We thank the reviewers for their positive comments on clarity, novelty, and convincing experiments. As R3 noted, one key feature of the method is that the measured data in a given direction is not directly used to estimate the denoised value: we literally replace each direction with its patch-wise prediction from the other directions using a linear model. It was unexpected to us too that such a simple method would work so well. This is a key advantage of self-supervised learning for medical imaging applications such as Diffusion-Weighted MRI (DWI).

We thank **R1** for pointing to the resources for denoising using dictionary learning and we will cite the mentioned papers. Those papers have two limitations: they are designed for densely sampled and long acquisitions (DSI) and also require learning the dictionary on high resolution data. Patch2Self in contrast works on any acquisition scheme and is unsupervised. Sparsity in a learned basis is an important approach distinct from our own, and we will mention it. The model is not trained in an online manner. We train one regressor per held-out volume (§2.2). To further clarify the training, we have included a descriptive implementation with examples on open source datasets.

The comments from the **R2** are constructive and we address the criticism herewith. Both Patch2Self and Marchenko-Pastur (MP) are patch based algorithms, assembling voxels from a patch across volumes into a matrix. MP uses a low-rank approximation of that matrix, with thresholds based on random matrix theory which depends on an assumption of homoskedastic noise. By learning a self-supervised regressor, Patch2Self relies on a weaker assumption of independent noise over different volumes and can learn a full rank model. We agree with R2 that smoother does not imply better, and therefore have provided extensive qualitative and quantitative comparisons on both synthetic and real data. We also agree that the qualitative comparisons are hard to make and therefore provide quantitative comparisons using goodness-of-fit measure for tensor and spherical harmonic models that downstream analyses depend on. Improved RMSE and R^2 scores on realistic synthetic data have been reported as more direct measures of performance. We would like to note that the tractography comparisons are in fact quantitative in nature, where a kernel density estimate is performed on each streamline to evaluate the spurious streamlines from the tracking (using the Fiber Bundle Coherency (FBC) Metric) and the number of streamlines is counted. R2 is correct that diffusion kurtosis (DKI) is a linear estimation problem. By degeneracy, we refer to obtaining a biophysically impossible parameter estimates. The microstructure model, here DKI, is unable to fit due to the low SNR of the acquired data. DKI is also a very widely used higher-order model following the initially proposed Diffusion Tensor (DTI) mapping, motivating us to include it as a part of our analysis. In the presence of more noise, one can imagine that the slope of the line fitted to the DWI data to be reversed, causing a degeneracy (seen as dark black voxels). Values close to zero are not plausible in a healthy brain and can therefore confound DKI analyses, classifying that voxel as disease/ abnormal (as in the case of tissue degeneration). As pointed out by the arrows, the number of these degenerate (black) voxels have been reduced in the data. We do agree that there are some more degeneracies that the denoising could not suppress, but may be indicative of other artefacts such as Gibbs oscillations/ ringing. With respect to impact, the paper is aimed for the brain imaging category of NeurIPS. DWI is currently the most powerful non-invasive way of assessing structural information, and denoising is an essential component of DWI analysis. The proposed approach of self-supervised learning can also be extended to other 4D modalities like fMRI.

We thank the **R3** for the positive feedback. The reviewer is correct in identifying part A, where patches corresponding to the direction j are held-out and used as targets for training a regression function Φ. The rest of the directions were used as features, and the model is trained on all voxels. In part B, the same training samples are given to the trained regressor, whose output is the denoised volume. The rationale behind doing so is that, since the noise across volumes is uncorrelated, the regressor will only be able to learn to predict the underlying signal and not the noise component. Collinearity is unlikely in DWI because of noise in the measurements. Moreover, the similar performance of OLS, L1- and L2- regularized regression implies that collinearity is not a problem (§S1). As for the comment on line line 150 by R3, an increase in dimensionality per volume will only increase the number of training samples, not the number of features, per volume (giving more information about the same object). Only an increase in the number of volumes gives the regressor access to different information about the held out volume. For Fig. 2, although direct comparisons cannot be done on real data, we provide a way to quantify the improvements due to denoising using R^2 and FBC measures on downstream tasks of microstructure analysis and tractography.

While we do agree with **R4** that the R^2 metric is indirect, we also provide the more direct RMSE scores for the synthetic phantom in Table §1. (This is not possible in the case of the real data due to the absence of ground truth.) We thank the reviewer for pointing out the mentioned references and will be included in the final document. For a justification of using MP as a baseline for comparison, we would like to point out that MP is the most cited method for DWI denoising (320+ citations since 2016) and is well received by the community from an application standpoint. Open source implementations are available in popular software packages such as MRtrix3 and DIPY. Without ground-truth for these datasets we are unable to train supervised methods, but we do compare against the state-of-the-art classical DWI algorithms such as AONLM and Local PCA in the supplement (see §S2 and §S3). The suggestion to compare to a deep neural net is useful. Our expectation is that it will not do much better because a shallow fully connected neural network (§S1) had a higher loss than the linear model.