practice of using lattice results, and drawing physics conclusions from them, without a critical assessment of the quality of the various calculations. We believe that, despite the risks, it is important to provide some compact information about the quality of a calculation. We stress, however, the importance of the accompanying detailed discussion of the results presented in the various sections of the present review.

2.1 Systematic errors and colour code

The major sources of systematic error are common to most lattice calculations. These include, as discussed in detail below, the chiral, continuum and infinite-volume extrapolations. To each such source of error for which systematic improvement is possible we assign one of three coloured symbols: green star, unfilled green circle (which replaced in Ref. [2] the amber disk used in the original FLAG review [1]) or red square. These correspond to the following ratings:
★ the parameter values and ranges used to generate the datasets allow for a satisfactory control of the systematic uncertainties;
○ the parameter values and ranges used to generate the datasets allow for a reasonable attempt at estimating systematic uncertainties, which however could be improved;
■ the parameter values and ranges used to generate the datasets are unlikely to allow for a reasonable control of systematic uncertainties.
The appearance of a red tag, even in a single source of systematic error of a given lattice result, disqualifies it from inclusion in the global average.

The attentive reader will notice that these criteria differ from those used in Refs. [1, 2]. In the previous FLAG editions we used the three symbols in order to rate the reliability of the systematic errors attributed to a given result by the paper's authors. This sometimes proved to be a daunting task, as the methods used by some collaborations for estimating their systematics are not always explained in full detail. Moreover, it is sometimes difficult to disentangle and rate different uncertainties, since they are interwoven in the error analysis. Thus, in the present edition we have opted for a different approach: the three symbols rate the quality of a particular simulation, based on the values and range of the chosen parameters, and its aptness to obtain well-controlled systematic uncertainties. They do not rate the quality of the analysis performed by the authors of the publication. The latter question is deferred to the relevant sections of the present review, which contain detailed discussions of the results contributing (or not) to each FLAG average or estimate. As a result of this different approach to the rating criteria, as well as changes of the criteria themselves, the colour coding of some papers in the current FLAG version differs from that of Ref. [2].

For most quantities the colour-coding system refers to the following sources of systematic errors: (i) chiral extrapolation; (ii) continuum extrapolation; (iii) finite volume. As we will see below, renormalization is another source of systematic uncertainties in several quantities. This we also classify using the three coloured symbols listed above, but now with a different rationale: they express how reliably these quantities are renormalized, from a field-theoretic

point of view (namely nonperturbatively, or with 2-loop or 1-loop perturbation theory).

Given the sophisticated status that the field has attained, several aspects, besides those rated by the coloured symbols, need to be evaluated before one can conclude whether a particular analysis leads to results that should be included in an average or estimate. Some of these aspects are not so easily expressible in terms of an adjustable parameter such as the lattice spacing, the pion mass or the volume. As a result of such considerations, it sometimes occurs, albeit rarely, that a given result does not contribute to the FLAG average or estimate, despite not carrying any red tags. This happens, for instance, whenever aspects of the analysis appear to be incomplete (e.g. an incomplete error budget), so that the presence of inadequately controlled systematic effects cannot be excluded. This mostly refers to results with a statistical error only, or results in which the quoted error budget obviously fails to account for an important contribution.

Of course any colour coding has to be treated with caution; we emphasize that the criteria are subjective and evolving. Sometimes a single source of systematic error dominates the systematic uncertainty and it is more important to reduce this uncertainty than to aim for green stars for other sources of error. In spite of these caveats we hope that our attempt to introduce quality measures for lattice simulations will prove to be a useful guide. In addition we would like to stress that the agreement of lattice results obtained using different actions and procedures provides further validation.

2.1.1 Systematic effects and rating criteria

The precise criteria used in determining the colour coding are unavoidably time-dependent; as lattice calculations become more accurate, the standards against which they are measured become tighter. For this reason, some of the quality criteria related to the light-quark sector have been tightened up between the first [1] and second [2] editions of FLAG.

In the second edition we have also reviewed quantities related to heavy quark physics [2]. The criteria used for light- and heavy-flavour quantities were not always the same. For the continuum limit, the difference was more a matter of choice: the light-flavour Working Groups defined the ratings using conditions involving specific values of the lattice spacing, whereas the heavy-flavour Working Groups preferred more data-driven criteria. Also, for finite-volume effects, the heavy-flavour groups slightly relaxed the boundary between ★ and ○, compared to the light-quark case, to account for the fact that heavy-quark quantities are less sensitive to the finiteness of the volume.

In the present edition we have opted for simplicity and adopted unified criteria for both light- and heavy-flavoured quantities.[1] The colour code used in the tables is specified as follows:

- Chiral extrapolation:
  - ★ $M_{\pi,\mathrm{min}} < 200$ MeV
  - ○ 200 MeV $\leq M_{\pi,\mathrm{min}} \leq 400$ MeV
  - ■ 400 MeV $< M_{\pi,\mathrm{min}}$

  It is assumed that the chiral extrapolation is performed with at least a three-point analysis; otherwise this will be explicitly mentioned. This condition is unchanged from Ref. [2].

[1] We note, however, that the data-driven criteria can be used by individual working groups in order to rate the reliability of the analyses for specific quantities.

- Continuum extrapolation:
  ★ at least 3 lattice spacings and at least 2 points below 0.1 fm and a range of lattice spacings satisfying $[a_{\rm max}/a_{\rm min}]^2 \geq 2$
  ○ at least 2 lattice spacings and at least 1 point below 0.1 fm and a range of lattice spacings satisfying $[a_{\rm max}/a_{\rm min}]^2 \geq 1.4$
  ■ otherwise
  It is assumed that the lattice action is $\mathcal{O}(a)$-improved (i.e. the discretization errors vanish quadratically with the lattice spacing); otherwise this will be explicitly mentioned. For unimproved actions an additional lattice spacing is required. This condition has been tightened compared to that of Ref. [2] by the requirements concerning the range of lattice spacings.

- Finite-volume effects:
  ★ $[M_{\pi,\rm min}/M_{\pi,\rm fid}]^2 \exp\{4 - M_{\pi,\rm min}[L(M_{\pi,\rm min})]_{\rm max}\} < 1$, or at least 3 volumes
  ○ $[M_{\pi,\rm min}/M_{\pi,\rm fid}]^2 \exp\{3 - M_{\pi,\rm min}[L(M_{\pi,\rm min})]_{\rm max}\} < 1$, or at least 2 volumes
  ■ otherwise
  It is assumed here that calculations are in the p-regime[2] of chiral perturbation theory, and that all volumes used exceed 2 fm. Here we are using a more sophisticated condition than that of Ref. [2]. The new condition involves the quantity $[L(M_{\pi,\rm min})]_{\rm max}$, which is the maximum box size used in the simulations performed at smallest pion mass $M_{\pi,\rm min}$, as well as a fiducial pion mass $M_{\pi,\rm fid}$, which we set to 200 MeV (the cutoff value for a green star in the chiral extrapolation).

  The rationale for this condition is as follows. Finite volume effects contain the universal factor $\exp\{-L\ M_\pi\}$, and if this were the only contribution a criterion based on the values of $M_{\pi,\rm min}L$ would be appropriate. This is what we used in Ref. [2] (with $M_{\pi,\rm min}L > 4$ for ★ and $M_{\pi,\rm min}L > 3$ for ○). However, as pion masses decrease, one must also account for the weakening of the pion couplings. In particular, 1-loop chiral perturbation theory [3] reveals a behaviour proportional to $M_\pi^2 \exp\{-L\ M_\pi\}$. Our new condition includes this weakening of the coupling, and ensures for example, that simulations with $M_{\pi,\rm min} = 135$ MeV and $L\ M_{\pi,\rm min} = 3.2$ are rated equivalently to those with $M_{\pi,\rm min} = 200$ MeV and $L\ M_{\pi,\rm min} = 4$.

- Renormalization (where applicable):
  ★ nonperturbative
  ○ 1-loop perturbation theory or higher with a reasonable estimate of truncation errors
  ■ otherwise
  In Ref. [1], we assigned a red square to all results which were renormalized at 1-loop in perturbation theory. In Ref. [2] we decided that this was too restrictive, since the error arising from renormalization constants, calculated in perturbation theory at 1-loop, is often estimated conservatively and reliably.

- Renormalization Group (RG) running (where applicable):
  For scale-dependent quantities, such as quark masses or B_K, it is essential that contact with continuum perturbation theory can be established. Various different methods are used for this purpose (cf. Appendix A.3): Regularization-independent Momentum

[2] We refer to Sec. 5.1 and Appendix A.4 in the Glossary for an explanation of the various regimes of chiral perturbation theory.

Subtraction (RI/MOM), the Schrödinger functional, and direct comparison with (resummed) perturbation theory. Irrespective of the particular method used, the uncertainty associated with the choice of intermediate renormalization scales in the construction of physical observables must be brought under control. This is best achieved by performing comparisons between nonperturbative and perturbative running over a reasonably broad range of scales. These comparisons were initially only made in the Schrödinger functional approach, but are now also being performed in RI/MOM schemes. We mark the data for which information about nonperturbative running checks is available and give some details, but do not attempt to translate this into a colour code.

The pion mass plays an important role in the criteria relevant for chiral extrapolation and finite volume. For some of the regularizations used, however, it is not a trivial matter to identify this mass.

In the case of twisted-mass fermions, discretization effects give rise to a mass difference between charged and neutral pions even when the up- and down-quark masses are equal: the charged pion is found to be the heavier of the two for twisted-mass Wilson fermions (cf. Ref. [4]). In early works, typically referring to $N_f = 2$ simulations (e.g. Refs. [4] and [5]), chiral extrapolations are based on chiral perturbation theory formulae which do not take these regularization effects into account. After the importance of keeping the isospin breaking when doing chiral fits was shown in Ref. [6], later works, typically referring to $N_f = 2 + 1 + 1$ simulations, have taken these effects into account [7]. We use $M_{\pi^\pm}$ for $M_{\pi,\mathrm{min}}$ in the chiral-extrapolation rating criterion. On the other hand, sea quarks (corresponding to both charged and neutral "sea pions" in an effective-chiral-theory logic) as well as valence quarks are intertwined with finite-volume effects. Therefore, we identify $M_{\pi,\mathrm{min}}$ with the root mean square (RMS) of M_{π^+}, M_{π^-} and M_{π^0} in the finite-volume rating criterion.[3]

In the case of staggered fermions, discretization effects give rise to several light states with the quantum numbers of the pion.[4] The mass splitting among these "taste" partners represents a discretization effect of $\mathcal{O}(a^2)$, which can be significant at large lattice spacings but shrinks as the spacing is reduced. In the discussion of the results obtained with staggered quarks given in the following sections, we assume that these artefacts are under control. We conservatively identify $M_{\pi,\mathrm{min}}$ with the root mean square (RMS) average of the masses of all the taste partners, both for chiral-extrapolation and finite-volume criteria.[5]

The strong coupling α_s is computed in lattice QCD with methods differing substantially from those used in the calculations of the other quantities discussed in this review. Therefore we have established separate criteria for α_s results, which will be discussed in Sec. 9.2.

2.1.2 Heavy-quark actions

In most cases, and in particular for the b quark, the discretization of the heavy-quark action follows a very different approach to that used for light flavours. There are several different

[3] This is a change from Ref. [2], where we used the charged pion mass when evaluating both chiral-extrapolation and finite-volume effects.

[4] We refer the interested reader to a number of good reviews on the subject [8–12].

[5] In Ref. [2], the RMS value was used in the chiral-extrapolation criteria throughout the paper. For the finite-volume rating, however, $M_{\pi,\mathrm{min}}$ was identified with the RMS value only in Secs. 4 and 6, while in Secs. 3, 5, 7 and 8 it was identified with the mass of the lightest pseudoscalar state.

methods for treating heavy quarks on the lattice, each with their own issues and considerations. All of these methods use Effective Field Theory (EFT) at some point in the computation, either via direct simulation of the EFT, or by using EFT as a tool to estimate the size of cutoff errors, or by using EFT to extrapolate from the simulated lattice quark masses up to the physical b-quark mass. Because of the use of an EFT, truncation errors must be considered together with discretization errors.

The charm quark lies at an intermediate point between the heavy and light quarks. In our previous review, the bulk of the calculations involving charm quarks treated it using one of the approaches adopted for the b quark. Many recent calculations, however, simulate the charm quark using light-quark actions, in particular the $N_f = 2+1+1$ calculations. This has become possible thanks to the increasing availability of dynamical gauge field ensembles with fine lattice spacings. But clearly, when charm quarks are treated relativistically, discretization errors are more severe than those of the corresponding light-quark quantities.

In order to address these complications, we add a new heavy-quark treatment category to the rating system. The purpose of this criterion is to provide a guideline for the level of action and operator improvement needed in each approach to make reliable calculations possible, in principle.

A description of the different approaches to treating heavy quarks on the lattice is given in Appendix A.1.3, including a discussion of the associated discretization, truncation, and matching errors. For truncation errors we use HQET power counting throughout, since this review is focused on heavy quark quantities involving B and D mesons rather than bottomonium or charmonium quantities. Here we describe the criteria for how each approach must be implemented in order to receive an acceptable (✓) rating for both the heavy quark actions and the weak operators. Heavy-quark implementations without the level of improvement described below are rated not acceptable (■). The matching is evaluated together with renormalization, using the renormalization criteria described in Sec. 2.1.1. We emphasize that the heavy-quark implementations rated as acceptable and described below have been validated in a variety of ways, such as via phenomenological agreement with experimental measurements, consistency between independent lattice calculations, and numerical studies of truncation errors. These tests are summarized in Sec. 8.

Relativistic heavy quark actions:

✓ at least tree-level $\mathcal{O}(a)$ improved action and weak operators

This is similar to the requirements for light quark actions. All current implementations of relativistic heavy quark actions satisfy this criterion.

NRQCD:

✓ tree-level matched through $\mathcal{O}(1/m_h)$ and improved through $\mathcal{O}(a^2)$

The current implementations of NRQCD satisfy this criterion, and also include tree-level corrections of $\mathcal{O}(1/m_h^2)$ in the action.

HQET:

✓ tree-level matched through $\mathcal{O}(1/m_h)$ with discretization errors starting at $\mathcal{O}(a^2)$

The current implementation of HQET by the ALPHA collaboration satisfies this criterion, since both action and weak operators are matched nonperturbatively through $\mathcal{O}(1/m_h)$. Calculations that exclusively use a static-limit action do not satisfy this criterion, since the static-limit action, by definition, does not include $1/m_h$ terms. We therefore consider static computations in our final estimates only if truncation errors (in $1/m_h$) are discussed and included in the systematic uncertainties.

Light-quark actions for heavy quarks:
✓ discretization errors starting at $\mathcal{O}(a^2)$ or higher
This applies to calculations that use the tmWilson action, a nonperturbatively improved Wilson action, or the HISQ action for charm quark quantities. It also applies to calculations that use these light quark actions in the charm region and above together with either the static limit or with an HQET inspired extrapolation to obtain results at the physical b quark mass. In these cases, the continuum extrapolation criteria described earlier must be applied to the entire range of heavy-quark masses used in the calculation.

2.1.3 Conventions for the figures

For a coherent assessment of the present situation, the quality of the data plays a key role, but the colour coding cannot be carried over to the figures. On the other hand, simply showing all data on equal footing would give the misleading impression that the overall consistency of the information available on the lattice is questionable. Therefore, in the figures we indicate the quality of the data in a rudimentary way, using the following symbols:
■ corresponds to results included in the average or estimate (i.e. results that contribute to the black square below);
□ corresponds to results that are not included in the average but pass all quality criteria;
□ corresponds to all other results;
■ corresponds to FLAG averages or estimates; they are also highlighted by a gray vertical band.
The reason for not including a given result in the average is not always the same: the result may fail one of the quality criteria; the paper may be unpublished; it may be superseded by newer results; or it may not offer a complete error budget.

Symbols other than squares are used to distinguish results with specific properties and are always explained in the caption.[6]

Often nonlattice data are also shown in the figures for comparison. For these we use the following symbols:
● corresponds to nonlattice results;
▲ corresponds to Particle Data Group (PDG) results.

2.2 Averages and estimates

FLAG results of a given quantity are denoted either as *averages* or as *estimates*. Here we clarify this distinction. To start with, both *averages* and *estimates* are based on results without any red tags in their colour coding. For many observables there are enough independent lattice calculations of good quality, with all sources of error (not merely those related to the colour-coded criteria), as analyzed in the original papers, appearing to be under control. In such cases it makes sense to average these results and propose such an *average* as the best current lattice number. The averaging procedure applied to this data and the way the error is obtained is explained in detail in Sec. 2.3. In those cases where only a sole result passes our rating criteria (colour coding), we refer to it as our FLAG *average*, provided it also displays adequate control of all other sources of systematic uncertainty.

[6] For example, for quark mass results we distinguish between perturbative and nonperturbative renormalization, for low-energy constants we distinguish between the p- and ϵ-regimes, and for heavy flavour results we distinguish between those from leptonic and semi-leptonic decays.

On the other hand, there are some cases in which this procedure leads to a result that, in our opinion, does not cover all uncertainties. Systematic error estimates are by their nature often subjective and difficult to estimate, and may thus end up being underestimated in one or more results that receive green symbols for all explicitly tabulated criteria. Adopting a conservative policy, in these cases we opt for an *estimate* (or a range), which we consider as a fair assessment of the knowledge acquired on the lattice at present. This *estimate* is not obtained with a prescribed mathematical procedure, but reflects what we consider the best possible analysis of the available information. The hope is that this will encourage more detailed investigations by the lattice community.

There are two other important criteria that also play a role in this respect, but that cannot be colour coded, because a systematic improvement is not possible. These are: *i)* the publication status, and *ii)* the number of sea-quark flavours N_f. As far as the former criterion is concerned, we adopt the following policy: we average only results that have been published in peer-reviewed journals, i.e. they have been endorsed by referee(s). The only exception to this rule consists in straightforward updates of previously published results, typically presented in conference proceedings. Such updates, which supersede the corresponding results in the published papers, are included in the averages. Note that updates of earlier results rely, at least partially, on the same gauge-field-configuration ensembles. For this reason, we do not average updates with earlier results. Nevertheless, all results are listed in the tables,[7] and their publication status is identified by the following symbols:

- Publication status:
  A published or plain update of published results
  P preprint
  C conference contribution

In the present edition, the publication status on the **30th of November 2015** is relevant. If the paper appeared in print after that date, this is accounted for in the bibliography, but does not affect the averages.

As noted above, in this review we present results from simulations with $N_f = 2$, $N_f = 2+1$ and $N_f = 2 + 1 + 1$ (except for $r_0\Lambda_{\overline{\rm MS}}$ where we also give the $N_f = 0$ result). We are not aware of an *a priori* way to quantitatively estimate the difference between results produced in simulations with a different number of dynamical quarks. We therefore average results at fixed N_f separately; averages of calculations with different N_f will not be provided.

To date, no significant differences between results with different values of N_f have been observed in the quantities listed in Tabs. 1 and 2. In the future, as the accuracy and the control over systematic effects in lattice calculations increases, it will hopefully be possible to see a difference between results from simulations with $N_f = 2$ and $N_f = 2 + 1$, and thus determine the size of the Zweig-rule violations related to strange-quark loops. This is a very interesting issue *per se*, and one which can be quantitatively addressed only with lattice calculations.

The question of differences between results with $N_f = 2 + 1$ and $N_f = 2 + 1 + 1$ is more subtle. The dominant effect of including the charm sea quark is to shift the lattice scale, an effect that is accounted for by fixing this scale nonperturbatively using physical quantities. For most of the quantities discussed in this review, it is expected that residual effects are small

[7] Whenever figures turn out to be overcrowded, older, superseded results are omitted. However, all the most recent results from each collaboration are displayed.

in the continuum limit, suppressed by $\alpha_s(m_c)$ and powers of Λ^2/m_c^2. Here Λ is a hadronic scale that can only be roughly estimated and depends on the process under consideration. Note that the Λ^2/m_c^2 effects have been addressed in Ref. [13]. Assuming that such effects are small, it might be reasonable to average the results from $N_f = 2+1$ and $N_f = 2+1+1$ simulations. This is not yet a pressing issue in this review, since there are relatively few results with $N_f = 2+1+1$, but it will become a more important question in the future.

2.3 Averaging procedure and error analysis

In the present report we repeatedly average results obtained by different collaborations and estimate the error on the resulting averages. We follow the procedure of the previous edition [2], which we describe here in full detail.

One of the problems arising when forming averages is that not all of the datasets are independent. In particular, the same gauge-field configurations, produced with a given fermion descretization, are often used by different research teams with different valence-quark lattice actions, obtaining results that are not really independent. Our averaging procedure takes such correlations into account.

Consider a given measurable quantity Q, measured by M distinct, not necessarily uncorrelated, numerical experiments (simulations). The result of each of these measurement is expressed as

$$Q_i \;=\; x_i \,\pm\, \sigma_i^{(1)} \pm \sigma_i^{(2)} \pm \cdots \pm \sigma_i^{(E)} \;\;, \tag{1}$$

where x_i is the value obtained by the $i^{\rm th}$ experiment ($i = 1, \cdots, M$) and $\sigma_i^{(k)}$ (for $k = 1, \cdots, E$) are the various errors. Typically $\sigma_i^{(1)}$ stands for the statistical error and $\sigma_i^{(k)}$ ($k \geq 2$) are the different systematic errors from various sources. For each individual result, we estimate the total error σ_i by adding statistical and systematic errors in quadrature:

$$\begin{aligned} Q_i \;&=\; x_i \,\pm\, \sigma_i \;\;, \\ \sigma_i \;&\equiv\; \sqrt{\sum_{k=1}^{E} \left[\sigma_i^{(k)}\right]^2} \;\;. \end{aligned} \tag{2}$$

With the weight factor of each total error estimated in standard fashion:

$$\omega_i \;=\; \frac{\sigma_i^{-2}}{\sum_{i=1}^{M} \sigma_i^{-2}} \;\;, \tag{3}$$

the central value of the average over all simulations is given by

$$x_{\rm av} \;=\; \sum_{i=1}^{M} x_i\, \omega_i \;. \tag{4}$$

The above central value corresponds to a $\chi^2_{\rm min}$ weighted average, evaluated by adding statistical and systematic errors in quadrature. If the fit is not of good quality ($\chi^2_{\rm min}/dof > 1$), the statistical and systematic error bars are stretched by a factor $S = \sqrt{\chi^2/dof}$.

Next we examine error budgets for individual calculations and look for potentially correlated uncertainties. Specific problems encountered in connection with correlations between different data sets are described in the text that accompanies the averaging. If there is reason

to believe that a source of error is correlated between two calculations, a 100% correlation is assumed. The correlation matrix C_{ij} for the set of correlated lattice results is estimated by a prescription due to Schmelling [14]. This consists in defining

$$\sigma_{i;j} \;=\; \sqrt{{\sum_{(k)}}' \left[\sigma_i^{(k)}\right]^2} \;, \tag{5}$$

with $\sum'_{(k)}$ running only over those errors of x_i that are correlated with the corresponding errors of measurement x_j. This expresses the part of the uncertainty in x_i that is correlated with the uncertainty in x_j. If no such correlations are known to exist, then we take $\sigma_{i;j} = 0$. The diagonal and off-diagonal elements of the correlation matrix are then taken to be

$$\begin{aligned} C_{ii} &= \sigma_i^2 \qquad\qquad (i = 1, \cdots, M) \;, \\ C_{ij} &= \sigma_{i;j}\,\sigma_{j;i} \qquad\quad (i \neq j) \;. \end{aligned} \tag{6}$$

Finally the error of the average is estimated by

$$\sigma_{\rm av}^2 \;=\; \sum_{i=1}^{M}\sum_{j=1}^{M} \omega_i\,\omega_j\,C_{ij} \;, \tag{7}$$

and the FLAG average is

$$Q_{\rm av} \;=\; x_{\rm av} \,\pm\, \sigma_{\rm av} \;. \tag{8}$$

References

[1] [FLAG 10] G. Colangelo, S. Dürr, A. Jüttner, L. Lellouch, H. Leutwyler et al., *Review of lattice results concerning low energy particle physics*, *Eur.Phys.J.* **C71** (2011) 1695, [1011.4408].

[2] [FLAG 13] S. Aoki, Y. Aoki, C. Bernard, T. Blum, G. Colangelo et al., *Review of lattice results concerning low-energy particle physics*, *Eur.Phys.J.* **C74** (2014) 2890, [1310.8555].

[3] G. Colangelo, S. Dürr and C. Haefeli, *Finite volume effects for meson masses and decay constants*, *Nucl. Phys.* **B721** (2005) 136–174, [hep-lat/0503014].

[4] [ETM 07A] Ph. Boucaud et al., *Dynamical twisted mass fermions with light quarks*, *Phys.Lett.* **B650** (2007) 304–311, [hep-lat/0701012].

[5] [ETM 09C] R. Baron et al., *Light meson physics from maximally twisted mass lattice QCD*, *JHEP* **08** (2010) 097, [0911.5061].

[6] O. Bär, *Chiral logs in twisted mass lattice QCD with large isospin breaking*, *Phys.Rev.* **D82** (2010) 094505, [1008.0784].

[7] [ETM 14] N. Carrasco et al., *Up, down, strange and charm quark masses with N_f = 2+1+1 twisted mass lattice QCD*, *Nucl. Phys.* **B887** (2014) 19–68, [1403.4504].

[8] S. Dürr, *Theoretical issues with staggered fermion simulations*, *PoS* **LAT2005** (2006) 021, [hep-lat/0509026].

[9] S. R. Sharpe, *Rooted staggered fermions: good, bad or ugly?*, *PoS* **LAT2006** (2006) 022, [hep-lat/0610094].

[10] A. S. Kronfeld, *Lattice gauge theory with staggered fermions: how, where, and why (not)*, *PoS* **LAT2007** (2007) 016, [0711.0699].

[11] M. Golterman, *QCD with rooted staggered fermions*, *PoS* **CONFINEMENT8** (2008) 014, [0812.3110].

[12] [MILC 09] A. Bazavov et al., *Full nonperturbative QCD simulations with 2+1 flavors of improved staggered quarks*, *Rev. Mod. Phys.* **82** (2010) 1349–1417, [0903.3598].

[13] [ALPHA 14A] M. Bruno, J. Finkenrath, F. Knechtli, B. Leder and R. Sommer, *Effects of Heavy Sea Quarks at Low Energies*, *Phys. Rev. Lett.* **114** (2015) 102001, [1410.8374].

[14] M. Schmelling, *Averaging correlated data*, *Phys.Scripta* **51** (1995) 676–679.