NEUROSCIENCE
RESEARCH ARTICLE

E. Neher, H. Taschenberger / Neuroscience 458 (2021) 182–202

Non-negative Matrix Factorization as a Tool to Distinguish Between Synaptic Vesicles in Different Functional States

Erwin Neher [a,c]* **and Holger Taschenberger** [b]

[a] *Emeritus Laboratory of Membrane Biophysics, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany*

[b] *Department of Molecular Neurobiology, Max Planck Institute of Experimental Medicine, 37075 Göttingen, Germany*

[c] *Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Germany*

Abstract—Synaptic vesicles (SVs) undergo multiple steps of functional maturation (priming) before being fusion competent. We present an analysis technique, which decomposes the time course of quantal release during repetitive stimulation as a sum of contributions of SVs, which existed in distinct functional states prior to stimulation. Such states may represent different degrees of maturation in priming or relate to different molecular composition of the release apparatus. We apply the method to rat calyx of Held synapses. These synapses display a high degree of variability, both with respect to synaptic strength and short-term plasticity during high-frequency stimulus trains. The method successfully describes time courses of quantal release at individual synapses as linear combinations of three components, representing contributions from functionally distinct SV subpools, with variability among synapses largely covered by differences in subpool sizes. Assuming that SVs transit in sequence through at least two priming steps before being released by an action potential (AP) we interpret the components as representing SVs which had been 'fully primed', 'incompletely primed' or undocked prior to stimulation. Given these assumptions, the analysis reports an initial release probability of 0.43 for SVs that were fully primed prior to stimulation. Release probability of that component was found to increase during high-frequency stimulation, leading to rapid depletion of that subpool. SVs that were incompletely primed at rest rapidly obtain fusion-competence during repetitive stimulation and contribute the majority of release after 3–5 stimuli.

Key words: transmitter release, quantal analysis, Short-term plasticity, calyx of Held.

*Correspondence to: E. Neher, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, D-37077 Göttingen, Germany.
E-mail address: eneher@gwdg.de (E. Neher).

Abbreviations: $[Ca^{2+}]_i$, concentration of intracellular Ca^{2+} ions; *BF*, basefunction, i.e. the normalized time course of quantal release during a stimulus train contributed by a specific fraction of SVs which were in a certain state immediately prior to stimulation, BF_S are arrays of $j = 1$ to J entries for trains consisting of J stimuli and are normalized to a cumulative sum of 1. The product of BF_j with the train quantal content M (see below) yields the release measured in SVs for stimulus j contributed by that specific fraction of SVs.; BF_{LS}, basefunction representing normalized time course of release contributed by those SVs that were in LS immediately prior to stimulation (SV_{LS}s) or other SVs with low release probability; $BF_{LS,RS}$, basefunction representing normalized time course of the release contributed by SV_{LS}s and SV_{RS}s as returned by a two-component NMF decomposition fit, $BF_{LS,RS}$ corresponds to the sum of BF_{LS} plus BF_{RS} after renormalization to a cumulative sum of 1; BF_{RS}, basefunction representing normalized time course of release contributed by those SVs that were neither in TS nor in LS immediately prior to stimulation but are newly recruited and released during a stimulus train (SV_{RS}s); BF_{TS}, basefunction representing normalized time course of release contributed by those SVs that were in TS immediately prior to stimulation (SV_{TS}s) or other SVs with high release probability; eEPSC, evoked excitatory postsynaptic current; ISI, inter-stimulus interval; LS, 'loosely docked state' of an SV or else a state of low release probability; *m*, quantal content (number of SVs) released during a single AP-evoked EPSC; *M*, quantal content of an entire eEPSC train i.e. the sum of all SVs contributing to release during the entire stimulus train usually defined as an array with as many entries as there are synapses in the data set; M_{LS}, fraction of *M* representing those SVs that had been in LS immediately prior to stimulation; $M_{LS,RS}$, fraction of *M* representing those SVs that had been either in LS immediately prior to stimulation or else are newly recruited and released during a stimulus train; M_{RS}, fraction of *M* representing those SVs that had been neither in TS nor in LS immediately prior to stimulation but are newly recruited and released during a stimulus train; M_{TS}, fraction of *M* representing those SVs that had been in TS prior to stimulation, M_{TS} represents the entire subpool of SV_{TS}s (SP_{TS}) available for release immediately prior to stimulation if a stimulus train is long enough such that BF_{TS} decays to 0.; $\boldsymbol{p_{LS}}$, apparent probability for a pre-existing SV_{LS}s of contributing to release for a given stimulus within a train given that it had not been released before, used in the context of a sequential SV priming scheme in which SV_{LS}s are not fusion competent but need to convert to SV_{TS}s before being capable of undergoing fusion; *p*, vesicular release probability i.e. the probability of a docked and fusion competent SV being released during a single AP; p_{LS}, probability for a pre-existing SV_{LS} of contributing to release for a given stimulus within a train given that it had not been released before, used in the context of a non-sequential SV priming scheme in which SV_{LS}s are fusion competent; p_{TS}, probability for a pre-existing SV_{TS} of contributing to release for a given stimulus within a train given that it had not been released before; SP_{LS}, subpool of SV_{LS}s i.e. all SVs in LS at any given moment; SP_{TS}, subpool of SV_{TS}s i.e. all SVs in TS at any given moment; STP, short-term plasticity; SV, synaptic vesicle; SV_{LS}, synaptic vesicle currently in LS; SV_{TS}, synaptic vesicle currently in TS; TS, 'tightly docked state' of an SV or else a state of high release probability.

https://doi.org/10.1016/j.neuroscience.2020.10.012
0306-4522/

INTRODUCTION

The traditional view on quantal release at presynaptic terminals posits that there is a certain number of synaptic vesicles (SVs) in a release-ready state, a certain – typically small – fraction of which is released upon the arrival of an action potential (AP) (Katz, 1969). Accordingly, electrophysiological methods for analyzing the release process aim at determining the size of the so-called 'readily releasable pool' (RRP) and the released fraction (F) (Rosenmund and Stevens, 1996; Schneggenburger et al., 1999), also termed 'vesicular release probability' (p). Most of the methods currently in use assume the SVs comprising the RRP to be homogeneous with respect to p. Accumulating recent evidence, however, established that time courses of neurotransmitter release during low- and high-frequency trains of presynaptic APs are not compatible with the concept of a single, homogeneous pool of SVs. To explain experimentally observed changes in quantal content (m) during repetitive presynaptic AP firing, often multiple SV subpools with distinct functional properties had to be assumed (Hanse and Gustafsson, 2001; Moulder and Mennerick, 2005; Schlüter et al., 2006; Hallermann et al., 2010a; Müller et al., 2010; Taschenberger et al., 2016; Doussau et al., 2017; Miki et al., 2018). Variability in the relative contributions of those functionally distinct SV subpools may be the cause for the large variability in synaptic strength and short-term plasticity (STP) observed at seemingly similar synapses (Taschenberger et al., 2016). For instance, detailed studies of the statistics of synaptic responses at hippocampal CA3–CA1 synapses led to the postulate of 'pre-primed' SVs, representing a variable fraction of SVs in the RRP that are preferentially released early during stimulus trains due to their high p. They are complemented by other SVs, which first need to be recruited to release sites during stimulation before being released with lower p (Hanse and Gustafsson, 2001). Likewise, studies on cultured hippocampal autapses (Schlüter et al., 2006) and at the calyx of Held (Müller et al., 2010; Taschenberger et al., 2016) identified so-called 'superprimed' SVs with properties similar to those of 'pre-primed' SVs. These findings call for more elaborate analysis methods to separate release components contributed by SVs residing in distinct functional states. Properties of these SV subpools, such as the respective p values, time courses of release during repetitive stimulation, and rates of SV recruitment to release sites need to be established. Here, an attempt is made to determine such functional synaptic parameters by using non-negative matrix factorization (NMF) or the related technique of non-negative tensor factorization (NTF) to analyze ensembles of eEPSC trains obtained from several calyx of Held synapses, each subjected to the same stimulus protocol at various stimulation frequencies.

NMF and NTF are two varieties of so-called 'blind source separation techniques' (Cichocki et al., 2009) which can be applied to complex data sets in order to reduce their dimensionality or else to describe a given data set as the sum of a small number of components (here normalized SV release time courses representing contributions by distinct SV subpools during trains). Subsets of the data (here responses of individual synapses) are described by coefficients, which indicate how much a given SV subpool contributes to the total release for that individual synapse, while the normalized time courses of release from given subpools are the same for all synapses. Like other blind source separation techniques, NMF and NTF provide both such coefficients and SV release time courses as the result of minimizing a cost function reflecting the quality of the fit. NMF and NTF use the relatively weak criterion of non-negativity to reduce the number of possible fits. As discussed below, this leads to degeneracy, which has to be removed by suitable further constraints. Other blind source techniques, such as principal component analysis (PCA), use stronger criteria, e.g. orthogonality of principal components, resulting in less degeneracy. PCA was recently applied to a set of data from glutamatergic synapses in the cerebellum (Dorgans et al., 2019), which provided a classification of synapses into four categories with different STP properties. However, the principal components included negative values and bore no direct relationship with the underlying release time courses during STP. The aim here is to find a set of functions providing the time courses of different release components. These time courses are postulated to be the same for all synapses for a given stimulation frequency and for all SVs in a given functional state. The relative contributions of distinct SVs subpools to the release time course can be widely different among synapses which accounts for synapse variability both in terms of their initial synaptic strength as well as their STP.

EXPERIMENTAL PROCEDURES

Electrophysiology

All experiments were performed on 200 μm thick brainstem slices of juvenile post-hearing (P13–16) Wistar rats at room temperature (22–24 °C) essentially as described before (Taschenberger and von Gersdorff, 2000). Whole-cell patch-clamp recordings were made from principle neurons of the medial nucleus of the trapezoid body. eEPSCs were elicited by afferent fiber stimulation in a bath solution containing 2 mM Ca^{2+} and 1 mM Mg^{2+} and recorded at room temperature in the presence of 1 mM kynurenic acid (kyn) in order to minimize AMPAR desensitization and saturation (Diamond and Jahr, 1997; Neher and Sakaba, 2001; Taschenberger et al., 2002; Wong et al., 2003). Bicuculline methiodide (25 μM) and strychnine (5 μM) were added to block inhibitory synaptic currents. Trains of 25 stimuli were applied at six different stimulation frequencies (5, 10, 20, 50, 100, 200 Hz), in some experiments preceded by 2 or more conditioning stimuli at 10 Hz (see section on NTF). The interval between successive stimulus trains was $\geq$15 s which was sufficient to allow for complete recovery from synaptic depression. Three or four repetitions of the whole protocol were recorded and eEPSC peak amplitudes obtained from corresponding traces were averaged. Peak eEPSC amplitudes were determined after off-line compensation for remaining

series resistance and offset correction. To further minimize effects of AMPAR desensitization and saturation on eEPSC peak amplitudes, only synapses with initial eEPSC amplitudes $\leq$2.6 nA (in the presence of 1 mM kyn) were selected for analysis. Part of the experimental data set was common with that of a recent publication (Taschenberger et al., 2016) and was reanalyzed.

NMF and NTF analysis

All data analysis was performed with Igor Pro (Wavemetrics). Mean eEPSC peak amplitudes from several synapses at a given frequency were assembled into a matrix, the rows of which represent the peaks during a stimulus train on a given synapse. The NMF data set analyzed here included 20 synapses, resulting in 20 $\times$ 25 data matrices (Fig. 1A), one matrix for each stimulation frequency. For NTF, only five synapses with sufficiently stable recordings were available, including trains of 100 Hz with two or five conditioning pre-pulses. Thus, the tensor for 100 Hz had three layers, namely a 5 $\times$ 25 matrix with amplitude values without conditioning, a second one with amplitudes from the 100 Hz episode preceded by two conditioning pulses, and a third one with values after five conditioning pulses. In addition, each of the other frequencies (without conditioning pre-pulses) provided another 5 $\times$ 25 matrix. In order to convert eEPSC peaks into quantal content m, we assumed an ''effective quantal size'' $q^* = -6.6$ pA in the presence of 1 mM kyn in the bath. This q^* value was derived by considering a measured average mEPSC size of about -60 pA for P13–16 rat calyx synapses, taking into account the experimentally determined eEPSC block by 1 mM kyn and applying a scaling factor that corrects eEPSC peak-based m estimates for the temporal dispersion of AP-evoked release (Taschenberger et al., 2005).

The source code for iterative non-negative fitting (available at request from the corresponding author) is written for NTF, based on the algorithm originally published by Lee and Seung (2001) and used by Neher et al. (2009). For this reason, NMF matrices, as discussed here, have to be declared as standard tensors with only one layer.

For a three-component fit, basefunctions (*BF*s; see *Results* for meaning and nomenclature) were initialized for all stimuli j running from 1 to J according to:

$$BF_{TS,j} = A_{TS} \times e^{-(j-1)/T}$$

$$BF_{LS,j} = A_{LS} \times \left(1 - (1 - S_{LS}) \times e^{-(j-1)/L1}\right) \times e^{-(j-1)/L2} + bL$$

$$BF_{RS,j} = A_{RS} \times \left(1 - e^{-(j-1)/R}\right)$$

where j is the index for stimulus number and the following standard values are used: $T = 1$, $L1 = 2$, $L2 = 7$, $R = 7$, $S_{LS} = 0.001$ (a parameter setting the starting point for BF_{LS}), $bL = 0.002$ (an offset to allow small changes at late times during iterations). A_{TS}, A_{LS} and A_{RS} are scaling factors chosen to yield a cumulative sum of 1 for initial guesses of BF_{TS}, BF_{LS} and BF_{RS}, respectively. Subtracting 1 from the running stimulus index j ensures starting values of very close to zero and zero for $BF_{LS,1}$ and $BF_{RS,1}$, respectively. For the two-component fit, BF_{LS} and BF_{RS} are replaced by

$$BF_{LS,RS,j} = A_{LS,RS} \times \left(1 - (1 - S_{LS}) \times e^{-(j-1)/LR}\right)$$

with $LR = 2$. For simulating alternative kinetic schemes of SV priming and fusion in which SV_{LS}s were allowed to directly undergo fusion, the parameter S_{LS} was varied between 0.001 and 1. NMF, as used here, always normalizes *BF*s to a cumulative sum of 1 during fit iterations. Such normalization is compensated by reciprocal changes in *M*s.

For a two-component fit, M values were initialized to 470 and 2110 pre-existing SV_{TS}s for M_{TS}, and the combined LS,RS pool, respectively, and 200 iterations were used. For three-component fits the initial M_{TS} was set to the final value of a preceding two-component fit. M_{TS} and M_{LS} values were set to 47% and 53% of the combined LS,RS pools of such a fit, respectively and 100 iterations were used.

The initial scaling factors SF_{TS}, SF_{LS} and SF_{RS} for NTF analysis were calculated as a 40% attenuation per conditioning eEPSC for the TS-component, 10% attenuation for the LS-component, and 4% for the RS-component.

For comparisons of goodness of fit between different conditions a measure of χ^2, denoted as χ^{2*}, was calculated as the mean squared deviation between data and fit divided by the fit (assuming Poisson statistics and statistical independence of data points). However, statistical independence is violated in some of the early responses in trains (Scheuss and Neher, 2001). Therefore, what is given as χ^{2*}, in the text can only be used for comparison between different fits. For calculating this quantity, only the first 10 values in a train were used.

During fitting, the χ^{2*} rapidly approached a stationary value within 5 to 10 iterations. Further changes during iterations depended on the strengths of amplitude and p constraints (see *Results* for details). They were small ($<$10%) for weak constraints, but tended to increase substantially after some 20 iterations for stronger constraints. A compromise had to be found for which the secondary increases were small, while avoiding variations of first amplitude and p values higher than those expected on the basis of Poisson statistics. It turned out that the 'fraction' of the correction, as explained in the *Results*, had to be set to 15% both for the M constraint and for the p values. For fine-tuning the decomposition into LS- and RS-components, a bias towards LS was applied in each iteration by increasing M_{LS} and decreasing M_{RS} by 0.34% each. However, in all fitting runs, constraints were removed during the last 15 iterations, such that a local minimum of χ^{2*} could be reached.

Estimation of vesicle pool sizes

In order to estimate the size of the pre-existing LS-pool (M_{LS}) the results of the two-component fit of the 200 Hz trains were used. The average release time course of

the second component (sum of LS- and RS-contributions) was calculated as:

$\overline{m}_j = \overline{M}_{LS,RS} \times BF_{LS,RS,j}$

where the overbar denotes averages over all synapses and j is the stimulus number. This time course was analyzed using standard methods of pool size estimation, such as the EQ- and the SMN-method

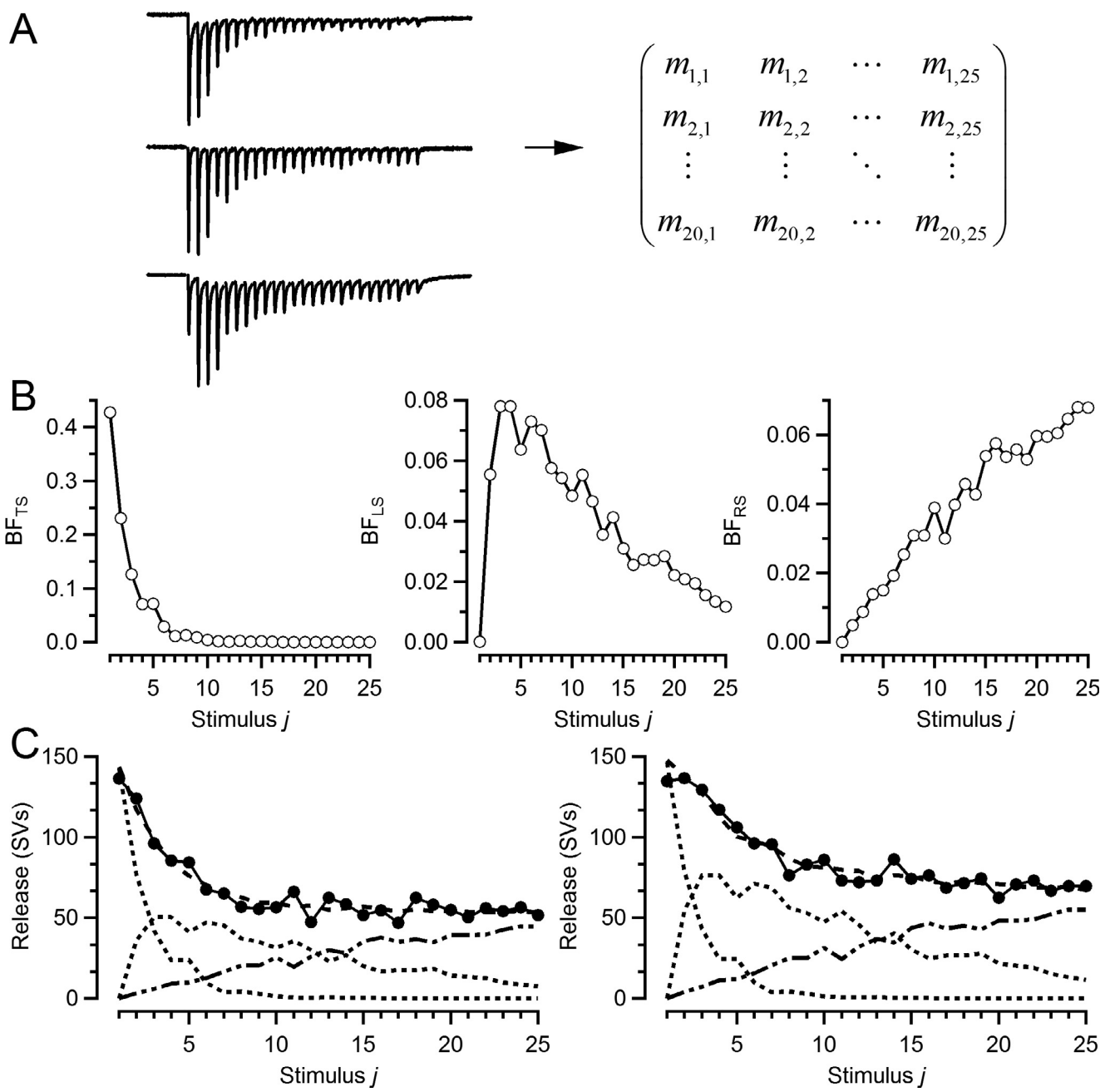


Fig. 1. Principle and examples of NMF decomposition of eEPSC train data. **(A)** Composition of the data set subjected to NMF analysis. Left, eEPSCs trains (exemplified by 200 Hz eEPSC trains from three different synapses) were recorded in response to afferent fiber stimulation from a total of 20 P13–16 rat calyx of Held synapses. Trains of 25 stimuli at various frequencies (5, 10, 20, 50, 100, 200 Hz) were applied, in some experiments preceded by 2 or more conditioning stimuli delivered at 10 Hz. Right, Matrix of quantal content estimates $m_{i,j}$ derived for synapse i = 1–20 and stimulus number j = 1–25. The data set for a complete NMF decomposition fit consisted of a total of six such matrices representing six different stimulation frequencies. An ''effective quantal size" of q^* = −6.6 pA was assumed to estimate m (see *Experimental procedures* for details). For mean quantal contents and example recordings see Fig. 2. **(B)** A set of three *BF*s for 20 Hz eEPSC trains that satisfy the criterion that all time courses of release from the given data set of 20 synapses and the given stimulation frequency of 20 Hz can be approximated as linear combinations of those same three basefunctions. *BF* values (always normalized to a cumulative sum of 1) are plotted against stimulus number. **(C)** Examples from two synapses exhibiting different STP (left: depressing; right: slightly facilitating) when stimulated at 20 Hz. Solid traces with filled circles: Experimental m values (such as those in a row of the matrix in **(A)**) plotted against stimulus number. Superimposed are NMF fits (dashed traces). Dotted traces: products of $BF \times M$ for three release components characterized by the *BF*s shown in **(B)**, the sum of these products provides the fit to the data of the given synapse. The main difference between the two synapses illustrated is a higher number of pre-existing SV_{LS}s prior to stimulation in the synapse on the right (1077 SVs) compared with the synapse on the left (729 SVs).

(detailed in Neher, 2015). Reported values are averages over the six frequencies employed and are given as mean ± SD.

Renormalization of basefunctions BF_{LS}s

In order to obtain accurate values for release probabilities from basefunctions, stimulus trains must be long enough to completely deplete the respective SV subpools. This was not quite fulfilled for the BF_{LS}s. In order to correct for this shortcoming, BF_{LS}s were renormalized before calculating apparent release probabilities $p'_{r,LS}$. This was performed the following way: The decaying part of a given BF_{LS} was fitted by an exponential with the baseline forced to zero. The area, A_{tail}, under this fit from the last point of the BF_{LS} to infinity was calculated, followed by division of the BF_{LS} by $(1 + A_{tail})$. Renormalizations were typically on the order of 10 to 20% and was only used for calculation of $p'_{r,LS}$. All BF_{LS}s, displayed in figures, were not renormalized, but shown, as they are returned by NMF (i.e. normalized for a cumulative sum of 1).

RESULTS

Concept of a three-component NMF decomposition fit to eEPSC trains

The strategy for decomposing release time courses into components contributed by distinct SVs subpools will be explained here by means of a data set which comprises quantal content (m) estimates for eEPSC trains recorded from calyx of Held synapses in acute brain slice preparations (Forsythe and Barnes-Davies, 1993). Presynaptic AP trains consisting of 25 APs each were elicited by afferent fiber stimulation at six frequencies (5–200 Hz). Peak eEPSC amplitudes were obtained for the same set of stimulation frequencies from a total of 20 individual synapses. Thus, the data set for a given stimulus frequency can be represented by a matrix of 20×25 eEPSC peak amplitudes (Fig. 1A). Entries were converted into quantal content by dividing eEPSC peaks by an ''effective quantal size'' q^* of −6.6 pA (see *Experimental procedures* for details). Mean time courses of quantal release are shown for all stimulus frequencies in Fig. 2A. Characteristically at the calyx of Held (Borst et al., 1995; Taschenberger and von Gersdorff, 2000; Sahara and Takahashi, 2001) and also at other glutamatergic synapses (Debanne et al., 1996; Dobrunz and Stevens, 1997), the number of SVs released by single APs or in response to the first AP in a stimulus train is quite variable among synapses. Correlated with this variability are pronounced differences in STP (Taschenberger et al., 2016; Fekete et al., 2019). Synapses with high initial synaptic strength typically display strong and fast synaptic depression during repetitive activation, while those with lower initial synaptic strength often facilitate for 2nd and 3rd eEPSCs, before they develop depression (Fig. 2B). The reason for this variability was described to reside in the relative abundance of two classes of functionally distinct readily releasable SVs present at resting calyx terminals: so-called 'superprimed' SVs, which fuse with high probability upon AP arrival and therefore are rapidly consumed during the onset of high-frequency AP trains, and a 2nd class of SVs, which need an additional step of maturation and therefore have low or zero p initially. The relative contribution of that latter class of SVs to the total release increases during stimulus trains because of their slower consumption (Taschenberger et al., 2016). A certain level of steady-state release is maintained by SVs that are newly recruited to release sites. They will be released during continued stimulation and constitute a 3rd class of released SVs.

A desirable objective of NMF would be to describe quantal release during AP trains as the sum of three components representing the release time courses contributed by distinct subpools of SVs, which had been in one of three functional states prior to stimulation. We call these time courses 'basefunctions' (*BF*s) and consider three such *BF*s: (1) BF_{TS} represents release contributed by SVs, which had been in the 'fully-primed' or '**t**ightly-docked' **s**tate prior to stimulation. (2) BF_{LS} represents release contributed by SVs, which likewise had been docked at release sites prior to stimulation, but were incompletely primed or in a '**l**oosely docked **s**tate. (3) BF_{RS} represents release contributed by SVs that were not yet docked prior to stimulation but are newly recruited to release sites and subsequently released during ongoing stimulation. As we will show, given the exemplar data set, the analysis goal can only be achieved by invoking additional information for the separation of components (2) and (3).

The quantal content $m_{i,j}$ for each stimulus j at a given synapse i for the case of three components is calculated as a linear combination of the three *BF*s according to:

$$m_{i,j} = M_{TS,i} \times BF_{TS,j} + M_{LS,i} \times BF_{LS,j} + M_{RS,i} \times BF_{RS,j} \quad (1)$$

where $M_{TS,i}$, $M_{LS,i}$ and $M_{RS,i}$ are quantal contents of the entire eEPSC train contributed by the respective pre-existing SV_{TS}s and SV_{LS}s and those SVs that are new recruited and subsequently released (SV_{RS}s) at a given synapse *i.* Importantly, $BF_{TS,j}$, $BF_{LS,j}$, *and* $BF_{RS,j}$ are the same for all synapses.

Alternatively, the release time course can be described in terms of two components (see below). In this case components (2) and (3) are merged and represented in the equivalent of Eq. (1) by the product $M_{LS,RS} \times BF_{LS,RS}$.

*BF*s are normalized to a cumulative sum of 1

$$\sum_{j=1}^{J} BF_{TS,j} = \sum_{j=1}^{J} BF_{LS,j} = \sum_{j=1}^{J} BF_{RS,j} = 1,$$

such that for each synapse i

$$\sum_{j=1}^{J} m_{i,j} = M_{TS,i} + M_{LS,i} + M_{RS,i}, \quad (2)$$

where J is the total number of stimuli in a train, in our data set always 25.

*BF*s which decay to near zero during sufficiently long stimulus trains indicate nearly complete consumption of those SVs of the respective SV subpool that pre-existed prior to stimulation. For such *BF*s, their corresponding *M*s therefore represent a size estimate of the respective

SV subpool. Variability among synapses originates from different relative abundances of SV_{TS}s versus SV_{LS}s prior to stimulation (Fig. 1C).

During the fitting procedure, the NMF algorithm iteratively updates both *BF*s and *M*s for the minimization of a cost function (see below). A data set obtained from I synapses, J stimuli per train, and R components, is described by $R \times (I + J)$ parameters for a given frequency, since for each component r, one M parameter per synapse and a BF consisting of J values common to all synapses is used in Eq. (1). Since *BF*s are normalized and constraints are applied, the number of fitting parameters is somewhat below $R \times (I + J - 1)$. The data set contains $I \times J$ amplitudes per frequency. Thus, in order for the number of observations to be larger than the number of fitting parameters, data from at least four synapses are required for $R = 3$ components and $J = 25$ stimuli per train.

Surprisingly, it is not difficult for an NMF algorithm to find a set of *BF*s and *M*s, which provides a good fit to all 20 calyx synapses (see Fig. 1C for two examples), in spite of their quite variable STP (Fig. 2B). The problem of NMF rather is that there are multiple solutions that satisfy the weak criterion of non-negativity. This is expected, since the data set is treated as a linear system and solutions are only confined by the restriction that both *BF*s and *M*s are non-negative. Thus, any linear combination of *BF*s from a given solution with correspondingly altered *M*s is again a solution to the decomposition problem, as long as the non-negativity constraints are not violated. A general experience with NMF is that multiple solutions exist unless the underlying matrices and vectors are sparse, i.e. unless they have a large number of zeros (Cichocki et al., 2009). This, unfortunately, is not the case for most of our data, as described below. Thus, one needs to find constraints which reduce the number of possible solutions to those that are compatible with a presumed kinetic scheme of SV priming and fusion. For this reason, an NMF decomposition fit result can rarely be considered as a unique solution to a given set of eEPSC train data. Rather, NMF analysis should be regarded as a tool to explore what release time courses and SV subpool sizes are compatible with the synapse-to-synapse variability within the given eEPSC train dataset, dependent on model assumptions.

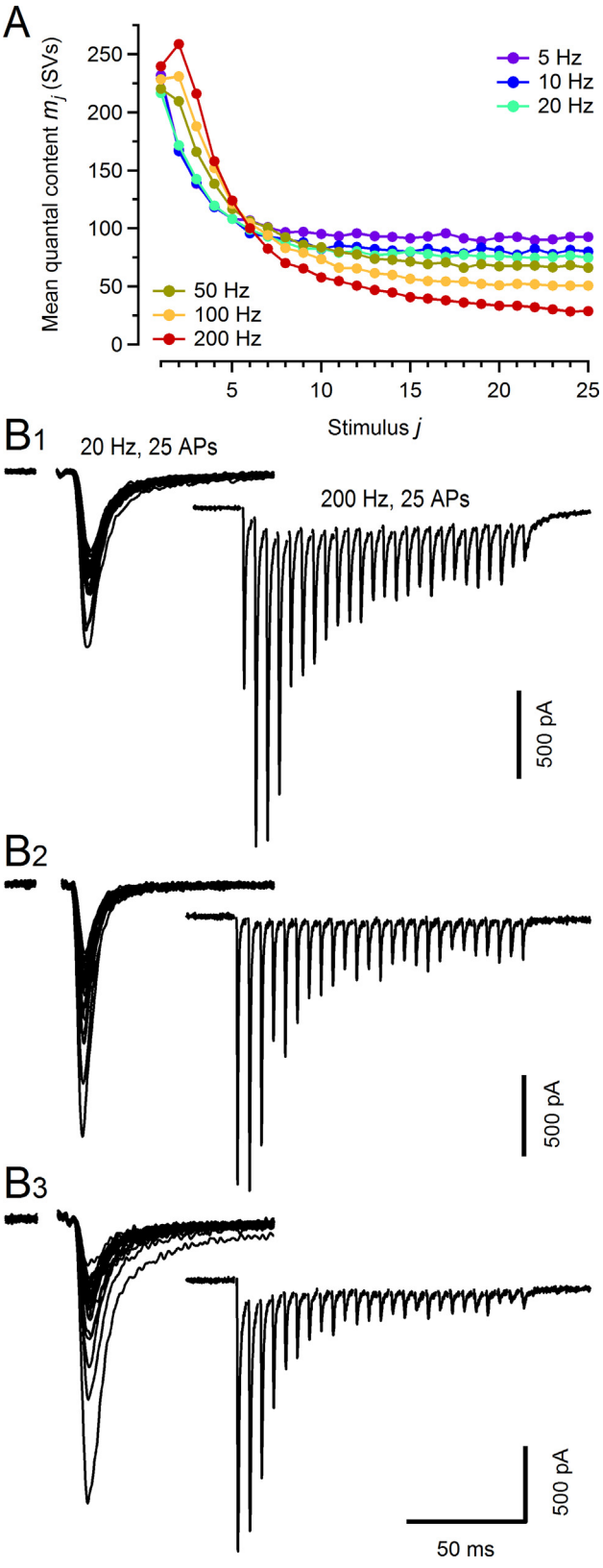


Fig. 2. eEPSC train data set subjected to NMF decomposition analysis. **(A)** Average quantal content estimates $\overline{m}_j$ for stimuli $j = 1$ to 25 of eEPSC trains over a total of twenty P13–16 rat calyx of Held synapses. Stimulation frequencies ranged from 5 to 200 Hz. For each synapse, ≥3 repetitions of the respective eEPSC train for a given frequency were acquired. eEPSC train peaks were obtained for each trial and subsequently averaged over repetitions. For a given stimulation frequency, the m estimates for each synapse contribute a single row to the NMF data set matrix (Fig. 1A right). For the $\overline{m}$ estimates over all synapses, initial synaptic facilitation is evident only for the highest frequencies (50, 100, 200 Hz). Steady state $\overline{m}$ estimates show a frequency-dependent degree of depression. Though for stimulation frequencies between 5 and 50 Hz, the steady state $\overline{m}$ is relatively similar. **(B)** Representative recordings from three different calyx synapses showing either synaptic facilitation (B1) or different degrees of depression (B2, B3) during the onset of 200 Hz eEPSC trains (right column). Stimulus-aligned eEPSCs from 20 Hz trains are show in the left column for comparison. The 200 Hz eEPSC trains are the same as those depicted in the schematic representation in Fig. 1A left.

Constraints derived from a presumed sequential SV priming and fusion scheme

For our analysis, we assume that SVs undergo a sequence of docking, priming and fusion as described by the kinetic scheme illustrated in Fig. 3 and discussed in detail in Neher and Brose (2018). Alternative assumptions about the pool configuration and their consequences on NMF fit results will be presented below. Support for the scheme of Fig. 3 comes from EM studies showing that SVs can exist in a tightly docked state characterized by a membrane-to-membrane distance in the range 1–5 nm (Imig et al., 2014; Chang et al., 2018). The presence of such tightly docked SVs depends on the integrity of a number of presynaptic proteins (for review see Südhof, 2012; Imig et al., 2014). Further, in view of molecular (Prinslow et al., 2019) and physiological (Zenisek et al., 2000; He et al., 2017) evidence that SV priming is a reversible process, it was proposed that SVs docked to release sites can exist in at least two functional states: (1) in a loosely docked state representing fusion incompetent SV_{LS}s with only minimally zippered SNARE complexes and (2) in a tightly docked state representing fusion competent SV_{TS}s with well-zippered SNARE complexes (Neher and Brose, 2018). At rest, SV_{TS}s are in a dynamic equilibrium with SV_{LS}s. A third class of SVs contributing to release during stimulus trains are those, which are newly recruited to release sites during ongoing stimulation and need to undergo the whole sequence of docking and two-step priming, before they can fuse.

The following four constraints for the temporal profiles of *BF*s are implied by such a kinetic scheme (see Fig. 4A, B and Fig. 6A, B for examples of *BF*s):

(1) BF_{TS} starts with a high value, which accounts for all SVs released during the first eEPSC in a train (Fig. 4A, B). BF_{TS} rapidly decays to zero because pre-existing SV_{TS}s are quickly consumed due to their high p.
(2) BF_{LS} starts at zero, since SV_{LS}s need to convert to SV_{TS}s before fusion can occur (BF_{LS}, Fig. 6A, B). Similarly to BF_{TS}, but more slowly, BF_{LS} approaches zero while pre-existing SV_{LS}s are progressively consumed.
(3) BF_{RS} starts at zero, increases during trains and eventually accounts for all release, when all SVs that had been either in LS or in TS prior to stimulation have been completely consumed (Fig. 6C, D).
(4) No negative values are allowed for *BF*s, since the contribution of any SVs to release needs to be positive. This constraint is intrinsic to NMF.

The constraint of $BF_{LS,1} = 0$ is specific to the kinetic scheme shown in Fig. 3. While it seems reasonable that only SVs with a fully assembled and tightly zippered release machinery can undergo exocytosis within the short episode of an AP-induced $[Ca^{2+}]_i$ transient, some published models (Trommershäuser et al., 2003; Hallermann et al., 2010a, 2010b) allow release from two or more states (or sites). We start the analysis here with all four constraints listed above applied and explore later the consequences of relaxing the constraint on $BF_{LS,1}$.

The constraints, listed so far, are readily enforced by the choice of initial guess values for the fit parameters. NMF is an iterative algorithm, minimizing a cost function, which usually is the mean squared deviation between experimental data and fit. The fit is a linear superposition of *BF*s multiplied by their respective *M*s (see Eq. (1)). Both, individual values of *BF*s and *M*s, are updated during each iteration cycle. The algorithm employed here, first described by Lee and Seung (2001), uses step sizes and gradients during iterations,

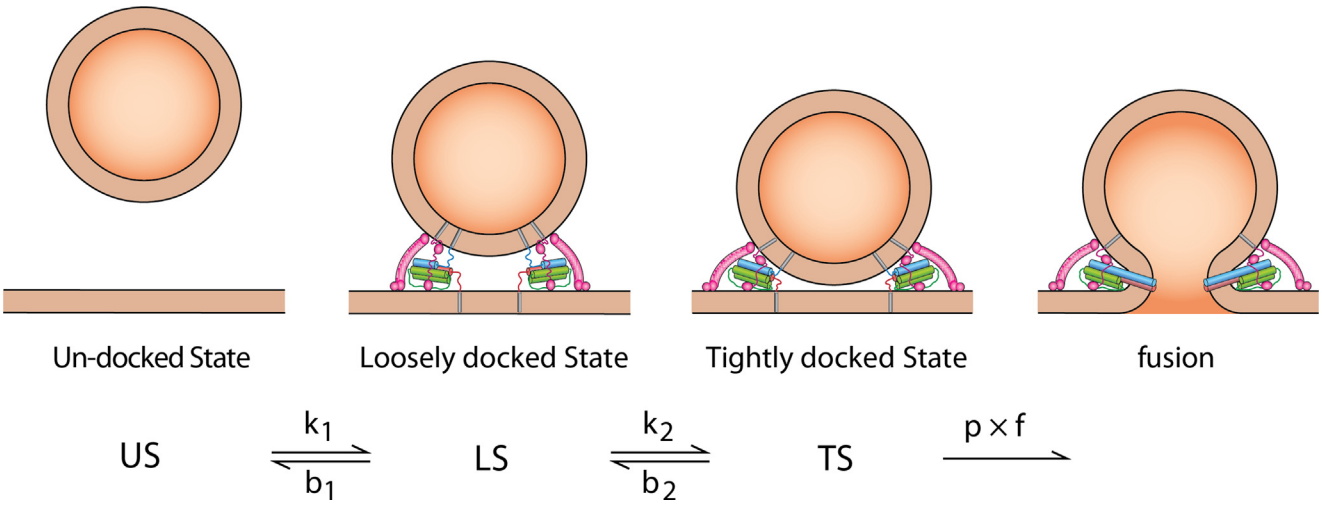


Fig. 3. A sequential model of SV priming and fusion with two functional SV docking states. Diagram of release sites and functional states (top) and the corresponding kinetic scheme (bottom). The total number of docking sites is assumed to be fixed. SVs available for docking (US) bind to an empty docking site and first remain in a loosely docked state (LS). SV docking is assumed to be a reversible step. Once an SV is docked, the LS is in dynamic equilibrium with the tightly docked state (TS), from where SVs can fuse upon action potential arrival with a probability p. The average rate constant of SV fusion during stimulus trains is given by the product of release probability times stimulation frequency $p \times f$. SVs are assumed to remain docked to a given release site, when undergoing LS ⇌ TS transitions, unlike other models, which assume separate release sites between which SVs migrate (Miki et al., 2018) or from which they can be released with site-specific release probabilities (Trommershäuser et al., 2003; Schlüter et al., 2006; Taschenberger et al., 2016).

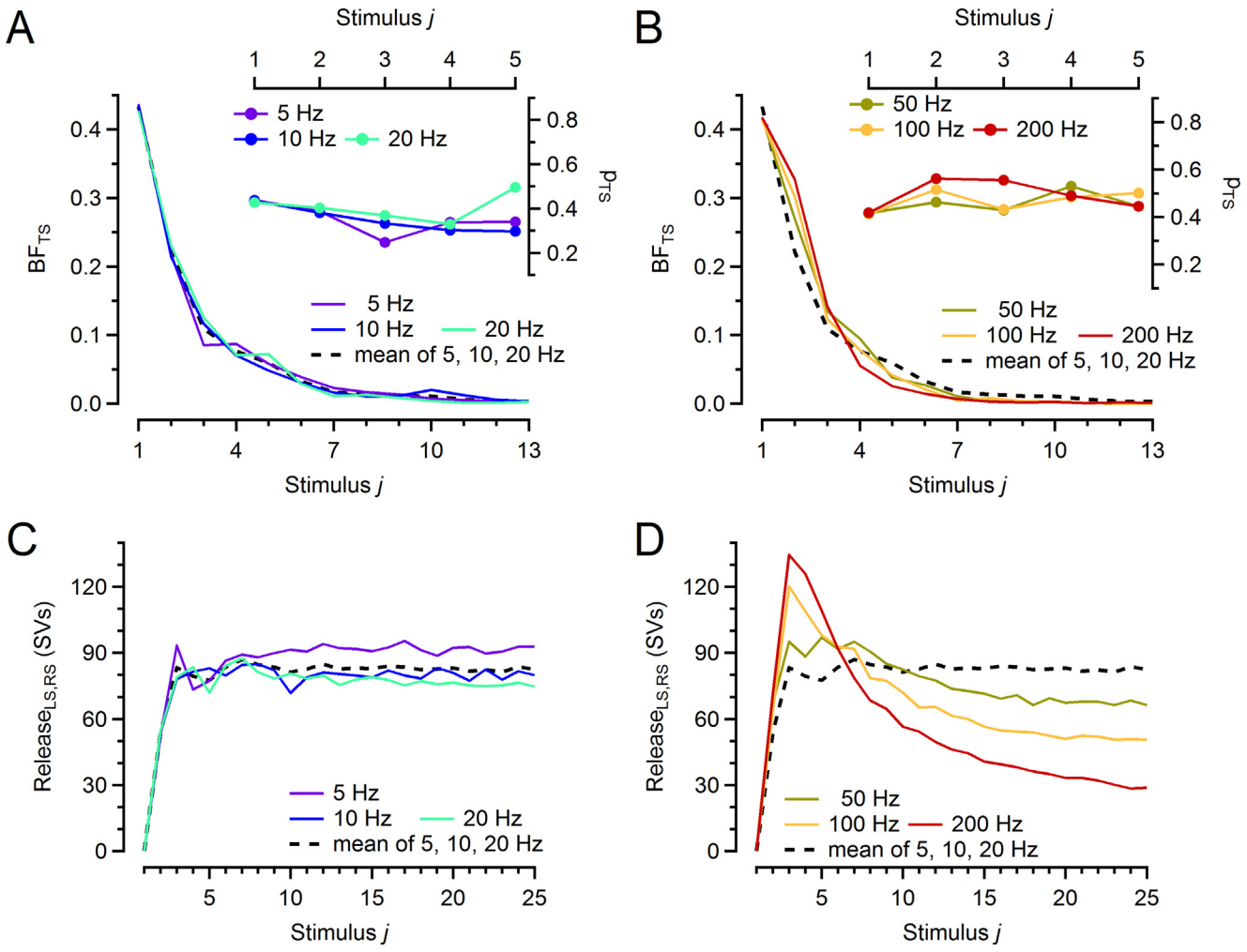


Fig. 4. Basefunctions determined by a two-component NMF decomposition fit. **(A, B)** *BF*s for release of SVs which had been in TS prior to stimulation (BF_{TS}s) are plotted against stimulus number for frequencies of 5–20 Hz **(A)** and 50–200 Hz **(B)**. The average time course of *BF*s for 5, 10, and 20 Hz is included in both panels (black dashed traces) in order to facilitate comparison of time courses at low and high frequencies. *BF*s are normalized to a cumulative sum of 1. $BF_{TS,1}$ is the fraction of pre-existing SV_{TS}s released during the 1st stimulus, which is the release probability p_{TS} at stimulus onset. P_{TS} values during the train are calculated according to Eq. (3) and shown as inserts in **(A, B)** for the first 5 stimuli. **(C, D)** Quantal release contributed by SVs which had been either in LS at stimulus onset or were recruited to release sites and released during the stimulus train (products of $BF_{LS,RS}$ and mean $M_{LS,RS}$ over all synapses for a given frequency) for 5–20 Hz **(C)** and 50–200 Hz **(D)**. The average time course of quantal release for 5, 10, and 20 Hz is included in both panels (black dashed traces). Quantal release rises rapidly to a plateau level, which is very similar for low frequency stimulation **(C)**. For higher frequencies **(D)** quantal release peaks at the 3rd and 4th stimulus and subsequently declines, presumably due to depletion of SV_{LS}s.

which together result in multiplicative updating rules for both *BF*s and *M*s. As long as positive starting values have been chosen, the zero-level cannot be crossed during iterations, preserving non-negativity. Also, parameters initialized to zero or very close to zero do not change noticeably during a finite number of iterations. As a result, the over-all time course of BF_{TS}s, which are initialized as a function decaying to zero during stimulus trains, will be preserved. Likewise, BF_{LS}s, which are initialized to start at zero and eventually decay to near zero will maintain this temporal profile during iterations.

Unfortunately, the constraints listed so far are usually not sufficient to arrive at a unique solution, i.e. different solutions are obtained when initializing *BF*s differently. In particular, the estimated *M*s can vary quite substantially – of course, only within the limitations that both *BF*s and *M*s need to be non-negative.

To further restrict the set of possible NMF fit solutions, additional restrictions can be introduced because data are available at several stimulation frequencies. Separate NMF runs are performed, one for each frequency (5, 10, 20, 50, 100 and 200 Hz), with the following additional three constraints applied to data originating from the same individual synapse:

(1) Values for M_{TS} of a given synapse need to be very similar for all stimulus frequencies. They represent the number of SV_{TS}s available prior to stimulation and are therefore expected to be invariant, regardless of the stimulus protocol a given synapse is sub-

jected to as long as all pre-existing SV_{TS}s are consumed during a given stimulus train.
(2) For the same reason, M_{LS} values should be similar for all stimulus frequencies for a given synapse.
(3) The values for $BF_{TS,1}$ should be very similar for all stimulus frequencies, since they represent the initial p_{TS}, corresponding to the first eEPSC in a train, which is independent of subsequent stimulation.

These constraints are implemented by an iterative cycle for minimizing the difference between measured quantal contents and the respective $m_{i,j}$ obtained by the NMF fit, which consists of the following four steps:

i) perform one iteration for each stimulus frequency,
ii) calculate the means over all stimulus frequencies of M_{TS}, M_{LS} and of $BF_{TS,1}$ (representing the mean p_{TS}),
iii) shift individual values towards their mean by adding a certain fraction of the deviation (for M_{TS}) or of the ratio between individual values and mean (for $BF_{TS,1}$) and
iv) use the updated values as initial parameters for the next NMF iteration.

This cycle is repeated until *BF*s and *M*s, as well as a measure of χ^2 become stationary (see *Experimental procedures* for details).

It is important to reiterate that these constraints are only valid when stimulus trains are long enough (25 APs in the present data set) to fully consume pre-existing SV_{TS}s and SV_{LS}s. Similar constraints for BF_{RS}s cannot easily be formulated, since the number of newly recruited SVs may be quite variable for different stimulation frequencies. In particular, SV recruitment per inter-stimulus interval (ISI) may be large for very low frequencies but decreases at higher frequencies that provide little time for reloading of SV at docking sites between successive APs.

Reliability of the three-component NMF decomposition fit

The constraints listed so far restrict the NMF decomposition of eEPSC train data to solutions that are compatible with the specific kinetic scheme illustrated in Fig. 3. This, however, does not guarantee that no other solutions exist, which satisfy all the constraints listed. Uniqueness of the NMF fit result depends partially on the structure of the data set. The more diverse individual synapses are in terms of their STP and their *M* values, the less degenerated is the solution. Notably, degeneracy is much reduced, if the data set contains synapses, which lack one component all together, since then any new linear combination of *BF*s is likely to require negative *M*s. Our experience with the present data set was that even with all the constraints listed so far, somewhat different solutions were obtained when parameters were initialized differently. In particular, it was found to be very difficult to separate unequivocally the contributions from pre-existing SV_{LS}s and SV_{RS}s.

ENHANCING FIT ROBUSTNESS WITH A PREPARATORY TWO-COMPONENT NMF DECOMPOSITION FIT

To address the ambiguity discussed above, the following two-step approach was implemented: In a first run on a given eEPSC train data set, a two-component NMF decomposition was performed with the initialization of fit parameters adjusted to be compatible with a BF_{TS} and a second *BF* representing the sum of BF_{LS} and BF_{RS}, denoted as $BF_{LS,RS}$. BF_{TS} was initialized to monotonically decay to zero while $BF_{LS,RS}$ was initialized to start at zero and to rise exponentially to a plateau. For each component, *BF*s for all stimulation frequencies were initialized with the same time course (but see an alternative procedure below). Two hundred iterations were used. Surprisingly, a two-component NMF fit could be obtained with a measure of χ^2 not more than 32% higher than that of a three-component NMF fit. This may indicate that M_{LS} and M_{RS} estimates are correlated with each other (which will be confirmed below). The NMF decomposition obtained this way is quite robust. For instance, a 2.5-fold change in the initial guess for the initial value for BF_{TS} changed the final value of that parameter by only 16%. The same relative change was observed in the rise time constant of $BF_{LS,RS}$ when the initial guess for this parameter was decreased 2.5-fold. Time courses for these *BF*s using standard initializations (see *Experimental procedures*) are shown in Fig. 4 (panel A for frequencies 5, 10 and 20 Hz and panel B for 50, 100 and 200 Hz). Strikingly, the BF_{TS}s for 5, 10 and 20 Hz trains are very similar. To facilitate comparison, the mean time course of BF_{TS}s over these three frequencies is included in Fig. 4A, B.

Given the robustness of the two-component NMF fit, results for BF_{TS} provide the first answer for the decomposition problem. The value of $BF_{TS,1} = 0.43 \pm 0.01$ is the release probability of pre-existing SV_{TS}s (p_{TS}) for a the first eEPSC in a train or a single eEPSC. The mean M_{TS} for the present eEPSC train data set is 536 ± 24 pre-existing SV_{TS}s (averaged across cells and stimulus frequencies). For low frequencies (5–20 Hz), BF_{TS}s decay in a nearly geometric fashion, which indicates a relatively constant *p* throughout the stimulus train. However, a closer look reveals more details. Release probability values $p_{TS,j}$ for any stimulus *j* within a train, i.e. the probability of a pre-existing SV_{TS} being released at stimulus *j*, given that it had not been released before, can be calculated as the ratio of a given value of the BF_{TS} at index *j* divided by the fraction of pre-existing SV_{TS}s still available for release at that stimulus:

$$p_{TS,1} = BF_{TS,1} \tag{3}$$

$$p_{TS,j} = BF_{TS,j} \Big/ \left(1 - \sum_{k=1}^{j-1} BF_{TS,k}\right); \quad \text{for } j = 2..J$$

Release probabilities calculated this way are shown for 5–20 Hz trains in the insert of Fig. 4A. Only the first five p_{TS} values are shown. For later stimuli, BF_{TS} values are quite small and $1 - \sum_{k=1}^{j-1} BF_{TS,k}$ approaches 0, such that p_{TS} is a ratio of two small quantities and dominated

by random fluctuations. The plots show that for 5–20 Hz trains, p_{TS} is quite constant or decreases slightly. Time courses for p_{TS} at 50–200 Hz stimulation are shown in the insert of Fig. 4B which illustrates that p_{TS} increases for higher stimulation frequencies. This trend is particularly conspicuous at 200 Hz where the 2nd and 3rd p_{TS} values are 34% larger than the value for the 1st stimulus.

The second component of the two-component NMF fit, $BF_{LS,RS}$, represents the sum of pre-existing SV_{LS}s plus newly recruited and released SV_{RS}s. Unfortunately, these basefunctions are not as readily interpreted in terms of SV subpool size and release probability. NMF returns $BF_{LS,RS}$ normalized to the cumulative sum of its 25 entries. Therefore, we plot in Fig. 4C, D for each stimulation frequency the product of $BF_{LS,RS}$ and $\overline{M}_{LS.RS}$. $\overline{M}_{LS.RS}$ denotes the mean quantal content over all synapses at that frequency of the combined release contributed by LS and RS SVs. For stimulus frequencies of 5, 10 and 20 Hz, these traces are very similar to each other, rising rapidly to a relatively constant plateau of about 80 SVs per stimulus, irrespective of stimulation frequency. Since these values are plotted here against stimulus number, similar values actually indicate increasing rates of release at increasing stimulus frequencies and, therefore, linearly rising rates of SV_{TS} resupply. Such a result is expected, if the LS → TS transition rate (k_2; Fig. 3) increases linearly with stimulus frequency while p and therefore SV consumption from TS remains relatively constant. Then, the number of newly available SV_{TS}s can keep up with their release, independent of stimulus frequency. A likely mechanism for such linear increase in k_2 is a linear increase in global $[Ca^{2+}]_i$. However, any other process by which a presynaptic AP triggers the conversion of a certain fraction of SVs from LS into TS would also lead to a frequency-independent plateau of release provided that the LS subpool (SP_{LS}) itself is not or only little depleted.

For stimulus frequencies $\geq$50 Hz, release initially increases well above 80–90 SVs per stimulus and later declines below this level. The decline most likely represents depletion of SP_{LS}. This feature will be discussed in the context of results from a three-component NMF fit described below.

The shape of the *BF*s explains the robustness of the two-component NMF fit. Any linear combination of BF_{TS} and $BF_{LS,RS}$ would in principle be eligible as a new set of *BF*s. However, it would have non-zero BF_{TS} values for late stimuli during the trains and, therefore, would disqualify as a candidate for a BF_{TS}. Likewise, such a combination would have a non-zero $BF_{LS,RS,1}$ value and therefore would violate the constraints for $BF_{LS,RS}$.

Separating three release components by NMF decomposition

The remaining problem of separating release carried by pre-existing SV_{LS}s from that contributed by SVs newly recruited during ongoing stimulation is very similar to the conventional problem of determining a 'readily releasable pool' of SVs on the background of ongoing pool refilling. Methods used to this end include cumulative plots of eEPSC amplitudes and curve fitting, based on pool models (see Neher, 2015 for discussion). The benefit of the NMF decomposition in the context of such methods is that after separating the TS-component, the remaining release is that of a single homogeneous SV pool (within the framework of NMF assumptions), thus satisfying better the assumptions of most of the traditional methods of pool size estimation utilizing eEPSC trains. An Elmqvist-Quastel type analysis (Elmqvist and Quastel, 1965) of the isolated mean release time course at 200 Hz (see *Experimental procedures* for details) yields a pool size estimate of 1071 pre-existing SV_{LS}s (Fig. 5A). This value may be considered as an upper bound to M_{LS} (Thanawala and Regehr, 2016). A similar estimate, based on a plot of cumulative release resulted in 732 pre-existing SV_{LS}s. It may be considered as a lower bound to M_{LS}. How does this compare to the result of a three-component NMF fit, or else – given the dependence of NMF fit results for M_{LS} and M_{RS} on their initial guess values – which choice of initial guess values would yield M_{LS} estimates within such upper and lower bounds? To answer this question, we performed three-component NMF fits, 100 iterations each, with BF_{TS} and M_{TS} constrained to the result of the two-component NMF fit described above. When varying the initial guess for the decay time constant of the BF_{LS} between 1 and 8 ISIs, estimates for M_{LS} between 300 and 1050 pre-existing SV_{LS}s were obtained. The reason for this wide variation becomes obvious, when examining the fit results more closely: Plotting the values of M_{LS} against M_{RS} for all synapses revealed a strong correlation between these two quantities (Fig. 5B). This renders separation of the two components by NMF quite ill-defined, because two components cannot be separated by NMF, if their magnitudes are proportional to each other due to the linear nature of the decomposition. Nevertheless, an NMF decomposition of the eEPSC train data set into three release time course components can provide interesting options for its interpretation.

When studying the dependence of the subdivision between LS- and RS-components on the choice of initial guess values, it was found that M_{LS} increased when increasing the initial guess for the decay time constant of BF_{LS}. However, time constant values $\geq$7 ISIs led to BF_{LS}s decaying towards a non-zero plateau, which is in conflict with the constraint that pre-existing SV_{LS}s should deplete completely during trains. When choosing 7 ISIs as the initial guess for the decay time constant, BF_{LS} decayed to near zero and a M_{LS} of 1028 pre-existing SV_{LS}s was obtained. This is well within the upper and lower bounds of 1071 and 732 pre-existing SV_{LS}s obtained above. As a compromise, we used 7 ISIs and trimmed M_{LS} somewhat towards the mean of lower and upper bounds by introducing a slight bias in the fitting routine (see *Experimental procedures*). The fit, which we accepted for further analysis resulted in a M_{LS} of 901 pre-existing SV_{LS}s. The features of this decomposition – many of which qualitatively agree with those using other initial values for the fit parameters – will be described in the following.

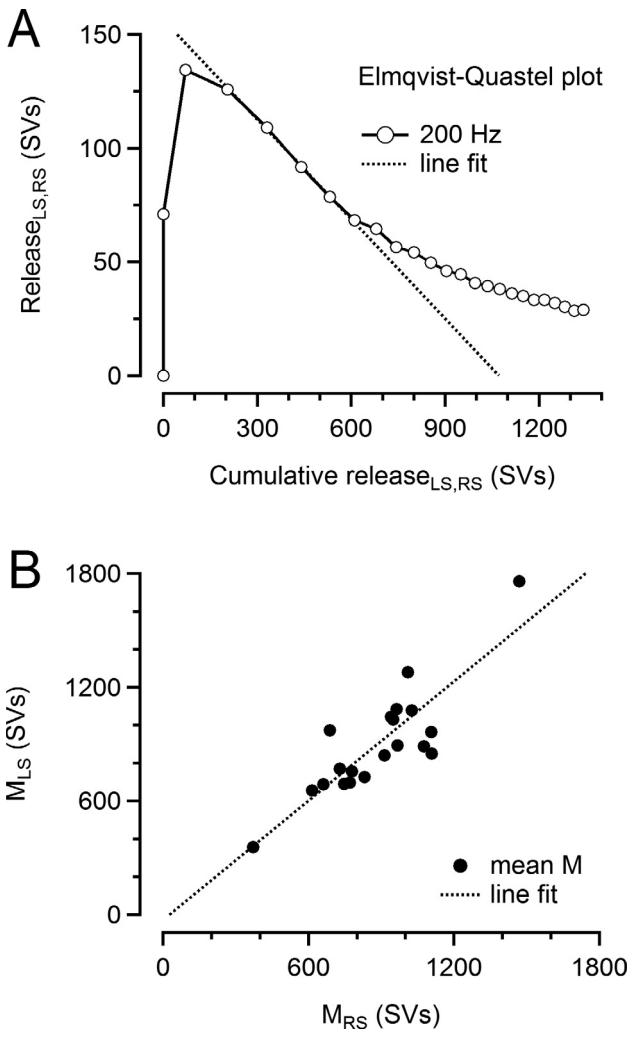


Fig. 5. Separating LS and RS components following a two-component NMF decomposition fit. **(A)** Total quantal release (averaged over all synapses) contributed by the combined LS,RS components during 200 Hz stimulation is plotted versus cumulative release of preceding stimuli. A line is fitted between values corresponding to the 4th and 7th stimulus. The line fit intersects the x-axis at a value of 1071 SVs, which, according to Elmqvist and Quastel (1965), represents an estimate for the size of an SV pool that depletes during high-frequency stimulation. **(B)** Correlation between M_{LS} and M_{RS} abundances (mean values over all stimulus frequencies) derived from a three-component NMF decomposition fit with the BF_{TS} fixed to the one obtained in a two-component NMF fit. Each data point represents one synapse. The correlation implies that such a three-component NMF fit is not well defined.

Fig. 6A, B show BF_{LS}s for 5, 10 and 20 Hz and for 50, 100 and 200 Hz, respectively. For low frequencies, BF_{LS}s are almost identical and, as in the case of BF_{TS}s, decay nearly exponentially. This is expected, if each AP triggers the conversion of a constant fraction of remaining SV_{LS}s into SV_{TS}s and the latter are released subsequently with constant p (Note that BF_{LS} represents only those SVs, which had been in LS prior to stimulation). The fraction of SVs lost per stimulus was determined as the reciprocal time constant of an exponential fit to the decaying section of the mean BF_{LS} over stimulation frequencies 5, 10 and 20 Hz (Fig. 6A, B, black dashed line). It was found to be 0.064 pools/ISI. This is the rate constant k_2 (in units of pools/ISI) for the LS → TS transition, when interpreted in terms of the sequential model of Fig. 3. While time courses of BF_{LS}s $\leq$ 20 Hz are very similar to each other, they change shape for stimulus frequencies $\geq$ 50 Hz (Fig. 6B). BF_{LS}s with initial overshoots are observed at such high stimulus frequencies. Overshoots reflect higher p values at these stimulation frequencies (see above), but may also result – in part – from enhanced recruitment. They are accompanied by faster decays of BF_{LS}s later during trains.

Fig. 6C, D show for each stimulation frequency release contributed by SVs that are newly recruited during trains, i.e. products of BF_{RS}s and $\overline{M}_{RS}$s (mean M_{RS}s over all synapses). It should be noted that BF_{RS}, in contrast to other BFs, does not represent release of SVs that had been in a specific state at stimulus onset. Rather it comprises release of newly recruited SVs from docking sites which either had been empty at stimulus onset or else became vacant during stimulation. Therefore, individual synapses may vary somewhat not only in their M_{RS} values, but also with respect to the time course of their BF_{RS}s. However, the data in Fig. 6D represent averages over synapses for which such differences are expected to be evened-out. For low stimulus frequencies (5–20 Hz), again very similar time courses are observed. Their means are plotted for comparison in both panels C, D of Fig. 6. The finding that the contribution of newly recruited SVs to release depends very little on stimulation frequency in this range (5–20 Hz) may indicate that each AP shifts a certain fraction of SVs from LS to TS and also loads SVs onto a certain fraction of empty docking sites. However, marked deviations from the low-frequency mean time course are observed for later responses at stimulus frequencies $\geq$50 Hz. In particular at 100 and 200 Hz (Fig. 6B), the amount of release plateaus after about 10 stimuli. Maximum rates are between 55 and 18 SV_{RS}s/ISI for 50 and 200 Hz respectively. When expressed as rates of recruitment per second the corresponding values are 2750–3600 SV_{RS}s/s. Given that the sum of M_{TS} plus M_{LS}, representing all SV_{TS}s and SV_{LS}s prior to stimulation is 1787 SVs, this corresponds to a rate constant between 1.5 and 2.0 release events per s and per site, assuming that for these stimulation frequencies the average number of empty sites at steady state is approximately equal to that of occupied sites at rest. Furthermore, at 200 Hz a slight rundown of recruitment is observed later in trains. This amounts to a decline by about 30% with respect to the early peak and possibly represents either depletion of a vesicle pool upstream of SP_{LS}, refractoriness of release sites (Hosoi et al., 2009; Hua et al., 2013) or is spuriously generated by residual AMPAR desensitization.

Fig. 7 compares the predictions of the three-component NMF decomposition yielding fit results as detailed above and experimental data for the time course of quantal release during stimulus trains of 5–20 Hz (Fig. 7A) and 50–200 Hz (Fig. 7B) for all 20 synapses in the data set. For each of the stimulus

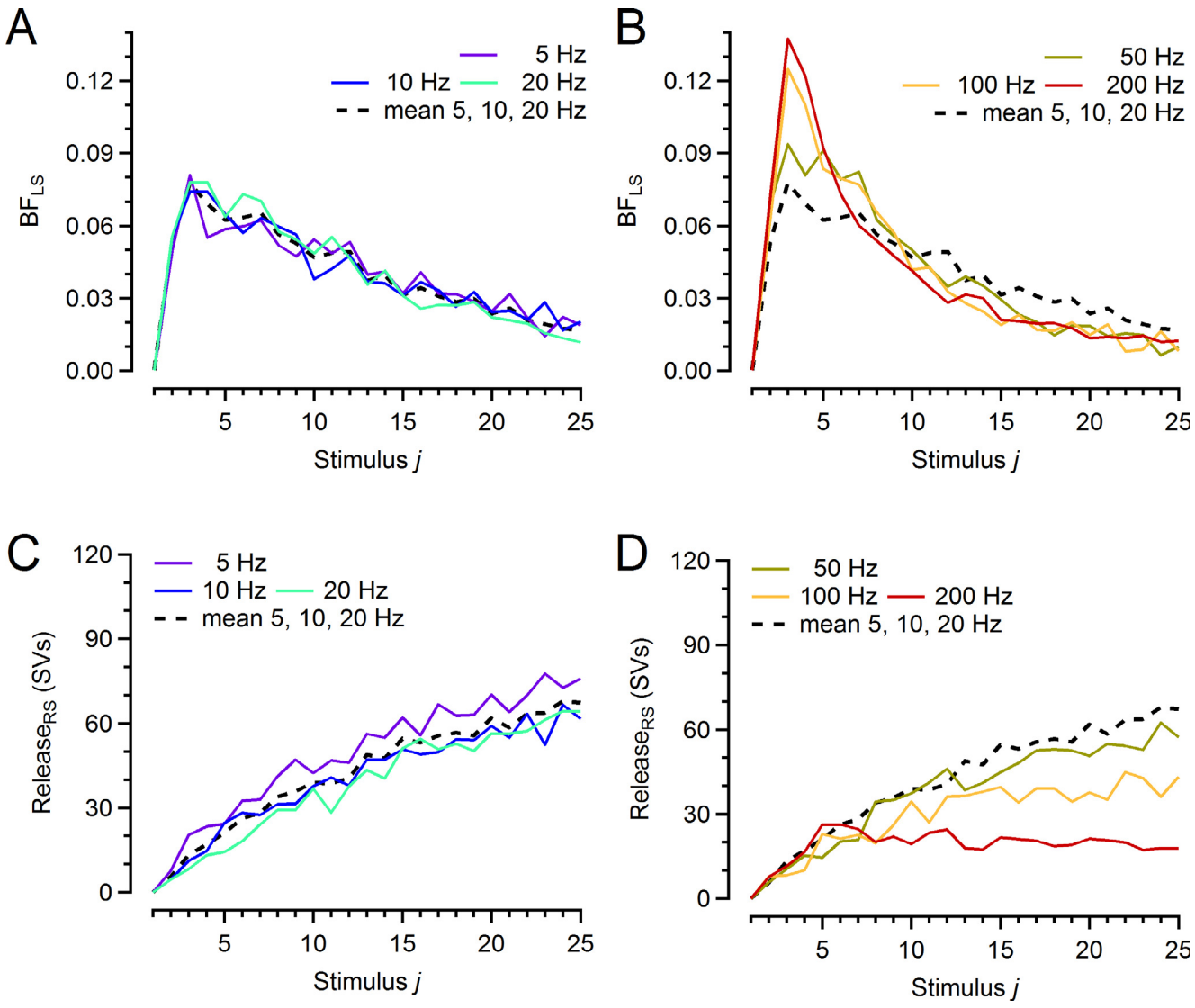
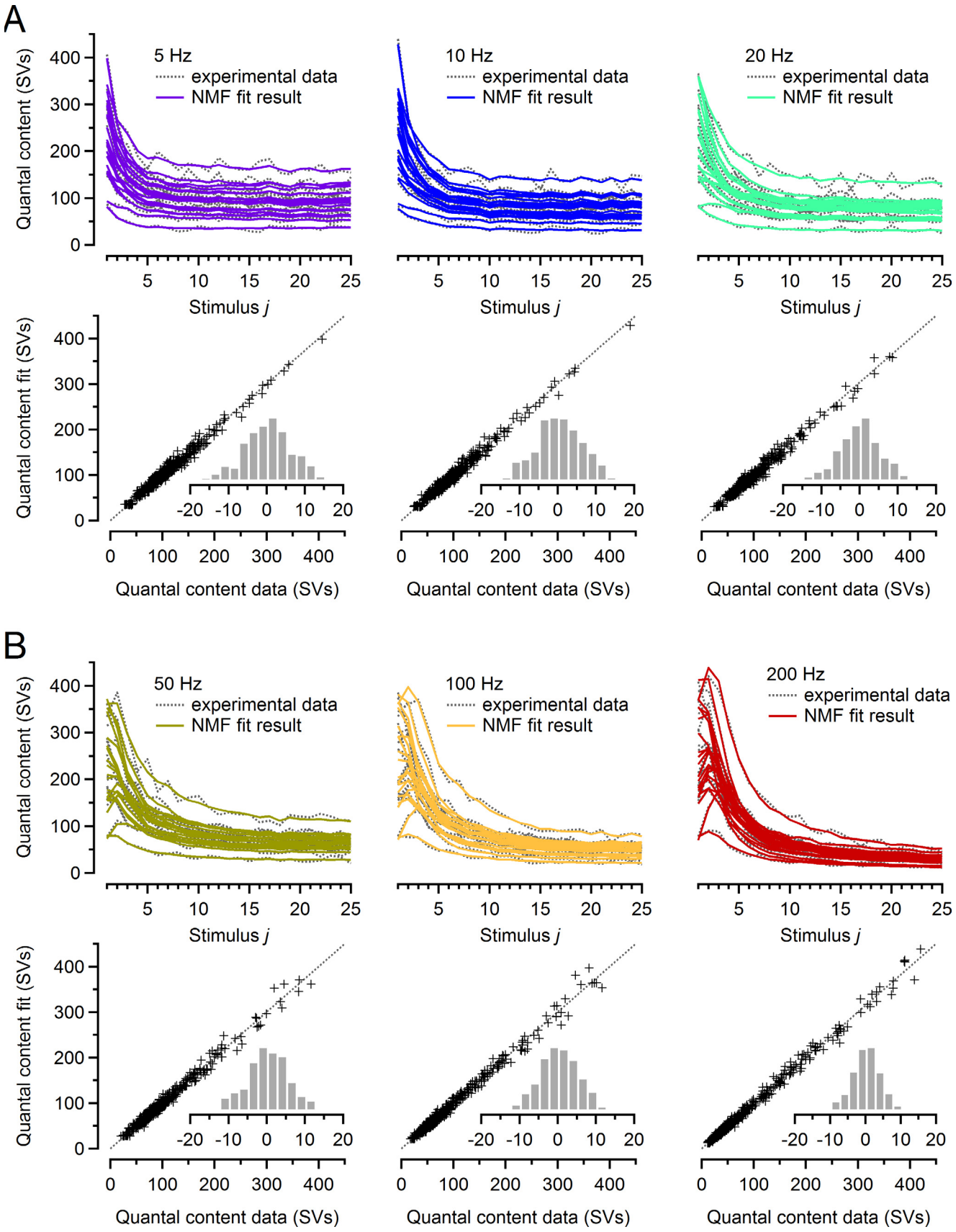


Fig. 6. Basefunctions determined by a three-component NMF fit with BF_{TS}s constrained to those obtained by a previous two-component NMF fit. BF_{TS}s are identical to those shown in Fig. 4A, B and therefore not presented here. **(A, B)** *BF*s for release of SVs which had been in LS prior to stimulation (BF_{LS}s) are plotted against stimulus number for 5–20 Hz **(A)** and 50–200 Hz **(B)**. BF_{LS}s for 5, 10, and 20 Hz are very similar. The black dashed traces in **(A, B)** represent the average time courses of BF_{LS}s for 5, 10 and 20 Hz. BF_{LS}s for 50, 100 and 200 Hz develop an early peak. **(C, D)** Contributions of newly recruited SVs (SV_{RS}s) to quantal release for stimulation frequencies of 5–20 Hz **(C)** and 50–200 Hz **(D)** are plotted against stimulus number. Again, traces for 5, 10, and 20 Hz are very similar. The black dashed traces in **(C, D)** represent the average release time courses of SV_{RS}s for 5, 10 and 20 Hz.

frequencies, the fit residuals are narrowly distributed around zero indicating that the NMF decomposition fit accounts well for the heterogeneity among the 20 synapses.

Some quantitative aspects of the findings reported here depend on the particular choice of initial guess values for the fit parameters (see above). However similar trends were observed over a wide range of parameter initializations. In particular, the similarity of BF_{LS}s in the range of 5–50 Hz, when plotted against stimulus number, was a robust finding, which is quite remarkable. It means that rates of release of pre-existing SV_{LS} and correspondingly the rates of LS → TS transition are actually 10 times higher at 50 Hz as compared to 5 Hz stimulation, when expressed per time unit. In order to assure that this property is not just a consequence of our choice of fit parameter initialization, we performed a second series of calculations: Instead of using the same initialization for *BF*s across all frequencies, we systematically varied rise times and decay time constants during the initialization of BF_{LS}s. Using eightfold faster values (measured in units of ISIs) at 5 Hz as compared to 50 Hz stimulation resulted in BF_{LS}s which had on average about 30% faster half-decay times than those obtained by an NMF fit using standard initialization. However, this speed-up was similar for all low frequencies, such that BF_{LS}s for different stimulus frequencies maintained a similar time course when plotted against stimulus number, in spite of having been initialized differently.


A
5 Hz
experimental data
NMF fit result
10 Hz
experimental data
NMF fit result
20 Hz
experimental data
NMF fit result
Quantal content (SVs)
Stimulus j
Quantal content fit (SVs)
Quantal content data (SVs)
B
50 Hz
experimental data
NMF fit result
100 Hz
experimental data
NMF fit result
200 Hz
experimental data
NMF fit result
Quantal content (SVs)
Stimulus j
Quantal content fit (SVs)
Quantal content data (SVs)

Apparent release probability of SVs in the loosely coupled state

Given a BF_{LS}, which represents the normalized release time course of those SVs that had been in LS prior to stimulation, one can ask the question, what is the probability of such an SV to be released at stimulus j, given it had not been released earlier? We call this the apparent release probability p'_{LS}. It can be calculated for all stimuli j analogous to p_{TS} (Eq. (3)):

$$p'_{LS,1} = 0 \tag{4}$$

$$p'_{LS,j} = BF_{LS,j} / \left(1 - \sum_{k=1}^{j-1} BF_{LS,k}\right); \text{ for } j = 2..J,$$

Fig. 8A shows the corresponding time courses of p'_{LS} for all frequencies (5–200 Hz) up to the 15th stimulus. Note that under our recording conditions, BF_{LS}s did not completely decay to zero during the trains and had, therefore, to be renormalized to allow for the missing tail beyond the 25th stimulus (see *Experimental procedures*). We found that p'_{LS} time courses for 5–20 Hz are very similar, reaching a value of 0.043 for the 2nd stimulus and increasing towards a plateau near 0.07. The value $p'_{LS,1} = 0$ reflects the assumption that SV_{LS}s have to undergo the LS → TS transition before they can be released. The plateau is reached when a constant ratio between TS and LS subpools is established while both pools decay. For 100 Hz and 200 Hz trains, values for p'_{LS} display an initial peak which mirrors the peaks of the BF_{LS}. The increase in release probability from TS may underlie this feature. For 200 Hz trains, p'_{LS} decreases later during stimulation. This may reflect an actual decrease in p or else a decrease in the rate of LS → TS transition. However, values of BFs in that regime are small, such that a small misassignment in the separation of LS- and RS-components might erroneously lead to this trend. Nevertheless, we can conclude that release contributed by SV_{LS}s displays strong facilitation during the initial phase of stimulation, which contributes importantly to synaptic facilitation – in particular at high frequencies and at those synapses that have abundant SV_{LS}s at rest.

A simple model of SV state transitions reproducing basefunctions for stimulation frequencies from 5–20 Hz

The striking similarity among BF_{LS}s in the range of 5–20 Hz suggests a very simple model for changes in subpools SP_{TS} and SP_{LS}. The model iteratively calculates subpool occupancies for each stimulus j by assuming that immediately following each AP a fraction p is lost from SP_{TS} due to SV fusion while a constant fraction α of SP_{LS} is converted to SP_{TS}. If $SP_{TS,1}$ and $SP_{LS,1}$ are initialized to 0 and 1, respectively, and pool occupancies are updated iteratively according to

$$SP_{TS,j+1} = SP_{TS,j} \times (1 - p_{TS,j}) + SP_{LS,j} \times \alpha, \tag{5}$$

and

$$SP_{LS,j+1} = SP_{LS,j} \times (1 - \alpha), \tag{6}$$

these time courses describe the state of a single SV that had been in LS prior to stimulation. The BF_{LS} value at a given stimulus index j, which by definition is the normalized release at j of such an SV, can then be calculated according to

$$BF_{LS,j} = SP_{TS,j} \times p_{TS,j} \tag{7}$$

where $p_{TS,j}$ is the release probability as obtained from an NMF fit according to Eq. (3).

Likewise, BF_{TS} can be calculated as

$$BF_{TS,j} = SP_{TS,j} \times p_{TS,j} \tag{8}$$

after reversely initializing $SP_{TS,1}$ and $SP_{LS,1}$ to 1 and 0, respectively. The value for α can be determined from an NMF fit by calculating the mean BF_{LS} for frequencies 5–20 Hz and fitting a single exponential to the decaying part of it. The resulting time constant τ yields $\alpha = 1 - e^{-1/\tau}$. Given such values for α and $p_{TS,j}$, one can simulate time courses for subpools, BFs as well as for p'_{LS} which can be derived from BF_{LS} according to Eq. (4). Fig. 8B compares such model predictions for SP_{TS} and p'_{LS} (solid traces) with the same quantities obtained from NMF fit results (broken traces). For the latter we calculated p'_{LS} from the average BF_{LS} for 5–20 Hz using Eq. (4). We obtained $SP_{TS,j}$ from $BF_{LS,j}/p_{TS,j}$ (see Eq. (8)).

This simple model, which has no explicit frequency dependence, describes the experimental data very well in the stimulus frequency range from 5 to 20 Hz. Its results deviate from experimental data for higher frequencies due to a frequency-dependent increase in p and a stronger than linear acceleration of the LS → TS transition. Also, the model does not reproduce pool sizes correctly at frequencies lower than 5 Hz since it does not consider basal priming and unpriming rates.

Decomposition by non-negative tensor factorization (NTF)

In general, non-negative factorization techniques are more likely to provide unique, non-degenerate solutions

Fig. 7. The three-component NMF decomposition fit reproduces variability among individual synapse with high accuracy. **(A)** Time courses of quantal release m_j derived from experimental data (solid colored traces) are compared to the predictions of the three-component NMF decomposition fit (dotted gray traces) for 20 individual calyx of Held synapses stimulated at 5 Hz (*left*), 10 Hz (*middle*) and 20 Hz (*right*) in the top row. The corresponding scatter graphs of fit results versus experimental data are illustrated in the bottom row. Each symbol represents one of the 25 m_j values of the trains obtained for one of the 20 synapses. Note that data points cluster tightly around the identity lines (dotted traces) indicating close correspondence between experimental data and NMF fit results for m_j. Insets show histograms of the fit residuals (bin width = 2.5 SVs). **(B)** Similar plots as shown in (**A**) but for stimulation frequencies 50 Hz (*left*), 100 Hz (*middle*) and 200 Hz (*right*). For each stimulation frequency, the basefunctions BF_{TS}s and BF_{LS}s underlying the fit illustrated here are shown in Figs. 4A, B and 6A, B, respectively. Release contributed by newly recruited SVs (SV_{RS}s) is shown in Fig. 6C, D.

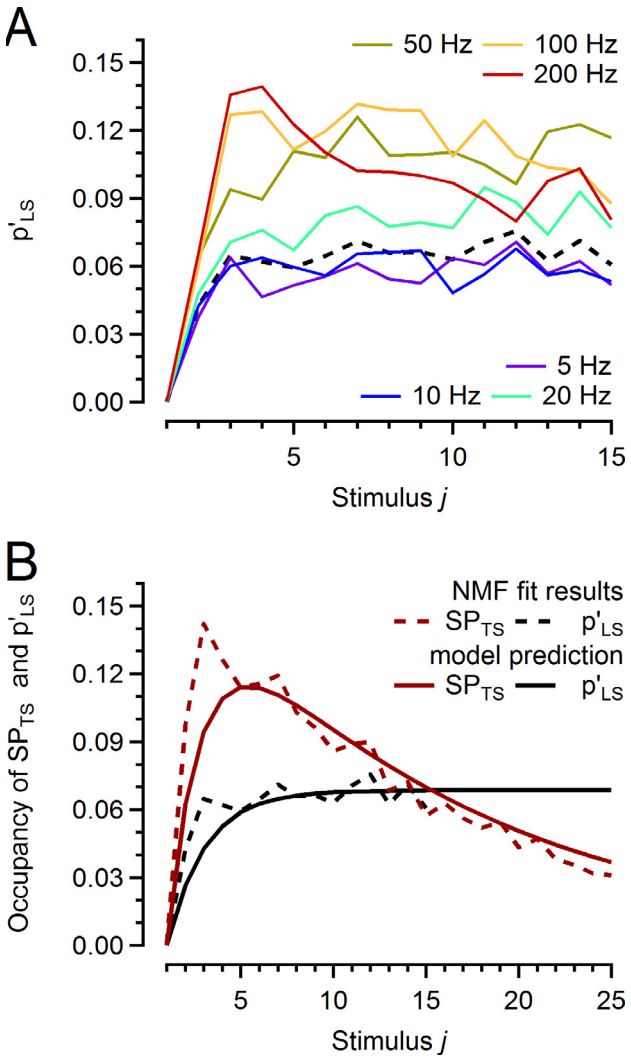


Fig. 8. Apparent release probabilities of the LS-component and a simple model for release at stimulus frequencies between 5 and 20 Hz. **(A)** Apparent release probabilities p'_{LS} of pre-existing SV_{LS}s plotted against stimulus number for several stimulation frequencies (see Eq. (4)). Very similar time courses are obtained for frequencies of 5–20 Hz with p'_{LS} reaching a plateau near 0.07. The average time course of p'_{LS} for 5, 10 and 20 Hz stimulation is shown as dashed black trace. For higher stimulation frequencies (50, 100 and 200 Hz), p'_{LS} values rise to much higher levels. **(B)** Time courses of p'_{LS} (black) and of the occupancy of the subpool of SV_{TS}s (SP_{TS}, dark red). For clarity, only SV_{TS}s generated by an LS → TS transition during the stimulus train are considered, i.e. the initial occupancy of SP_{TS} at the onset of stimulation is assumed to be zero. Comparison of NMF fit results (broken traces) with predictions of a simple model (solid traces). Model *BF*s and model time courses were calculated according to Eqs. (5)–(7), as explained in the text. The model time course for p'_{LS} (black) was calculated from the model BF_{LS} according to Eq. (4). Values for p'_{LS} and SP_{TS} derived from NMF fit result represent means obtained from the average BF_{LS} (5–20 Hz) using Eq. (4) and an equation analogous to Eq. (6). For simplicity, constant values of $p_{TS} = 0.40$ and $\alpha = 0.0685$ were used for model calculations.

if variations in more dimensions are considered. In fact, for a three-dimensional data set it can be shown that degeneracy disappears under a wide range of conditions (Kruskal, 1977). Our variant of NMF, described so far, considers two dimensions, one along the time axis, represented by the stimulus number *j*, the other one represented by the individual synapses *i* in the eEPSC train data set. The data set subjected to NMF analysis, therefore, is a two-dimensional matrix with rows representing synapses and as many columns, as there are stimuli. The properties of *BF*s suggest an option for introducing a third dimension for high-frequency trains by building tensors for 100 and 200 Hz eEPSC trains, the first layers of which are matrices as used so far. Additional layers represent eEPSC trains obtained at the same stimulus frequencies, which however are preceded by a few conditioning eEPSCs elicited at a low frequency. It has been shown that few stimuli at low frequency (e.g. 10 Hz) deplete the so-called 'superprimed' SV pool (Taschenberger et al., 2016) which in the context of the present analysis corresponds to the pre-existing SV_{TS}s. Conditioning 10 Hz stimulation, therefore, reduces the number of SV_{TS}s that remain available for release during subsequent high-frequency stimulation. Other properties, however, in particular *BF*s are not expected to be influenced in a major way. Thus, one can include in the analysis the high-frequency sections of such trains. Assuming that preceding low-frequency stimulation does not change the *BF*s for high-frequency eEPSC trains and that the relative depletion of the SP_{TS} by conditioning stimulation is similar for different synapses, one can then calculate the response $m_{i,j,l}$ of the synapse *i* to stimulus *j*, after *l* conditioning eEPSCs as a sum (over components *k*) of the product of three quantities: the *M*s for the three components of synapse *i* before conditioning, the value of the *BF* of a given component at stimulus *j*, and the scaling factor *SF* of a given component, which is the fraction of SVs of a given type remaining after conditioning pulses:

$$m_{i,j,l} = M_{TS,i} \times BF_{TS,j} \times SF_{TS,l} + M_{LS,i} \times BF_{LS,j} \times SF_{LS,l} + M_{RS,i} \times BF_{RS,j} \times SF_{RS,l} \quad (9)$$

Trains at 100 Hz and 200 Hz after 2 and 5 conditioning stimuli were found to be suitable for this kind of analysis. In the example discussed here, the tensor for 100 Hz consists of three layers, the first one being a matrix just like the one used for NMF, corresponding to non-conditioned trains, plus two additional layers, similar to the first, but containing *m*-estimates for trains following two or five conditioning eEPSCs. The NTF algorithm iteratively calculates values of $m_{i,j,l}$ according to the above equation and minimizes the fitting error by updating all terms in Eq. (9) in a way, which preserves non-negativity (Lee and Seung, 2001). It uses the same set of *BF*s for all cells and layers, the same *M*s for a given cell in all layers, and the same *SF*s in a given layer. In the example described below, *SF*s for non-conditioned trains are fixed to one. For conditioned trains, scaling factors for the TS-component are expected to be the smaller, the more conditioning eEPSCs were evoked. Those for the LS- and RS-components are expected to remain near 1, since these components are affected very little by conditioning. These expectations are confirmed experimentally (see below). A tensor for a given frequency is readily combined with matrices and tensors at other frequencies, provided that the high-frequency sections of trains have equal numbers of stimuli.

When subjecting eEPSC data to NTF, it was found that special care has to be taken to avoid influences of amplitude run-down or run-up. The experimental protocol for eEPSC recordings on a given synapse involves numerous stimulation episodes over an extended period of time. This may cause eEPSCs to run-down. One may also observe the opposite – run-up –, since the total number of stimuli is close to that of protocols used for the induction of post-tetanic potentiation at the calyx of Held (Habets and Borst, 2005; Korogod et al., 2005; Lee et al., 2008). Such eEPSC amplitude trends need to be carefully monitored and data should be discarded, if changes exceed 10%. Non-conditioned and conditioned eEPSC trains should therefore be acquired relatively close to each other during the recording session in order to prevent run-down or run-up of synaptic responses which would otherwise strongly influence scaling factors.

An NTF analysis using non-conditioned eEPSC trains obtained in response to 5, 10, 20, 50, 100 and 200 Hz stimulation and conditioned 100 Hz eEPSC trains (2 and 5 conditioning eEPSCs at 10 Hz) was performed on data from 5 calyx synapses. The same sequence of analysis as used for NMF (two-component analysis, followed by three-component analysis with BF_{TS}s constrained to those of the preceding two-component NMF fit), the same amplitude constraints and the same parameters for initialization were employed. This resulted in *BF*s very similar to those of the NMF analysis described above. The mean initial p of SV_{TS}s, however, was somewhat lower (0.351 ± 0.012). During 200 Hz stimulus trains p_{TS} increased substantially, as observed in NMF. Both BF_{TS}s, as well as BF_{LS}s were very similar to each other for low stimulus frequencies (5–20 Hz). As in the case of the NMF analysis, BF_{LS}s developed an early peak for 100 and 200 Hz. M_{TS} was found to be 940 pre-existing SV_{TS}s.

The quality of the NTF fit was comparable to that achieved by NMF decomposition. The measure of goodness of fit χ^{2^*} (see *Experimental procedures*) increased by only 2%. Thus, it is reassuring that experimental data obtained from only five calyx synapses can reveal the most important features of STP during stimulus trains.

Alternative kinetic schemes of SV priming and fusion

Accumulating evidence since the beginning of synaptic research (Katz, 1969) shows that presynaptic terminals are endowed with discrete sites to which SVs need to bind ('dock') before fusion can be triggered and neurotransmitter is released. As detailed in the Introduction, there is good reason to postulate at least two distinct priming states of docked SVs. So far, we assumed that there is a single kind of release site, which can however exist in the three states: (*i*) empty, (*ii*) occupied by an SV_{LS}, or (*iii*) occupied by an SV_{TS}. Furthermore, we postulated that only SV_{TS}s are fusion competent and that during stimulation $\geq$5 Hz, SVs progress unidirectionally along a sequence of docking in LS, followed by the LS → TS transition and fusion. Alternative models assume different kinds of sites, between which SVs can migrate and from which they are released with site-specific p values. Such reaction paths can be in parallel (Trommershäuser et al., 2003; Schlüter et al., 2006; Taschenberger et al., 2016) or sequential (Pan and Zucker, 2009; Hallermann et al., 2010a, 2010b; Doussau et al., 2017; Miki et al., 2018).

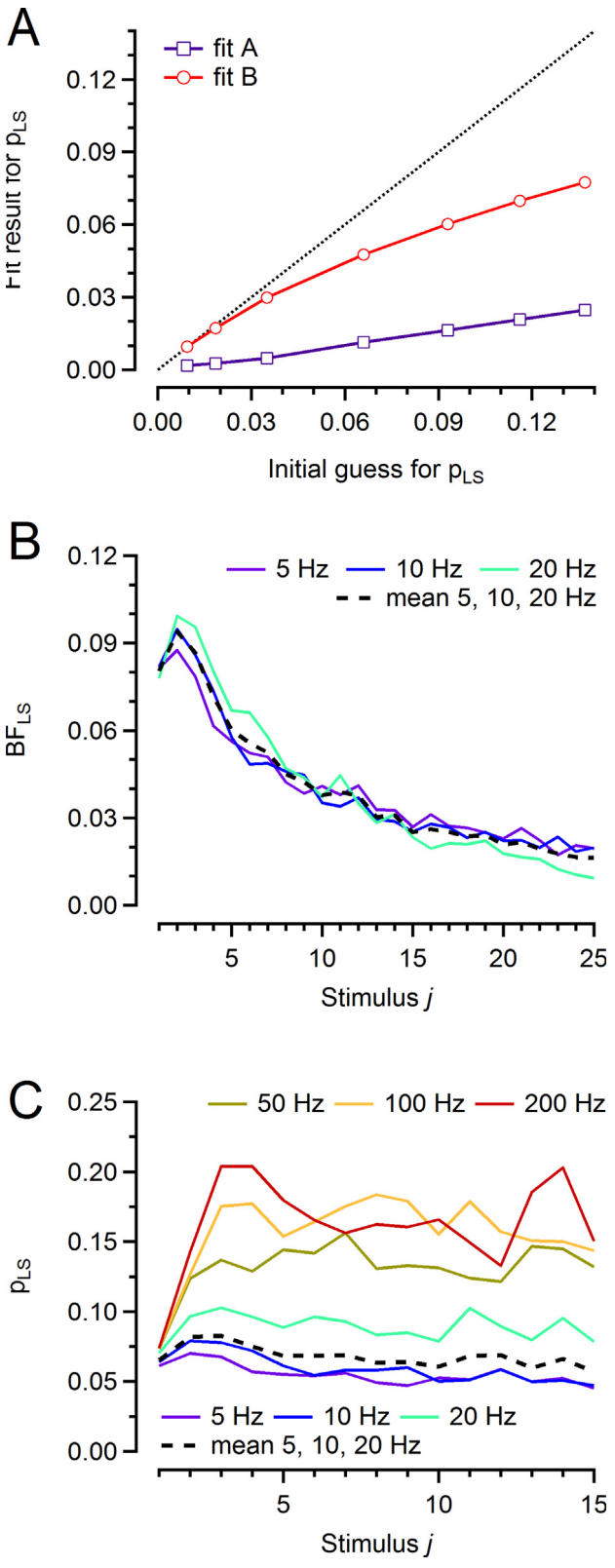
NMF analysis can, of course, be applied to a given data set irrespective of the underlying model. However, constraints and initial guesses for *BF*s need to be suitably adjusted. For instance, for a kinetic scheme with two kinds of sites at which release can occur in parallel, two kinds of *BF*s would have to be assumed, both of which either monotonically decay to zero or else undergo facilitation, followed by depression. However, allowing both SV pools to undergo fusion in parallel will make the separation of the two components more difficult, since one constraint used so far ($BF_{LS,1} = 0$) cannot be applied. On the other hand, an additional constraint can be introduced for the 2nd SV pool imposing frequency independence of the 1st value of its *BF* in analogy to the constraint imposed on $BF_{TS,1}$. However, we found, in agreement with the general rule that sparsity of the *BF*s is most effective in reducing degeneracy of solutions, that these new constraints are less stringent than our standard ones.

In order to explore alternative kinetic schemes, we performed a number of NMF analysis runs assuming a kinetic scheme in which SV_{LS}s were allowed to directly undergo fusion. Following the analysis paradigm described earlier, each run consisted of an initial two-component NMF fit followed by a three-component NMF fit with BF_{TS} held fixed. A constraint was added, which forced $BF_{LS,1}$ (initial p_{LS}) to be similar for all frequencies. The impact of this constraint on the final NMF fit result depended critically on the handling of M_{TS}. If it was constrained to that of the two-component NMF fit, results for p_{LS} were quite robust with respect to changes in initial guess values for this parameter. Raising the initial guess up to 0.14 increased its final value to only 0.025 (Fig. 9A, fit A). Thus, the NMF algorithm tended to favor low-p_{LS} solutions. If, however, M_{TS} was allowed to vary (while still fixing BF_{TS} to the result obtained by the preceding two-component NMF fit), the final p_{LS} followed its initial guess almost linearly up to about 0.035 (Fig. 9A, fit B) with little change in the quality of the fit. Thus, relaxing the BF_{TS} constraint increases dramatically the degeneracy of possible solutions. On the other hand, it allows one to study the properties of solutions with substantial direct release from the SP_{LS}.

Calculating p_{LS} values by applying Eq. (3) to the renormalized BF_{LS} (see *Experimental procedures*), we found that they remained quite constant throughout trains for stimulus frequencies $\leq$20 Hz. They increased during trains at higher frequencies very much like p'_{LS} (Fig. 8A). We, therefore, asked the question which choice of $BF_{LS,1}$ initialization would yield an NMF fit result with constant p_{LS} during low-frequency trains. When allowing M_{TS} to vary, we found that an NMF fit obtained after initializing $BF_{LS,1}$ with a value of about 0.1 resulted in a final value of about 0.06 (close to the

inverse of the decay time constant of the BF_{LS} of Fig. 6A). Following a small initial increase, BF_{LS}s decayed almost exponentially, similar to those of Fig. 6A, but now starting at the elevated level (see Fig. 9B). Low-frequency p_{LS} values, calculated on the basis of these BF_{LS}s, were almost constant during trains, and the mean $p_{LS,1}$ calculated over all stimulus frequencies (5–200 Hz) was 0.065 (Fig. 9C). At higher frequencies, p_{LS} increased strongly during the first few stimuli, similar to the results obtained for the strictly sequential model (Fig. 8A). In fact, release probability time courses in Fig. 8A are very similar to those of Fig. 9C, except for the 1st point, which is fixed to near zero in the former case. Release probability p_{TS} was slightly higher (0.44 versus 0.43 for the parallel versus the strictly sequential model, respectively). However, M_{TS} was substantially lower (357 SVs versus 536 SVs) because quantal contents for the first eEPSC of the trains are now split up into contributions from SP_{TS} and those from SP_{LS}.

These results show that our eEPSC train data set is also well compatible with a model consisting of two parallel pathways for SV fusion – one representing high p (0.44) SVs, which are rapidly depleted during repetitive stimulation, and a second one representing quite low p (0.065) SVs. Release probabilities of both types of SVs increase during high-frequency stimulation. It should be noted, though, that the mean release probability – averaged over all high p and low p SVs released per stimulus – decreases during stimulus trains because of the preferential depletion of SV_{TS}s.



DISCUSSION

Non-negative matrix factorization (NMF), as applied here to ensembles of eEPSC trains recorded at calyx of Held synapses, is an attempt to obtain insight into the mechanisms of synaptic short-term plasticity from the variability among individual synapses.

Functional heterogeneity within a well-defined population of synapses is often regarded as a nuisance and sought to be eliminated by averaging over multiple trials obtained from many individual synapses. Here we actually make use of the information contained in such synapse-to-synapse variability postulating that it arises from differences in abundances of SVs that pre-existed in distinct functional states prior to stimulation. If so, NMF analysis allows one to approximate the time course of quantal release observed at any individual

Fig. 9. Estimates for release probabilities p_{LS} for a parallel SV priming and fusion scheme featuring fusion competent SV_{LS}s. **(A)** Relationship between the final value for p_{LS} after 100 NMF fit iterations and the initial guess values for p_{LS}. For fit A both BF_{TS}s and M_{TS} values were fixed to those of a preceding two-component NMF fit. For fit B only time courses of BF_{TS}s were fixed to those of a preceding two-component NMF fit, while corresponding M_{TS} values were updated during iterations of the three-component NMF fit. The dotted line represents the identity line. **(B)** The NMF fit prediction for BF_{LS} when using an initial guess (selected on the basis of the relationship shown in **(A)**) which results in p_{LS} = 0.065. **(C)** Time courses of conditional release probability p_{LS} resulting from the BF_{LS}s shown in **(B)** and the respective BF_{LS}s for higher stimulus frequencies. The discrepancy between the 1st value of BF_{LS} (0.08) and the 1st value of p_{LS} (0.065) arises from the renormalization of BFs (see *Experimental procedures*) which was applied for calculating p values at increased accuracy. The difference is negligible for BF_{TS}s which decay rapidly to zero within 25 stimuli. It is on the order of 20% for BF_{LS}s which do not decay completely within 25 stimuli.

synapse in response to repetitive stimulation as a linear superposition of contributions from distinct SV subpools. Quantal contents of such subpools are specific for a given synapse, while time courses for a given component are identical for all synapses. We call such time courses 'basefunctions' (*BF*s). *BF*s are normalized to a cumulative sum of 1, such that the response j during a stimulus train applied to a given synapse i is the sum over contributions by subpools, each given by the product of $M_i \times BF_j$ (Eq. (1)). Surprisingly, relatively good NMF fits can be obtained by postulating only two or three SV subpools: (i) a subpool of SVs which are released rapidly during repetitive stimulation (SP_{TS}), (ii) a subpool of SVs which have to undergo a final step of priming before being released (SP_{LS}), and (iii) a subpool of SVs which are recruited, primed, and released during stimulation (SP_{RS}). In two-component NMF fits, SP_{LS} and SP_{RS} are treated as one common subpool ($SP_{LS,RS}$). The NMF algorithm, similar to other 'blind source separation techniques', provides estimates for both *BF*s and *M*s.

Unfortunately, solutions provided by NMF decomposition are not necessarily unique for typical electrophysiological data sets. Rather, depending on the choice of parameter initialization, the final estimates for *BF*s and their associated *M*s can be quite variable. Thus, a second goal of the work, presented here, is to identify appropriate constraints and suitable initial guess values for fit parameters to select those solutions that are mechanistically compatible with knowledge about functional states of SVs and/or kinetic schemes for SV priming and fusion. Of particular interest to us are solutions compatible with a two-step priming process, as suggested for the calyx of Held synapse (Neher and Brose, 2018) and neuromuscular synapses of C. elegans (Michelassi et al., 2017) (see also *Introduction*). Performance of the NMF algorithm, its reliability, insights about mechanisms of SV priming and synaptic STP derived from its application, and options for consolidation of the findings in future work will be discussed here.

Sparsity of the data set and constraints

Solutions of non-negative factorization techniques are best defined by high sparsity and high dimensionality of the data set (Cichocki et al., 2009). Unfortunately, data sets considered in this work are not sparse, since substantial release is observed at all synapses, even late during high-frequency eEPSC trains. Extending the data set to a third dimension (see paragraph on NTF) should alleviate the problem, however requires great care in data acquisition and very stable recording conditions. Thus, without further constraints, multiple solutions of the NMF fitting process are possible. The approach proposed here is to formulate constraints derived from a particular two-step SV priming scheme (Fig. 3). Given this model, the NMF fit results provide a number of interesting features (discussed below). However, NMF fit solutions are not guaranteed to actually reflect SV states as assumed for the formulation of constraints. Rather, NMF should be considered as a tool to explore the consequences of assumptions made in the formulation of constraints. For example, it allows one to explore what time courses of components are compatible with the observed variability among synapses when assuming 2 or 3 subpools of SVs with specific properties as listed above. NMF decomposition offers such time courses, together with parameters, such as p or M values for the individual components, which may serve as initial guesses for fitting average release time courses to kinetic schemes of SV priming and fusion. Such models can invoke additional data from the given set of synapses, such as recovery from depression in order to validate the assumed scheme. For a reverse approach one may calculate basefunctions, as predicted by a given model and use NMF to test, whether these are compatible with the observed variations among synapses.

The choice of kinetic schemes of SV priming and fusion

In principle, NMF decomposition of eEPSC train data can be applied in the absence of a tangible underlying kinetic scheme of quantal release. It is only the choice of constraints, which requires specific assumptions regarding the sequence of transitions that eventually leads to SV fusion. The view of two structurally different docking states (TS and LS), adopted here, implies that only SV_{TS}s can be released synchronously during the short lifetime of a local $[Ca^{2+}]_i$ domain. We used this argument to justify the constraint of $BF_{LS,1} = 0$. However, recently a detailed study of the sub-millisecond time course of AP-evoked release at glutamatergic synapses of the cerebellum showed tight synchronization of release only for single APs and early APs in stimulus trains. Following repetitive stimulation, individual AP-evoked release transients decayed bi-exponentially with fast and slow time constants of 0.49 ms and 1.87 ms, respectively (Miki et al., 2018). The slowly decaying release component was interpreted as 'two-step-release', representing fusion of SVs that undergo a final step of priming followed by exocytosis during the short lifetime of an AP-induced $[Ca^{2+}]_i$ transient. If such two-step release occurred also at the calyx of Held, its contribution would be interpreted to be generated by the SP_{LS}, since it would vary among synapses with the size of the SP_{LS}. Thus, the basic postulate of $BF_{LS,1} = 0$ would be violated if two-step release occurred already during the first stimulus in a train. Exploring how direct release from SP_{LS} influences *BF*s revealed some interesting features, although these depended strongly on the choice of initial guess values for fitting parameters. In particular, a combination of such parameters, which resulted in $p_{LS} = 0.065$ provided *BF*s with almost exponential decays for low stimulus frequencies (5–20 Hz, Fig. 9B). Characteristically, release probabilities of the two postulated SV pools were quite distinct under these conditions, with 0.44 and 0.065 for high-p and low-p SVs, respectively. They were almost constant during low-frequency trains for both pools, but increased during high-frequency trains by up to a factor of two (Fig. 9C).

Comparison of *p* estimates derived from NMF/NTF fits to those obtained by 'traditional' methods

Because of the simplicity of the analysis, reported estimates for average release probability ($\overline{p}$) are often based on the measured release during depleting stimulus trains. For a synapse at rest, $\overline{p}$ is conveniently obtained by calculating the ratio of the initial eEPSC size over some measure of the total *RRP* size, with the latter being derived from the total number of quanta that can be released during high-frequency trains consisting of several tens of stimuli (see Neher, 2015 for discussion). How do p_{TS} estimates obtained by NMF analysis compare to such $\overline{p}$ values? We consider here the sequential SV priming and fusion scheme illustrated in Fig. 3 and note that virtually all SV_{TS}s and SV_{LS}s that pre-existed prior to stimulation are depleted during such stimulus trains (Fig. 3A, 5B). Therefore, the total *RRP* size when estimated by cumulative release methods corresponds to the sum of the TS and LS SV subpools. Because we postulate that only a smaller fraction of all SVs belonging to total *RRP* is in a fusion competent state, i.e. only SVs of the TS subpool, the relation between $\overline{p}$ and p_{TS} is expected to be $\overline{p} = p_{TS} \times M_{TS}/(M_{TS} + M_{LS}) = 0.43 \times 436/(436 + 901) = 0.14$. In other words, the 43% of all SV_{TS} that are consumed during a single eEPSC or $eEPSC_1$ in a train correspond to approximately 14% of the sum of all SV_{TS} and SV_{LS} combined that pre-existed prior to stimulation.

Alternatively, one may estimate $\overline{p}$ by analyzing stochastic properties of the release process, for example by measuring the relationship between eEPSC variance and their mean amplitude (Oleskevich et al., 2000; Meyer et al., 2001; Scheuss and Neher, 2001; Koike-Tani et al., 2008). Under the assumption that SV priming is also a stochastic process, such methods yield $\overline{p}$ estimates that represent the product of two probabilities: the probability of a SV docking site being occupied by a release-ready SV and the probability of a docked SV being released in response to an AP ($\overline{p} = p_{occ} \times p_r$) (Vere-Jones, 1966). Within the frame work of our NMF analysis and given the SV priming and fusion scheme illustrated in Fig. 3, p_r corresponds to p_{TS}. With all docking sites occupied at rest and assuming that only SV_{TS}s are fusion competent (Fig. 3), p_{occ} amounts to $M_{TS}/(M_{TS} + M_{LS})$ and, thus, $\overline{p}$ is again 0.14. This should, however, be regarded as an upper bound to the estimate by variance and mean, because SV docking is a reversible process and therefore some docking sites may be empty even at rest, which reduces p_{occ}.

It is noteworthy that 'traditional' $\overline{p}$ estimates for calyx of Held synapses at a comparable developmental stage (p = 0.13 in Taschenberger et al., 2002; p = 0.15 in Taschenberger et al., 2005; p = 0.2 in Koike-Tani et al., 2008) are very similar to the average release probability derived from NMF. In conclusion, the relatively high release probability p_{TS} reported here for a small subpool of tightly-docked SVs is well in line with the lower estimates for the average release probability $\overline{p}$ reported in previous studies that consider the sum of all primed SV (SV_{TS}s and SV_{LS}s) as representing the total *RRP*.

Properties of basefunctions

We restricted our analysis to eEPSC trains elicited by stimulus trains at frequencies of 5 Hz and higher because of evidence that priming states of SVs are in a dynamic equilibrium with each other also at rest (discussed in Brose and Neher, 2018). The timescale of such fluctuations is given by the recovery time course of release after pool-depleting stimuli, which is in the 1–10 s range for presynaptic $[Ca^{2+}]_i$ close to resting values. Thus, one can expect that for frequencies of 5 Hz and higher, when release and recruitment rates are high, SVs progress unidirectionally along the priming pathway. For lower frequencies (<5 Hz), however, one would have to consider that SVs can undergo multiple back and forward transitions during long-lasting interstimulus intervals, before being released by an AP. *BF*s for stimulation frequencies <5 Hz are therefore expected to be complex and are unlikely to report release of SVs that had been in a specific state prior to stimulation.

A surprising result of NMF decomposition is that *BF*s in a stimulus frequency range of 5–20 Hz are very similar when plotted against stimulus number. This applies to both BF_{TS}s (Fig. 4A) and BF_{LS}s (Fig. 6A). Here these findings are discussed, considering the kinetic scheme shown in Fig. 3. BF_{TS}s in this stimulus frequency range decay almost exponentially indicating that p of SV_{TS}s is quite constant during repetitive stimulation, unless ISIs are <20 ms. Deviations observed at such short ISIs indicate an increase in p (Fig. 4B insert). The increase in p_{TS} during high-frequency stimulation, has features similar to those of Ca^{2+} current facilitation (Borst and Sakmann, 1998; Cuttle et al., 1998; Lin et al., 2011). This facilitation contributes importantly to paired-pulse facilitation (PPF) of quantal release typically observed for such short ISIs at the calyx of Held (Inchauspe et al., 2004; Ishikawa et al., 2005). BF_{LS}s represent pre-existing SV_{LS}s, which require a LS → TS transition before being able to fuse. Assuming TS to be a well-defined molecular SV state, our model implies that these SVs eventually undergo fusion with the same p as those SVs that had been in TS already prior to stimulation. The finding that BF_{LS}s for stimulus frequencies of 5–20 Hz decay with very similar time constants (when measured in units of ISIs, Fig. 6A) indicates that each stimulus causes both release and LS → TS transition of a constant fraction of all pre-existing SV_{LS}s still remaining available at that stimulus. The length of the ISIs does not seem to affect the size of this fraction, unless it is $\leq$20 ms. As illustrated in Fig. 8B, a simple model implementing these features (Eqs. (5)–(7)) describes very well the mean time course of BF_{LS}s in the range 5–20 Hz. At very short ISIs, the fraction of LS → TS transitions per AP increases (Fig. 6B), contrary to the expectation for a rate-limited process.

Arguments very similar to those for the interpretation of the BF_{LS}s also apply to the contributions of the newly recruited SVs, represented by the BF_{RS}s. Their similarity between 5 and 50 Hz again indicates that a certain fraction of release sites is being refilled after an AP,

largely independent of the actual length of the ISI. Assuming a $[Ca^{2+}]_i$-dependent process of recruitment this is readily understood for low stimulus frequencies, since global $[Ca^{2+}]_i$ transients after single APs decay with a time constant of ~25 ms (Müller et al., 2007), such that there is not much buildup of $[Ca^{2+}]_i$ during repetitive stimulation at ≤10 Hz. Rather, each AP in a train will cause a very similar global $[Ca^{2+}]_i$ transient and therefore very similar amounts of enhanced SV recruitment above the basal replenishment rate, even if the intrinsic relationship between the rate constant of SV recruitment and $[Ca^{2+}]_i$ is non-linear. Only at stimulus frequencies that cause a buildup of global $[Ca^{2+}]_i$ during repetitive stimulation would a non-linear $[Ca^{2+}]_i$ dependency of SV recruitment become relevant. Further options for explaining the enhanced docking and priming reactions during high-frequency bouts of synapse activity arise, when considering that such processes might depend not only on global $[Ca^{2+}]_i$ levels, but rather on short-lived local $[Ca^{2+}]_i$ domains in the vicinity of release sites. Furthermore, any other model, in which an AP causes constant amounts of release and reloading of SVs will result in *BF*s, which are similar when plotted against stimulus number. Irrespective of its molecular mechanism, the early acceleration of the LS → TS transition at stimulus frequencies ≥100 Hz provides a contribution to PPF in addition to that caused by an increase in p_{TS}. This may account for the remaining facilitation of quantal release observed under experimental conditions that eliminate presynaptic $I_{Ca(V)}$ facilitation (Müller et al., 2008, their Fig. 8).

Given these features of *BF*s, one can expect to obtain valuable insight about mechanisms of STP by NMF analysis. Heterogeneity among synapses, similar to that described here, has been observed in a variety of glutamatergic synapses. In such studies, fast components of short-term depression have been described to be mediated either by so-called 'pre-primed' SVs (Hanse and Gustafsson, 2001) or else by 'superprimed' SVs (Schlüter et al., 2006; Taschenberger et al., 2016), or more generally as rapidly releasing SVs in two-step priming models (Doussau et al., 2017). It will be interesting to see whether the interpretations suggested by NMF apply to all glutamatergic synapses and their STP. However, one should also be aware of the *caveat* that the constraints applied in the NMF analysis may not be sufficient to guarantee a unique solution. Rather, NMF results should be considered as an exploratory tool providing results that can serve as a source of ideas for further validation with models of synaptic release, which take into account additional data, such as time courses of recovery and/or effects of pharmacological and genetic interference with the release machinery.

ACKNOWLEDGEMENTS

We thank Drs. Suk-Ho Lee and Takeshi Sakaba for valuable discussions and comments on the manuscript and I. Herfort for excellent technical assistance. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2067/1-390729940 and the DFG Collaborative Research Center 1286 "Quantitative Synaptology" (E.N.).

REFERENCES

Borst JG, Helmchen F, Sakmann B (1995) Pre- and postsynaptic whole-cell recordings in the medial nucleus of the trapezoid body of the rat. J Physiol 489:825–840.
Borst JG, Sakmann B (1998) Facilitation of presynaptic calcium currents in the rat brainstem. J Physiol 513:149–155.
Chang S, Trimbuch T, Rosenmund C (2018) Synaptotagmin-1 drives synchronous Ca. Nat Neurosci 21(1):33–40.
Cichocki A, Zdunek R, Phan AH, Amari S (2009) Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. Chichester: Wiley.
Cuttle MF, Tsujimoto T, Forsythe ID, Takahashi T (1998) Facilitation of the presynaptic calcium current at an auditory synapse in rat brainstem. J Physiol 512:723–729.
Debanne D, Guerineau NC, Gahwiler BH, Thompson SM (1996) Paired-pulse facilitation and depression at unitary synapses in rat hippocampus: quantal fluctuation affects subsequent release. J Physiol 491(Pt 1):163–176.
Diamond JS, Jahr CE (1997) Transporters buffer synaptically released glutamate on a submillisecond time scale. J Neurosci 17(12):4672–4687.
Dobrunz LE, Stevens CF (1997) Heterogeneity of release probability, facilitation, and depletion at central synapses. Neuron 18 (6):995–1008.
Dorgans K, Demais V, Bailly Y, Poulain B, Isope P, Doussau F (2019) Short-term plasticity at cerebellar granule cell to molecular layer interneuron synapses expands information processing. eLife:8.
Doussau F, Schmidt H, Dorgans K, Valera AM, Poulain B, Isope P (2017) Frequency-dependent mobilization of heterogeneous pools of synaptic vesicles shapes presynaptic plasticity. Elife 6.
Elmqvist D, Quastel DM (1965) A quantitative study of end-plate potentials in isolated human muscle. J Physiol 178(3):505–529.
Fekete A, Nakamura Y, Yang YM, Herlitze S, Mark MD, DiGregorio DA, Wang LY (2019) Underpinning heterogeneity in synaptic transmission by presynaptic ensembles of distinct morphological modules. Nat Commun 10(1):826.
Forsythe ID, Barnes-Davies M (1993) The binaural auditory pathway: excitatory amino acid receptors mediate dual timecourse excitatory postsynaptic currents in the rat medial nucleus of the trapezoid body. Proc Biol Sci 251(1331):151–157.
Habets RL, Borst JG (2005) Post-tetanic potentiation in the rat calyx of Held synapse. J Physiol 564(Pt 1):173–187.
Hallermann S, Fejtova A, Schmidt H, Weyhersmuller A, Silver RA, Gundelfinger ED, Eilers J (2010a) Bassoon speeds vesicle reloading at a central excitatory synapse. Neuron 68(4):710–723.
Hallermann S, Heckmann M, Kittel RJ (2010b) Mechanisms of short-term plasticity at neuromuscular active zones of Drosophila. HFSP J 4(2):72–84.
Hanse E, Gustafsson B (2001) Vesicle release probability and pre-primed pool at glutamatergic synapses in area CA1 of the rat neonatal hippocampus. J Physiol 531(Pt 2):481–493.
He E, Wierda K, van Westen R, Broeke JH, Toonen RF, Cornelisse LN, Verhage M (2017) Munc13-1 and Munc18-1 together prevent NSF-dependent de-priming of synaptic vesicles. Nat Commun 8:15915.
Hosoi N, Holt M, Sakaba T (2009) Calcium dependence of exo- and endocytotic coupling at a glutamatergic synapse. Neuron 63 (2):216–229.
Hua Y, Woehler A, Kahms M, Haucke V, Neher E, Klingauf J (2013) Blocking endocytosis enhances short-term synaptic depression under conditions of normal availability of vesicles. Neuron 80 (2):343–349.
Imig C, Min SW, Krinner S, Arancillo M, Rosenmund C, Südhof TC, Rhee J, Brose N, et al. (2014) The morphological and molecular

nature of synaptic vesicle priming at presynaptic active zones. Neuron 84(2):416–431.
Inchauspe CG, Martini FJ, Forsythe ID, Uchitel OD (2004) Functional compensation of P/Q by N-type channels blocks short-term plasticity at the calyx of held presynaptic terminal. J Neurosci 24 (46):10379–10383.
Ishikawa T, Kaneko M, Shin HS, Takahashi T (2005) Presynaptic N-type and P/Q-type Ca^{2+} channels mediating synaptic transmission at the calyx of Held of mice. J Physiol 568(Pt 1):199–209.
Katz B (1969) The release of neural transmitter substances. Liverpool, England: Liverpool University Press.
Koike-Tani M, Kanda T, Saitoh N, Yamashita T, Takahashi T (2008) Involvement of AMPA receptor desensitization in short-term synaptic depression at the calyx of Held in developing rats. J Physiol 586(9):2263–2275.
Korogod N, Lou X, Schneggenburger R (2005) Presynaptic Ca^{2+} requirements and developmental regulation of posttetanic potentiation at the calyx of Held. J Neurosci 25(21):5127–5137.
Kruskal JB (1977) Three-way arrays - rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear Algebra Appl 18(2):95–138.
Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. Adv Neur In 13:556–562.
Lee JS, Kim MH, Ho WK, Lee SH (2008) Presynaptic release probability and readily releasable pool size are regulated by two independent mechanisms during posttetanic potentiation at the calyx of Held synapse. J Neurosci 28(32):7945–7953.
Lin KH, Oleskevich S, Taschenberger H (2011) Presynaptic Ca^{2+} influx and vesicle exocytosis at the mouse endbulb of Held: a comparison of two auditory nerve terminals. J Physiol 589(Pt 17):4301–4320.
Meyer AC, Neher E, Schneggenburger R (2001) Estimation of quantal size and number of functional active zones at the calyx of held synapse by nonstationary EPSC variance analysis. J Neurosci 21(20):7889–7900.
Michelassi F, Liu H, Hu Z, Dittman JS (2017) A C1–C2 module in Munc13 inhibits calcium-dependent neurotransmitter release. Neuron 95(3):577–590 e575.
Miki T, Nakamura Y, Malagon G, Neher E, Marty A (2018) Two-component latency distributions indicate two-step vesicular release at simple glutamatergic synapses. Nat Commun 9 (1):3943.
Moulder KL, Mennerick S (2005) Reluctant vesicles contribute to the total readily releasable pool in glutamatergic hippocampal neurons. J Neurosci 25(15):3842–3850.
Müller M, Felmy F, Schneggenburger R (2008) A limited contribution of Ca^{2+} current facilitation to paired-pulse facilitation of transmitter release at the rat calyx of Held. J Physiol 586(Pt 22):5503–5520.
Müller M, Felmy F, Schwaller B, Schneggenburger R (2007) Parvalbumin is a mobile presynaptic Ca^{2+} buffer in the calyx of held that accelerates the decay of Ca^{2+} and short-term facilitation. J Neurosci 27(9):2261–2271.
Müller M, Goutman JD, Kochubey O, Schneggenburger R (2010) Interaction between facilitation and depression at a large CNS synapse reveals mechanisms of short-term plasticity. J Neurosci 30(6):2007–2016.
Neher E (2015) Merits and limitations of vesicle pool models in view of heterogeneous populations of synaptic vesicles. Neuron 87 (6):1131–1142.
Neher E, Brose N (2018) Dynamically primed synaptic vesicle states: key to understand synaptic short-term plasticity. Neuron 100 (6):1283–1291.
Neher E, Sakaba T (2001) Combining deconvolution and noise analysis for the estimation of transmitter release rates at the calyx of Held. J Neurosci 21(2):444–461.
Neher RA, Mitkovski M, Kirchhoff F, Neher E, Theis FJ, Zeug A (2009) Blind source separation techniques for the decomposition of multiply labeled fluorescence images. Biophys J 96 (9):3791–3800.
Oleskevich S, Clements J, Walmsley B (2000) Release probability modulates short-term plasticity at a rat giant terminal. J Physiol 524(Pt 2):513–523.
Pan B, Zucker RS (2009) A general model of synaptic transmission and short-term plasticity. Neuron 62(4):539–554.
Prinslow EA, Stepien KP, Pan YZ, Xu J, Rizo J (2019) Multiple factors maintain assembled trans-SNARE complexes in the presence of NSF and alphaSNAP. Elife 8.
Rosenmund C, Stevens CF (1996) Definition of the readily releasable pool of vesicles at hippocampal synapses. Neuron 16 (6):1197–1207.
Sahara Y, Takahashi T (2001) Quantal components of the excitatory postsynaptic currents at a rat central auditory synapse. J Physiol 536(Pt 1):189–197.
Scheuss V, Neher E (2001) Estimating synaptic parameters from mean, variance, and covariance in trains of synaptic responses. Biophys J 81(4):1970–1989.
Schlüter OM, Basu J, Südhof TC, Rosenmund C (2006) Rab3 superprimes synaptic vesicles for release: implications for short-term synaptic plasticity. J Neurosci 26(4):1239–1246.
Schneggenburger R, Meyer AC, Neher E (1999) Released fraction and total size of a pool of immediately available transmitter quanta at a calyx synapse. Neuron 23(2):399–409.
Südhof TC (2012) The presynaptic active zone. Neuron 75(1):11–25.
Taschenberger H, Leao RM, Rowland KC, Spirou GA, von Gersdorff H (2002) Optimizing synaptic architecture and efficiency for high-frequency transmission. Neuron 36(6):1127–1143.
Taschenberger H, Scheuss V, Neher E (2005) Release kinetics, quantal parameters and their modulation during short-term depression at a developing synapse in the rat CNS. J Physiol 568(Pt 2):513–537.
Taschenberger H, von Gersdorff H (2000) Fine-tuning an auditory synapse for speed and fidelity: developmental changes in presynaptic waveform, EPSC kinetics, and synaptic plasticity. J Neurosci 20(24):9162–9173.
Taschenberger H, Woehler A, Neher E (2016) Superpriming of synaptic vesicles as a common basis for intersynapse variability and modulation of synaptic strength. Proc Natl Acad Sci U S A 113(31):E4548–4557.
Thanawala MS, Regehr WG (2016) Determining synaptic parameters using high-frequency activation. J Neurosci Methods 264:136–152.
Trommershäuser J, Schneggenburger R, Zippelius A, Neher E (2003) Heterogeneous presynaptic release probabilities: functional relevance for short-term plasticity. Biophys J 84(3):1563–1579.
Vere-Jones D (1966) Simple stochastic models for the release of quanta of transmitter from a nerve terminal. Austr J Stat 8 (2):53–63.
Wong AY, Graham BP, Billups B, Forsythe ID (2003) Distinguishing between presynaptic and postsynaptic mechanisms of short-term depression during action potential trains. J Neurosci 23 (12):4868–4877.
Zenisek D, Steyer JA, Almers W (2000) Transport, capture and exocytosis of single synaptic vesicles at active zones. Nature 406 (6798):849–854.

(Received 1 August 2020, Accepted 7 October 2020)
(Available online 13 January 2021)