Neuron, Volume *83*

Supplemental Information

Uniquantal Release through a Dynamic Fusion Pore Is a Candidate Mechanism of Hair Cell Exocytosis

Nikolai M. Chapochnikov, Hideki Takago, Chao-Hua Huang, Tina Pangršič, Darina Khimich, Jakob Neef, Elisabeth Auge, Fabian Göttfert, Stefan W. Hell, Carolin Wichmann, Fred Wolf, and Tobias Moser

Supplemental Figures 1-6, Movies 1-2, Tables 1-4, Experimental Procedures, and References

Content:

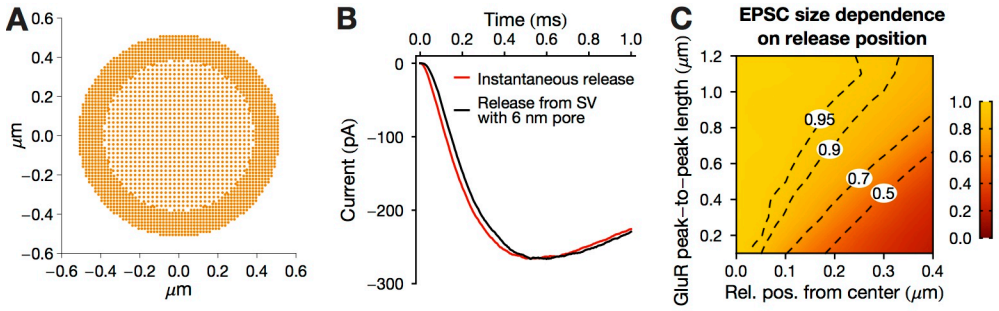

Figure S1. Model AMPAR organization in the PSD; difference between instantaneous release and release through a 6 nm pore; EPSC size as a function of release position for a disk AMPAR topography

Related to Figure 3.

(**A**) AMPAR arrangement for a peak-to-peak cluster size of 900 nm. The AMPAR density was 2,000 μm^{-2} in the inner disk and 4,000 μm^{-2} in the annulus. The resulting average density was 2,897 μm^{-2}.

(**B**) Difference in EPSC shape between instantaneous release of glutamate in the synaptic cleft and release from the SV with a fusion pore width of 6 nm. The amplitude of the average EPSCs (depicted here) were -265.77 pA and -266.4 pA, respectively. And the mean EPSC amplitudes were -280.75 pA and -280.3 pA, respectively (t test: p = 0.87).

(**C**) Normalized monophasic EPSC amplitude as a function of release position for different GluR cluster sizes. AMPAR scheme from Häusser and Roth (1997) and 3,000 glutamates released. The AMPAR cluster profile was a disk with a uniform density of 3,000 μm^{-2}. For each GluR cluster length, the EPSC amplitude was set to 1 for a release in the synapse center. To allow convenient comparison with **Figure 3G**, a disk profile with a given "GluR peak-to-peak length" had the same diameter as the overall diameter of the profile from **Figure 3C** with the same GluR peak-to-peak length. For example, a "GluR peak-to-peak length" of 0.8 μm corresponds to an outer diameter of 0.92 μm for both the disk profile and the profile from **Figure 3C**.

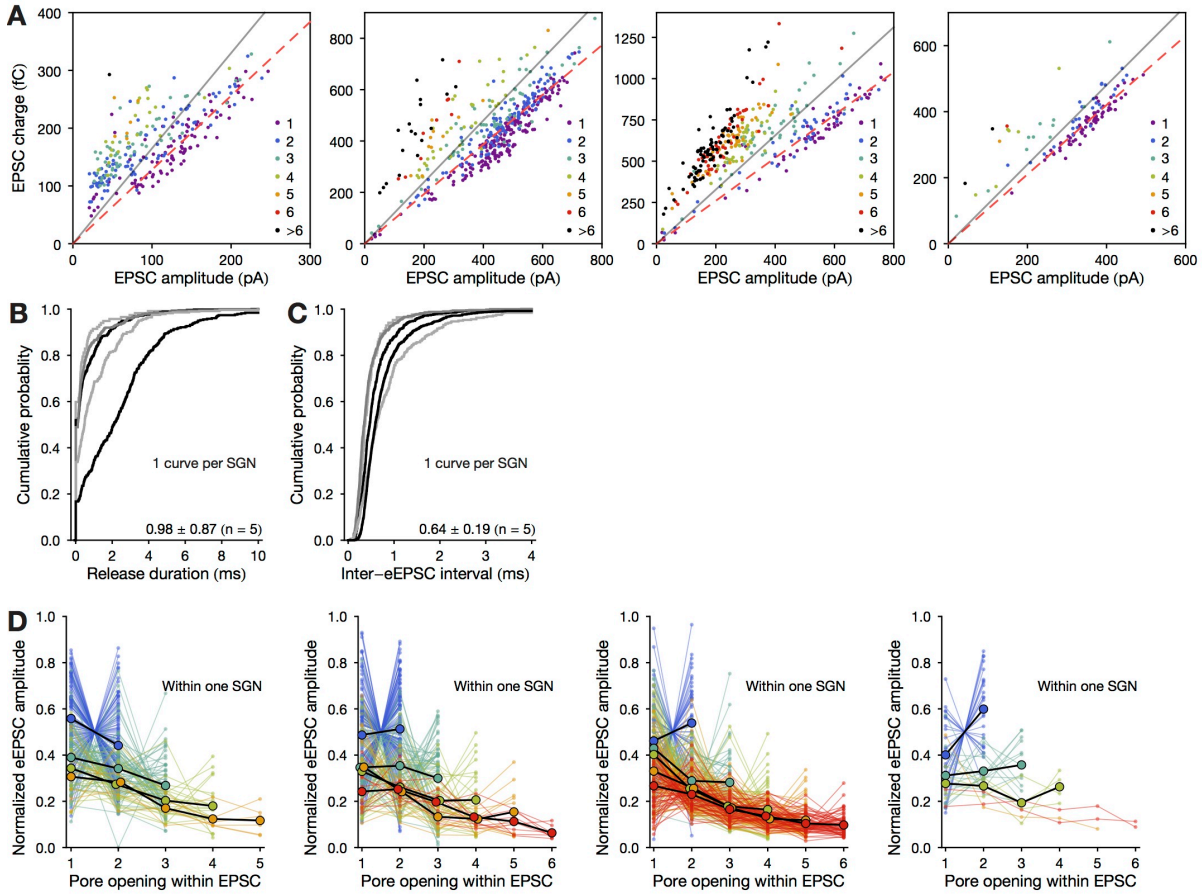

Figure S2. Number of underlying elementary release events per EPSC displayed in the EPSC charge vs. amplitude scatter plot for the four other SGNs; Release duration and inter-eEPSC-intervals; eEPSC sizes within EPSCs for the four other SGNs

Related to Figure 4.

(**A**) Same as **Figure 4C** for four other SGNs.

(**B**) Release duration of EPSCs. We defined the release duration as the interval between the first and the last underlying eEPSC in an EPSC. In 4 SGNs the distribution of durations was approximately exponential and all SGNs had EPSCs with release duration > 7 ms.

(**C**) Inter-eEPSCs intervals within EPSCs. The mean interval among SGNs was 0.64 ± 0.19 ms.

(**D**) Same as **Figure 4E** for four other SGNs.

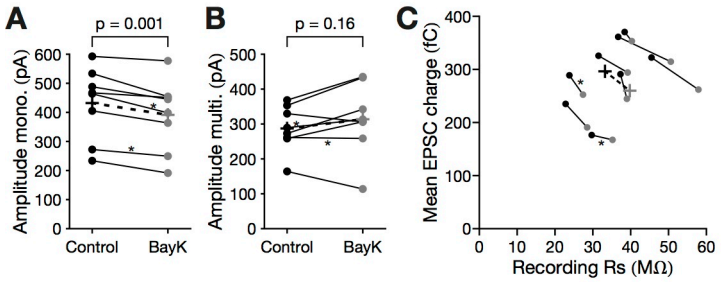

Figure S3. EPSC amplitude and charge decrease in BayK8644 can be explained by the increase in series resistance (Rs)

Related to Figure 6.

(**A**) Monophasic EPSC amplitude in control and BayK8644.

(**B**) Multiphasic EPSC amplitude in control and BayK8644.

(**C**) EPSC charge and recording series resistance (Rs) in control and BayK8644.

The data points denoted with a star (*), correspond to recordings performed at – 60 mV holding potential. Their mean amplitude and charge was multiplied by 90/60 to be more easily comparable with other recordings performed at -90 mV holding potential.

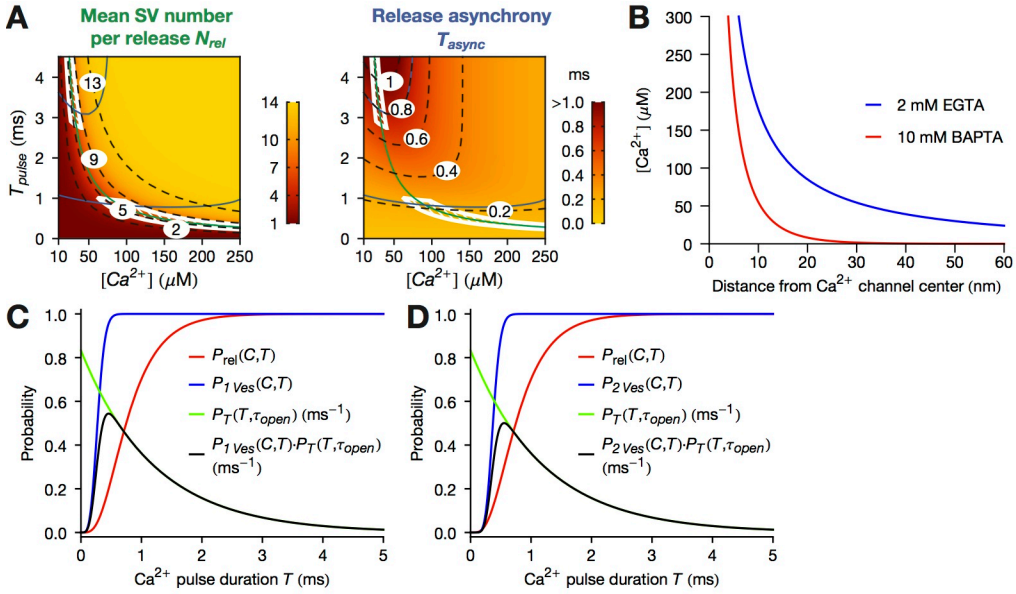

Figure S4. Screening the temporal precision and efficiency of "Ca^{2+}-synchronized MQR" model in terms of $[Ca^{2+}]$ and T_{pulse}; steady-state Ca^{2+} concentrations around an open Ca^{2+} channel; quantities used for the calculations of $N_{rel}([Ca^{2+}], \tau_{open})$ and $T_{async}([Ca^{2+}], \tau_{open})$

Related to Figure 7 and Methods.

(**A**) We calculated for which combination of pulse duration T_{pulse} and concentration $[Ca^{2+}]$ the "Ca^{2+}-synchronized MQR" model model could generate monophasic and multiphasic EPSCs in terms of the mean number of SVs released per release event N_{rel} (*left*) and of the release asynchrony T_{async} (*right*) for ranges of $[Ca^{2+}]$ (10 to 250 μM) and T_{pulse} (0 to 4.5 ms), potentially resulting from the opening of a single Ca^{2+} channel. The two blue and one green lines correspond to $T_{async,\ mono}$, $T_{async,\ multi}$, and $N_{rel,\ data}$, respectively. N_{rel} increased with $[Ca^{2+}]$ and T_{pulse}. In the region where $N_{rel} \approx N_{tot}$ (i.e., $P_{rel} \approx 1$), T_{async} mainly decreased with increasing $[Ca^{2+}]$. In the region where $N_{rel} \approx 1$ (i.e., $P_{rel} \approx 0$), T_{async} mainly decreased with T_{pulse}. The left consistency region (white stripes) corresponds to multiphasic EPSCs, and the right consistency region to monophasic EPSCs.

(**B**) Steady-state Ca^{2+} concentrations around an open Ca^{2+} channel calculated using the linear approximation (Naraghi and Neher, 1997) assuming a Ca^{2+} channel current of 0.5

pA and a single buffer species (see legend, with k_{on} and k_{off} rates from (Naraghi and Neher, 1997)). For example, a $[Ca^{2+}]$ = 180 μM is reached at approximately 10 nm from the open Ca^{2+} channel center, assuming mild Ca^{2+} buffering (2 mM Ca^{2+} binding sites with EGTA-like kinetics).

(**C**) Quantities used for the calculation of $N_{rel}(C,\tau_{open})$, where C is the Ca^{2+} concentration and τ_{open} is the Ca^{2+} channel mean open time. T is the Ca^{2+} pulse duration.

Red: SV release probability $P_{rel}(C,T)$. *Blue*: probability $P_{1Ves}(C,T)$ that at least one SV is released as a result of the Ca^{2+} pulse, calculated with equation (11) from $P_{rel}(C,T)$. *Green*: probability $P_T(T,\tau_{open})$ that the Ca^{2+} channel opens for a duration T. *Black*: $P_{1Ves}(C,T) \cdot P_T(T,\tau_{open})$ used in equation (9) for the calculation of $N_{rel}(C,\tau_{open})$.

(**D**) Quantities used for the calculation of $T_{async}(C,\tau_{open})$.

Red: SV release probability P_{rel} (C,D). *Blue*: probability $P_{2Ves}(C, T)$ that least two SVs are released, calculated with equation (12) from $P_{rel}(C,T)$. *Green*: probability $P_T(T,\tau_{open})$ that the Ca^{2+} channel opens for a duration T. *Black*: $P_{2Ves}(T,\tau_{open}) \cdot P_T(C,T)$ used in equation (10) for the calculation of $T_{async}(C,\tau_{open})$.

Parameters: τ_{open} = 1.2 ms, C = 120 μM, N_{tot} = 14, (Beutner et al., 2001) secretion model with γ = 10 ms^{-1}.

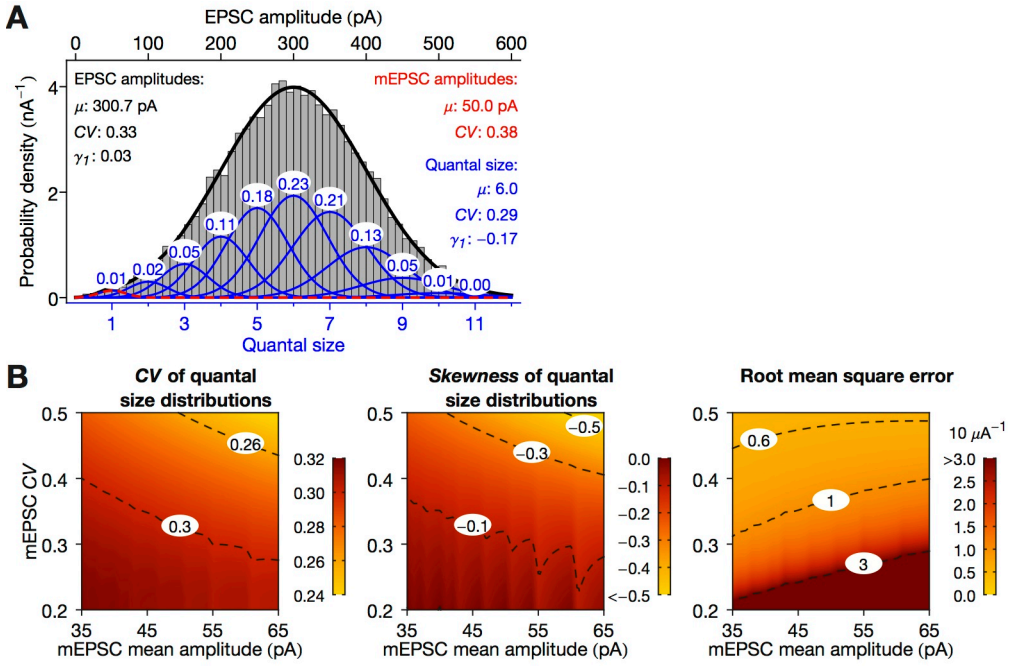

Figure S5. Quantal size distributions underlying the EPSC distributions in MQR Related to Figure 8.

(**A**) Artificially generated EPSC amplitude histogram with amplitudes drawn from a normal distribution with mean of 300 pA and SD = 100 pA, corresponding to an average SGN (Grant et al., 2010) (**Figure 2**). 20,000 EPSCs were randomly generated from which all EPSCs with negative amplitude were discarded. The characteristics of the resulting EPSC amplitude histogram are written in black in the legend. The coefficient of variation, CV, is 0.33 and the skewness is close to 0 (γ_1 = 0.03). The cumulative distribution of amplitudes was fitted by a sum of 12 cumulative distribution functions of Gaussians with means = *i* * 50 pA and SD = Sqrt(*i*) * 19 pA with *i* ranging from 1 to 12. The composing distributions are shown in blue, the mini amplitude distribution is shown in red dashed (mean 50 pA and CV = 0.38). The numbers at the top of each distribution are their relative contribution. The quantal size is the number of neurotransmitter quanta (i.e., number of uniquantal SV) an EPSC is made of. In blue we give the characteristics of the quantal size distribution. Note the negative skewness (γ_1 = -0.17) of the quantal size distribution, regardless of the fact the EPSC amplitude distribution had a skewness $\gamma_1 \sim 0$.

(**B**) In the following, we describe how we determined the range of the CV and the skewness of the experimental quantal size distributions: $0.22 < CV_{data} < 0.34$, $-0.6 < Skewness_{data} < 0.1$. Color maps: characteristics of the possible quantal size distribution underlying the EPSC distribution of the IHC of (**A**). Fitting was done by a sum of Gaussian functions as in (**A**), using mEPSC distributions with different mean and CV. The root mean square error was obtained by integrating the squared difference between the EPSC Gaussian and the fit, then dividing it by 600 pA (the largest EPSC size in this distribution) and taking the square root. Where the error of the fitting is relatively small, the CV covers the range from 0.24 to 0.32 whereas the skewness from -0.5 to 0.0. Thus we took a slight expansion of these ranges as limits of experimental quantal size distributions: $0.22 < CV_{data} < 0.34$, $-0.6 < Skewness_{data} < 0.1$.

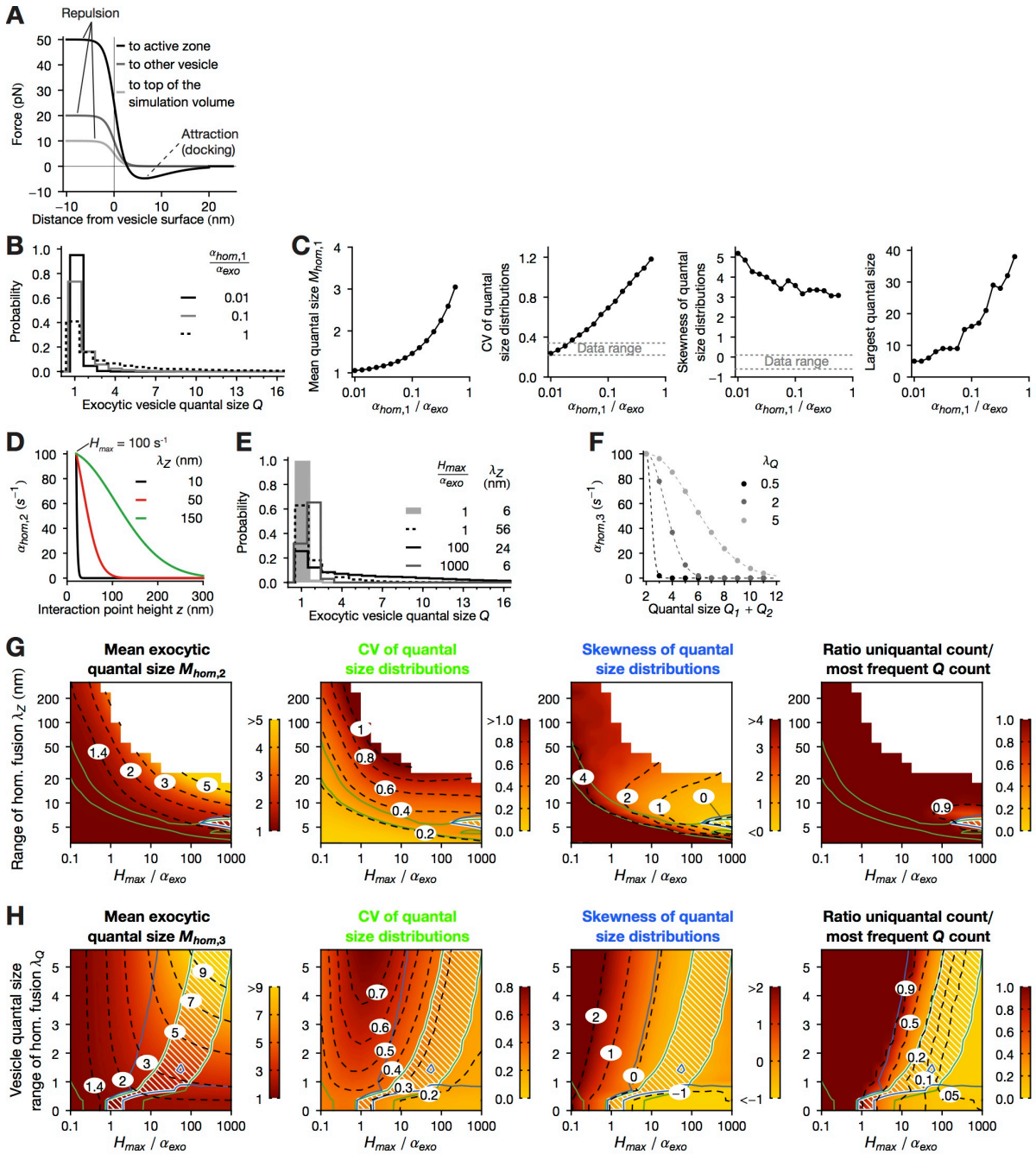

Figure S6. Parameter scan for 3 models of homotypic fusion

Related to Figure 8.

(**A**) Forces present in the simulation. The AZ attracts SVs that are in proximity (attraction of 4 pN), but hinders AZ crossing (repulsion of 50 pN). SVs repel each other

with a maximal force of 20 pN. The top of the simulation space pulled SVs into the space with a maximal force of 10 pN.

(**B**) Homotypic fusion model 1. In this model, the critical parameter in shaping the exocytic quantal size distributions was the ratio $\alpha_{hom,1}/\alpha_{exo}$. Histograms of the number of SVs per EPSC (i.e., exocytic quantal size) for different values of $\alpha_{hom,1}/\alpha_{exo}$.

(**C**) Key characteristics of the exocytic quantal size distributions as a function of $\alpha_{hom,1}/\alpha_{exo}$, for $\alpha_{hom,1}$ between $0.1\cdot\alpha_{exo}$ to $10\cdot\alpha_{exo}$. Note the logarithmic scale of the *x*-axis. The mean exocytic quantal size, the CV, and the largest exocytic quantal size rose with increasing $\alpha_{hom,1}/\alpha_{exo}$, while the skewness decreased. For $\alpha_{hom,1}/\alpha_{exo} < 1$ the distributions were strongly positively skewed with the highest proportion of uniquantal SVs (**Movie S1**). For $\alpha_{hom,1}/\alpha_{exo} \gtrsim 1$, the distributions were very broad and extremely large SVs ($Q >$ 40, diameter > 256 nm) appeared (**Movie S2**). Horizontal dashed lines delimit the range of experimentally observed exocytic distributions ("Data range," $0.22 < CV_{data} < 0.34$, $-0.6 < Skewness_{data} < 0.1$, **Figure S5**) and reveal that this model could not reproduce experimental distributions.

(**D**) Homotypic fusion model 2. Examples of homotypic fusion rate as a function of *z*, for different values of λ_Z and with $H_{max} = 100\ s^{-1}$. A small λ_Z assures that homotypic fusion is mainly restricted to docked SVs.

(**E**) Exocytic quantal size histograms from the homotypic fusion model 2 for different values of H_{max}/α_{exo} and λ_Z, sampled from different parameter space regions.

(**F**) Homotypic fusion model 3. Examples of homotypic fusion rate as a function of the sum of the quantal sizes Q_1 and Q_2 of the "parent" SV, for different values of λ_Q and with $H_{max} = 100\ s^{-1}$.

(**G**) Homotypic fusion model 2. Key characteristics of the exocytic quantal size distributions as a function of H_{max}/α_{exo} (H_{max} ranged from $0.1\cdot\alpha_{exo}$ to $1{,}000\cdot\alpha_{exo}$) and λ_Z (range: ~3 to ~300 nm). Note the logarithmic scale in H_{max}/α_{exo} and λ_Z. The mean exocytic quantal size ($M_{hom,2}$) increased from 1 to ≈ 7, both with H_{max}/α_{exo} and with λ_Z, thereby only slightly reaching the M_{data} of 6. The region of "runaway" homotypic fusion

(white space, SVs created with $Q > 40$) was very prominent, covering about one-third of this parameter space. The CV of the distributions had a similar dependence on H_{max}/α_{exo} and λ_Z as $M_{hom,2}$, and was higher than 0.4 for $M_{hom.2} > 2$, whereas $CV_{data} < 0.34$. The skewness mainly decreased with H_{max}/α_{exo} and was predominantly positive, comparing to $-0.6 < Skewness_{data} < 0.1$. The uniquantal-dominated region covered almost the whole parameter space. The tiny consistency region (white stripes at $H_{max}/\alpha_{exo} > 300$ and $\lambda_Z \approx 6$ nm) indicates where the model can reproduce the values of CV_{data} (falling between the green lines) and $Skewness_{data}$ (falling between the blue lines) of the exocytic quantal size distributions. However in this region the mean exocytic quantal size $M_{hom,2}$ did not match the experimental $N_{rel,data} = 6$. Thus, this model could not reproduce experimental exocytic quantal size distributions.

(**H**) Homotypic fusion model 3. Key characteristics of the exocytic quantal size distributions as a function of H_{max}/α_{exo} (H_{max} ranged from $0.1 \cdot \alpha_{exo}$ to $1{,}000 \cdot \alpha_{exo}$) and λ_Q (range: 0 to 5.4). Note the logarithmic scale for H_{max}/α_{exo}. Same conventions as in (**D**). $M_{hom,3}$ increased from 1 to > 9 both with increasing H_{max}/α_{exo} and λ_Q, covering the $M_{data} = 6$. No runaway fusion region was present in this parameter range, demonstrating that the self-amplification of homotypic fusion was effectively prevented. The CV of the distribution exhibited a complex dependence on H_{max}/α_{exo} and λ_Q. However, for $M_{hom,3} > 2$, the CV decreased both with increasing H_{max}/α_{exo} and decreasing λ_Q, and had values in the range 0.2 – 0.3 (overlapping with CV_{data}) for all $M_{hom,3}$. The skewness decreased with increasing H_{max}/α_{exo} and with decreasing λ_Z, and had values between -1 and 0 overlapping with $Skewness_{data}$ for all $M_{hom,3}$. Finally the ratio of uniquantal to most frequent events could be close to 0 for all $M_{hom,3}$. The consistency region in terms of CV and skewness was thus extended (white stripes), showing that this model could robustly reproduce experimental quantal size distribution shapes. Inside this region, the range $M_{hom,3} \approx 6$ represents the region consistent with experimental data. λ_Z was 100 nm.

See also **Figure S5 and Movies S1 and S2**.

Movie S1. Simulation run for the homotypic fusion model 1 for low homotypic fusion rate $\alpha_{hom,1}$, $\alpha_{hom,1}$ = 0.1 α_{exo}

Related to Figure 8.

First 500 ms of the beginning of a simulation. 1 frame per 1 ms of simulation. 1 ms represents 100 simulations steps (δt = 0.01 ms).

Movie S2. Simulation run for the homotypic fusion model 1 for high homotypic fusion rate $\alpha_{hom,1}$, $\alpha_{hom,1}$ = 10 α_{exo}, leading to the creation of extremely large SVs

Related to Figure 8.

Same as in **Movie S1**. The simulations stops at 410 ms, just before the creation of a SV with $Q > 40$ (not shown).

Table S1. Abbreviations, symbols, and parameters used

Related to Main text and Methods.

(In separate Excel file)

Table S2. EPSC charge, amplitude and CV for different shapes

Related to the EPSC shapes in Figure 3I.

SV Glu. content	EPSC 1	EPSC 2	EPSC 3	EPSC 4	EPSC 5
Mean charges (fC)					
2,000	290	249	243	219	232
3,000	532	467	461	422	440
4,000	796	710	702	648	678
Mean amplitudes (pA)					
2,000	155	131	125	93	99
3,000	281	240	230	173	181.6
4,000	412	362	347	264	273
Amplitude CV					
2,000	0.11	0.11	0.092	0.11	0.10
3,000	0.064	0.08	0.081	0.082	0.072
4,000	0.048	0.055	0.06	0.06	0.055

Obtained from 100 repetitions from different random seeds.

Shapes EPSC 1 and EPSC 2 would be in the “compact” category, because of the small deviation in the shape in EPSC 2. The decrease in EPSC charge of shapes 3-5 in comparison to the average charge of 1 and 2:

	Average charge	Charge reduction		
SV Glu. content	**EPSCs 1 and 2 (fC)**	**EPSC 3**	**EPSC 4**	**EPSC 5**
2,000	269.5	0.1	0.2	0.14
3,000	499.5	0.08	0.15	0.12
4,000	753	0.07	0.14	0.1

Table S3. Summary of recording properties in "1.3 Ca^{2+}" and "0 Ca^{2+}" conditions. Related to Figure 5.

	Leak (pA)	R_s (MΩ)	R_m (MΩ)	C_m (pF)	Events #	Frequency (Hz)	% mono	Amp. mono (pA)	Amp. multi (pA)	Charge (fC)	10-90% rise time mono (ms)	Decay time const. mono (ms)	Time to peak multi (ms)	FWHM multi (ms)
"1.3 Ca^{2+}", n = 6, P9-P11 mice														
Mean	115	39.8	686	2.6	139	0.75	55.4	182	132	272	0.39	1.52	1.49	2.25
SD	152	6.1	680	0.5	64	0.45	7.6	66	33	111	0.05	0.34	0.21	0.60
"0 Ca^{2+}", n = 4, P9-P10 mice														
Mean	503	39.6	247	2.9	65	0.071	81.1	108	72	124	0.41	1.25	1.19	1.59
SD	317	3.4	151	0.9	22	0.008	14.8	19	22	13	0.04	0.21	0.23	0.30
t test*	0.03	0.96	0.18	0.58		0.01	0.01	0.07	0.01	0.02	0.38	0.20	0.06	0.08

*Mann-Whitney U test for EPSC frequency and the percentage of monophasic EPSCs.

P: postnatal age

FWHM: full width at half maximum

Mono: monophasic EPSCs

Multi: multiphasic EPSCs

Amp.: amplitude

Table S4. Summary of recording properties in control and BayK8644 conditions. Related to Figure 6.

n = 8, P15-18 rats

	R_s (MΩ)	C_m (pF)	Events #	Frequency (Hz)	% mono	Amp. mono (pA)	Amp. multi (pA)	Charge (fC)	10-90% rise time mono (ms)	Decay time const. mono (ms)	Time to peak multi (ms)	FWHM multi (ms)
Control												
Mean	33.2	3.3	790	5.6	79.2	432	287	296	0.22	0.74	1.0	1.28
SD	7.7	1.7	579	4.7	12.4	124	65	65	0.05	0.15	0.1	0.31
BayK8644												
Mean	39.8	2.7	806	6.9	93.3	391	314	260	0.18	0.74	0.98	1.13
SD	10.3	1.0	808	6.2	4.6	123	102	61	0.04	0.16	0.18	0.33
Paired t test*	0.005	0.15		0.5	0.01	10^{-3}	0.16	$<10^{-3}$	0.002	0.95	0.6	0.006

*Wilcoxon signed-rank test for the percentage of monophasic EPSCs and the frequency.

The charges and amplitudes of EPSC in cells recorded at -60 mV were multiplied by 90/60 to compensate for the driving force differences with other cells recorded at – 90 mV.

P: postnatal age

FWHM: full width at half maximum

Mono: monophasic EPSCs

Multi: multiphasic EPSCs

Amp.: amplitude

SUPPLEMENTAL METHODS

Patch-clamp recordings and analysis

Preparation

P9-11 C57BL6 mice and P10-18 Wistar rats of either sex were used. Animal handling and experiments compiled with national animal care guidelines were approved by the University of Göttingen Board for animal welfare and the animal welfare office of the state of Lower Saxony. After mice/rats were decapitated after deep anesthesia by carbon dioxide, their cochleae were removed from the temporal bones and immersed into the standard extracellular solution (for its composition, see below). The apical turn of cochlea was carefully excised, and placed under a grid in the recording chamber. The postsynaptic boutons of type I SGNs innervating the IHCs were visually identified on a monitor using a 40x/63x water immersion objective lens attached to an upright microscope with differential interference contrast optics (Axioskop FS2, Zeiss), and a 4x/2.5x magnification camera (TILL Photonics).

Postsynaptic patch-clamp recording

Whole-cell voltage-clamp recordings from postsynaptic boutons of mouse/rat type I spiral ganglion neurons of the apical cochlear coil were performed as previously described for rats (Glowatzki and Fuchs, 2002; Grant et al., 2011; Rutherford et al., 2012) and for mice (Jing et al., 2013; Pangrsic et al., 2010). The recording pipettes, fabricated with a puller (P-97 and P-2000, Sutter) from borosilicate grass with an outer diameter of 1 mm (1B100F-3 or 4 and TW100F-3. World Precise Instruments), had a resistance of 6-15 MΩ following pressure polishing (Goodman and Lockery, 2000) with a custom-built microforge. The intracellular solution contained (in mM): 150 CsCl (rat: 137 KCl and 1.25 Na_2GTP), 3.5 $MgCl_2$, 0.1 $CaCl_2$, 5 EGTA, 5 K-HEPES, and 2.5 Na_2ATP, pH 7.2 (adjusted with KOH, osmolarity approx. 290 mOsm). The extracellular solution for both dissection and recording contained (in mM): 5.8 KCl, 155 NaCl (rat: 142-144), 0.9 $MgCl_2$, 1.3 $CaCl_2$, 0.7 NaH_2PO_4, 5.6 D-glucose, and 10 Na-HEPES, pH

7.4 (rat: 7.3) adjusted with NaOH, osmolarity approx. 300 mOsm. In the "0 Ca^{2+}" experiment, no $CaCl_2$ but instead 2 mM of the Ca^{2+} chelator EGTA was added to the extracellular solution. Note that the cochleae were immersed in the Ca^{2+}-free extracellular solution for at least 30 min before the beginning of the recordings. In most recordings, tetrodotoxin (1-5 µM) was added to block voltage-gated Na^+ channels. All chemicals were purchased from Sigma-Aldrich (St. Louis, MO), except for tetrodotoxin (Tocris). EPC-10 amplifier controlled by Patchmaster software (both HEKA Electronics, Lambrecht, Germany) was used to sample and filter currents at 20-50 kHz and 5-10 kHz, respectively. EPSCs were recorded at a holding potential of -90 mV (liquid junction potential not corrected) at room temperature (21-24 °C). Two SGNs were recorded at -60 mV for the BayK8644 experiment.

Data analysis mouse ("0 Ca^{2+}" dataset, Figure 5)

For detection and analysis of EPSCs, Mini Analysis software (Synaptosoft) was used with a detection threshold set at 3-5 times greater than the root mean square (rms) of the baseline noise. To classify EPSCs into mono- or multiphasic, the method described by (Grant et al., 2010) was employed. For plotting, IGOR Pro (Wavematrics), Sigmaplot (HULINKS), and Mathematica (Wolfram Research) were used. The series resistance (R_s) was calculated from the height of capacitive transient in response to a 10 mV voltage step. The membrane capacitance (C_m) was derived from single-exponential curve fitting of the decay phase of the capacitive transient. The membrane input resistance (R_m) was calculated from steady-state currents in response to the voltage step. Recording parameters of postsynaptic boutons are summarized in **Table S3**. A recording was discarded, if R_s > 50 MΩ or if it contained less than 30 EPSCs. Data is shown as mean ± SD.

Data analysis rat (BayK8644 dataset, Figure 6)

For detection and analysis of EPSCs, Mini Analysis software (Synaptosoft) was used with a detection threshold set at 15-30 pA. To classify EPSCs into mono- or multiphasic, the method described by (Grant et al., 2010) was employed. For plotting,

IGOR Pro (Wavematrics) and Mathematica (Wolfram Reseach) were used. The series resistance (R_s) was calculated from the double exponential fit of the capacitive transient in response to a 10 mV voltage step. The membrane capacitance (C_m) was derived from single-exponential curve fitting of the decay phase of the capacitive transient. Recording parameters of postsynaptic boutons are summarized in **Table S4**. Recordings were discarded if $R_s > 60$ MΩ or if R_s changed by more than 10 MΩ during the recording. Data is shown as mean ± SD.

Detailed EPSC analysis (for recordings from rat, Figures 2 and 4)

The analysis was performed in Mathematica (Wolfram Research) on data from 5 rat SGN recordings with abundant spontaneous activity driven by the IHC (more than 110 EPSCs per recordings, rate range: 1 to 8 Hz). Data is shown as mean ± SD.

To allow accurate EPSC detection and analysis, all current traces were baseline-subtracted. We first chose by eye several flat and EPSC-free segments of each recording to get an upper estimate of the recording background noise standard deviation σ_{BG}. σ_{BG} was within the range 1.6-3.6 pA. The recordings were then divided into 6 to 10 ms intervals, and selected for baseline calculations if their SD was smaller than 0.8 to 0.95 times σ_{BG}. Between 7-20% of the intervals were thus chosen. The mean value of these intervals was used as the baseline value in the center of the interval. The entire baseline was obtained as the linear interpolation between these selected points. EPSCs were detected using a threshold at -15 to -20 pA. The beginning and the end of EPSCs were set as the points where the current trace crossed a threshold value set between 0 and 2 pA. EPSC charge was calculated as the area between the beginning and the end of the EPSC. All events were inspected visually. Events were discarded when there was an unstable baseline, or when the detected event was noise or a clear superposition of multiple independent EPSCs.

To estimate the mode of the non-compact EPSC charge distributions (**Figure 2J**), we first constructed a smoothed version of the charge distribution, using a Gaussian smoothing kernel and Silverman's rule (Silverman, 1986) to determine the

kernel's bandwidth. Then the mode was obtained by finding the maximum of smoothed EPSC charge distribution.

To estimate the fraction of non-compact EPSCs arising from random EPSC superposition, we first measured the EPSC rate: $rate_{EPSC}$. We assumed that two releases occurring with an interval smaller than 3 ms will be counted as one non-compact EPSC. Since the inter-EPSC-intervals are exponentially distributed (Rutherford et al., 2012), the fraction F of inter-EPSC-intervals smaller than 3 ms is predicted to be $F = 1 - \mathrm{Exp}(-3\ \mathrm{ms} * rate_{EPSC})$. Multiplying F by the total number of EPSCs in the recording gives us approximately the expected number of non-compact EPSCs occurring from random EPSC superposition. This number varied from 1 to 13 across recordings.

EPSC deconvolution analysis (for recordings from rat, Figure 4)

We performed a deconvolution of EPSC shapes into underlying "elementary EPSC" (eEPSCs). An eEPSC was defined as an EPSC triggered by a virtually instantaneous neurotransmitter liberation into the synaptic cleft. First, we constructed an "ideal EPSC" shape for each recording, which was the basis for the eEPSC shape. In the EPSC charge vs. amplitude plot we placed, by visual inspection, a line passing through the origin and through the middle of the right elongated high-density region. The EPSCs to the right of this line were normalized, aligned at their half-maximal point on their rising phase, and averaged. Normalizing the resulting shape, we obtained the "ideal EPSC" ($EPSC_{ideal}(t)$). To characterize its decay kinetics, we fitted its decay phase by a double exponential $B_1 \exp(-t/\tau_1) + B_2 \exp(-t/\tau_2)$ (e.g., **Figure 4A**, time constants τ_1 = 0.63 ± 0.1 ms and τ_2 = 2.6 ± 1.2 ms with relative contributions 0.83 ± 0.12 and 0.17 ± 0.12, respectively, n = 5). The beginning of the fit was at t_{decay}, chosen where $EPSC_{ideal}(t_{decay})$ was between 70% to 80% of the peak amplitude. The kernel of the deconvolution algorithm, the eEPSC shape, was constructed in the following way:

$$eEPSC(t,\tau_{1,fit},\tau_{2,fit}) = \begin{cases} EPSC_{ideal}(t) & t < t_{decay} \\ B_1 \exp(-t/\tau_{1,fit}) + B_2 \exp(-t/\tau_{2,fit}) & t \geq t_{decay} \end{cases} \quad (1)$$

where $\tau_{1,fit}$ and $\tau_{2,fit}$ were time constants of the double exponential decay of *eEPSC*; they were allowed to vary around τ_1 and τ_2, respectively: $0.65 \cdot \tau_1 < \tau_{1,fit} < 1.5 \cdot \tau_1$, $0.85 \cdot \tau_2 < \tau_{2,fit} < 1.2 \cdot \tau_2$. This enabled the eEPSC to account for some stochasticity of the EPSC decay phase. For convenience we shifted the function $eEPSC(t, \tau_{1,fit}, \tau_{2,fit})$ so that $eEPSC(0, \tau_{1,fit}, \tau_{2,fit}) = 1$, i.e., its peak is at $t = 0$.

Using the following iterative deconvolution/fitting algorithm our goal was to obtain an optimal linear superposition of eEPSCs, $f(n,t)$, that reproduced the EPSC shape, without overfitting:

$$f(n,t) = \sum_{i=1}^{n} a_i \cdot eEPSC(t - t_i, \tau_{1,fit,i}, \tau_{2,fit,i}) \quad (2)$$

where t was time, n was the number of eEPSCs composing the fitting function, a_i and t_i were the amplitudes and the times of the underlying eEPSCs, respectively, $\tau_{1,fit,i}$ and $\tau_{2,fit,i}$ were the fast and slow time constants of the double exponential decay of the eEPSC i. We used an iterative approach to find the optimal deconvolution. Starting at $n = 1$ and then increasing n, we fitted the EPSC with $f(n,t)$ using the Levenberg–Marquardt algorithm (damped least-squares method) with a_i, t_i, $\tau_{1,fit,i}$, $\tau_{2,fit,i}$ for $i = 1$ to n as free parameters. We stopped at the n for which the fit satisfied two criteria: (1) the mean error E_{mean} of the fit was smaller than $\sigma_{threshold}$; (2) the maximum deviation of the fit was smaller than Maximum(20 pA, 0.1 · EPSC amplitude). The mean error E_{mean} was defined as:

$$E_{mean}(n) = \left(\int_{t_{begin}}^{t_{end}} \left(f(n,t) - EPSC(t) \right)^2 dt \right)^{1/2} \quad (3)$$

where $EPSC(t)$ is the EPSC shape, t_{begin} and t_{end} are the beginning and the end of the EPSC, respectively. $f(n,t)$ was the fit function. $\sigma_{threshold}$ was defined as $\sigma_{threshold}^2 = (2 \cdot \sigma_{BG})^2 + \sigma_{AMPAR}^2$ and incorporated the two principal components of noise in the

recording: σ_{BG} the recording background noise, and σ_{AMPAR} the noise due to the stochastic gating of the AMPAR. Assuming 20 pS as AMPAR conductance (Budisantoso et al., 2013; Sahara and Takahashi, 2001) we expect a single channel current of 1.8 pA at -90 mV holding potential. This allowed us to calculate the number of open AMPAR channels at the EPSC peak, $N_{AMPAR,open}$. During the EPSC decay phase channels close stochastically, and when the open probability reaches 0.5 for these channels, the variance is maximal. The standard deviation of the current at that time can be computed from binomial statistics: $\sigma_{AMPAR,max} = (N_{AMPAR,open} \cdot 0.5^2)^{0.5} \cdot 1.8$ pA. We assumed that no additional channels opened during the decay phase of the EPSC. We chose $\sigma_{AMPAR} = \sigma_{AMPAR,max}/2$, which corresponds to 5.8 pA for an EPSC of 300 pA amplitude.

Additionally, the iteration stopped at n if increasing the number of eEPSCs to $n + 1$ did not reduce the mean error E_{mean} by more than 5%, i.e., if $E_{mean}(n + 1) > 0.95 \cdot E_{mean}(n)$, or if the newly positioned eEPSC was closer to an other eEPSC by less than a sampling distance, or if the additional eEPSC had a smaller amplitude than $\sigma_{threshold}$. For the fit at $n = 1$, starting values for the fit for a_1 and t_1 where the EPSC amplitude and the peak time, respectively. For the fits at $n > 1$, the starting values for a_i and t_i (i from 1 to $n - 1$) were the values provided by the fit at $n - 1$, and the starting values for a_n and t_n were the amplitude and the time of the maximum error of the $n - 1$ fit.

Characterization of the AMPAR clusters

Immunohistochemistry and superresolution microscopy of immunolabeled hair cells

Freshly dissected apical cochlear turns from p16 Wistar rats were fixed with 4% formaldehyde (FA) in phosphate buffered saline (PBS) for 1 hour on ice. Thereafter, the tissue was washed 3x10 min in PBS and incubated for 1 hour in goat serum dilution buffer (GSDB: 16% normal goat serum, 450 mM NaCl, 0.3% Triton X-100, 20 mM phosphate buffer, pH 7.4) in a wet chamber at room temperature. Primary antibodies (mouse anti-bassoon, Abcam, 1:200 and rabbit anti-GluR2/3, Chemicon, 1:200) were dissolved in GSDB and applied overnight at +4°C in a wet chamber. After washing 10 min in wash buffer (450 mM NaCl, 20 mM phosphate buffer, 0.3% Triton X-100) and 3x10 min in 20% goat serum in PBS, the tissue was incubated with secondary antibodies (goat anti-mouse and sheep anti-rabbit, labeled with Atto590 and KK1212, 1:60) in 2-3% goat serum in PBS in a wet light-protected chamber for 1 hour at room temperature. Then the preparations were washed 3x10 min in 20% goat serum in PBS, placed onto glass microscope slides with a drop of Mowiol mounting medium and covered with thin glass coverslips. Images were acquired on a custom-built 2 color STED microscope with a lateral resolution of <50 nm.

GluR cluster length and line profile estimations

For each synapse an area of interest was defined, usually 82 by 82 pixels, corresponding to 1,600 nm x 1,600 nm. To remove the image background, the most frequently occurring value in the image was subtracted from everywhere in the image and all values below 0 were set to 0. Then the image values were normalized from 0 to 1.

From side-imaged clusters one line profile was taken, oriented in the long axis of the AMPAR cluster (**Figure 3**, *right column, white dashed lines*). For each en-face imaged clusters, 1 to 3 line profiles were taken, in order to obtain a good sampling of the profile in the long axis orientation (**Figure 3**, *first 2 columns, white dashed lines*).

For each line profile, the most external peaks were found (**Figure 3A**, *white crosses on the dashed lines*) and for each synapse one average peak-to-peak distance was obtained and reported in **Figure 3B**.

To obtain the “grand-average” line profile, first one average line profile was calculated for each AMPAR cluster. All line profiles in each synapse were aligned and stretched so that their peak positions were at -1 and 1 and averaged. Each line profile $f(x)$ was taken twice, once in each orientation (i.e., $f(x)$ and in the reversed orientation $f(-x)$). The obtained average line profile in each synapse was normalized to 1. The “grand-average” line profile among synapses was calculated as the average of all the individual synaptic line profiles.

Molecular model of glutamate release, diffusion, AMPAR binding and gating

The simulations were done in MCell 3.2.1 (Kerr et al., 2008; Stiles et al., 1996; Stiles and Bartol, 2001) and were typically run on a cluster of Linux cores.

The simulation space was composed of a square synaptic contact of 1,600 nm wide separated by an effective synaptic cleft height δ = 15 nm (Savtchenko and Rusakov, 2007). The model effective synaptic cleft height is smaller than the true cleft height, as it simulates the actual volume available for glutamate diffusion. The AMPARs were placed on grids of different densities according to the density profile of **Figure 3C**. The density had circular symmetry and was scaled to match the observed peak-to-peak cluster length (**Figure S1A**). The average AMPAR density of the whole cluster was ~3,000 receptors/μm^2 (Saito, 1990). To calculate the synaptic currents, the AMPAR conductance was set to 20 pS (Budisantoso et al., 2013; Sahara and Takahashi, 2001) and the holding potential to -90 mV. We used both a fast AMPAR kinetic scheme derived from cerebellar Purkinje cells (Häusser and Roth, 1997) and an AMPAR scheme obtained from the calyx of Held (Budisantoso et al., 2013).

The diffusion coefficient of glutamate D_{Glu} was set to 0.33 μm^2/ms (Nielsen et al., 2004). The lateral openings of the synaptic aposition absorbed glutamate, mimicking the escape of glutamate into the extracellular medium. Glutamate was either released 1 nm below the presynaptic membrane (for **Figures 3E-G**) or from a SV (**Figures 3H,I**). The SV had a cube shape with a 26 nm edge. This corresponds to a volume of 17,576 nm^3, a SV with an inner radius of 16 nm has a volume of 17,157 nm^3. Since the volume and the dimensions of the model and real SVs are very similar (only the shape is different), the non-spherical geometry of the model vesicle induces negligible differences in the dynamics of transmitter release. The pore was a column with a square cross section of various widths. The pore length was 8 nm, corresponding to the thickness of two lipid bilayers. The EPSC amplitude resulting from instantaneous release of the entire glutamate content placed into the synaptic cleft and from release through a pore of 6 nm width was virtually identical (**Figure S1B**).

For **Figure 3H**, only the glutamate that reached the synaptic cleft was considered as released. The glutamate inside the SV and the SV fusion pore was considered as not released.

The simulation time step was chosen 10 ns for simulations of release through a SV fusion pore and 50 ns for simulations with release modeled as neurotransmitter placement into the synaptic cleft. Data was sampled at 5 μs time intervals. For **Figures 3E,F**, 100 repetitions with different random seeds were performed to obtain the mean EPSC amplitude. For **Figure 3G**, 400 repetitions were performed. For **Figure 3H**, 10 repetitions were performed. To calculate the mean EPSC amplitude we averaged the peak amplitudes of individual EPSCs.

"Ca^{2+}-synchronized MQR" model

Each of the N_{tot} SVs' Ca^{2+} sensors was exposed to a rectangular Ca^{2+} pulse of concentration $[Ca^{2+}]$ for a duration T_{pulse}, mimicking the opening of an adjacent Ca^{2+} channel. The concentration time course was:

$$[Ca^{2+}](t) = \begin{cases} 0 & t < 0 \text{ ms} \\ [Ca^{2+}] & 0 \text{ ms} \le t < T_{pulse} \\ 0 & t \ge T_{pulse} \end{cases} \quad (4)$$

We calculated the mean number of SVs released during such pulses (N_{rel}) and the mean release asynchrony (T_{async}), defined as the mean time lag between the fusion of two SVs. This measure is a lower bound of the mean time lag between the release of first and the last SV during a release event composed of more than two SVs. Based on the Ca^{2+}-dependent release scheme at the IHC (Beutner et al., 2001), we numerically solved the Master equation for the sensors state occupancy $S_i(t)$, where each state is related to the number of Ca^{2+} bound to the sensor and S_6 is the released state. All simulations started from the initial condition $S_0(0 \text{ ms}) = 1$ and $S_i(0 \text{ ms}) = 0$ for $i = 1$ to 6.

We solved the system of differential equations for $S_i(t)$ with NDSolve in Mathematica 7.0.1.0 (Wolfram Research) on the two segments $0 \le t \le T_{pulse}$ and $T_{pulse} \le t \le T_{pulse}$ +10 ms separately. The discontinuity of $[Ca^{2+}](t)$ at $t = T_{pulse}$ was thus treated by taking the states of the end of the first segment as the initial condition for the second segment. Error tolerance in NDSolve was set to 10^{-20}.

The SV release probability P_{rel} was taken as $P_{rel} = S_6(T_{pulse} + 10 \text{ ms})$. We derived N_{rel} using binomial statistics, neglecting non-release events:

$$N_{rel}(N_{tot}, P_{rel}) = \sum_{k=1}^{N_{tot}} k\, b(k; N_{tot}, P_{rel}) \Big/ \sum_{k=1}^{N_{tot}} b(k; N_{tot}, P_{rel}) = \frac{N_{tot} P_{rel}}{1-(1-P_{rel})^{N_{tot}}} \quad (5)$$

where $b(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^k$ is the probability mass function of the binomial distribution with n trials and probability p.

The release probability density function $f_{rel}(t)$, which is also the SV release rate, was obtained by taking the temporal derivative of $S_6(t)$, $f_{rel}(t) = S_6'(t)$. The probability density $P_{lag}(l)$ that two SVs exocytose with an absolute time lag $l > 0$ ms can be determined from:

$$P_{lag}(l) \propto \int_{-\infty}^{+\infty} f_{rel}(t) \cdot f_{rel}(t+l)\, dt + \int_{-\infty}^{+\infty} f_{rel}(t) \cdot f_{rel}(t-l)\, dt = 2\int_{-\infty}^{+\infty} f_{rel}(t) \cdot f_{rel}(t+l)\, dt \qquad (6)$$

The mean release asynchrony T_{async} is then:

$$T_{async} = \int_{0}^{+\infty} l \cdot P_{lag}(l)\, dl \qquad (7)$$

We evaluated (6) and (7) numerically. We sampled $f_{rel}(t)$ in the interval 0 to T_{pulse} + 10 ms with a step size of 5 μs and performed an auto-correlation (using FFT) of this list of numbers, by padding 0 on the right (maximal overhang at the right-hand end). Due to the numerical integrations scheme, when $S_6(t)$ saturated, $f_{rel}(t)$ could slightly oscillate around 0. In that case, the interval of interest was limited up to the first negative value of $f_R(t)$. This calculation gave us the probability $P_{lag}(l)$ of a certain time lag l, up to normalization. T_{async} was calculated using the discrete version (Riemann sum) of (7).

We also calculated the mean number of SVs released N_{rel} and the mean release asynchrony T_{async} for different [Ca^{2+}] and Ca^{2+} channel mean open time τ_{open}. Ion channels have in the simplest case exponentially distributed open times. For a given τ_{open}, the probability density $P_T(T_{pulse}, \tau_{open})$ for the ion channel to open for a duration T_{pulse} is:

$$P_T(T_{pulse}, \tau_{open}) = \exp(-T_{pulse} / \tau_{open}) / \tau_{open} \qquad (8)$$

In the following, [Ca^{2+}] and T_{pulse} are abbreviated as C and T, respectively. For given C and τ_{open}, the mean number of SVs released N_{rel} is:

$$N_{rel}(C, \tau_{open}) = \frac{\int_{0}^{\infty} N_{rel}(C,T) \cdot P_T(D, \tau_{open}) \cdot P_{1Ves}(C,T) \cdot dT}{\int_{0}^{\infty} P_T(T, \tau_{open}) \cdot P_{1Ves}(C,T) \cdot dT} \qquad (9)$$

while the release asynchrony T_{async} is:

$$T_{async}(C,\tau_{open}) = \frac{\int_0^\infty T_{async}(C,T)\cdot P_T(T,\tau_{open})\cdot P_{2Ves}(C,T)\cdot dT}{\int_0^\infty P_T(T,\tau_{open})\cdot P_{2Ves}(C,T)\cdot dT} \quad (10)$$

where $P_{1Ves}(C,T)$ and $P_{2Ves}(C,T)$ are the probabilities that at least one and at least two SVs get released during a Ca^{2+} pulse of duration T and concentration C, respectively:

$$P_{1Ves}(C,T) = 1 - \left(1 - P_{rel}(C,T)\right)^{N_{tot}} \quad (11)$$

$$P_{2Ves}(C,T) = 1 - \left(1 - P_{rel}(C,T)\right)^{N_{tot}} - N_{tot}\cdot P_{rel}(C,T)\cdot\left(1 - P_{rel}(C,T)\right)^{N_{tot}-1} \quad (12)$$

It is important to notice that $N_{rel}(C,T)$ is a distinct function from $N_{rel}(C,\tau_{open})$ and $T_{async}(C,T)$ is distinct from $T_{async}(C, \tau_{open})$. To compute (9) and (10) for given C and τ_{open} we first calculated the functions $P_{rel}(C,T)$, $P_{1Ves}(C,T)$, $P_{2Ves}(C,T)$, $N_{rel}(C,T)$, and $T_{async}(C,T)$ for T from 0.01 to 20 ms with a maximum step size of 0.1 ms (up to 10 ms and 0.5 step form 10 to 20 ms). Because the largest changes in $P_{rel}(C,T)$ and $T_{async}(C,T)$ occurred for small T, from 0.01 to 0.31 ms we used finer logarithmic steps in T: $T = 10^{dx}$ for dx from -2 to -0.5 in step of 0.1. Then, we performed a third order interpolation of all the points and resampled the interpolated function with a step size of 10 μs. For the point T = 0 ms, we took $P_{rel}(C,T = 0\text{ ms}) = 0$ (as we found that $P_{rel}(C,T = 0.01\text{ ms}) < 10^{-6}$) and $T_{async}(C,T = 0) = T_{async}(C,T = 0.01\text{ ms})$. Finally, we calculated the integrals as a Riemann sum using the resampled points. An ion channel with τ_{open} = 3 ms has >99% of its openings shorter 20 ms. Terminating the calculations at T = 20 ms thus only slightly underestimated $T_{async}(C,\tau_{open})$ and $N_{rel}(C,\tau_{open})$.

In **Figures S4C,D**, we show $P_T(T,\tau_{open})$, $P_{rel}(C,T)$, $P_{1Ves}(C,T)$, $P_{2Ves}(C,T)$, $P_{1Ves}(C,T)\cdot P_T(T,\tau_{open})$ and $P_{1Ves}(C,T)\cdot P_{2Ves}(T,\tau_{open})$ for a representative choice of parameters.

The IHC Ca^{2+}-dependent release scheme (Beutner et al., 2001):

$$S_0 \underset{1k_{off}\cdot b^0}{\overset{5k_{on}[Ca^{2+}]}{\rightleftarrows}} S_1 \underset{2k_{off}\cdot b^1}{\overset{4k_{on}[Ca^{2+}]}{\rightleftarrows}} S_2 \underset{3k_{off}\cdot b^2}{\overset{3k_{on}[Ca^{2+}]}{\rightleftarrows}} S_3 \underset{4k_{off}\cdot b^3}{\overset{2k_{on}[Ca^{2+}]}{\rightleftarrows}} S_4 \underset{5k_{off}\cdot b^4}{\overset{1k_{on}[Ca^{2+}]}{\rightleftarrows}} S_5 \xrightarrow{\gamma} S_6 \quad (13)$$

And the corresponding differential Master equation of the sensor occupancy:

$$\begin{cases} \dot{S}_0(t) = & - & 5k_{on}[Ca^{2+}](t)\cdot S_0(t) & + & 1k_{off}\cdot b^0\cdot S_1(t) \\ \dot{S}_1(t) = 5k_{on}[Ca^{2+}](t)\cdot S_0(t) & - & (1k_{off}\cdot b^0 + 4k_{on}[Ca^{2+}](t))\cdot S_1(t) & + & 2k_{off}\cdot b^1\cdot S_2(t) \\ \dot{S}_2(t) = 4k_{on}[Ca^{2+}](t)\cdot S_1(t) & - & (2k_{off}\cdot b^1 + 3k_{on}[Ca^{2+}](t))\cdot S_2(t) & + & 3k_{off}\cdot b^2\cdot S_3(t) \\ \dot{S}_3(t) = 3k_{on}[Ca^{2+}](t)\cdot S_2(t) & - & (3k_{off}\cdot b^2 + 2k_{on}[Ca^{2+}](t))\cdot S_3(t) & + & 4k_{off}\cdot b^3\cdot S_4(t) \\ \dot{S}_4(t) = 2k_{on}[Ca^{2+}](t)\cdot S_3(t) & - & (4k_{off}\cdot b^3 + 1k_{on}[Ca^{2+}](t))\cdot S_4(t) & + & 5k_{off}\cdot b^4\cdot S_5(t) \\ \dot{S}_5(t) = 1k_{on}[Ca^{2+}](t)\cdot S_4(t) & - & (5k_{off}\cdot b^4 + \gamma)\cdot S_5(t) & & \\ \dot{S}_6(t) = \gamma\cdot S_5(t) & & & & \end{cases} \quad (14)$$

with b = 0.4, k_{on} = 27.6 $\mu M^{-1}s^{-1}$, k_{off} = 2,150 s^{-1}, and γ = 10,000 s^{-1} (instead of 1,695 s^{-1} as in the publication).

Electron microscopy

Apical cochlear coils from p14 C57BL6/J wild type mice were explanted and either stimulated or inhibited for 15 min as described previously (Pangrsic et al., 2010). Subsequently samples were placed in 200 µm/100 µm aluminum planchettes covered with an according lid and prepared for high-pressure freezing and freeze-substitution essentially as described previously (Wong et al., 2014). Briefly, samples were frozen under high pressure (HPM 010; Baltec), followed by freeze-substitution. Substitution was performed in a Leica AFS2 in acetone with 0.1% tannic acid at −90°C for 4 d, followed by acetone with 2% osmium during the last 7 h. The samples were warmed (5°C/h) to −20°C and incubated for 16 additional hours before being warmed (10°C/h) to 4°C. At 4°C, the samples were washed in acetone and warmed to room temperature and Epon embedded. After trimming and ultrathin sectioning with a Leica EM UC6 ultramicrotome slices were placed on 1 x 2 mm copper slot grids and poststained with uranyl acetate/lead citrate following standard protocols. Micrographs were taken with a JEOL electron microscope (JEM 1011) equipped with a Gatan Orius 1200A camera using the Digital Micrograph software package or a ZEISS EM 902A with a 1,024 x 1,024 charge coupled device detector (HSS 512/1,024, Proscan Electronic systems).

For quantification random 70 nm AZ sections were taken. The SV diameter of all SVs in the first row around the ribbon has been determined using the ImageJ software package. In the stimulated condition there were 151 SVs from 16 sections using 2 organs of Corti. In the inhibited condition there were 572 SVs from 37 sections using 2 organs of Corti.

We derived the predicted SV diameter distribution by convolving the model SV quantal size distributions with the experimental uniquantal SV diameter distribution. As uniquantal SV diameter distribution we used the Gaussian fit to the experimental diameter distribution of membrane-proximal SVs in the inhibited condition (**Figure 8F**). We chose this distribution because it contained the smallest proportion of large SVs, and is the best candidate for the true uniquantal distribution. We stochastically calculated the diameters of SVs with quantal size *Q*:

$$\rho_Q = \sqrt{\rho_{1,1}^2 + \rho_{1,2}^2 + ... + \rho_{1,Q}^2} \quad (15)$$

where $\rho_{1,i}$ are independent identically distributed random variables representing the uniquantal SV diameter, and ρ_Q is the random variable representing the diameter of a SV with quantal size *Q*. This equation satisfies the conservation of the total membrane surface upon homotypic fusion. To obtain the model curve in **Figure 8M**, 200,000 SVs were randomly drawn.

Data are expressed as mean ± SD.

Homotypic fusion model of MQR

To examine the homotypic fusion scenario of MQR, we simulated the dynamics of SVs on the synaptic ribbon surface. Our goal was first to construct a homotypic fusion model that would reproduce experimental exocytic quantal size distributions and then to compare predicted SV size distributions on the ribbon with ultrastructural data. The model, implemented as a stochastic simulation, incorporated the Brownian motion of SVs in a fixed two-dimensional (2D) geometry, homotypic fusion, SV docking to the AZ, exocytosis, and SV replenishment. At each simulation time step, the order of processes was the following: replenishment, homotypic fusion, exocytosis, and diffusion of SVs. At the end of each homotypic fusion and diffusion event, SVs extending below the AZ were translated directly above it.

Simulation space and SV pool replenishment

The simulation space represented the whole 2D ribbon surface (**Figure 4A**). SV centers moved on the surface of the ribbon, modeled as a cylinder with a perimeter *Per* = 500 nm and height *H* = 300 nm (Nouvian et al., 2006). The bottom of the simulation space corresponded to the AZ. Simulations were initialized with 0 SVs on the ribbon surface.

A SV packing density $\phi = A_{ves} / A_{ribbon}$ was set and maintained by adding SVs to the simulation space, modeling trapping of SVs by the ribbon. $A_{ribbon} = H \cdot Per$ was the total ribbon surface and A_{ves} was the area filled by SVs, i.e., the sum of surfaces of the maximal SV cross-sections: $A_{ves} = \sum_{i=1}^{n} \pi R^2_i$, where R_i is the radius of SV *i* and *n* is the number of SVs in the simulation. As soon as ϕ decreased below a fixed ϕ_{min} (i.e., $\phi < \phi_{min}$), a new SV with quantal size *Q* = 1 and *R* = 20 nm was added and placed with its center *Z* at *H* + 20 nm, directly above the simulation space. The horizontal coordinate *X* was chosen randomly among 10 equally spaced positions. Due to forces at the upper boundary, the SVs were pushed into the simulation space. For the model 4 (**Figure 8M**), SVs were added on the whole ribbon surface, down to 40 nm above the AZ. The theoretical maximum packing density in 2D is $\phi \approx 90\%$ (hexagonal packing). SVs were

added at a maximal rate of 2 kHz and at least three SVs were always maintained in the simulation space.

Brownian motion of SVs

To simulate SV movement, we used an "overdamped" Langevin equation, i.e. a first order stochastic differential equation. In this simplified "position" Langevin equation (Allen and Tildesley, 1989), the evolution equation for a SV *i*, subject to deterministic and stochastic forces is:

$$\frac{d\vec{x}_i(t)}{dt} = \frac{D_i}{k_B T}\left(\sum_{j \neq i} \vec{F}(\vec{x}_j, R_j, \vec{x}_i, R_i) + \vec{F}_B(\vec{x}_i, R_i) \right) + \sqrt{2D_i}\,\vec{\xi}_i(t) \qquad (16)$$

where $\vec{x}_i(t)$ is the position of the SV *i* at time *t*, R_i its radius, D_i its diffusion coefficient, $\vec{F}(\vec{x}_j, R_j, \vec{x}_i, R_i)$ the force of SV *j* on SV *i*, $\vec{F}_B(\vec{x}_i, R_i)$ the force acting on SV *i* due to the ribbon area boundaries, k_B is the Boltzmann constant, and *T* the absolute temperature (295 K = 22 °C used). $\vec{\xi}_i(t)$ is a zero-mean Gaussian white noise, satisfying the auto-correlation relation $< \vec{\xi}_i(t_1) \bullet \vec{\xi}_j(t_2) > = d\,\delta(t_1 - t_2)\delta_{ij}$, where $\delta(t)$ is the Dirac function, δ_{ij} is the Kronecker delta, and *d* the number of dimensions in the system, here 2.

We did not consider the evolution of velocities, as we were interested in time scales significantly longer than the decay time constant of the velocity auto-correlation function $\tau_i = m_i/\zeta_i \approx 10^{-9}$ s = 1 ns, where ζ_i is the drag constant (i.e., the inverse of the mobility, unit: kg/s) due to the surrounding medium and m_i is the SV mass (Allen and Tildesley, 1989). The following consideration shows that this approximation is well justified for our system. The Stokes' law for small Reynolds number and spherical objects is:

$$\zeta_i = 6\pi R_i \eta \qquad (17)$$

where η the medium viscosity. We can rewrite τ_i as:

$$\tau_i = m_i / \zeta_i = \frac{4}{3}\pi R_i^3 \rho \Big/ 6\pi R_i \eta = 2R_i^2 \rho / 9\eta \qquad (18)$$

where ϱ is the density of a SV, which is well approximated by the density of water (ϱ = 1,000 kg/m^3) (Takamori et al., 2006). For a rough estimation of τ, we took the viscosity of water η = 1 mPa s (viscosity of water at 22°C is ~1 mPa s, the cytoplasm has higher viscosity and is non-Newtonian) and the largest SV we have in the simulations: 253 nm of diameter (quantal size Q = 40). With these values $\tau = 3.6 \cdot 10^{-9}$ s.

To simulate these equations we used a stochastic 2nd order Runge-Kutta method (Honeycutt, 1992) with a time step δt = 10 μs.

The diffusion coefficient

The Einstein-Stokes equation gives the diffusion coefficient for a spherical object of radius R as a function of temperature and viscosity η of a medium:

$$D_i = \frac{k_B T}{6\pi R_i \eta} \tag{19}$$

Thus, the predicted diffusion coefficient of a 20 nm radius SV in water is around 10,800 nm^2/ms. Due to the higher viscosity of the cytoplasm, its probable non-Newtonian nature, and the crowding due to the presence of other macromolecules, the free diffusion coefficient of a 20 nm radius SV is expected to be around 500 nm^2/ms (Luby-Phelps, 2000). The impact of ribbon tethering on SV mobility is not well understood - depending on the tethering force properties, SV motion might or might not be slowed down. We assumed a SV diffusion coefficient of D = 50 nm^2/ms for a 20 nm radius SV, which is 10 times smaller than predicted for the cytoplasm. Measurements of the diffusion of SVs in ribbon synaptic terminals indicated values from 10 or 30 nm^2/ms (Holt et al., 2004) to 110 nm^2/ms (Rea et al., 2004). These apparent diffusion coefficients are smaller than the expected free diffusion coefficient partly because of SV packing (Cichocki and Hinsen, 1990; Gaffield et al., 2006). We used (19) to account for the dependence of the diffusion coefficient on the SV radius. For a uniquantal SV, the mean square root displacement was 1 nm in a given direction, according to $\delta x^2 = 2D\delta t$. The total mean square displacement is $\delta x^2 = 2dD\delta t$, where d is the dimensionality of the system, here 2.

SV repulsion forces

We modeled the elastic repulsive forces between SVs so that they do not overlap and, if they do, rapidly repel each other. The force of a SV *j* on a SV *i* was (**Figure S6A**):

$$\vec{F}(\vec{x}_j,R_j,\vec{x}_i,R_i) = F_V \cdot S\left(\left|\vec{x}_i-\vec{x}_j\right| - R_i - R_j\right) \cdot \frac{\vec{x}_i-\vec{x}_j}{\left|\vec{x}_i-\vec{x}_j\right|} \tag{20}$$

with F_V = 20 pN and where $\vec{x}_i$ and R_i are the position and the radius of a SV *i*, respectively. $S(d)$ is a sigmoid function:

$$S(d) = \begin{cases} \dfrac{1}{1+\exp\left(\dfrac{d}{\Delta x}\right)} & d \le 7\,\text{nm} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

with Δx = 1 nm. For a point positioned at the border of a force field *F*, the mean penetration depth due to random diffusion into the force field is given by $k_B \cdot T/F$. At room temperature and for F = 20 pN, this is only 0.2 nm. Thus, two SVs diffusing randomly would hardly overlap, as already when touching, the force they exert on each other is $F_V/2$ = 10 pN. If a diffusing SV penetrates by a distance *Y* into the force field *F*, the average time to diffuse out is: $Y/v_0 = Y \cdot k_B \cdot T/(F \cdot D)$, where v_0 is the steady-state velocity resulting from the force. Thus two 20 nm radius SVs completely overlapping would be separated after only 41 μs, when D = 50 nm^2/ms.

In atomic force microscopy experiments, SVs were shown to be less stiff during small deformation, with a stiffness of around 0.2 pN/nm (Awizio et al., 2007). In the linear regime a 10 nm deformation would thus produce a force of only 2 pN. But for larger deformation the force increased up to 400 pN. We chose the force to grow only up to 20 pN to guarantee numerical stability in the simulations. The force amplitudes were independent of SV sizes.

Boundary repulsion and attraction forces

We modeled boundary forces so that SVs stayed inside the simulation space and added a short-range attracting force to the AZ to mimic docking. With *1* the coordinate

perpendicular to the AZ and $\hat{e}^1$ its unitary vector oriented towards the cytoplasm, the boundary force acting on a SV *i* was defined as (**Figure S6A**):

$$\vec{F}(\vec{x}_i, R_i) = F_M S(\vec{x}_i^1 - R_i)\hat{e}^1 + F_A A(\vec{x}_i^1 - R_i)\hat{e}^1 - F_B S(H - \vec{x}_i^1 - R_i)\hat{e}^1 \quad (22)$$

where F_M = 50 pN, F_B = 10 pN, F_A = 1 pN, and the attraction force:

$$A(d) = \begin{cases} -\left(\frac{d}{\Delta x_1}\right)^2 \exp\left(\frac{d}{\Delta x_2}\right) & 0\,\text{nm} \le d \le 20\,\text{nm} \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

with Δx_1 = 1 nm and Δx_2 = 3 nm. The AZ repulsion force was thus assumed stronger (50 pN) than the SV force (20 pN). The maximal attraction force was about 10 times weaker than the maximal force of repulsion. The forces canceled each other when the SV was at a distance of about 3 nm from the AZ. The time a SV stayed docked with no other SV in the simulation space was 15 ± 15 s (mean ± SD).

Homotypic fusion

The homotypic fusion reaction was implemented with the kinetic scheme:

$$\text{Vesicle}_i + \text{Vesicle}_j \xrightarrow{\alpha_{hom}} \text{Vesicle}_k \quad (24)$$

which could only happen if the membranes of the 2 SVs *i* and *j* were closer than 15 nm. This is motivated by the length of the synaptobrevin molecules (Takamori et al., 2006). α_{hom} was fixed or depended on the positions and/or the sizes of the 2 SVs.

At each simulation time step, for every pair (*i*, *j*) of SVs for which:

$$\left|\vec{x}_i - \vec{x}_j\right| - R_i - R_j < 15\,\text{nm} \quad (25)$$

homotypic fusion occurred stochastically with probability $p_{hom} = \delta t\ \alpha_{hom}$. This probability was always much smaller than 1. When required, the height *z* of SVs interaction was calculated as following:

$$z = \vec{x}_i + \frac{\vec{x}_j - \vec{x}_i}{\left|\vec{x}_j - \vec{x}_i\right|}\left(R_i + \frac{\left|\vec{x}_j - \vec{x}_i\right| - R_i - R_j}{2}\right) \quad (26)$$

When homotypic fusion occurred, the two "parent" SVs *i* and *j* disappeared and a new SV *k* emerged in their center of mass:

$$\vec{x}_k = \frac{\vec{x}_i \cdot R_i^3 + \vec{x}_j \cdot R_j^3}{R_i^3 + R_j^3} \tag{27}$$

with quantal size $Q_k = Q_i + Q_j$ and radius $R_k = \sqrt{R_i^2 + R_j^2} = \sqrt{Q_k}\, R_0$ with R_0 = 20 nm. Thus the total lipid bilayer of the "parent" SVs was preserved and their neurotransmitter content was summed. The newly created SV could potentially overlap with an already existing SV. Due to the SV forces, such SVs were rapidly pushed apart.

Exocytosis

The exocytosis reaction was implemented with the kinetic scheme:

$$\text{Vesicle} + \text{AZ} \xrightarrow{\alpha_{exo}} \text{AZ} \tag{28}$$

which could happen when the SV membrane was closer than 15 nm from the AZ surface. At each time step when it was the case, exocytosis occurred with probability $p_{exo} = \delta t\ \alpha_{exo}$. This probability was always much smaller than 1. After exocytosis, the SV was removed from the simulation space. The exocytosis rate was fixed to 3 Hz. It yielded a total maximum release rate of about 36 Hz (maximally 12 uniquantal SVs of 40 nm diameter were docked at the AZ) and allowed time for SVs to perform homotypic fusion before exocytosis.

4 models of homotypic fusion

In homotypic fusion model 1, we assumed that the homotypic fusion rate $\alpha_{hom,1}$ is constant on the whole ribbon (**Figures S6B,C**).

In homotypic fusion model 2, we assumed that the homotypic fusion rate $\alpha_{hom,2}$ decayed with the distance *z* of the 2 SVs interaction point from the AZ, mimicking for example a Ca^{2+} concentration gradient from the AZ (He et al., 2009; Roberts, 1994) (**Figures S6D,E,G**):

$$\alpha_{hom,2}(z) = H_{max}\, e^{-\left(z/\lambda_Z\right)^2} \,/\, e^{-\left(z_0/\lambda_Z\right)^2} \tag{29}$$

where z_0 was set to 20 nm (a uniquantal SV radius) and corresponded to the most AZ-proximal interaction point for SVs. H_{max} was the maximum homotypic fusion rate, found at $z = z_0$. λ_Z was the characteristic decay length of the homotypic fusion rate from the AZ.

In homotypic fusion model 3, the homotypic fusion rate $\alpha_{hom,3}$ decreased both with the distance z from the AZ and with Q_1 and Q_2 of the interacting SVs (**Figures 8K, S6F,H**):

$$\alpha_{hom,3}(z, Q_1, Q_2) = \alpha_{hom,2}(z) \cdot e^{-\left(\frac{Q_1+Q_2-2}{\lambda_Q}\right)^2} \quad (30)$$

where $\alpha_{hom,2}(z)$ is defined in equation (29) and λ_Q quantifies how fast homotypic rates decrease with Q. This model would still reproduce experimental exocytic quantal size distributions for different λ_Z, making the z-dependence dispensable in the third model.

In homotypic fusion model 4, SVs were replenished everywhere on the ribbon, down to 40 nm from the AZ. The homotypic fusion rate function was as in model 3, and the exocytosis rate increased with SV quantal size:

$$\alpha_{exo,4}(Q) = 10 * \alpha_{exo}(Q-1) + \alpha_{exo} \quad (31)$$

Summary of parameters used for simulations

Table S1 summarizes the parameters used for the simulations. For each point in the parameter space (**Figures S6**), 25 simulation repetitions were performed. If one of the simulations led to the creation of a SV with $Q > 40$, this parameter point was considered as “runaway fusion”. Otherwise, from 1,200 exocytosis events, the last 1,000 were taken to construct the exocytic size histograms. Simulations were performed in C++ and analyzed in Mathematica (Wolfram Research).

Supplemental References

Allen, M.P., and Tildesley, D.J. (1989). Computer simulation of liquids (Oxford University Press, USA).

Awizio, A.-K., Onofri, F., Benfenati, F., and Bonaccurso, E. (2007). Influence of synapsin I on synaptic vesicles: an analysis by force-volume mode of the atomic force microscope and dynamic light scattering. Biophys J *93*, 1051–1060.

Cichocki, B., and Hinsen, K. (1990). Dynamic computer simulation of concentrated hard sphere suspensions. I. Simulation technique and mean square displacement data. Physica A *166*, 473–491.

Gaffield, M.A., Rizzoli, S.O., and Betz, W.J. (2006). Mobility of synaptic vesicles in different pools in resting and stimulated frog motor nerve terminals. Neuron *51*, 317–325.

Goodman, M.B., and Lockery, S.R. (2000). Pressure polishing: a method for re-shaping patch pipettes during fire polishing. J Neurosci Methods *100*, 13–15.

Grant, L., Yi, E., Goutman, J.D., and Glowatzki, E. (2011). Postsynaptic recordings at afferent dendrites contacting cochlear inner hair cells: monitoring multivesicular release at a ribbon synapse. J Vis Exp.

Holt, M., Cooke, A., Neef, A., and Lagnado, L. (2004). High mobility of vesicles supports continuous exocytosis at a ribbon synapse. Curr Biol *14*, 173–183.

Honeycutt, R. (1992). Stochastic Runge-Kutta algorithms. I. White noise. Phys Rev A *45*, 600–603.

Jing, Z., Rutherford, M.A., Takago, H., Frank, T., Fejtova, A., Khimich, D., Moser, T., and Strenzke, N. (2013). Disruption of the presynaptic cytomatrix protein bassoon degrades ribbon anchorage, multiquantal release, and sound encoding at the hair cell afferent synapse. J Neurosci *33*, 4456–4467.

Luby-Phelps, K. (2000). Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. Int Rev Cytol *192*, 189–221.

Naraghi, M., and Neher, E. (1997). Linearized buffered Ca^{2+} diffusion in microdomains and its implications for calculation of $[Ca^{2+}]$ at the mouth of a calcium channel. J Neurosci *17*, 6961–6973.

Nielsen, T.A., DiGregorio, D.A., and Silver, R.A. (2004). Modulation of glutamate mobility reveals the mechanism underlying slow-rising AMPAR EPSCs and the diffusion coefficient in the synaptic cleft. Neuron *42*, 757–771.

Nouvian, R., Beutner, D., Parsons, T.D., and Moser, T. (2006). Structure and function of the hair cell ribbon synapse. J Membr Biol *209*, 153–165.

Rea, R., Li, J., Dharia, A., Levitan, E.S., Sterling, P., and Kramer, R.H. (2004). Streamlined synaptic vesicle cycle in cone photoreceptor terminals. Neuron *41*, 755–766.

Sahara, Y., and Takahashi, T. (2001). Quantal components of the excitatory postsynaptic currents at a rat central auditory synapse. J Physiol (Lond) *536*, 189–197.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis (CRC Press).

Stiles, J.R., Van Helden, D., Bartol, T.M., Salpeter, E.E., and Salpeter, M.M. (1996). Miniature endplate current rise times less than 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle. Proc Natl Acad Sci USA *93*, 5747–5752.

Stiles, J.R., and Bartol, T.M. (2001). Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. In Computational Neuroscience: Realistic Modeling for Experimentalists, E. De Schutter, ed. (CRC Press, Boca Raton, FL), pp. 87–127.