Supporting Information to

Estimation of Migration-time and Mobility Distributions in Organelle Capillary Electrophoresis with Statistical-Overlap Theory

by

Joe M. Davis*

Department of Chemistry and Biochemistry

Southern Illinois University at Carbondale

Carbondale, IL 62901 USA

and

Edgar A. Arriaga

Department of Chemistry

University of Minnesota

Minneapolis, MN 55455 USA

*corresponding author

Current address: 733 Schloss Street

Wrightsville Beach, NC 28480

Phone: 910 256 4235

e-mail: chimicajmd@ec.rr.com

This Supporting Information contains five parts. Each part identifies the relevant section of the main article. The terminology is the same as the main article.

Part 1. Conversion between α and α_e

The conversions between saturation α and effective saturation α_e are mentioned in the paragraph of the main article after eq 2b. They are appropriate for an exponential distribution of peak heights. Over the range, $0 \le \alpha_e \le 25$, α_e can be converted to α using the empirical expression[1]

$$\alpha = \frac{0.725\alpha_e}{1+\delta_1\alpha_e^{\delta_2}} \tag{S-1}$$

with $\delta_1 = 0.1942 \pm 0.0005$ and $\delta_2 = 0.930 \pm 0.001$. Over the range, $0 \le \alpha \le 3.85$, α is converted to α_e on dividing α by R_s^*, as expressed by eq 4c in the main article.

Part 2. Standard deviation σ_{m_e} of m_e distribution

The standard deviation σ_{m_e} is first mentioned in the paragraph of the main article containing eq 4.

Theory. Consider a separation ensemble containing peaks of constant density and width, with peak numbers and migration times obeying Poisson statistics. Each member of the ensemble contains m peaks and p observed peaks. Consider equating p to

$$\overline{p} = \overline{m}\exp(-\alpha) \tag{S-2}$$

where

$$\alpha = 4\overline{m}\sigma R_s^* / X \tag{S-3}$$

and solving for $\overline{m}$. (Eqs S-2 and S-3 are eqs 1b and 1a, respectively, in the main article.) Here, $\overline{m}$ and $\overline{p}$ are the mean numbers of peaks and observed peaks in the ensemble, α is the saturation, σ is the peak standard deviation, R_s^* is the average minimum resolution (which depends on α), and X is the duration of the separation. We assume $\overline{m}$ is the only unknown, with σ, X, and R_s^* being known or calculable. The determined $\overline{m}$ is interpreted as the estimated number of peaks, m_e.

We assume that m_e is distributed randomly about $\overline{m}$, just as p is distributed randomly about $\overline{p}$. The difference $\Delta m = m_e - \overline{m}$ is related to the difference $\Delta p = p - \overline{p}$ by the first-order Taylor series

$$\Delta m \approx \frac{\partial \overline{m}}{\partial \overline{p}} \Delta p = \frac{\partial \overline{m}}{\partial \alpha} \frac{\partial \alpha}{\partial \overline{p}} \Delta p \qquad \text{(S-4)}$$

where the final identity results from the chain rule (the derivatives are partials because σ / X is constant). The theory of the propagation of errors[2] is applied to J independent members of the ensemble by adding the squares of the left- and right-hand sides of eq S-4 and dividing by J

$$J^{-1} \sum_{i=1}^{J} (m_{e,i} - \overline{m})^2 \approx J^{-1} \left[\frac{\partial \overline{m}}{\partial \alpha} \frac{\partial \alpha}{\partial \overline{p}} \right]^2 \sum_{i=1}^{J} (p_i - \overline{p})^2 \qquad \text{(S-5)}$$

where p_i and $m_{e,i}$ are the p and m_e values of the i^{th} ensemble member. The derivatives are factored out of the sum in eq S-5, because they are the same for every ensemble member. As J approaches infinity, eq S-5 has the limit

$$\sigma_{m_e}^2 = \left[\frac{\partial \overline{m}}{\partial \alpha} \frac{\partial \alpha}{\partial \overline{p}} \right]^2 \sigma_p^2 \qquad \text{(S-6)}$$

where $\sigma_{m_e}^2$ and σ_p^2 are the variances of m_e and p, respectively. For peaks that are Poisson distributed, σ_p^2 is

$$\sigma_p^2 = \overline{m} \exp(-\alpha)[1 - 2\alpha \exp(-\alpha)] \qquad \text{(S-7)}$$

(Eq S-7 is eq 3 in the main article.)

The derivatives in eq S-6 are evaluated simply. In accordance with eqs S-2 and S-3, $\overline{m}$ and $\overline{p}$ equal

$$\overline{m} = \alpha / (zR_s^*); \;\; \overline{p} = \alpha \exp(-\alpha) / (zR_s^*) \qquad \text{(S-8)}$$

with $z = 4\sigma / X$. For constant peak widths (i.e., constant z)

$$\frac{\partial \overline{m}}{\partial \alpha} = \overline{m}\left[\alpha^{-1} - d\ln R_s^* / d\alpha\right] \tag{S-9}$$

and

$$\frac{\partial \overline{p}}{\partial \alpha} = \overline{m}\exp(-\alpha)\left(\alpha^{-1} - 1 - d\ln R_s^* / d\alpha\right) \tag{S-10}$$

where $d\ln R_s^* / d\alpha$ is $[dR_s^* / d\alpha] / R_s^*$.

We want the reciprocal of eq S-10, $\partial\alpha / \partial\overline{p}$, which in combination with eqs S-6 and S-9 gives

$$\sigma_{m_e}^2 = [g(\alpha)]^2 \sigma_p^2 \tag{S-11a}$$

with

$$g(\alpha) = \frac{\alpha^{-1} - d\ln R_s^* / d\alpha}{\exp(-\alpha)[\alpha^{-1} - 1 - d\ln R_s^* / d\alpha]} \tag{S-11b}$$

where $g(\alpha)$ is the reciprocal of the slope, $\partial\overline{p} / \partial\overline{m}$, of the $\overline{p}$ vs $\overline{m}$ curve. When combined with eq S-7, the square root of eq S-11a gives σ_{m_e} as a function of $\overline{m}$ and α

$$\sigma_{m_e} = \left(\overline{m}[1 - 2\alpha\exp(-\alpha)]\right)^{1/2} \frac{\exp(\alpha / 2)[\alpha^{-1} - d\ln R_s^* / d\alpha]}{\alpha^{-1} - 1 - d\ln R_s^* / d\alpha} \tag{S-12}$$

The coefficient of variation (CV), $100\sigma_{m_e} / \overline{m}$, can be calculated from eq S-12. On substituting the left-hand side of eq S-8 for $\overline{m}$ in eq S-12 and the CV, we obtain

$$\sigma_{m_e}(\sigma / X)^{1/2} = \frac{1}{2}\left(\frac{\alpha}{R_s^*}[1 - 2\alpha\exp(-\alpha)]\right)^{1/2} \frac{\exp(\alpha / 2)[\alpha^{-1} - d\ln R_s^* / d\alpha]}{\alpha^{-1} - 1 - d\ln R_s^* / d\alpha} \tag{S-13a}$$

$$\frac{CV}{(\sigma / X)^{1/2}} = 200\left(\frac{R_s^*}{\alpha}[1 - 2\alpha\exp(-\alpha)]\right)^{1/2} \frac{\exp(\alpha / 2)[\alpha^{-1} - d\ln R_s^* / d\alpha]}{\alpha^{-1} - 1 - d\ln R_s^* / d\alpha} \tag{S-13b}$$

which are eqs 4a and 4b in the main article. Both R_s^* and $d\ln R_s^*/d\alpha$ can be evaluated from eq 4c in that article.

The denominator of the final factor in eq S-13 is the negative of the bracketed term in eq 5 of the main article. The latter is proportional to the slope, $\partial\overline{p}/\partial\overline{m}$. Thus, σ_{m_e} and the CV approach infinity as $\partial\overline{p}/\partial\overline{m}$ approaches zero.

In accordance with Poisson statistics, the standard deviation σ_m of the number of peaks in the ensemble is $\sqrt{\overline{m}}$. It differs from σ_{m_e}, which is the standard deviation of the *estimated* numbers of peaks m_e. The ratio σ_{m_e}/σ_m is calculated from eq S-12 as

$$\frac{\sigma_{m_e}}{\sigma_m} = \left(1-2\alpha\exp(-\alpha)\right)^{1/2}\frac{\exp(\alpha/2)[\alpha^{-1}-d\ln R_s^*/d\alpha]}{\alpha^{-1}-1-d\ln R_s^*/d\alpha} \tag{S-14}$$

and depends only on α.

Results and discussion. Figure S-1a is a graph of the ratio σ_{m_e}/σ_m vs saturation α, as calculated from eq S-14. As the saturation approaches zero, the ratio approaches one. This is expected, since at zero saturation all peaks are resolved and $p = m = m_e$ for each ensemble member. As α increases, σ_{m_e}/σ_m rapidly increases (e.g, $\sigma_{m_e} = 3.40\sigma_m$ at $\alpha = 1$) and the precision of m_e decreases.

As discussed in the main article, the poor precision of m_e at high saturation results from the decreasing slope $\partial\overline{p}/\partial\overline{m}$, which maps small random fluctuations of p into large random fluctuations of m_e. Figure S-1b is a graph of $g(\alpha)$ vs α, where $g(\alpha)$, eq S-11b, is the reciprocal of this slope. It resembles Figure S-1a but increases with α even more rapidly. Values of $g(\alpha)$ agree with the reciprocal slope determined numerically from the graph of $\overline{p}$ vs $\overline{m}$.

Part 3. Monte-Carlo simulation of m_e distribution

The Monte-Carlo simulations are mentioned in the main article after eq 4c.

Procedures. To characterize the m_e distribution, 2 x 10^5 Monte-Carlo simulations of p and m, and calculations of m_e, were made. In each simulation, a Poisson distributed number m of Gaussian peaks having constant standard deviation σ and exponentially random heights spanned an interval of duration X, with peak overlap producing p observed peaks (maxima). This p then determined m_e via eqs S-2 and S-3 in Part 2 of the Supporting Information. Discrete distributions were built from the p, m, and m_e values. Further details are given in the Procedures section of the main article.

Results and Discussion. The panels in Figure S-2 are graphs of probability vs p, m, and m_e at different saturations α, as determined for σ / X = 8 x 10^{-5}. All distributions are discrete but are shown as continuous functions for simplicity. At low saturation (e.g., $\alpha = 0.2$), the m and m_e distributions are almost identical. As α increases, the m_e distribution becomes broader than the m distribution, and its average shifts slightly downward from the mean of the m distribution, $\overline{m}$. The shift occurs, because eq S-2 slightly underestimates peak overlap as α increases. Values of σ_{m_e} calculated from eq S-12 in Part 2 of the Supporting Information and from moments analysis of the m_e distributions are reported in the panels. At low α, excellent agreement is found. At higher α, eq S-12 overpredicts the standard deviation of m_e. In all cases, σ_{m_e} exceeds the standard deviation of the m distribution, which is $\sqrt{\overline{m}}$ (and calculable from the $\overline{m}$'s reported in the panels). Further verification of eq S-12 is provided in the main article.

Part 4. Equations for migration-time distributions $f(\zeta)$

The equations are mentioned in the main article at the end of the first paragraph in the section, "Analysis of Migration-Time Distributions", under Procedures. All migration-time distributions $f(\zeta)$ are models, have unit area, and are bound by the reduced times, $\zeta = 0$ and $\zeta = 1$. Normalization coefficients were obtained by integrating $f(\zeta)$ between these bounds. Various coefficients were selected by trial and error to obtain the desired appearance of $f(\zeta)$. Equations are given for the reduced migration times ζ_c of peaks. All random numbers R are uniform and bound by the integers, 0 and 1.

Gaussian $f(\zeta)$. The Gaussian migration-time distribution is

$$f(\zeta) = (2\pi\sigma_G^2)^{-1} \exp[-(\zeta - \mu)^2 / (2\sigma_G^2)] \tag{S-15a}$$

with $\mu = 0.5$ and $\sigma_G = 0.125$. A peak migration time ζ_c was calculated with the Box-Muller transform[3]

$$\zeta_c = \mu + \sigma_G \left[\sqrt{-2 \ln R_1} \sin(2\pi R_2) \right] \tag{S-15b}$$

where R_1 and R_2 are independent random numbers (the radicand in eq S-15b is positive because $\ln R_1$ is negative).

Strictly, eq S-15a is not normalized between $\zeta = 0$ and $\zeta = 1$. However, the area between these bounds is erf$(4/\sqrt{2}) \approx 0.9999^+$, where erf is the error function. The very small fraction (< 0.01%) of peak migration times generated by eq S-15b outside the bounds was not used; its discard had negligible effect on results.

Bimodal $f(\zeta)$. The bimodal migration-time distribution is

$$f(\zeta) = A_1 \left\{ \zeta \exp(-\kappa_1 \zeta) + (1 - \zeta) \exp(-\kappa_2 [1 - \zeta]^2) \right\} \tag{S-16a}$$

with κ_1 and κ_2 equaling constants (in the main article, $\kappa_1 = 6.5$ and $\kappa_2 = 8.5$). The normalization constant A_1 is

$$A_1 = \left\{ [1 - \exp(-\kappa_2)] / (2\kappa_2) - \exp(-\kappa_1) / \kappa_1 + [1 - \exp(-\kappa_1)] / \kappa_1^2 \right\}^{-1} \tag{S-16b}$$

A peak migration time ζ_c was determined by solving

$$\begin{gathered} (\kappa_1)^{-2} - \exp(-\kappa_2) / (2\kappa_2) - \exp(-\kappa_1 \zeta_c)(\zeta_c + \kappa_1^{-1}) / \kappa_1 \\ + \exp(-\kappa_2 [1 - \zeta_c]^2) / (2\kappa_2) - R / A_1 = 0 \end{gathered} \tag{S-16c}$$

using bisection, with R equaling a random number. Eq S-16c is based on a transformation of random numbers into an arbitrary distribution[4] (here, into eq S-16a).

Asymmetric $f(\zeta)$. The asymmetric migration-time distribution is

$$f(\zeta) = A_2 \zeta \exp(-\kappa_3 \zeta) \tag{S-17a}$$

with κ_3 equaling a constant (in the main article, $\kappa_3 = 3.5$). The normalization constant A_2 is

$$A_2 = \left\{[1-\exp(-\kappa_3)]/\kappa_3^2 - \exp(-\kappa_3)/\kappa_3\right\}^{-1} \tag{S-17b}$$

A peak migration time ζ_c was determined by solving

$$[1-\exp(-\kappa_3\zeta_c)]/\kappa_3^2 - \zeta_c \exp(-\kappa_3\zeta_c)/\kappa_3 - R/A_2 = 0 \tag{S-17c}$$

using bisection, with R equaling a random number. Eq S-17c is based on the same transformation as eq S-16c.

Constant $f(\zeta)$. The constant migration-time distribution is

$$f(\zeta) = 1 \tag{S-18a}$$

A peak migration time ζ_c was determined as

$$\zeta_c = R \tag{S-18b}$$

with R equaling a random number.

Part 5. Least square fits to graphs of α_t, $\alpha_{e,t}$, and $\log(\bar{m}_t)$ vs $\log(\sigma / X)$

The fits are mentioned in the main article at the end of the section, " Threshold Values of $\bar{p}$ versus $\bar{m}$ Curve", in the Results and Discussion. Let $s = \log(\sigma / X)$. The graphs in Figure 3a of the main article can be estimated from the polynomial fits

$$\alpha_t = -0.9530 - 0.9848s - 0.1442s^2 - 0.007830s^3 \tag{S-19}$$

$$\alpha_{e,t} = -1.423 - 1.287s - 0.0877s^2 \tag{S-20}$$

$$\log(\bar{m}_t) = -1.986 - 1.949s - 0.1718s^2 - 0.01102s^3 \tag{S-21}$$

with correlation coefficients of 0.99998 or larger.

References

1. Davis, J.M.; Carr, P.W. *Anal. Chem.* **2009**, *81*, 1198-1207.
2. Bevington, P.R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw Hill Book Company: New York, 1969, pp. 56-60.
3. Dahlquist, G.; Bjorck, A. *Numerical Methods*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1974, p. 453.
4. *Ibid.*, p. 452.

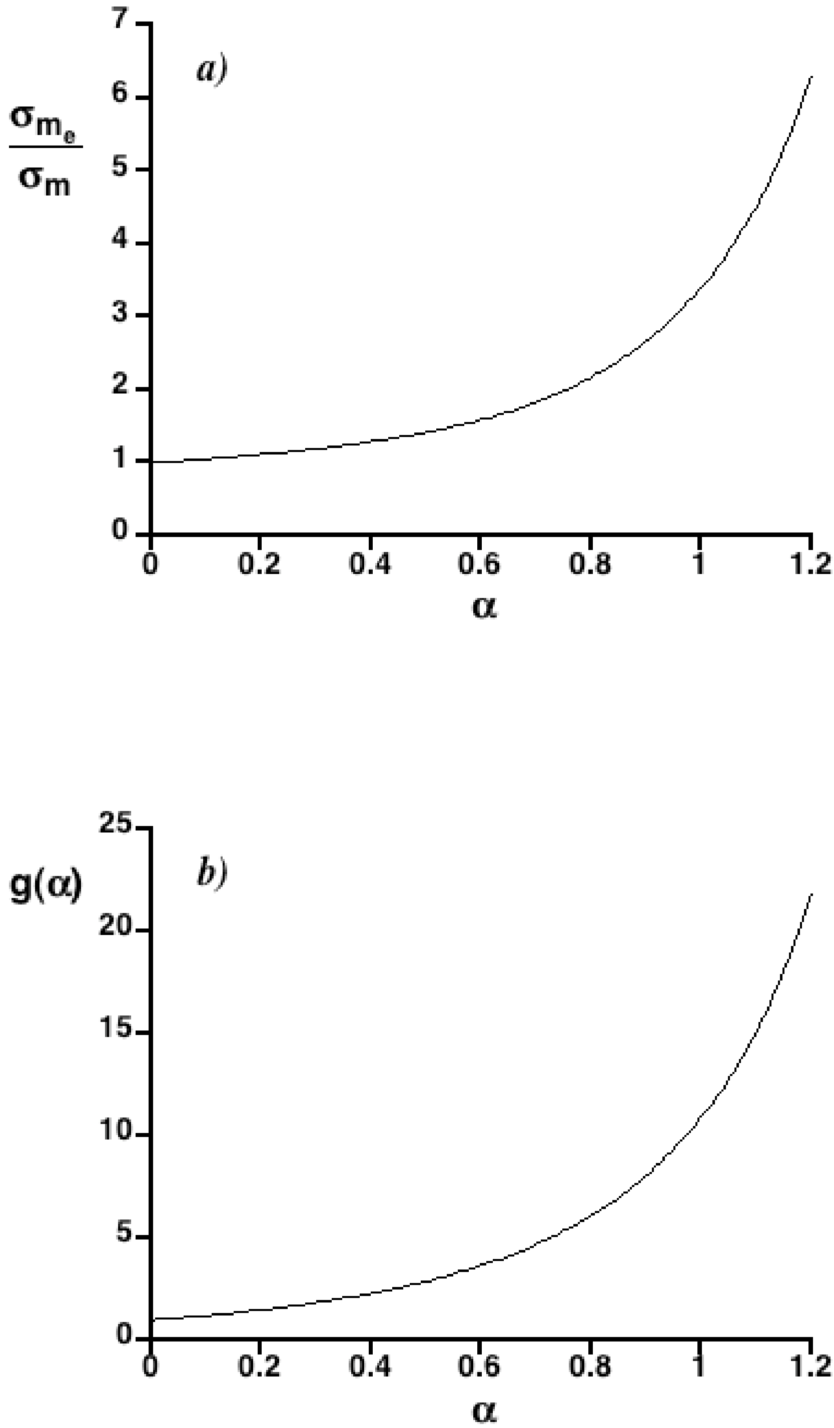
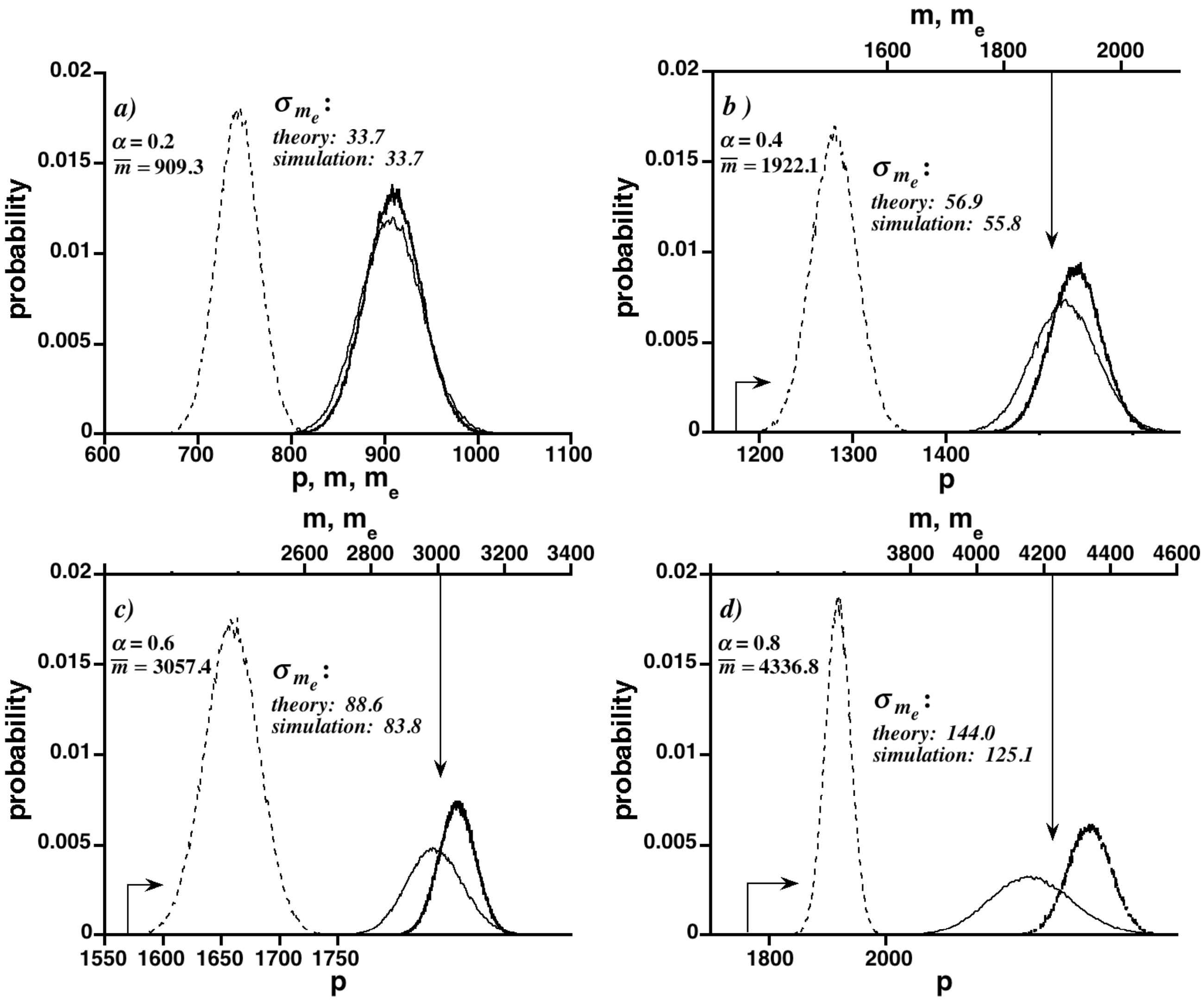


Figure S-1. a) Graph of σ_{m_e} / σ_m vs saturation α (eq S-14). b) Graph of $g(\alpha)$ vs α (eq S-11b), with R_s^* equal to eq 4c of the main article.



Caption on next page

Figure S-2. Graphs of discrete probability distributions vs p, m, and m_e, as determined by Monte-Carlo simulations (p, m) and solutions to eqs S-2 and S-3 (m_e) in Part 2 of the Supporting Information. Dashed, bold, and normal-weight curves are the p, m, and m_e distributions, respectively. In b) – d) different abscissas are used for p, and for m and m_e, to reduce unused space. σ / X = 8 x 10^{-5}. a) $\alpha = 0.2$. b) $\alpha = 0.4$. c) $\alpha = 0.6$. d) $\alpha = 0.8$.