Supplementary Information: Computational Processing and Error Reduction Strategies for Standardized Quantitative Data in Biological Networks

Marcel Schilling,[1,2] Thomas Maiwald,[3,2] Sebastian Bohl,[1] Markus Kollmann,[3] Clemens Kreutz,[3] Jens Timmer,[3] and Ursula Klingmüller[1]

[1] *German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.*
[2] *These authors contributed equally.*
[3] *FDM, Freiburg Center for Data Analysis and Modeling, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany.*

(Dated: October 25, 2005)

I. ERROR CLASSIFICATION

Immunoblotting as a technique for quantitative analysis of protein concentrations has several sources of noise, which can be divided into three classes:

(i) pipetting errors $f(j)$ change the amount of each protein in the j-th lane by the same factor, reflecting, e.g., the different amount of lysate loaded on each gel lane

(ii) blotting errors $g(j,m)$, observed as brighter and darker areas, arise from inhomogeneities of the gel and transfer to the membrane. They are highly correlated for neighboring lanes, j, and rows, m.

(iii) contributions independent from the loaded protein concentrations. They are modeled as Gaussian noise, $\eta(j,m)$.

As shown in the main text, error contributions from pipetting are small compared to the highly correlated errors of the blotting technique (see Fig. 2b, main text). Figure 7 gives an example of a blotting error and Figure 1 displays the mean autocorrelation function (ACF) for 13 gels of the normalizer protein βActin and for 8 gels of the calibrator GST-EpoR depending on the lane distance.

The effects on the concentration $x^*(t_j)$ of a given protein in the lysate at time t_j, resulting in the measured concentration $x(t_j)$, can be described by

$$x(t_j) = [1 + g(j,m)]\,[1 + f(j)]\,x^*(t_j) + \eta(j,m). \quad (1)$$

Here, the time point t_j after stimulation corresponds to lane j, the smooth systematic error g depends on the lane index j and on the molecular weight m, measured in kD. The pipetting error f depends only on the lane number j, since it changes the amount of all proteins in a lane by the same factor. Errors arising from (i) and (iii) are uncorrelated among different lanes, resulting in $\langle f(j)f(j')\rangle = 0$ and $\langle \eta(j,m)\eta(j',m)\rangle = 0$ for $j \neq j'$.

In the following we use normalizers and calibrators and a randomized, non-chronological loading of the lanes, to identify and reduce the highly correlated blotting errors, $g(j,m)$.

II. ELIMINATION OF THE BLOTTING ERROR

The highly correlated errors arising from the blotting technique vary gradually over the lanes and make it difficult to extract the actual values. To eliminate the correlations among the lanes we employ non-chronological gel loading. Here, the subsequent time-points after stimulation are loaded randomized on the gel under the condition that consecutive time points are separated by minimum number of 4 lanes for 20 time points (compare loading example in Figure 7). By applying this method, the errors between consecutive time points are uncorrelated. We are able to estimate the true time-course from the data by rearranging the time points in chronological order. As shown in Figure 2C, we employ a cubic spline whose smoothness is determined by generalized cross-validation. This technique demands statistical independent errors as generated by the randomized gel loading. The estimation of a time-course from noisy data by smoothing splines has been worked out in detail in Refs. [1–4]. We emphasize that a sufficiently dense grid of time-points is necessary to keep the bias of this method small.

For the case that a normalizer protein, $x_n(j)$, can be measured with a similar molecular weight as the protein of interest, it is possible to estimate the blotting error $g(j,m)$ as $x_n^*(j) = const$ by definition. The true signal is then given by

$$\hat{x}^*(t_j) \approx \frac{x_i(t_j)}{\bar{x}_n(t_j)}. \quad (2)$$

Here, $\bar{x}_n(t)$, denotes the smoothing spline generated from the date set $\{x_n(t_j)\}$ by keeping the *lane ordering* of the randomly loaded gels. Smoothing of the data is performed in oder to average over error contributions arising from pipetting, $f(j)$, and other sources of noise, $\eta(j,m)$. We further denote by $\tilde{x}^*(t)$ and $\tilde{x}(t)$ the time-courses of the smoothing splines generated from

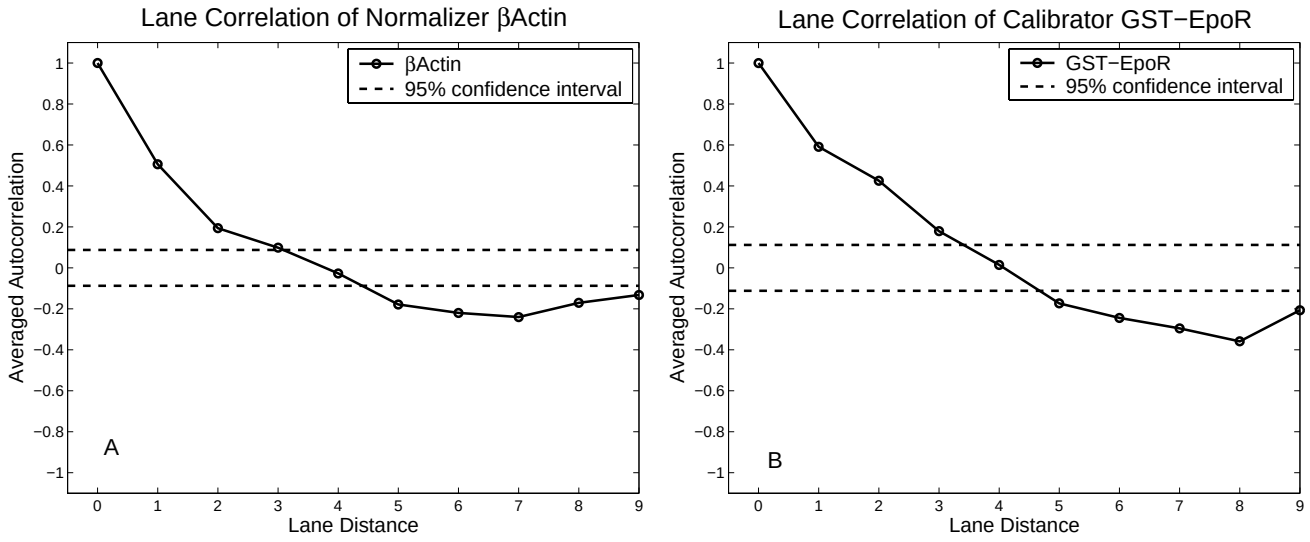

FIG. 1: Autocorrelation of the normalizer protein βActin (A) and the calibrator protein GST-EpoR (B) in the gel domain. The two-sided 95% confidence interval for the averaged autocorrelation function of a purely random process is in both cases not preserved, indicating a strong correlation of neighbored gel lanes.

the *chronological* ordered data sets $\{\hat{x}^*(t_j)\}$, $\{x(t_j)\}$. The residuals $|\tilde{x}^*(t_j) - \hat{x}^*(t_j)|$, are expected to be significantly smaller than the residuals without employing a normalizer, $|\tilde{x}(t_j) - x(t_j)|$, as we have readily accounted for the blotting errors in the first case. This in turn gives us a reliable measure for the quality of the used normalizer protein.

III. ERROR REDUCTION VIA RANDOMIZATION AND NORMALIZERS - SIMULATION STUDY

Simulations of typical immunoblotting experiments were performed by generating a simulated signal with quadratic rise and exponential decay and a maximum at half lane number, equidistantly sampled (Figure 2B). This simulates a typical time-course experiment after stimulation with a hormone. The true signal $x^*(t_j)$ was processed with the two main sources of errors as described in the previous section, a pipetting and a blotting error. In detail:

1. A multiplicative, uncorrelated pipetting error was applied as shown in Figure 2A representing errors derived from unequal cell number or errors in pipetting the cellular lysates:

   $$x'(t_j) = x^*(t_j) \cdot (1 + \sigma\varepsilon(j)) \quad \varepsilon(j) \in N(0,1)).$$

2. A multiplicative, strongly correlated blotting error was applied, representing errors from differences in migration in the SDS polyacrylamide gel or unequal transfer to the membrane:

   $$x(t_j) = x'(t_j) \cdot (1 + g(j)),$$

   with the blotting error $g(j)$ represented by a sine function with mean zero and phase, amplitude and frequency consistent with experimental observations.

The processing was applied to a chronological and to a randomized true signal, x^*_{rand} and x^*_{chron}, respectively, leading to "measurements" like in Figure 2B. Note that the chronological signal is rather smooth but changes the characteristic of the true signal: The maximum occurs earlier and a new minimum is observed at t = 15. The randomized signal on the other hand is very noise, but does not introduce systematic effects. The smoothed processed randomized signal $\tilde{x}_{rand}$ is very close to the true time-course, whereas the smoothed processed chronological signal $\tilde{x}_{chron}$ still keeps correlated deviations from the true signal (Figure 2C). The correlation structure of the deviations can be investigated via the autocorrelation function (Figure 2D). For uncorrelated errors, the autocorrelation function should drop from 1 at $\tau = 0$ into the 95% confidence interval for $\tau > 0$. This is not the case for the processed chronologically signal, which can lead to misleading conclusions if methods are applied which assume uncorrelated noise.

Besides visual inspection of the autocorrelation function, the improvement of data quality by means of a randomized gel loading can be quantified by the *error reduc-*

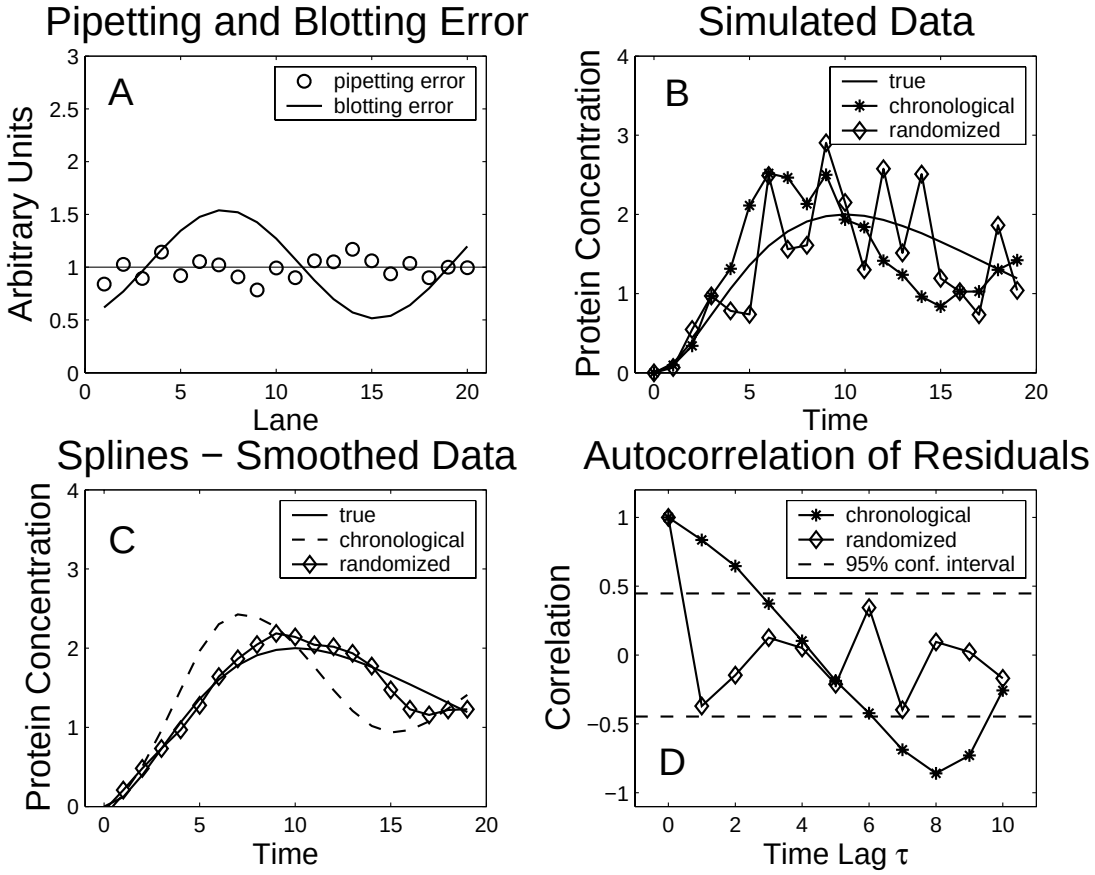

FIG. 2: Effect of randomization on immunoblotting data. (A) Simulated uncorrelated pipetting error and highly correlated, sine-like blotting error. (B) Simulated signal perturbed with the pipetting and blotting error in a chronological and randomized manner. Only the randomized procedure does not change the characteristics of the true signal, as the smoothed data show (C). (D) The residuals of the perturbed to the true signal exhibit a strong autocorrelation for the chronological procedure, which is not agreeable with white noise. This can be achieved by randomizing.

tion factor:

$$\alpha = \frac{\sqrt{\sum_j \left(\tilde{x}_{rand}(t_j) - x^*(t_j)\right)^2}}{\sqrt{\sum_j \left(\tilde{x}_{chron}(t_j) - x^*(t_j)\right)^2}}$$

For the illustrated data set the achieved reduction of the standard deviation was $\alpha \approx 0.4$. The reduction can only be quantified when the actual values are available, which is not the case in experimental measurements. Hence, the question arises whether a general error reduction factor can be established by randomizing or whether it depends on experimental parameters like the number of lanes, strength of signal maximum, blotting error or pipetting error. A simulation study showed that for small pipetting errors an error reduction factor of 0.45 ± 0.1 could be established independently from other parameters. Details of the study follow in the next section.

A. Quantifying the Error Reduction using Randomization and Calibrators

To determine the usefulness of randomization for the improvement of data quality, several parameters were varied including the number of lanes (10 to 100, Fig. 3A), the number of sine periods of the blotting error (0.8 to 2.2, Fig. 3B), the strength of the blotting error (ratio of smallest to largest value ranging from 1.5 to 10, Fig. 3C), the maximum signal strength (0.1 to 20, Fig. 3D), and the strength of the pipetting error (σ ranging from 0 to 1, Fig. 4). During the variation of one parameter, the other parameters were fixed:

- Number of lanes: 20
- Number of sine periods of the blotting error: 1
- Strength of the blotting error (max/min): 3
- Maximum signal strength: 2

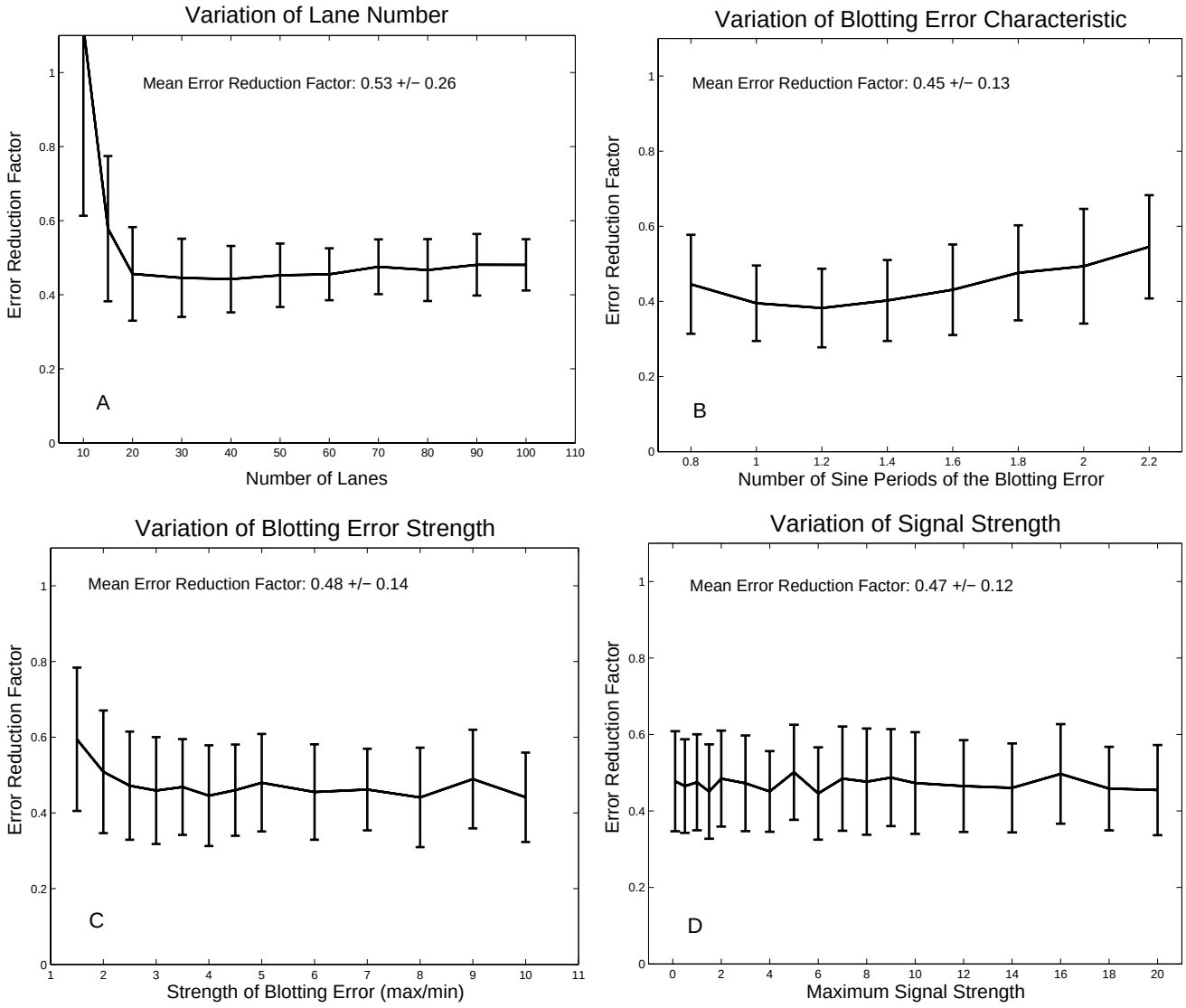

FIG. 3: Reduction of standard deviation of smoothed simulated blotting data by randomization. (A) For gels with more than 15 gel lanes randomization reduces robustly standard deviation by 50%. This holds also for variation of blotting error characteristics like number of periods (B) or error strength (C) and for a wide range of signal maximum (D).

- Standard deviation of the pipetting error: 0.1

Figures 3 and 4 display the error reduction factor for all parameter variations. At least 20 lanes should be used to achieve an optimal improvement. For the other investigated parameter ranges no strong effect is observed for all variations except for the strength of the pipetting error (Fig. 4). Since pipetting errors are uncorrelated, they cannot be reduced by randomization - if the fraction of the pipetting errors increases, the randomization takes less effects. In general, randomization decreases the standard deviation in quantitative immunoblotting to ca. 0.45 of the value without randomization, as long as the pipetting error is not too large. An approach to control the pipetting error in experiments is sampling the same number of cells for each time point or measuring and adjusting total protein concentration.

B. Criteria for Employing Normalization with Normalizers and Calibrators

Calibrators and normalizers possess a constant concentration. Fluctuations occur only as measurement errors. Since the blotting error changes slowly from lane to lane and other errors like the pipetting error are rather uncorrelated, the blotting error can be estimated by smoothing the calibrator or normalizer signal, e.g. with a smoothing spline. Based on this blotting error estimate, the protein of interest can be normalized.

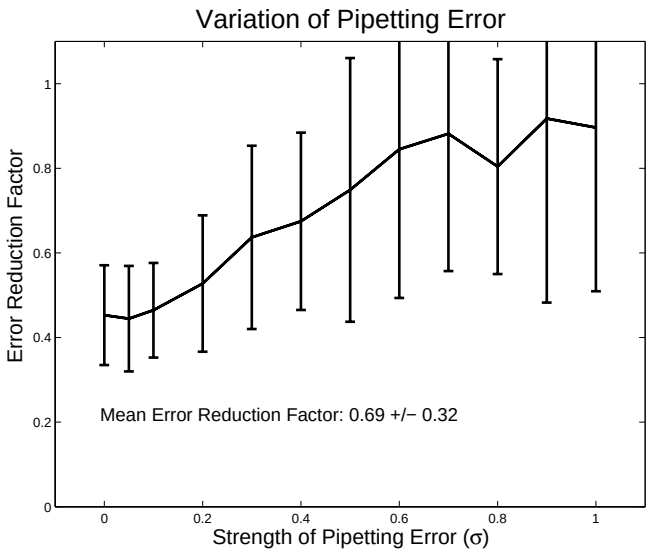

FIG. 4: Reduction of standard deviation of smoothed simulated blotting data by randomization for variation of the pipetting error strength. Increasing the pipetting error – blotting error ratio decreases the error reduction factor, since only blotting errors are tackled by randomization.

However, since the blotting error is a local property of the gel, normalizers and calibrators are required with a similar molecular weight as the protein of interest. If the molecular weight of the normalizer is different it does not reflect the blotting error for the protein of interest. We therefore developed criteria for employing normalizers and calibrators.

Figure 5A shows a simulated blotting error and a good estimation, corresponding to the smoothed signal of an appropriate normalizer in a real experiment. Smoothing the processed randomized signal leads to an acceptable estimation of the true signal. Smoothing the normalized signal yields virtually the true signal itself, as shown in Figure 5C. Even the correlation structure of the estimation error in the gel domain is improved.

The estimation of the blotting error displayed in Figure 6A is inaccurate: A strong phase shift can be observed, corresponding to a skewed gradient of the blotting error depending on the position on the blot. In this situation, normalizing the data increases the deviation of the estimated signal from the true signal. Hence, a criterion whether a normalization is applicable would be a decreased standard deviation of the estimated signal. This, though, requires knowledge of the true time course, which is not available. Instead, the smoothed curve of the randomized but not yet normalized signal is used as preliminary estimator of the true signal. If the normalizer is applicable, a new estimate can be calculated based on the randomized and normalized data, otherwise the former estimate is kept.

The shown simulated data sets have the following standard deviations:

- Figure 5:
  - Randomized (true): 0.533
  - Randomized (estimation): 0.722
  - Randomized, normalized (true): 0.157
  - Randomized, normalized (estimation): 0.515
- Figure 6:
  - Randomized (true): 0.533
  - Randomized (estimation): 0.722
  - Randomized, normalized (true): 1.208
  - Randomized, normalized (estimation): 1.068

The estimated error decreases in case of Figure 5 and increases in case of Figure 6 if a normalizer is used. Hence, the normalization procedure is only applicable in the first case, reducing the true standard deviation from 0.533 to 0.157. In the other case it would increase the standard deviation from 0.533 to 1.208. This procedure works robustly for normalizers and calibrators, as long as randomized gel loading is applied.

C. Application to Stimulation Experiment

The randomizing and normalization procedure was applied on an erythropoietin (Epo)-induced time-course experiment resulting in phosphorylation of ERK1 and ERK2. Samples were loaded randomized and separated on 17.5% SDS polyacrylamide gel, and transferred to membranes that were developed with chemiluminescent substrates and quantified with the Lumi-Imager (Figure 7). We calculated the standard deviation of the signals to their spline approximation to 2.524 for pERK1 and 0.455 for pERK2. Normalization with βActin reduced the standard deviation to the spline approximation to 1.878 for pERK1 and 0.262 for pERK2. The reduced lane-correlation for the normalized data confirms the quality of data processing. In this case the correlation structure of the systematical blotting error could be disrupted validating the normalization.

IV. CALCULATION OF MOLECULES PER CELL

A. Linearity of imaging unit

Quantification of a protein P measured by immunoblotting is performed via chemiluminescence detection yielding total intensities P_{blu} which are proportional

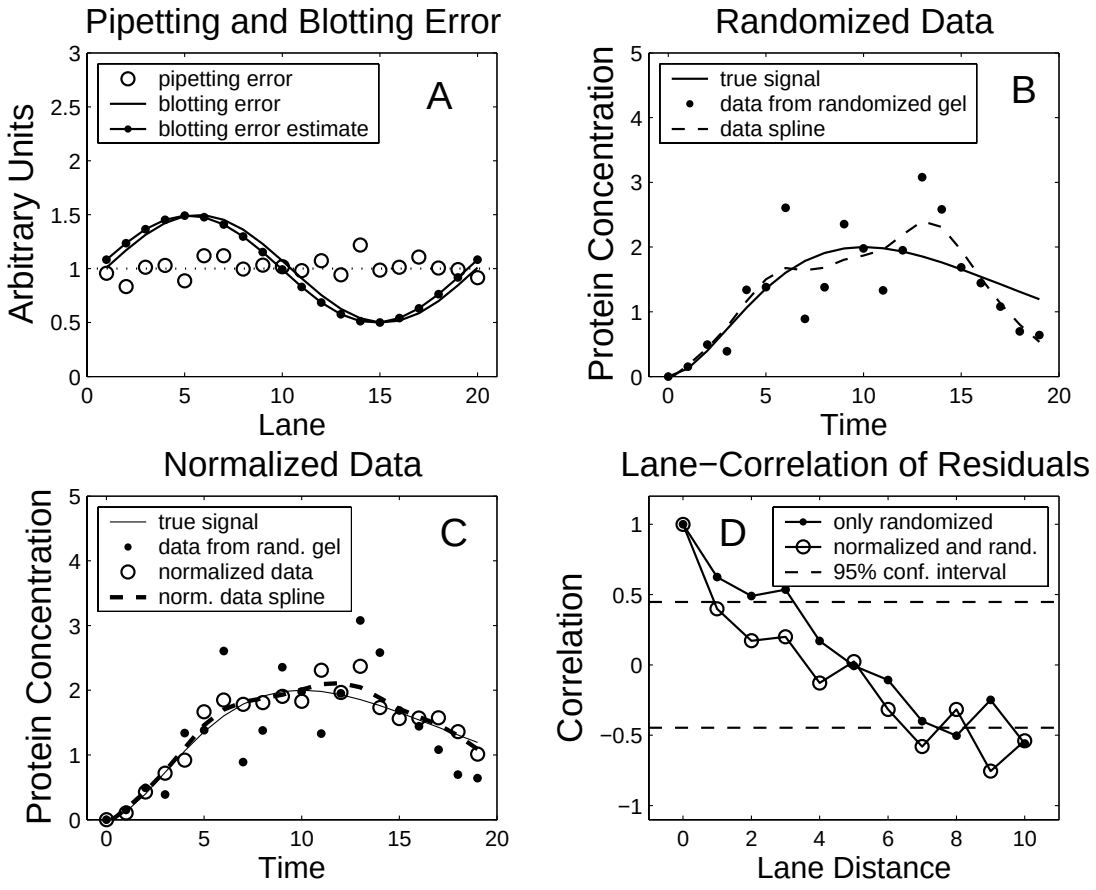

FIG. 5: Normalization of simulated time-course data depicting a valid procedure according to our criteria. (A) The blotting error is very well estimated corresponding to a suitable normalization protein. The perturbation of a simulated signal (B) is strongly reduced after normalization (C). Even the correlation in gel domain of the residuals is improved (D). The autocorrelation, i.e., the correlation in time domain, agrees for both randomized signals with white noise (not shown).

to the total number of molecules P_{tmlc} on the blot (Fig. 8. The linear relationship reads

$$P_{tmlc} = a\, P_{blu}$$

with a proportionality factor a and 0 y-axis interception. The factor a has to be determined for each protein species and for every blot, since the amount of antibody added varies for different blots and the antibody affinity differs for different proteins.

B. Requirements for the standard/calibrator protein

The reference protein R realized by a standard or calibrator protein should

- contain the same epitope binding to the antibody as the protein of interest,
- have a known molecular weight R_{mw},
- be added to the lysate with a known amount R_g.

C. Calculation of the proportionality factor

The total number of reference proteins in the lysate is given as

$$R_{tmlc} = \frac{N_A\, R_g}{R_{mw}},$$

with Avogadro constant $N_A = 6.022 \cdot 10^{23}$. If the imaging unit measures the intensity R_{blu}, the proportionality factor can be calculated as

$$a = \frac{R_{tmlc}}{R_{blu}} = \frac{R_{tmlc}\, N_A\, R_g}{R_{blu}\, R_{mw}}.$$

If possible, one should measure the reference protein several times and estimate a by linear regression. This provides also a standard deviation for a.

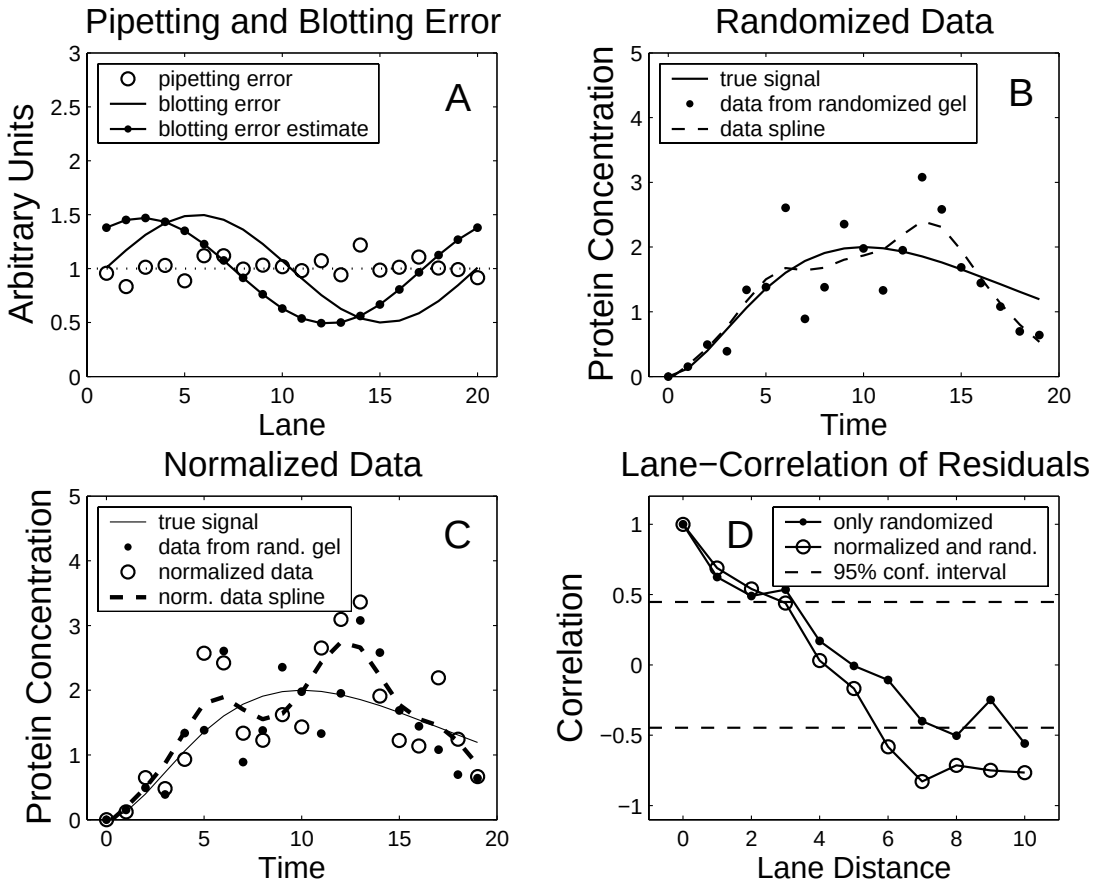

FIG. 6: Normalization of simulated time-course data depicting a rejected procedure according to our criteria. (A) The blotting error estimate is phase shifted, corresponding with a too distant normalization protein. The perturbation of a simulated signal (B) cannot be reduced with normalization (C) and the lane-to-lane correlation is not improved (D).

D. Calibrators

Calibrator proteins harbor the same antibody epitope as the protein of interest, P, yet possess a different molecular weight than P resulting in a distinct band in the immunoblot analysis. If analysis of total cellular lysates are performed, a few lanes of the immunoblot have to be used for the standard protein to facilitate parallel detection.

E. Calculation of molecules per cell

P_{tmlc} is the total number of molecules of the investigated lysate. If the number of cells in the lysate is available, the molecule number per cell can be calculated as

$$P_{mlc} = \frac{P_{tmlc}}{\#\ \text{cells}}.$$

[1] P. Green and B. Silverman, *Nonparametric Regression and Generalized Linear Models* (Chapman and Hall, London, 1994).
[2] A. Buja, T. Hastie, and R. Tibshirani, Ann. Stat. **17**, 453 (1989).
[3] P. Craven and G. Wahba, Numer. Math. **31**, 377 (1979).
[4] S. Wood, J. Royal Statistical Soc. B. **65(1)**, 95 (2003).

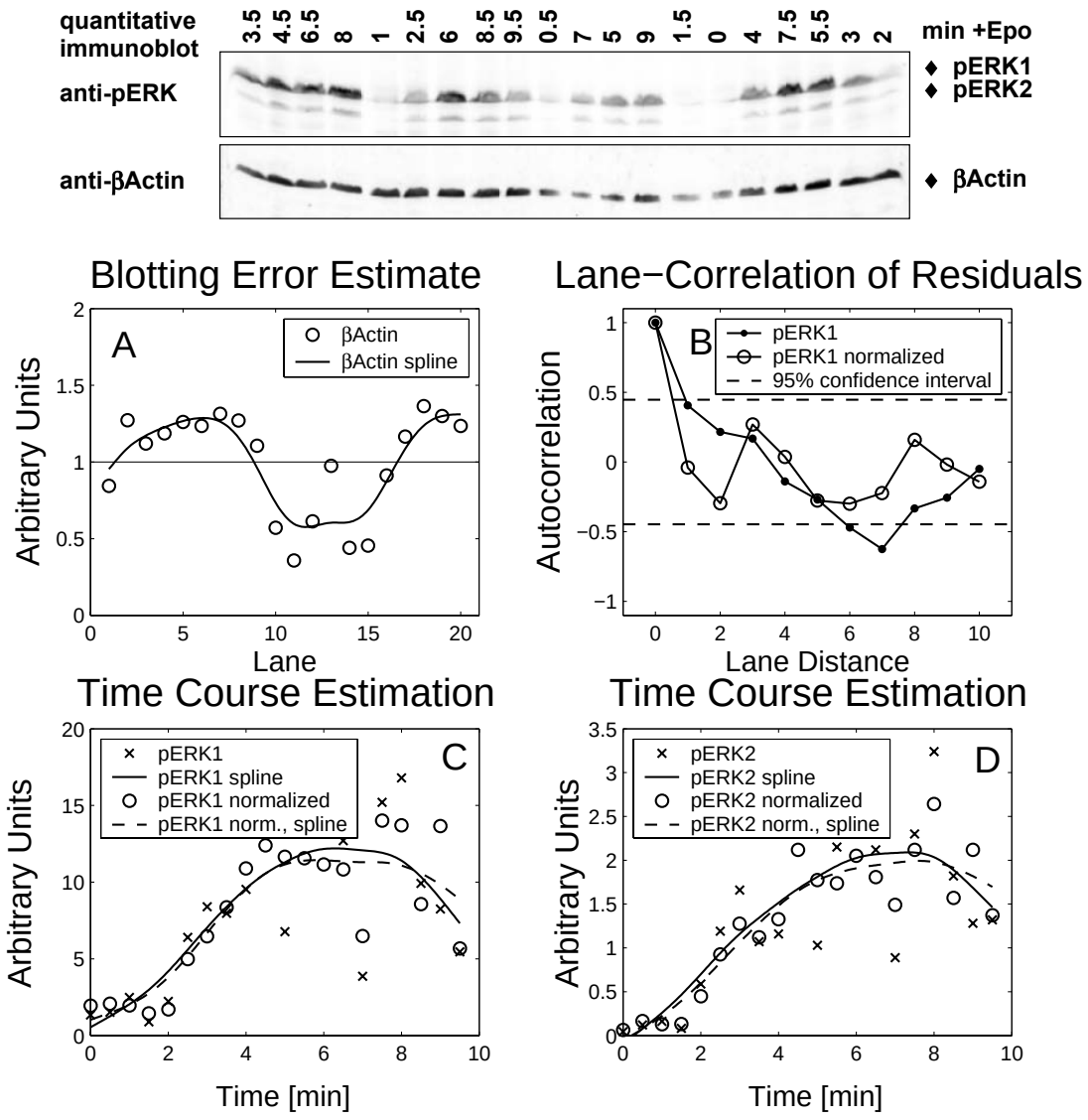

FIG. 7: Randomization and normalization of an Erythropoietin-induced time-course experiment. BaF3-EpoR cells are stimulated with 50 units/ml Epo resulting in ERK phosphorylation. Gel electrophoresis has been applied with a randomized, non-chronological gel loading with βActin as normalizer protein (upper panel). (A) Smoothed measurements of βActin serve as estimate of the strong, sine-like blotting error. (C, D) Normalization reduces significantly standard deviation of pERK1/2 measurements compared to a spline-smoothed pERK1/2 signal (cont. line), which serves as first estimate of the true signal.

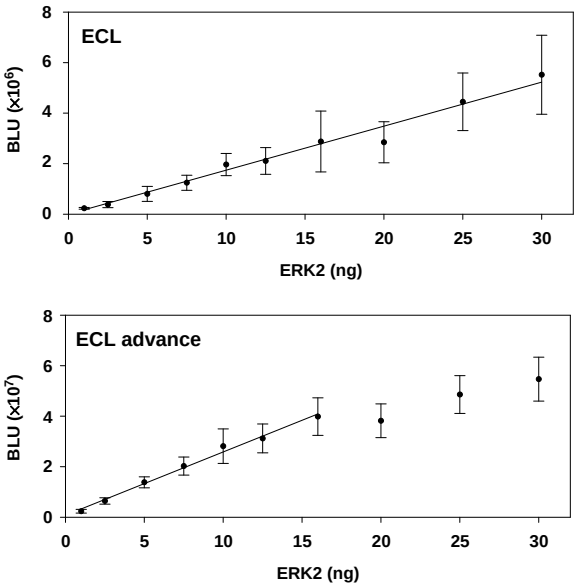

FIG. 8: A dilution series of purified ERK2 was separated eight times by a 10% SDS polyacrylamide gel and transferred to a membrane that was probed with anti-ERK antibody and subsequently developed with ECL or ECL advance. To determine linearity, the amount of ERK2 was plotted versus measured signal strength. Signals were linear up to 4 x 10^7 BLU.