We sincerely thank all reviewers for their detailed feedback and the positive comments indicating that we present a novel idea (R2, R4) of great relevance for the community (R2) with an empirical evaluation that demonstrates the performance of the method on both images and graphs (R1). We will clarify the issues raised by the reviewers in our response below.

Firstly R1, R2, R3 are interesteed in runtime of the convexp. We agree that this can be valueable and we will include a runtime comparison experiment with a table of the different linear flows in our updated manuscript. In short, the increase in computation is straightforward: we typically need 6 convolutional calls (Fig. 3), but this is somewhat balanced by a cheap determinant. Our tests show that the runtime of the convexp-flow utilizes 10.9% more computation time during training than a flow using 1x1 convolutions.

R1, R4 note that the section on equivariance is lacking clarity, because the notation in this section is somewhat different. We agree, and we will clean up this notation and remove terms that are not necessary to understand the section such as capsules and feature field. The main idea of this section is to prove that the exponential convolution preserves the equivariant properties of its underlying convolutions. *R1* asks why $[K, M] = 0$ implies $[K^n, M^m]$. The result can be derived as follows: $[A, BC] = ABC - BCA = ABC - BAC + BAC - BCA = [A, B]C + B[A, C]$. Higher powers follow from induction. Our claim $[K, \exp M] = 0$ can also be derived straightforwardly: Define $\exp_n$ as the exponential taking only the first n terms of the series. Since $0 = \lim_{n\to\infty}[\exp_n M, K] = [\lim_{n\to\infty} \exp_n M, K] = [\exp M, K]$ by continuity of $[..., ...]$. These intermediate steps will also be clarified in the manuscript. We will also make a note of the connection with Lie algebras and Lie groups. By request of R4 we will also focus more on the intuitive implications.

R1 is asking whether the determinant of Sylvester flows could be derived without Sylvester's identity, and asks whether Sylvester flows always have dimensionality reduction. Note that the original Sylvester flows come in three flavours: Although O-SNFs reduce dimensions, H-SNFs and T-SNFs do not. Further, there are indeed multiple derivations that give the determinant in Eq. 9. Notice that the derivation suggested by R1 is very similar to our "Remark II, App. A" which gives an alternate proof for invertibility (which also applies to H/T-SNFs). Due to the similarity between our extension and H/T-SNFs, we decided to call them generalized Sylvester Flows. To clarify our manuscript, we will include the derivation suggested by R1 and write Thm. 1 more succinctly using (Papamakarios, 2019).

R3 questions whether f_{AR} is indeed L-Lipschitz. Firstly, note that this is for an arbitrarily high constant L (which makes it a rather weak constraint). The reason why we require this, is so that $\gamma^t \cdot L \to 0$ eventually, and the FPI converges. R3 is correct that in theory on $\mathbb{R}$ the function may not be Lipschitz, caused by the product $s_1(u) \cdot u$. However, computer signals generally have bounded domains, and f_{AR} is already L-Lipschitz on these bounded domains. The theoretical issue can be solved by altering the transformation slightly: we can clip/threshold the variable u which is multiplied with s_1, but only for *very high* magnitudes. Since s_1, s_2 are bounded by a $\tanh$ and u is now also bounded, the function f_{AR} is now Lipschitz even for $\mathbb{R}$. Note that this modifications leaves all experimental results valid, as these large values were practically never reached. Although it may be expensive to compute L explicitly, it can be steered by limiting the Lipschitz continuity of s_1, s_2, t_1, t_2. Further, γ should be seen as an upper bound on the magnitude of the diagonal of $\mathbf{J}_{f_{\mathrm{AR}}}$. Since s_1, s_2 are bounded $(-1, 1)$ and are strictly triangular functions, the γ in the main paper is an upper bound of this diagonal. This discussion will be added to the inverse analysis in the appendix.

R3 asks about the expressitivity of the exponential. The output of the exponential ($\exp M$) can be any matrix that is the solution to the linear ODE of the form $\dot{\mathbf{x}} = \mathbf{M}\mathbf{x}$ from $t = 0$ to $t = 1$. It cannot model all invertible matrices, for instance, matrices with negative determinants cannot be modelled. Further, the spectral normalization constraints the matrix $\mathbf{M}$ in the possible linear ODEs. This will be included in the manuscript.

Other comments/questions | *R2 asks about replacing spectral norm with weight norm.* Although weight norm could also improve the series convergence, spectral norms give theoretically guaranteed convergence behaviour as shown in Fig. 3. | *R3 asks for a discussion on connections related work.* We will explain the connection to computing the logarithm of a matrix (which in contrast with the exp cannot always be computed in a stable fashion and converges slowly) from (Behrmann et al.) and the orthogalization procedure in (Li et al. 2019). | *R1 notes that Emerging convs and Sylvester flows both need to be solved iteratively.* Correct, but in L217-218 different linear flows are compared. This iterative inverse of emerging convs would make them impractical to use as basis change inside Sylvester flows as both inverse and forward of the linear flow are required during optimization. | *R2 asks how methods were compared.* In image experiments, we adjusted the size of the coupling layers to ensure a roughly equal parameter budget. For graph experiments, the coupling layers were kept the same as the convexp only added a negligible number of parameters ($< 0.01\%$). | *R3 is concerned that convexp are combined with 1x1 convs.* Note that 1x1 convs can be made cheap, especially when modelled using Householder transformations (as we proposed for Sylvester Flows). We will include a discussion of the tendency of convexp to remain close to the identity in the main text and connect it to the limited induced matrix norm. | *R3 suggests to add details on spectral normalization.* We will extend the discussion of spectral norms using (Gouk et al.) | *R1, R2, R3, R4:* Beside the already mentioned changes we also fixed all minor issues that reviewers spotted in the paper (e.g. typos, tips to better phrase some sentences and references to the appendix.)