1 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 possibility 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.

1.1 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 data sets allow for a satisfactory control of the systematic uncertainties;

○ the parameter values and ranges used to generate the data sets allow for a reasonable attempt at estimating systematic uncertainties, which however could be improved;

■ the parameter values and ranges used to generate the data sets 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.

Note that in the first two editions [1, 2], FLAG used the three symbols in order to rate the reliability of the systematic errors attributed to a given result by the paper's authors. Starting with FLAG 16 [3] the meaning of the symbols has changed slightly—they now 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.

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.

1.1.1 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 FLAG reassesses criteria with each edition and as a result some of the quality criteria (the one on chiral extrapolation for instance) have been tightened up over time [1–4].

In the following, we present the rating criteria used in the current report. While these criteria apply to most quantities without modification, there are cases where they need to be amended or additional criteria need to be defined. For instance, the discussion of the strong coupling constant in Sec. 9 requires tailored criteria for renormalization, perturbative behaviour, and continuum extrapolation. Finally, in the section on nuclear matrix elements, Sec. 10, the chiral extrapolation criterion is made slightly stronger, and a new criterion is adopted for excited-state contributions. In such cases, the modified criteria are discussed in the respective sections. Apart from only a few exceptions the following colour code applies in the tables:

- Chiral extrapolation:

  ★ $M_{\pi,\mathrm{min}} < 200$ MeV, with three or more pion masses used in the extrapolation
  or two values of M_π with one lying within 10 MeV of 135 MeV (the physical neutral pion mass) and the other one below 200 MeV

  ○ 200 MeV $\leq M_{\pi,\mathrm{min}} \leq$ 400 MeV, with three or more pion masses used in the extrapolation
  or two values of M_π with $M_{\pi,\mathrm{min}} < 200$ MeV
  or a single value of M_π, lying within 10 MeV of 135 MeV (the physical neutral pion mass)

  ■ otherwise

  This criterion is unchanged from FLAG 19. In Sec. 10 the upper end of the range for $M_{\pi,\mathrm{min}}$ in the green circle criterion is lowered to 300 MeV, as in FLAG 19.

- Continuum extrapolation:

  ★ at least three lattice spacings and at least two points below 0.1 fm and a range of lattice spacings satisfying $[a_{\mathrm{max}}/a_{\mathrm{min}}]^2 \geq 2$

  - ○ at least two lattice spacings and at least one point below 0.1 fm and a range of lattice spacings satisfying $[a_{\text{max}}/a_{\text{min}}]^2 \geq 1.4$
  - ■ 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 is unchanged from FLAG 19.

- Finite-volume effects:
  The finite-volume colour code used for a result is chosen to be the worse of the QCD and the QED codes, as described below. If only QCD is used the QED colour code is ignored.

  – *For QCD:*

  - ★ $[M_{\pi,\text{min}}/M_{\pi,\text{fid}}]^2 \exp\{4 - M_{\pi,\text{min}}[L(M_{\pi,\text{min}})]_{\text{max}}\} < 1$, or at least three volumes
  - ○ $[M_{\pi,\text{min}}/M_{\pi,\text{fid}}]^2 \exp\{3 - M_{\pi,\text{min}}[L(M_{\pi,\text{min}})]_{\text{max}}\} < 1$, or at least two volumes
  - ■ otherwise

  where we have introduced $[L(M_{\pi,\text{min}})]_{\text{max}}$, which is the maximum box size used in the simulations performed at the smallest pion mass $M_{\pi,\text{min}}$, as well as a fiducial pion mass $M_{\pi,\text{fid}}$, which we set to 200 MeV (the cutoff value for a green star in the chiral extrapolation). It is assumed here that calculations are in the p-regime of chiral perturbation theory, and that all volumes used exceed 2 fm. The rationale for this condition is as follows. Finite-volume effects contain the universal factor $\exp\{-M_\pi L\}$, and if this were the only contribution a criterion based on the values of $M_{\pi,\text{min}}L$ would be appropriate. However, as pion masses decrease, one must also account for the weakening of the pion couplings. In particular, 1-loop chiral perturbation theory [5] reveals a behaviour proportional to $M_\pi^2 \exp\{-M_\pi L\}$. Our condition includes this weakening of the coupling, and ensures, for example, that simulations with $M_{\pi,\text{min}} = 135$ MeV and $M_{\pi,\text{min}}L = 3.2$ are rated equivalently to those with $M_{\pi,\text{min}} = 200$ MeV and $M_{\pi,\text{min}}L = 4$.

  – *For QED (where applicable):*

  - ★ $1/([M_{\pi,\text{min}}L(M_{\pi,\text{min}})]_{\text{max}})^{n_{\text{min}}} < 0.02$, or at least four volumes
  - ○ $1/([M_{\pi,\text{min}}L(M_{\pi,\text{min}})]_{\text{max}})^{n_{\text{min}}} < 0.04$, or at least three volumes
  - ■ otherwise

  Because of the infrared-singular structure of QED, electromagnetic finite-volume effects decay only like a power of the inverse spatial extent. In several cases like mass splittings [6, 7] or leptonic decays [8], the leading corrections are known to be universal, i.e., independent of the structure of the involved hadrons. In such cases, the leading universal effects can be directly subtracted exactly from the lattice data. We denote n_{min} the smallest power of $\frac{1}{L}$ at which such a subtraction cannot be done. In the widely used finite-volume formulation QED$_L$, one always has $n_{\text{min}} \leq 3$ due to the nonlocality of the theory [9]. The QED criteria are used here only in Sec. 4. Both QCD and QED criteria are unchanged from FLAG 19.

- Isospin-breaking effects (where applicable):

  - ★ all leading isospin-breaking effects are included in the lattice calculation
  - ○ isospin-breaking effects are included using the electro-quenched approximation

- 🟥 otherwise

  This criterion is used for quantities which are breaking isospin symmetry or which can be determined at the sub-percent accuracy where isospin-breaking effects, if not included, are expected to be the dominant source of uncertainty. In the current edition, this criterion is only used for the up- and down-quark masses, and related quantities (ϵ, Q^2 and R^2). The criteria for isospin-breaking effects are unchanged from FLAG 19.

- 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 FLAG 13 [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. These criteria have remained unchanged since then.

- 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 in FLAG 19 [4]): Regularization-independent Momentum 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. [10]). In early works, typically referring to $N_f = 2$ simulations (e.g., Refs. [10] and [11]), chiral extrapolations are based on chiral perturbation theory formulae which do not take these regularization effects into account. After the importance of accounting for isospin breaking when doing chiral fits was shown in Ref. [12], later works, typically referring to $N_f = 2+1+1$ simulations, have taken these effects into account [13]. We use $M_{\pi^\pm}$ for $M_{\pi,\mathrm{min}}$ in the chiral-extrapolation rating criterion. On the other hand, 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.

In the case of staggered fermions, discretization effects give rise to several light states with the quantum numbers of the pion.[1] 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 artifacts 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.

In some of the simulations, the fermion formulations employed for the valence quarks are different from those used for the sea quarks. Even when the fermion formulations are the same, there are cases where the sea- and valence-quark masses differ. In such cases, we use the smaller of the valence-valence and valence-sea $M_{\pi_{\mathrm{min}}}$ values in the finite-volume criteria, since either of these channels may give the leading contribution depending on the quantity of interest at the 1-loop level of chiral perturbation theory. For the chiral-extrapolation criteria, on the other hand, we use the unitary point, where the sea- and valence-quark masses are the same, to define $M_{\pi_{\mathrm{min}}}$.

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. 9.2.1.

In Sec. 10 on nuclear matrix elements, an additional criterion is used. This concerns the level of control over contamination from excited states, which is a more challenging issue for nucleons than for mesons. In response to an improved understanding of the impact of this contamination, the excited-state contamination criterion has been made more stringent compared to that in FLAG 19.

1.1.2 Data-driven criteria

For some time, the FLAG working groups have been considering using a ‘data-driven’ criterion in assessing how well the continuum limit is controlled. The quantity $\delta(a)$ is defined as

$$\delta(a) \equiv \frac{|Q(a) - Q(0)|}{\sigma_Q}\,, \tag{1}$$

were $Q(a)$ is the quantity under consideration with lattice spacing a, $Q(0)$ is the extrapolated continuum-limit value, and σ_Q is its error in the continuum limit. If a_{min} is the smallest lattice spacing used, there is concern if $\delta(a_{\mathrm{min}})$ is very large. That is, the results at the finest lattice spacing should not be too many standard deviations from the continuum limit in order for the extrapolation to be considered reliable.

The following is adopted for the current edition of the review: (1) Each working group attempts to determine $\delta(a_{\mathrm{min}})$ for each calculation that contributes to a FLAG average. However, it is not currently used as a criterion for inclusion in the averages. (2) The text of the report includes these values for calculations contributing to FLAG averages. (3) For the current edition of FLAG it is at the discretion of each working group to decide whether they wish to inflate the error of contributions to the average for calculations with large values of $\delta(a_{\mathrm{min}})$. If this is done, the inflation factor will be

$$s(\delta) = \max[1, 1 + 2(\delta - 3)/3]. \tag{2}$$

The inflation of the error is not displayed in tables or plots. It is only used to evaluate FLAG averages.

[1] We refer the interested reader to a number of reviews on the subject [14–18].

1.1.3 Heavy-quark actions

For the b quark, the discretization of the heavy-quark action follows a very different approach from that used for light flavours. There are several different methods for treating heavy quarks on the lattice, each with its own issues and considerations. Most 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 earlier reviews, the calculations involving charm quarks often treated it using one of the approaches adopted for the b quark. Since FLAG 16 [3], however, most calculations simulate the charm quark using light-quark actions. 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, the heavy-quark section adds an additional, bipartite, 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 can be found in Appendix A.1.3 of FLAG 19 [4]. 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 include 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 twisted-mass Wilson action, a nonperturbatively improved Wilson action, domain-wall fermions 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 combined list of lattice spacings used for the data sets with $m_h > 0.5 m_{h,\text{phys}}$ must satisfy the continuum-extrapolation criteria.

1.1.4 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 might 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.[2]

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.

1.2 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.

[2] For example, for quark-mass results we distinguish between perturbative and nonperturbative renormalization, and for heavy-flavour results we distinguish between those from leptonic and semileptonic 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 errors 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,[3] 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 April 2024** is relevant. If the paper appeared in print after that date, this is accounted for in the bibliography, but does not affect the averages.[4]

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 are not provided.

To date, no significant differences between results with different values of N_f have been observed in the quantities listed in Tabs. 1, 2, 3, and 4. In particular, differences between results from simulations with $N_f = 2$ and $N_f = 2+1$ would reflect Zweig-rule violations related to strange-quark loops. Although not of direct phenomenological relevance, the size of such violations is an interesting theoretical issue *per se*, and one that can be quantitatively addressed only with lattice calculations. It remains to be seen whether the status presented here will change in the future, since this will require dedicated $N_f = 2$ and $N_f = 2+1$ calculations, which are not a priority of present lattice work.

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

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

[4] As noted above in footnote 1, two exceptions to this deadline were made, Refs. [19, 20].

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 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 Refs. [21–25], and were found to be small for the quantities considered. Assuming that such effects are generically small, it might be reasonable to average the results from $N_f = 2+1$ and $N_f = 2+1+1$ simulations, although we do not do so here.

1.3 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. Here we provide details on how averages are obtained.

1.3.1 Averaging — generic case

We continue to follow the procedure of FLAG 13 and FLAG 16 [2, 3] which we describe here in full detail.

One of the problems arising when forming averages is that not all of the data sets are independent. In particular, the same gauge-field configurations, produced with a given fermion discretization, 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

$$Q_i \;=\; x_i \pm \sigma_i^{(1)} \pm \sigma_i^{(2)} \pm \cdots \pm \sigma_i^{(E)} \;, \tag{3}$$

where x_i is the value obtained by the $i^{\rm th}$ experiment ($i = 1, \cdots, M$) and $\sigma_i^{(\alpha)}$ (for $\alpha = 1, \cdots, E$) are the various errors. Typically $\sigma_i^{(1)}$ stands for the statistical error and $\sigma_i^{(\alpha)}$ ($\alpha \geq 2$) 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:

$$\begin{aligned} Q_i \;&=\; x_i \pm \sigma_i \;, \\ \sigma_i \;&\equiv\; \sqrt{\sum_{\alpha=1}^{E} \left[\sigma_i^{(\alpha)}\right]^2} \;. \end{aligned} \tag{4}$$

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

$$\omega_i \;=\; \frac{\sigma_i^{-2}}{\sum_{i=1}^{M} \sigma_i^{-2}} \;, \tag{5}$$

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

$$x_{\rm av} \;=\; \sum_{i=1}^{M} x_i \, \omega_i \;. \tag{6}$$

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 ($\chi^2_{\rm min}/{\rm dof} > 1$), the statistical and systematic error bars are stretched by a factor $S = \sqrt{\chi^2/{\rm dof}}$.

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 covariance matrix C_{ij} for the set of correlated lattice results is estimated by a prescription due to Schmelling [26]. This consists in defining

$$\sigma_{i;j} \;=\; \sqrt{{\sum_\alpha}' \left[\sigma_i^{(\alpha)}\right]^2} \;\;, \tag{7}$$

with ${\sum}'_\alpha$ running only over those errors of x_i that are correlated with the corresponding errors of the 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 covariance matrix are then taken to be

$$\begin{aligned} C_{ii} &= \sigma_i^2 \qquad && (i = 1, \cdots, M)\ , \\ C_{ij} &= \sigma_{i;j}\,\sigma_{j;i} \qquad && (i \neq j)\ . \end{aligned} \tag{8}$$

Finally, the error of the average is estimated by

$$\sigma^2_{\rm av} \;=\; \sum_{i=1}^{M}\sum_{j=1}^{M} \omega_i\,\omega_j\,C_{ij}\ , \tag{9}$$

and the FLAG average is

$$Q_{\rm av} \;=\; x_{\rm av} \,\pm\, \sigma_{\rm av}\ . \tag{10}$$

1.3.2 Nested averaging

We have encountered one case where the correlations between results are more involved, and a nested averaging scheme is required. This concerns the B-meson bag parameters discussed in Sec. 8.2. In the following, we describe the details of the nested averaging scheme. This is an updated version of the section added in the web update of the FLAG 16 report.

The issue arises for a quantity Q that is given by a ratio, $Q = Y/Z$. In most simulations, both Y and Z are calculated, and the error in Q can be obtained in each simulation in the standard way. However, in other simulations only Y is calculated, with Z taken from a global average of some type. The issue to be addressed is that this average value $\overline{Z}$ has errors that are correlated with those in Q.

In the example that arises in Sec. 8.2, $Q = B_B$, $Y = B_B f_B^2$ and $Z = f_B^2$. In one of the simulations that contribute to the average, Z is replaced by $\overline{Z}$, the PDG average for f_B^2 [27] (obtained with an averaging procedure similar to that used by FLAG). This simulation is labeled with $i = 1$, so that

$$Q_1 = \frac{Y_1}{\overline{Z}}. \tag{11}$$

The other simulations have results labeled Q_j, with $j \geq 2$. In this set up, the issue is that $\overline{Z}$ is correlated with the Q_j, $j \geq 2$.[5]

We begin by decomposing the error in Q_1 in the same schematic form as above,

$$Q_1 = x_1 \pm \frac{\sigma_{Y_1}^{(1)}}{\overline{Z}} \pm \frac{\sigma_{Y_1}^{(2)}}{\overline{Z}} \pm \cdots \pm \frac{\sigma_{Y_1}^{(E)}}{\overline{Z}} \pm \frac{Y_1 \sigma_{\overline{Z}}}{\overline{Z}^2}. \tag{12}$$

Here the last term represents the error propagating from that in $\overline{Z}$, while the others arise from errors in Y_1. For the remaining Q_j $(j \geq 2)$ the decomposition is as in Eq. (3). The total error of Q_1 then reads

$$\sigma_1^2 = \left(\frac{\sigma_{Y_1}^{(1)}}{\overline{Z}}\right)^2 + \left(\frac{\sigma_{Y_1}^{(2)}}{\overline{Z}}\right)^2 + \cdots + \left(\frac{\sigma_{Y_1}^{(E)}}{\overline{Z}}\right)^2 + \left(\frac{Y_1}{\overline{Z}^2}\right)^2 \sigma_{\overline{Z}}^2, \tag{13}$$

while that for the Q_j $(j \geq 2)$ is

$$\sigma_j^2 = \left(\sigma_j^{(1)}\right)^2 + \left(\sigma_j^{(2)}\right)^2 + \cdots + \left(\sigma_j^{(E)}\right)^2. \tag{14}$$

Correlations between Q_j and Q_k $(j, k \geq 2)$ are taken care of by Schmelling's prescription, as explained above. What is new here is how the correlations between Q_1 and Q_j $(j \geq 2)$ are taken into account.

To proceed, we recall from Eq. (9) that $\sigma_{\overline{Z}}$ is given by

$$\sigma_{\overline{Z}}^2 = \sum_{i',j'=1}^{M'} \omega[Z]_{i'} \omega[Z]_{j'} C[Z]_{i'j'}. \tag{15}$$

Here the indices i' and j' run over the M' simulations that contribute to $\overline{Z}$, which, in general, are different from those contributing to the results for Q. The weights $\omega[Z]$ and covariance matrix $C[Z]$ are given an explicit argument Z to emphasize that they refer to the calculation of this quantity and not to that of Q. $C[Z]$ is calculated using the Schmelling prescription [Eqs. (7)–(9)] in terms of the errors, $\sigma[Z]_{i'}^{(\alpha)}$, taking into account the correlations between the different calculations of Z.

We now generalize Schmelling's prescription for $\sigma_{i;j}$, Eq. (7), to that for $\sigma_{1;k}$ $(k \geq 2)$, i.e., the part of the error in Q_1 that is correlated with Q_k. We take

$$\sigma_{1;k} \;=\; \sqrt{\frac{1}{\overline{Z}^2} \sideset{}{'}\sum_{(\alpha)\leftrightarrow k} \left[\sigma_{Y_1}^{(\alpha)}\right]^2 + \frac{Y_1^2}{\overline{Z}^4} \sum_{i',j'}^{M'} \omega[Z]_{i'} \omega[Z]_{j'} C[Z]_{i'j'\leftrightarrow k}}\;. \tag{16}$$

The first term under the square root sums those sources of error in Y_1 that are correlated with Q_k. Here we are using a more explicit notation from that in Eq. (7), with $(\alpha) \leftrightarrow k$ indicating that the sum is restricted to the values of α for which the error $\sigma_{Y_1}^{(\alpha)}$ is correlated

[5] There is also a small correlation between Y_1 and $\overline{Z}$, but we follow the original Ref. [28] and do not take this into account. Thus, the error in Q_1 is obtained by simple error propagation from those in Y_1 and $\overline{Z}$. Ignoring this correlation is conservative, because, as in the calculation of B_K, the correlations between $B_B f_B^2$ and f_B^2 tend to lead to a cancellation of errors. By ignoring this effect we are making a small overestimate of the error in Q_1.

with Q_k. The second term accounts for the correlations within $\overline{Z}$ with Q_k, and is the nested part of the present scheme. The new matrix $C[Z]_{i'j'\leftrightarrow k}$ is a restriction of the full covariance matrix $C[Z]$, and is defined as follows. Its diagonal elements are given by

$$C[Z]_{i'i'\leftrightarrow k} = (\sigma[Z]_{i'\leftrightarrow k})^2 \qquad (i' = 1, \cdots, M') \ , \tag{17}$$

$$(\sigma[Z]_{i'\leftrightarrow k})^2 = \sum_{(\alpha)\leftrightarrow k}{}' (\sigma[Z]_{i'}^{(\alpha)})^2, \tag{18}$$

where the summation $\sum'_{(\alpha)\leftrightarrow k}$ over (α) is restricted to those $\sigma[Z]_{i'}^{(\alpha)}$ that are correlated with Q_k. The off-diagonal elements are

$$C[Z]_{i'j'\leftrightarrow k} = \sigma[Z]_{i';j'\leftrightarrow k}\,\sigma[Z]_{j';i'\leftrightarrow k} \qquad (i' \neq j') \ , \tag{19}$$

$$\sigma[Z]_{i';j'\leftrightarrow k} = \sqrt{\sum_{(\alpha)\leftrightarrow j'k}{}' (\sigma[Z]_{i'}^{(\alpha)})^2}, \tag{20}$$

where the summation $\sum'_{(\alpha)\leftrightarrow j'k}$ over (α) is restricted to $\sigma[Z]_{i'}^{(\alpha)}$ that are correlated with *both* $Z_{j'}$ and Q_k.

The last quantity that we need to define is $\sigma_{k;1}$.

$$\sigma_{k;1} = \sqrt{\sum_{(\alpha)\leftrightarrow 1}{}' \left[\sigma_k^{(\alpha)}\right]^2} \ , \tag{21}$$

where the summation $\sum'_{(\alpha)\leftrightarrow 1}$ is restricted to those $\sigma_k^{(\alpha)}$ that are correlated with one of the terms in Eq. (13).

In summary, we construct the covariance matrix C_{ij} using Eq. (8), as in the generic case, except the expressions for $\sigma_{1;k}$ and $\sigma_{k;1}$ are now given by Eqs. (16) and (21), respectively. All other $\sigma_{i;j}$ are given by the original Schmelling prescription, Eq. (7). In this way, we extend the philosophy of Schmelling's approach while accounting for the more involved correlations.

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] [FLAG 16] S. Aoki et al., *Review of lattice results concerning low-energy particle physics*, *Eur. Phys. J.* **C77** (2017) 112 [1607.00299].

[4] [FLAG 19] S. Aoki et al., *FLAG Review 2019: Flavour Lattice Averaging Group (FLAG)*, *Eur. Phys. J. C* **80** (2020) 113 [1902.08191].

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

[6] [BMW 14] Sz. Borsanyi et al., *Ab initio calculation of the neutron-proton mass difference*, *Science* **347** (2015) 1452 [1406.4088].

[7] Z. Davoudi and M.J. Savage, *Finite-Volume Electromagnetic Corrections to the Masses of Mesons, Baryons and Nuclei*, *Phys. Rev.* **D90** (2014) 054503 [1402.6741].

[8] V. Lubicz, G. Martinelli, C.T. Sachrajda, F. Sanfilippo, S. Simula and N. Tantalo, *Finite-Volume QED Corrections to Decay Amplitudes in Lattice QCD*, *Phys. Rev.* **D95** (2017) 034504 [1611.08497].

[9] Z. Davoudi, J. Harrison, A. Jüttner, A. Portelli and M.J. Savage, *Theoretical aspects of quantum electrodynamics in a finite volume with periodic boundary conditions*, *Phys. Rev.* **D99** (2019) 034510 [1810.05923].

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

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

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

[13] [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 [1403.4504].

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

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

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

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

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

[19] [RBC/UKQCD 24] P. Boyle, Felix, J.M. Flynn, N. Garron, J. Kettle, R. Mukherjee et al., *Kaon mixing beyond the standard model with physical masses*, *Phys. Rev. D* **110** (2024) 034501 [2404.02297].

[20] N. Brambilla, V. Leino, J. Mayer-Steudte and A. Vairo, *Static force from generalized Wilson loops on the lattice using the gradient flow*, *Phys. Rev. D* **109** (2024) 114517 [2312.17231].

[21] [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].

[22] [ALPHA 17A] F. Knechtli, T. Korzec, B. Leder and G. Moir, *Power corrections from decoupling of the charm quark*, *Phys. Lett. B* **774** (2017) 649 [1706.04982].

[23] A. Athenodorou, J. Finkenrath, F. Knechtli, T. Korzec, B. Leder, M.K. Marinkovic et al., *How perturbative are heavy sea quarks?*, *Nucl. Phys.* **B943** (2019) 114612 [1809.03383].

[24] S. Cali, F. Knechtli and T. Korzec, *How much do charm sea quarks affect the charmonium spectrum?*, *Eur. Phys. J. C* **79** (2019) 607 [1905.12971].

[25] [ALPHA 21A] S. Cali, K. Eckert, J. Heitger, F. Knechtli and T. Korzec, *Charm sea effects on charmonium decay constants and heavy meson masses*, *Eur. Phys. J. C* **81** (2021) 733 [2105.12278].

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

[27] J.L. Rosner, S. Stone and R.S. Van de Water, *Leptonic Decays of Charged Pseudoscalar Mesons, in Review of Particle Physics [29] 2015 update*, 1509.02220.

[28] [FNAL/MILC 16] A. Bazavov et al., *$B^0_{(s)}$-mixing matrix elements from lattice QCD for the Standard Model and beyond*, *Phys. Rev.* **D93** (2016) 113016 [1602.03560].

[29] Particle Data Group collaboration, *Review of Particle Physics*, *Chin. Phys.* **C38** (2014) 090001 and 2015 update.