

Patch2Self2: Self-supervised Denoising on Coresets via Matrix Sketching

Shreyas Fadnavis*
 Johnson and Johnson R&D
 Cambridge, MA
 sfadnavi@its.jnj.com

Agniva Chowdhury
 Oak Ridge National Laboratory
 Oak Ridge, TN
 chowdhurya@ornl.gov

Joshua Batson
 Anthropic
 San Francisco, CA
 joshb@anthropic.com

Petros Drineas
 Purdue University
 West Lafayette, IN
 pdrineas@purdue.edu

Eleftherios Garyfallidis
 Indiana University Bloomington
 Bloomington, IN
 elef@iu.edu

Abstract

*Diffusion MRI (dMRI) non-invasively maps brain white matter, yet necessitates denoising due to low signal-to-noise ratios. Patch2Self (P2S), employing self-supervised techniques and regression on a Casorati matrix, effectively denoises dMRI images and has become the new de-facto standard in this field. P2S however is resource intensive, both in terms of running time and memory usage, as it uses all voxels (n) from all-but-one held-in volumes ($d - 1$) to learn a linear mapping $\Phi : \mathbb{R}^{n \times (d-1)} \mapsto \mathbb{R}^n$ for denoising the held-out volume. The increasing size and dimensionality of higher resolution dMRI acquisitions can make P2S infeasible for large-scale analyses. This work exploits the redundancy imposed by P2S to alleviate its performance issues and inspect regions that influence the noise disproportionately. Specifically, this study makes a three-fold contribution: (1) We present Patch2Self2 (P2S2), a method that uses matrix sketching to perform self-supervised denoising. By solving a sub-problem on a smaller sub-space, so called, **coreset**, we show how P2S2 can yield a significant speedup in training time while using less memory. (2) We present a **theoretical analysis** of P2S2, focusing on determining the **optimal sketch size** through rank estimation, a key step in achieving a balance between denoising accuracy and computational efficiency. (3) We show how the so-called statistical **leverage scores** can be used to **interpret** the denoising of dMRI data, a process that was traditionally treated as a black-box. Experimental results on both simulated and real data affirm that P2S2 maintains denoising quality while significantly enhancing speed and memory efficiency, achieved by training on a reduced data subset.*

1. Introduction

Patch2Self (P2S) [23] is currently a leading denoising technique for diffusion MRI (dMRI), mainly due to its broad

applicability and parameter-free nature. This method is increasingly used in applications, ranging from strongly-encoded acquisitions [52] and imaging of small (low SNR) structures such as the spinal cord [59] to delineating and measuring treatment responses in cancer [15] and high-field imaging [56]. Additionally, it is also useful in relaxometry for myelin water imaging [9], pediatric/neonatal imaging [61], and neuroimaging in animals, including mice and macaques [8]. Patch2Self serves as a preliminary step in super-resolution processes [10, 33, 41] and is a building-block component in generative diffusion model-based denoising techniques [68]. Not surprisingly, Patch2Self has been integrated into widely-used open-source packages like DIPY [25], QSIPrep [11], and Spinal Cord Toolbox [14].

Diffusion MRI is a 4D acquisition method that generates a series of 3D volumes each corresponding to different gradient directions [4, 34]. Each 3D volume provides unique information about the underlying diffusion processes in the brain. This information is used to probe the tissue microstructure in the living brain by modeling the signal per voxel using a variety of biophysical models [50, 51]. This derived information can however be corrupted due to low signal-to-noise ratio (SNR). Multiple sources of noise are apparent in dMRI that reduce SNR. Furthermore, with new acquisition schemes, high-field MR gradients [47, 62] and multi-dimensional diffusion encoding strategies [30, 55] the effect of noise sources is exaggerated and affects image conspicuity.

Self-supervised learning, as a sub-domain of unsupervised learning algorithms, has been rapidly gaining traction over the past years. Novel strategies of self-supervision are being developed and employed for different problems such as multi-modal learning [48, 69], self-labeling [36, 44], learning semantic context [16] and contrastive predictive coding [3]. Denoising strategies based on self-supervision

have revolutionized image denoising performance across different scientific domains[29, 35]. Leveraging statistical independence of noise, first introduced in the work of Noise2Noise (N2N) [40] was given a theoretical grounding in the work of Noise2Self by posing it under a self-supervised framework [6]. Other approaches similar to N2S and N2N (such as Noise2Void, etc.) were also proposed around the same time [37, 38]. P2S belongs to this same family of algorithms and leverages the statistical independence of noise. In a way, P2S combines N2N and N2S and recasts the 4D self-supervised denoising of dMRI as an image in-painting problem [7]. Instead of holding out a part of the same 3D volume, P2S holds out the entire 3D volume itself and learns to predict a denoised version of the held-out volume as a linear combination of the remaining 3D volumes. P2S relies on the theory of \mathcal{J} -invariance [6, 23] to perform the denoising, which can be seen as a way of performing 4D image in-painting in the q -space [5]. In the case of dMRI data, denoising is typically done on a single subject and P2S showed how one can get state-of-the-art denoising performance using linear regression as its backbone. Self-supervised denoisers (typically making assumptions on the noise) outperform traditional methods of denoising, but can be computationally expensive as they do not place assumptions on signal properties such as sparsity, compressibility, repetition, etc. In pursuit of acquiring data at a higher resolution and to extract detailed diffusion information, the dimensionality of the data increases rapidly (per-volume and per-scan). Moreover, due to the advent of high-field scanners, more noise is induced in the signal due to the use of stronger magnetic gradients. Along with the dimensionality of a single scan, the number of scans acquired and released for analyses is also increasing rapidly, requiring fast and efficient denoising algorithms.

In this work we introduce **Patch2Self2 (P2S2)**, which proposes the sketching of the large matrix \mathcal{A} [23] constructed for training the denoiser via P2S to create a coreset. \mathcal{A} is constructed by vectorizing each 3D volume and concatenating it along the columns of \mathcal{A} . This is called a *Casorati matrix* where each independent measurement forms a column of \mathcal{A} [47]. So, instead of performing the self-supervised denoising on the over-determined set of constraints ($\mathcal{A} \in \mathbb{R}^{n \times d}$, where $n \gg d$), P2S2 samples and rescales the matrix to a much smaller subset of the constraints ($\mathcal{A} \in \mathbb{R}^{s \times d}$, where $n \gg s \approx d$). By training the self-supervised denoiser on this much smaller induced subproblem, we show that one can achieve the same level of denoising performance as P2S with a highly reduced time complexity and a much smaller memory footprint. We show the speedup gains both via the theoretical complexity analysis and the empirical comparisons on simulated and real datasets of different dimensionalities. To ensure that P2S2 does not hamper the denoising performance, we compare

P2S2 against P2S on both simulated and real data using the root mean squared error (RMSE) and the R^2 metrics. We also evaluate the performance on the downstream tasks of microstructure modeling and tractography. While the sketch size required for sketching and solving the linear system within P2S2 may vary, our experiments suggest at least a 60% redundancy in the training set. We compared the performance of P2S2 via different sketching methods such as CountSketch, leverage score sampling, and the Subsampled Randomized Fourier Transform (SRFT). Our results show that leverage score sampling yielded the best performance. We discuss how leverage scores can be used for interpretability of P2S2, revealing which regions of the data have a higher influence on the denoising algorithm. This enables interpretability (crucial to medical imaging) of the self-supervised denoising, which is otherwise treated as a black-box approach. With the help of the Rank Revealing QR (RRQR) decomposition, we also show how one can calibrate the optimal sketch size to construct the coreset for P2S2 via a self-supervised loss.

2. Preliminaries and Approach

Leveraging the fact that each 3D volume of the 4D data can be assumed to be an independent measurement of the same underlying object, P2S proposes constructing a large Casorati matrix wherein each column can be assumed to be linearly independent of the other columns. P2S sets up the self-supervised regression task so that each column is denoised by representing it as a combination of the remaining columns. P2S relies on the self-supervised loss: $\mathcal{L}(\Phi_J) = \mathbb{E} \|\Phi_J(\mathcal{A}_{-j}) - \mathcal{A}_j\|_2^2$. Here \mathcal{A}_j refers to the voxels corresponding to the volume that were held-out and used as target for training the rest of the voxels from the remaining 3D volumes \mathcal{A}_{-j} . P2S showed that this \mathcal{J} -invariant function was in fact a linear map $\Phi_J : \mathcal{A}_{-j} \mapsto \mathcal{A}_j$ that achieved the optimal denoising performance. This allows re-writing the P2S as a linear regression problem:

$$\min_{\mathbf{w}} \|\mathcal{A}_{-j}\mathbf{w} - \mathcal{A}_j\|_2, \quad (1)$$

where $\mathcal{A}_{-j} = [\mathbf{a}_1 \dots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times d-1}$ is the design matrix without the held-out j -th column, denoted by $\mathcal{A}_j = [a_1, \dots, a_n]^\top \in \mathbb{R}^n$. \mathcal{A}_j is the target 3D volume for learning Φ_J . Here $\mathbf{a}_i \in \mathbb{R}^{d-1}$ and $a_i \in \mathbb{R}$, for $i = 1, \dots, n$.

2.1. Coresets for Regression via Matrix Sketching

The self-supervised denoising performed in P2S relies on least squares regression as described in Sec. 2. Linear regression typically performed via Cholesky, SVD or QR decomposition needs $\mathcal{O}(nd^2)$ time. In the case of P2S, the problem setup via the Casorati matrix \mathcal{A} is massively over-constrained, i.e., $n \gg d$ with full-column rank¹, and has an

¹Similar results also apply to a rank-deficient system.

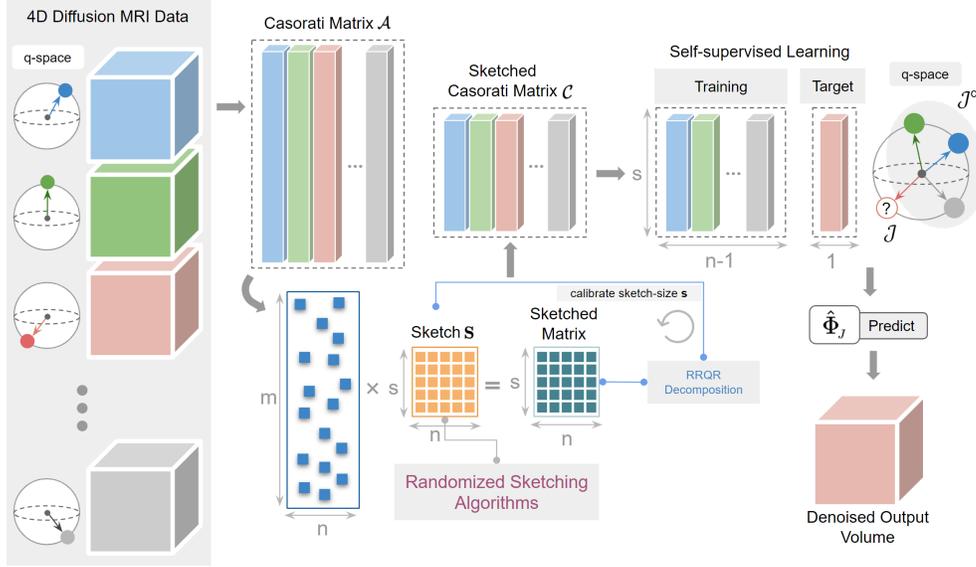


Figure 1. We show how Patch2Self2 works in the case of Diffusion MRI data. The Casorati Matrix \mathcal{A} is constructed by flattening each gradient direction. \mathcal{A} is then sketched via a randomized algorithm to generate the smaller sketched Casorati Matrix \mathcal{C} . The \mathcal{J} -invariant training is then performed on \mathcal{C} to learn $\hat{\Phi}_{\mathcal{J}}$ and predict the denoised volume.

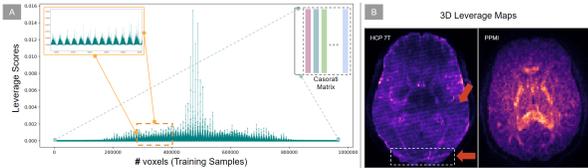


Figure 2. (A) An exemplary point-plot of row leverage scores from the Casorati matrix of a real dataset is depicted. (B) Axial slices of 3D leverage score maps on two datasets are shown, highlighting structured artifacts (marked by red arrows) across gradient directions in the HCP 7T data.

added time complexity, since the regression is performed d times given that each volume needs to be separately denoised. Therefore, in P2S2, instead of computing the exact solution vector $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} \|\mathcal{A}_{-j}\mathbf{w} - \mathcal{A}_j\|_2^2$, we propose to approximate it using tools from randomized matrix multiplication and subspace embeddings. The key idea here is to solve a sub-problem $\tilde{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{S}\mathcal{A}_{-j}\mathbf{w} - \mathbf{S}\mathcal{A}_j\|_2^2$ such that:

$$\|\mathcal{A}_{-j}\tilde{\mathbf{w}} - \mathcal{A}_j\|_2 \leq (1 + \epsilon)\|\mathcal{A}_{-j}\mathbf{w}^* - \mathcal{A}_j\|_2, \quad (2)$$

where ϵ is the desired level of accuracy and $\mathbf{S} \in \mathbb{R}^{s \times n}$ with $d \approx s \ll n$ is the so-called *sketching matrix*². Given the linear sketch $\mathbf{S}\mathcal{A}_{-j}$ of \mathcal{A}_{-j} , note that computing $\tilde{\mathbf{w}}$ takes $\mathcal{O}(sd^2)$ time, which is indeed much faster than the classical computation of \mathbf{w}^* . Therefore, our goal is to work with a suitable \mathbf{S} such that the sketch $\mathbf{S}\mathcal{A}_{-j}$ can be computed ef-

²For ease of exposition we use \mathbf{S} in place of \mathbf{S}_j .

ficiently and satisfies eqn. (2) with high probability. There are several such choices for \mathbf{S} :

Count-sketch [12]: In this case, \mathbf{S} is a sparse embedding matrix with $s = \mathcal{O}(d^2/\epsilon)$ rows and has exactly one non-zero entry per column, which is chosen randomly and set to ± 1 independently. The product $\mathbf{S}\mathcal{A}_{-j}$ can be computed in time $\mathcal{O}(\operatorname{nnz}(\mathcal{A}_{-j}))$. Assuming the failure probability to be a constant, the overall running time to compute $\tilde{\mathbf{w}}$ is given by $\mathcal{O}(\operatorname{nnz}(\mathcal{A}_{-j})) + \operatorname{poly}(d/\epsilon)$. Here, $\operatorname{nnz}(\cdot)$ denotes the sparsity of the underlying matrix.

Fast Johnson-Lindenstrauss transformations [13, 21]: Other choices for the sketching matrix \mathbf{S} include structured random matrices such as the subsampled randomized Fourier transform (SRFT) or the subsampled randomized Hadamard transform (SRHT). In these cases, the sketching matrix \mathbf{S} is typically of the form $\mathbf{S} = \sqrt{\frac{n}{s}}\mathbf{R}\mathbf{H}\mathbf{D}$, where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a random diagonal matrix with the entries set to ± 1 independently; $\mathbf{R} \in \mathbb{R}^{s \times n}$ is a subset of s rows of the $n \times n$ identity matrix chosen uniformly at random, independently, without replacement; and $\mathbf{H} \in \mathbb{R}^{n \times n}$ is either a normalized discrete Fourier transform (for SRFT) or a normalized Walsh-Hadamard matrix (for SRHT). Note that both SRFT and SRHT are based on randomized linear transformations, which can be applied rapidly to arbitrary vectors. Indeed, we can compute the matrix-matrix product $\mathbf{S}\mathcal{A}_{-j}$ in $\mathcal{O}(nd \log n)$ time exploiting the structure of the underlying Fourier/Hadamard matrix; if $s = \mathcal{O}(d + \log^{1/\epsilon} \log d/\epsilon)$, then the resulting sketching matrix satisfies eqn. (2). The overall running time is $\mathcal{O}(nd \log n) + \tilde{\mathcal{O}}(d^3/\epsilon^2)$.

Sampling-based sketching [20]: A third way of achiev-

ing a subspace embedding that satisfies eqn. (2) is data-dependent, and can be obtained by sampling rows of a matrix proportional to their leverage scores $\ell_i = \|\mathbf{U}_{i*}\|_2^2$ for $i = 1 \dots n$, where $\mathbf{U}_{i*} \in \mathbb{R}^n$ is the i -th row of the matrix of the left singular vectors of \mathcal{A}_{-j} that are computed using the thin SVD of \mathcal{A}_{-j} . In this context, the sketching matrix \mathbf{S} is the so-called *sampling-and-rescaling* matrix of [19] with the sampling probabilities $p_i = \ell_i/d$, $i = 1 \dots n$, and the sample complexity is given by $s = \mathcal{O}(d \log d/\epsilon^2)$. We note that computing ℓ_i s exactly needs access to the matrix \mathbf{U} which is expensive. Therefore, in practice, approximate leverage scores also suffice and they can be efficiently computed without computing the matrix \mathbf{U} [12, 22]. We also note that there is another line of work [2, 58] that used sketching as a randomized preconditioner to come up with high precision solutions for overconstrained regression problems. However, in context of P2S2, we can achieve the desired accuracy in denoising performance, even with a *sketch-and-solve* approach as discussed above. Finally, we refer the interested reader to the surveys [17, 18, 42, 46, 67] for background on Randomized Linear Algebra.

2.2. Statistical Leverage for Interpretability

In the previous section, we described how leverage scores can be used to perform importance sampling in order to generate the coreset for P2S2. Here we show how to use the statistical leverage scores of the underlying linear model for interpreting areas of the data that influence the noise. From a statistical perspective, an alternative formulation of leverage scores are the diagonal entries of the projection matrix constructed to solve the linear regression [17, 43]. Typically, one looks at the standard deviation of the noise derived from the models used to perform the denoising [63]. Since P2S and consequently P2S2 are set up to denoise in a predictive setting via self-supervision, we can use statistical leverage to get a more detailed view of the factors that would influence the denoising performance.

In Fig. 2A, we show the leverage scores computed on a subject from the PPMI dataset [45]. As one can see in the plot, some voxels in the data exhibit considerably higher statistical leverage. Voxels where an anatomical signal from the brain is captured exhibit higher leverage when compared to the voxels in the background, which have very small leverage scores. In Fig. 2B, we show the spatial map of leverage scores for two example datasets, PPMI and HCP 7T [62]. In the case of the PPMI data, we note that the regions of the white matter, such as the corpus callosum, have higher statistical leverage compared to the rest of the brain. Strikingly, we also see that a slanting structured pattern (indicated with red arrows) appears in the leverage scores map of the HCP 7T data. This structure can also be seen in the noise map of Fig. 3A in the residual map. Leverage score maps of the same HCP 7T subject also highlighted a struc-

tured artefact at the bottom of the axial slice (depicted as a white-dotted bounding box). It could be a ghost or a structured artefact added to the 7T data.

3. Patch2Self2 Algorithm

P2S2 extends the idea of P2S (see Sec. 2) by performing self-supervised training on coresets. In case of dMRI, we have d number of 3D volumes each with dimensionality: $l \times w \times h$. Each of the volumes is flattened to a 1D array ($n = l \times w \times h$) to form a column of the Casorati matrix $\mathcal{A} \in \mathbb{R}^{n \times d}$. Next, we sketch this matrix \mathcal{A} using the sketching matrix $\mathbf{S} (\in \mathbb{R}^{s \times n})$ to get a sketched Casorati matrix $\mathcal{C} = \mathbf{S}\mathcal{A}$ (see Sec. 2.1). We perform a self supervised denoising on this sketched matrix \mathcal{C} by solving the sub-problem: $\min_{\tilde{\mathbf{w}}} \|\mathcal{C}_{-j}\mathbf{w} - \mathcal{C}_j\|_2$. As shown in 1, this denoising is performed on a volume-by-volume basis as proposed in P2S, where j corresponds to the volume held out for denoising³.

Thus, P2S2 learns a linear map $\Phi_J : \mathbb{R}^{s \times (d-1)} \mapsto \mathbb{R}^s$, which is a much smaller sub-problem to solve since $s \ll n$. After the training is done, the approximate solution vector $\tilde{\mathbf{w}}$ learned via Φ_J is used to predict the held out volume of \mathcal{A}_j . In order to predict the denoised volume, the full Casorati matrix \mathcal{A}_{-j} with all rows is given as input to the function Φ_J . In P2S2, we allow switching between different sketching methods such as SRFT, leverage scores, and CountSketch (see Sec. 2.1 for details of each). Our results show that sketching via leverage scores outperforms other sketching methods (detailed comparison in Sec. 5).

As per the above procedure, P2S2 introduces using a new hyperparameter - the sketch size that needs to be tuned. We propose a \mathcal{J} -invariant self-supervised calibration procedure to find an optimal sketch size s based on the QR decomposition and leverage score sketching. Leverage scores are a univariate statistic. When two rows of \mathcal{A} have identical or similar leverage scores, it implies that the rows are highly correlated and therefore redundant in the construction of the sketched Casorati matrix \mathcal{C} . Thus, a redundancy removal step is often useful to reduce the sketch size s while retaining denoising performance. Towards that end, P2S2 employs the Rank Revealing QR (RRQR) factorization [28], to flag such redundancies by ranking the rows of the matrix in order of independence. In other words, highly linearly independent rows are given priority after the RRQR has been computed. In P2S2, to calibrate the sketch size, we start with $s < 50\%$ of the number of rows in \mathcal{A} and compute the rank-revealing QR (RRQR) decomposition [28] of \mathcal{C} . The RRQR ranks the rows of \mathcal{C} in order of importance and can be used to remove redundant rows from the sketched matrix. Next, as shown in Fig. 4A, we select the top k -ranked rows from \mathcal{C} and compute the self-supervised loss. Gradually, by

³In Algorithm 1, we use * to represent matrix dimensions for clarity.

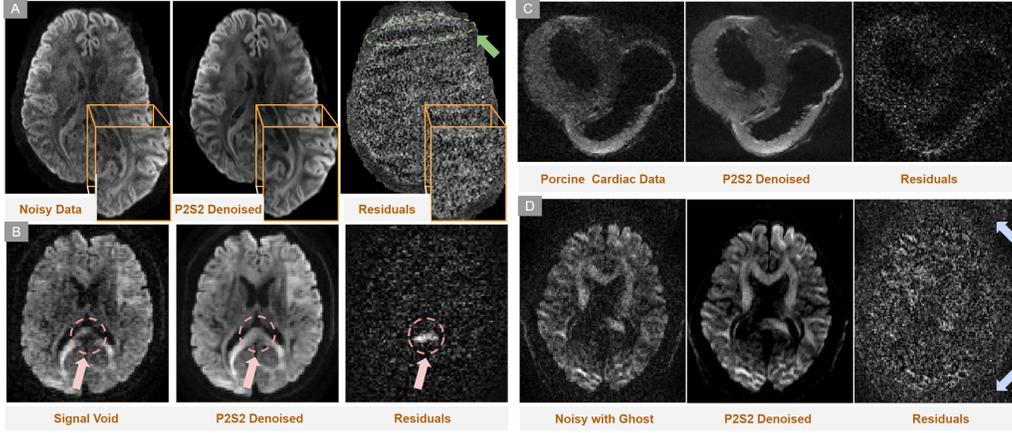


Figure 3. Depicts (A) Suppression of noisy band-like patterns in residual maps by P2S2 (green arrow). (B) In-painting of signal voids in original data and predicted signal in residual map (red arrows). (C) Application of P2S2 to other organs like the heart (Porcine Cardiac Data). (D) Reduction of ghosting artifacts (blue arrows).

Algorithm 1 Patch2Self2

- 1: Input 4D data X of dimension $l \times w \times h \times d$
- for volume $j = 1, 2, \dots, d$ [where d is the number of volumes] do
 - 2: Flatten volume v_j into an n -sized feature vector [$m = l \times w \times h$].
- 3: Stack feature vectors into a $n \times d$ matrix to form the Casorati matrix \mathcal{A} .
- for volume $j = 1, 2, \dots, d$ do
 - 4: Generate $\mathbf{S}_j \in \mathbb{R}^{s \times n}$ from row leverage scores of $\mathcal{A}_{*, -j}$.
 - 5: Compute a sketched Casorati matrix $\mathcal{C}_{*, -j} = \mathbf{S}_j \mathcal{A}_{*, -j} \in \mathbb{R}^{s \times (d-1)}$
 - 6: $[\mathbf{Q}, \mathbf{R}, \text{rank}] = \text{RRQR}(\mathcal{C}_{*, -j})$.
 - 7: Select top k rows from $\mathcal{C}_{*, -j}$ post-RRQR for $\hat{\mathcal{C}}_{*, -j} \in \mathbb{R}^{k \times (d-1)}$.
 - 8: Hold-out features of volume j to form a $k \times (d-1)$ matrix $\hat{\mathcal{C}}_{*, -j}$.
 - 9: Select the target volume j to be denoised $\hat{\mathcal{C}}_{*, j}$ of dimension $k \times 1$.
 - 10: Train a linear regressor $\Phi: \hat{\mathcal{C}}_{*, -j} \mapsto \hat{\mathcal{C}}_{*, j}$.
 - 11: Set the denoised volume $\hat{X}_{*, *, *, j}$ to the unraveled output $\hat{\Phi}(\mathcal{A}_{*, -j})$.
- 12: Return denoised 4D data \hat{X}

increasing the size of k in each iteration we compute the self-supervised loss for each iteration. The self-supervised loss for each value of k is computed as: $\|\Phi_j(\mathcal{A}_{-j}) - \mathcal{A}_j\|_2$. As shown in Fig. 4B, the loss eventually converges to a minimum with minimal change in the denoising performance. At that point, the elbow in the error plot (shown by the red box) reveals the number of redundant rows of the data, which allows us to estimate the (approximately) optimal s .

Leverage Score Sampling Strategies: We can obtain the sketched matrix \mathcal{C} via leverage score sampling using the following two procedures: (1) Deterministically choosing top s leverage scores; (2) Randomized sampling based on leverage scores. From Fig. 2A, it is evident that only a few rows of the Casorati matrix have a very high leverage score in comparison with the remaining ones. These rows seem to have a higher influence on the denoising performance. As shown in Fig. 6A, the deterministic selection of the highest leverage scores consistently performs worse at all sketch sizes when compared to randomly sampling leverage scores. To investigate this effect, we compared the distribution of the rows corresponding to deterministically chosen

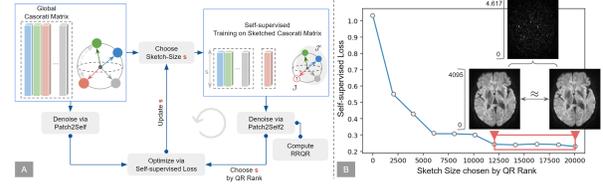


Figure 4. (A) Iterative self-supervised calibration procedure to obtain the optimal sketch size s to be used within P2S2 using a RRQR decomposition. the self-supervised loss is computed, by iteratively reducing the size of s until convergence. (B) We depict an example loss plot of the self-supervised calibration with exemplary PPMI data.

top 20K leverage scores against the randomly sampled 20K rows. From the joint plot obtained by fitting a kernel density estimation shown in Fig. 6B, it can be seen that the distribution of the randomly sampled leverage scores forms a bi-modal distribution as opposed to the uni-modal distribution obtained from a deterministic selection of the top 20K leverage scores. This implies that the randomization in the sampling procedure helps denoising by using values that do not have a "high leverage".

4. Theoretical Analysis of Sketch Size

As mentioned in Sec. 2.1, if ρ represents the rank of \mathcal{A}_{-j} with $\rho \leq d-1 \ll n$, eqn. (2) holds with high probability through leverage score-based sampling with a sample size of $k = \mathcal{O}(\rho \log \rho / \epsilon^2)$. Utilizing standard notations, let the *mean squared errors* (MSEs) resulting from the exact solution \mathbf{w}^* and the sampling-based solution $\tilde{\mathbf{w}}$ be expressed as $\text{MSE}^* = \|\mathcal{A}_{-j} \mathbf{w}^* - \mathcal{A}_j\|_2^2 / n$ and $\widehat{\text{MSE}} = \|\mathcal{A}_{-j} \tilde{\mathbf{w}} - \mathcal{A}_j\|_2^2 / n$, respectively. From eqn. (2), we can straightforwardly derive the following:

Proposition 1 *If the sketching matrix \mathbf{S} satisfies eqn. (2), then $\widetilde{\text{MSE}} \leq (1 + \epsilon) \text{MSE}^*$.*

Proof Squaring and dividing both sides of eqn. (2) by n readily yields $\widetilde{\text{MSE}} \leq (1 + \epsilon)^2 \text{MSE}^*$. Given that $\epsilon \leq 1$, this further simplifies to $\widetilde{\text{MSE}} \leq (1 + 3\epsilon) \text{MSE}^*$. Therefore, with a simple change of variables (using $\epsilon/3$ instead of ϵ) and sample size $k = \mathcal{O}(\rho \log \rho / \epsilon^2)$, we can indeed observe that the $\widetilde{\text{MSE}}$ closely approximates MSE^* . Now, the question is how to obtain a precise estimate of ρ to determine an accurate sketch size k that approximates the optimal MSE.

To this end, the following discussion aims to narrow the *theory-practice gap* and validate that Steps 5-7 in Algorithm 1 are indeed a theoretically relevant approach to determining the optimal k . Let $\mathcal{A}_{-j} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ represent the thin SVD of \mathcal{A}_{-j} , where the matrices $\mathbf{U} \in \mathbb{R}^{n \times \rho}$, $\mathbf{V} \in \mathbb{R}^{d-1 \times \rho}$, and $\mathbf{\Sigma} \in \mathbb{R}^{\rho \times \rho}$ follow standard notational convention. For our purpose, we can start with an initial sketch size $s = \mathcal{O}(d \log d / \epsilon^2)$ (which is always greater than k) and compute \mathcal{C}_{-j} from Step 5 in Algorithm 1. Now, we already have the following result:

Proposition 2 (Lemma 4.1 of [20]) *If \mathbf{S} satisfies the ℓ_2 -subspace embedding property, we have $\text{rank}(\mathbf{S}\mathbf{U}) = \rho$.*

Similar to eqn. (2), it can also be shown that our leverage score-based sampling matrix \mathbf{S} indeed satisfies the aforementioned ℓ_2 -subspace embedding property with high probability using $\mathcal{O}(\rho \log \rho / \epsilon^2)$ samples (e.g., see [67]). Next result is a well-known, provable fact from linear algebra literature about matrix products and their rank.

Fact 3 *For any two matrices \mathbf{A} and \mathbf{B} , $\text{rank}(\mathbf{A}\mathbf{B}) = \text{rank}(\mathbf{A})$ if \mathbf{B} is of full row-rank.*

Now, note that $\mathbf{\Sigma}$ is non-singular and \mathbf{V} is of full-column-rank. Therefore, applying Fact 3 sequentially on $\mathbf{S}\mathbf{U}$ and combining with Proposition 2, we have $\text{rank}(\mathcal{C}_{-j}) = \text{rank}(\mathbf{S}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top) = \text{rank}(\mathbf{S}\mathbf{U}) = \rho$. This implies that we can actually determine ρ directly through \mathcal{C}_{-j} using RRQR in Step 6, and consequently, obtain the optimal sketch size k from Step 7 of Algorithm 1. Here note that as $s = \mathcal{O}(d \log d)$, the worst case complexity of Step 6 is given by $\mathcal{O}(d^3 \log d)$. When $\rho \ll d - 1$, the RRQR algorithm can be even more efficient, and the actual computational cost may be much lower than $\mathcal{O}(d^3 \log d)$ in practice. In this context, it's also worth mentioning that approximate leverage scores are sufficient for satisfying eqn. (2) as well as the aforementioned ℓ_2 -subspace embedding property. Their computation can be efficiently done without the need to compute \mathbf{U} , achieving a time complexity of $\mathcal{O}(\text{nnz}(\mathcal{A}_{-j}) \log n + d^3 \log^2 d + d^2 \log n)$ due to [12].

5. Evaluation on Real Data

To evaluate the performance on real data, we qualitatively and quantitatively compare the effects of P2S2 denoising on the residual maps, microstructure modeling [50] and tractography [31]. We quantify the effect of different sketch

sizes on the speed and accuracy of the approximate solution ($\tilde{\mathbf{w}}$) in comparison with the P2S solution (\mathbf{w}^*).

5.1. Noise suppression and artefact removal

Visual conspicuity of the data (i.e. image quality) is crucial to any form of medical imaging, especially dMRI where the images are inherently limited by SNR. While thermal noise is known to dominate the sources of noise that corrupt the underlying signal [63], different acquisition strategies tend to induce different types of artefacts that hamper the signal structure. We show that the self-supervised setup of P2S2 deals with these artefacts without loss of signal corresponding to anatomical structure. In Fig. 3A, we denoised a subject **HCP 7T** dataset [62] using only 50K out of 61M (0.083%) training samples obtained via leverage score sampling. Noise mapping from this type of high-field imaging data (acquired using 7 Tesla scanner) is still under-investigated. We show that P2S2 suppressed band-like structured noise, which may be correlated across some volumes, but is largely uncorrelated across all 3D volumes. One of the main motives of high-field 7T scanners is to acquire data at a much higher resolution. With a zoomed-in cross-section in Fig. 3A, we show how P2S2 uncovers much more anatomical detail without loss of information.

Signal voids are a common issue in MRI that occur due to certain voxels not emitting any radio-frequency signal due to a lack of activated protons in that region [65]. Since P2S, and consequently P2S2, are similar to image inpainting [7], where an entire 3D volume is predicted as a combination of the rest of the volumes, this signal void can be imputed with context learned from the rest of the volumes. This setup resolves a unique issue for dMRI data which was not addressed by any other denoising algorithm in the past. In Fig. 3B we show how P2S2 fills the signal void present in the Stanford HARDI [57] data in gradient direction 33, without removing or smoothing the signal in the rest of the image. In Fig. 3C,D we show how P2S2 does not cause any signal loss even in the presence of physiological noise (porcine cardiac data) [24, 49] and ghost artefacts [39, 54], which are ubiquitous in MRI acquisitions. Note that in either case, P2S2 strictly only suppresses noise and does not lead to signal loss or smoothing. [In the supplement, we compare the P2S2 residuals with P2S].

Data	SNR 5		SNR 10		SNR 15		SNR 20		SNR 25	
	MSE	R ²	MSE	R ²	MSE	R ²	MSE	R ²	MSE	R ²
Raw	4.45	0.01	2.39	0.15	1.65	0.37	1.32	0.56	1.16	0.69
P2S	2.98	0.17	1.43	0.57	1.14	0.78	1.07	0.84	0.98	0.88
P2S2	2.82	0.10	1.42	0.52	1.21	0.75	1.15	0.83	0.99	0.87

Table 1: Quantitative comparison via MSE and R^2 metrics.

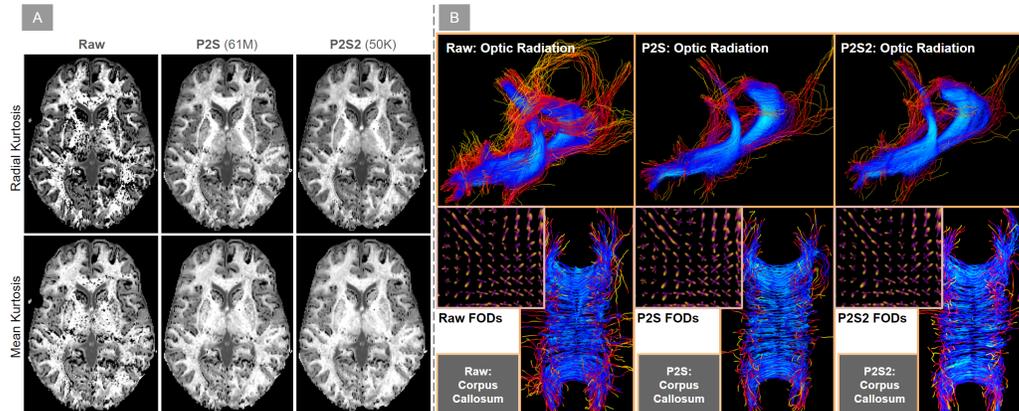


Figure 5. (A) Demonstrates the impact of P2S and P2S2 denoising on HCP 7T data using Diffusion Kurtosis Imaging (DKI), comparing Radial Kurtosis and Mean Kurtosis metrics for both. (B) Compares P2S2, P2S, and noisy data performance through fiber-to-bundle coherence in optic radiation and corpus callosum, including cross-sections of spherical harmonic representations (FODs).

5.2. Comparison of Sketching Methods

The P2S2 algorithm follows a sketch-and-solve [20, 64] approach and therefore the sketch size can affect the denoising performance. In order to quantify this effect, we chose a random subject from the PPMI dataset which was acquired with a widely used 64-directions DTI protocol. A random volume (here vol. #11) from this data was first denoised with P2S and the solution \mathbf{w}^* obtained from it was treated as the optimal solution. Each volume of this subject contained around 960K voxels. Starting with a sketch size of 500 samples, P2S2 denoising was performed on the data with sketches computed using CountSketch, leverage score sampling and SRFT algorithms explained in Sec. 2.1. The sketch size was then increased iteratively until the approximate solution $\tilde{\mathbf{w}}$ was numerically very close to \mathbf{w}^* . The relative error for each iteration and method was computed as: $\frac{\|\mathbf{w}^* - \tilde{\mathbf{w}}\|_2}{\|\mathbf{w}^*\|_2}$. This procedure was repeated ten times to capture the variance of denoising performance since the underlying algorithms used to approximate the solution are randomized. The variance with a 95% confidence interval was plotted at each iteration (i.e., for each sketch size).

We also compared the performance of the sketching methods with uniform sampling and with the deterministic choice of the rows corresponding to the top leverage scores, for the same sketching sizes. As shown in Fig. 6A, the variance of all the sampling algorithms is reduced as the sketch size increases. Uniform sampling and deterministic leverage scores perform worse than the randomized algorithms at each sketch size. While CountSketch and leverage score sampling perform approximately the same, leverage score sampling performs slightly better and offers the added advantage of interpretability. In Fig. 6C, we also empirically compare the speedup obtained from P2S2 in comparison with P2S. This supplements our theoretical complexity analysis in Sec. 2.1. With experiments on three different

datasets we note that the speedup obtained increases as the sketch size reduces. We also find that the speedup obtained via P2S2 increases in proportion to the dimensionality of the data. As one can see in Fig. 6C, the speedup on the Stanford HARDI data (shape: $81 \times 106 \times 76 \times 160$) is much more than simulated (shape: $256 \times 256 \times 4 \times 63$) and PPMI data (shape: $116 \times 116 \times 72 \times 65$). The QR decomposition computed as a part of self-supervised calibration in P2S2 does not add a significant computational overhead. The wall-clock time on an i7 CPU with 16GB RAM for the QR decomposition took 0.0904s on a sketch-size of 20k which amounts to 20% of the PPMI data [45] that the calibration was run on (see Fig. 4). If repeated 10 times for calibration, the QR computation would take $< 1s$ due to fewer than 20k rows in subsequent runs.

5.3. Impacts on Microstructure and Tractography

To estimate the underlying tissue **microstructure** in the living brain one typically fits a biophysical model to each voxel of the dMRI data to capture tissue heterogeneity. Diffusion kurtosis imaging (DKI) [32] is one such modeling scheme that quantifies the degree of non-Gaussian diffusion. DKI is however sensitive to noise and can often lead to fitting degeneracy in the derived maps. In Fig. 5A, we show that P2S alleviates this issue by significantly reducing the failures of model fitting in the data. We demonstrate that both P2S and P2S2 methods offer comparable estimates of radial (RK) and mean (MK) kurtosis derived from DKI metrics, using the same HCP 7T dataset. This data, which exhibited band-like structured noise, showed that noise affects the DKI metrics, particularly visible in the raw noisy data. To assess the impact of denoising on these metrics, RK and MK were analyzed. Notably, P2S2, trained on just 0.083% of the 61M raw data samples, delivered results similar to P2S. The process involves modeling the signal for each voxel in dMRI data, aiding in white matter tract recon-

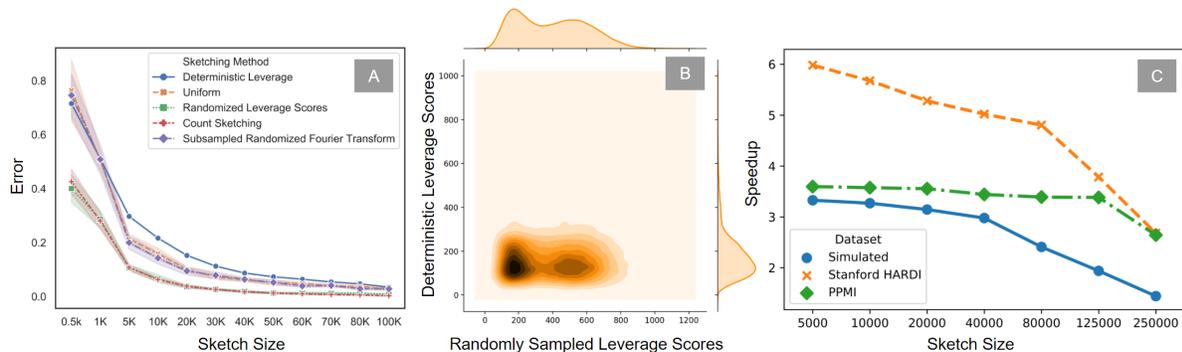


Figure 6. (A) Compares error variance in P2S2’s solution approximation to P2S using three randomized algorithms (CountSketch, leverage score sampling, SRFT) against uniform sampling and deterministic selection of highest leverage score rows. (B) Presents a joint kernel density estimate plot for 20K voxels sampled randomly by leverage scores versus deterministic choice. (C) Displays empirical speedup comparisons with P2S across three datasets, varying the sketch size.

struction (tractography).

However, noise in the data can lead to inaccurate tractography, creating spurious tracts. Denoising helps in obtaining more accurate signal representations, thereby reducing the occurrence of false-positive tracts. In order to evaluate this, probabilistic tractography was performed on the Stanford HARDI dataset [57]. The tracking was performed using Constrained Spherical Deconvolution (CSD) model [60] (implementation using DIPY [26]). In Fig. 5B we show the effect of denoising via P2S and P2S2 on the fiber orientation distribution (FOD) plotted via the CSD model. To denoise via P2S2, only 50K samples out of the 6M (8.3%) samples were used in the training process. Note that both P2S and P2S2 suppress noisy lobes uncovering the underlying fiber crossings. The generalized fractional anisotropy obtained from the constrained solid angle algorithm [1] was used as the stopping criterion of the probabilistic tracking. Next, the streamlines tracking the optic radiation and corpus callosum bundles, obtained from noisy, P2S denoised and P2S2 denoised data, were quantified using the fiber-to-bundle coherence metric [53] shown in Fig. 5B. Red-yellow streamlines show erratic flows; blue streamlines indicate consistent flows. Despite the probabilistic tracking algorithm’s stochastic nature, P2S and P2S2 yield similar results.

5.4. Evaluation on Simulated Data

In a performance comparison on simulated data, following the P2S methodology [23, 27, 66], data with 2 b0 (non-dMRI) and 60 diffusion-weighted dMRI volumes (30 at 1000 s/mm^2 and 30 at 2000 s/mm^2) was generated. Gaussian noise was added using an 8-channel coil sensitivity map to mimic realistic Rician noise. Six datasets were simulated: one noise-free and others with SNR of 10, 15, 20, 25, and 30. Both P2S and P2S2 were used for denoising. Their performance, as seen in Fig. 7A, was visually similar. Quantitative analysis using root mean squared error and the

R^2 metric (Table 1) showed both methods closely matched in denoising efficiency, improving with higher SNR. Scatter plots at SNRs of 15 and 20 (Fig. 7B) further demonstrated the near-perfect overlap between P2S and P2S2 results.

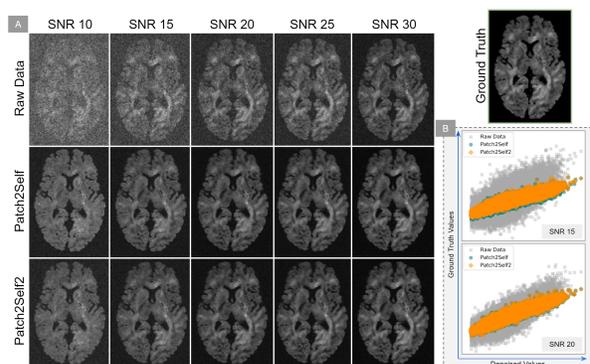


Figure 7. Shows (A) Qualitative comparison of denoising performance between P2S and P2S2 (trained on 20K samples, i.e. 7% of the data). (B) P2S2 closely approximates P2S via scatter plots at SNRs 15&20.

6. Conclusion

In this work, we introduced Patch2Self2, a self-supervised denoising technique using coresets created through matrix sketching, achieving notable speedups and reduced memory usage. We found that sampling-based sketching with leverage scores yielded the best results. Interestingly, leverage scores can also serve as a statistic to identify influential regions affecting denoising performance. Furthermore, we provided theoretical evidence demonstrating that Patch2Self2 can automatically detect the optimal sketch size with high probability. Patch2Self2 will be released as part of DIPY (dipy.org).

References

- [1] Iman Aganj, Christophe Lenglet, Guillermo Sapiro, Essa Yacoub, Kamil Ugurbil, and Noam Harel. Reconstruction of the orientation distribution function in single- and multiple-shell Q-ball imaging within constant solid angle. *Magnetic resonance in medicine*, 64(2):554–566, 2010. [8](#)
- [2] Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging LAPACK’s least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010. [4](#)
- [3] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. [1](#)
- [4] P.j. Basser, J. Mattiello, and D. Lebihan. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1):259–267, 1994. [1](#)
- [5] Peter J Basser. Relationships between diffusion tensor and q-space mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 47(2):392–397, 2002. [2](#)
- [6] Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, pages 524–533. PMLR, 2019. [2](#)
- [7] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. [2](#), [6](#)
- [8] Tobias J Buscham, Maria A Eichel-Vogel, Anna M Steyer, Olaf Jahn, Nicola Strenzke, Rakshit Dardawal, Tor R Memhave, Sophie B Siems, Christina Müller, Martin Meschkat, et al. Progressive axonopathy when oligodendrocytes lack the myelin protein cmtm5. *bioRxiv*, pages 2021–11, 2021. [1](#)
- [9] Kwok-Shing Chan, Maxime Chamberland, and José P Marques. On the performance of multi-compartment relaxometry for myelin water imaging (mcr-mwi)–test-retest repeatability and inter-protocol reproducibility. *NeuroImage*, 266:119824, 2023. [1](#)
- [10] Geng Chen, Yafeng Wu, Dinggang Shen, and Pew-Thian Yap. Xq-nlm: denoising diffusion MRI data via x-q space non-local patch matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 587–595. Springer, 2016. [1](#)
- [11] Matthew Cieslak, Philip A Cook, Xiaosong He, Fang-Cheng Yeh, Thijs Dhollander, Azeez Adebimpe, Geoffrey K Aguirre, Danielle S Bassett, Richard F Betzel, Josiane Bourque, et al. Qsiprep: an integrative platform for preprocessing and reconstructing diffusion mri data. *Nature methods*, 18(7):775–778, 2021. [1](#)
- [12] Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017. [3](#), [4](#), [6](#)
- [13] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal Approximate Matrix Product in Terms of Stable Rank. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, pages 11:1–11:14, 2016. [3](#)
- [14] Benjamin De Leener, Simon Lévy, Sara M Dupont, Vladimir S Fonov, Nikola Stikov, D Louis Collins, Virginie Callot, and Julien Cohen-Adad. Sct: Spinal cord toolbox, an open-source software for processing spinal cord mri data. *Neuroimage*, 145:24–43, 2017. [1](#)
- [15] Robin den Boer, Kelvin Ng Wei Siang, Mandy Yuen, Alicia Borggreve, Ingmar Defize, Astrid van Lier, Jelle Ruurda, Richard van Hillegersberg, Stella Mook, and Gert Meijer. A robust semi-automatic delineation workflow using denoised diffusion weighted magnetic resonance imaging for response assessment of patients with esophageal cancer treated with neoadjuvant chemoradiotherapy. *Physics and Imaging in Radiation Oncology*, 28:100489, 2023. [1](#)
- [16] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2015. [1](#)
- [17] Petros Drineas and Michael W. Mahoney. RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016. [4](#)
- [18] Petros Drineas and Michael W. Mahoney. *Lectures on randomized numerical linear algebra*. American Mathematical Society, 2018. [4](#)
- [19] Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006. [4](#)
- [20] Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136, 2006. [3](#), [6](#), [7](#)
- [21] Petros Drineas, Michael W Mahoney, Shan Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011. [3](#)
- [22] Petros Drineas, Malik Magdon-Ismael, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(111):3475–3506, 2012. [4](#)
- [23] Shreyas Fadnavis, Joshua Batson, and Eleftherios Garyfallidis. Patch2self: Denoising diffusion MRI with self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 16293–16303. Curran Associates, Inc., 2020. [1](#), [2](#), [8](#)
- [24] Pedro F Ferreira, Sonia Nilles-Vallespin, Andrew D Scott, Ranil de Silva, Philip J Kilner, Daniel B Ennis, Daniel A Auger, Jonathan D Suever, Xiaodong Zhong, Bruce S Spottiswoode, et al. Evaluation of the impact of strain correction on the orientation of cardiac diffusion tensors with in vivo and ex vivo porcine hearts. *Magnetic resonance in medicine*, 79(4):2205–2215, 2018. [6](#)
- [25] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, Ian Nimmo-Smith, and Dipy Contributors. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8, 2014. [1](#)

- [26] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian and Nimmo-Smith. Dipy, a library for the analysis of diffusion MRI data. *Frontiers in Neuroinformatics*, 8, 2014. 8
- [27] Mark S. Graham, Ivana Drobnjak, and Hui Zhang. Realistic simulation of artefacts in diffusion MRI for validating post-processing correction techniques. *NeuroImage*, 125: 1079–1094, 2016. 8
- [28] Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996. 4
- [29] Allard Adriaan Hendriksen, Daniël Maria Pelt, and K Joost Batenburg. Noise2inverse: Self-supervised deep convolutional denoising for tomography. *IEEE Transactions on Computational Imaging*, 6:1320–1335, 2020. 2
- [30] Rafael N Henriques, Marco Palombo, Sune N Jespersen, Noam Shemesh, Henrik Lundell, and Andrada Ianuş. Double diffusion encoding and applications for biomedical imaging. *Journal of Neuroscience Methods*, page 108989, 2020. 1
- [31] Saad Jbabdi and Heidi Johansen-Berg. Tractography: where do we go from here? *Brain connectivity*, 1(3):169–183, 2011. 6
- [32] Jens H. Jensen and Joseph A. Helpert. MRI quantification of non-gaussian water diffusion by kurtosis analysis. *NMR in Biomedicine*, 23(7):698–710, 2010. 7
- [33] Changhao Jiang, Xuanyu Tian, Yanbin Li, Jiangjie Wu, Xin Mu, Lei Zhang, and Yuyao Zhang. Self-supervised high-dimensional magnetic resonance image denoising using super-resolved single noisy image. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 1
- [34] Derek K Jones. *Diffusion MRI*. Oxford University Press, 2010. 1
- [35] Wesley Khademi, Sonia Rao, Clare Minnerath, Guy Hagen, and Jonathan Ventura. Self-supervised poisson-gaussian denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2131–2139, 2021. 2
- [36] Lisa Sophie Kölln, Omar Salem, Jessica Valli, Carsten Gram Hansen, and Gail McConnell. Label2label: Using deep learning and dual-labelling to retrieve cellular structures in fluorescence images. *bioRxiv*, 2020. 1
- [37] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images, 2018. 2
- [38] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising, 2019. 2
- [39] Denis Le Bihan, Cyril Poupon, Alexis Amadon, and Franck Lethimonnier. Artifacts and pitfalls in diffusion MRI. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 24(3):478–488, 2006. 6
- [40] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *ICML*, pages 2971–2980, 2018. 2
- [41] Matthew Lyon, Paul Armitage, and Mauricio A Álvarez. Spatio-angular convolutions for super-resolution in diffusion mri. *arXiv preprint arXiv:2306.00854*, 2023. 1
- [42] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. 4
- [43] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009. 4
- [44] Devraj Mandal, Shrisha Bharadwaj, and Soma Biswas. A novel self-supervised re-labeling approach for training with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390, 2020. 1
- [45] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 4, 7
- [46] Per-Gunnar Martinsson and Joel Tropp. Randomized numerical linear algebra: Foundations & algorithms. *arXiv preprint arXiv:2002.01387*, 2020. 4
- [47] Steen Moeller, Pramod Pisharady Kumar, Jesper Andersson, Mehmet Akcakaya, Noam Harel, Ruoyun Ma, Xiaoping Wu, Essa Yacoub, Christophe Lenglet, and Kamil Ugurbil. Diffusion imaging in the post hcp era. *Journal of Magnetic Resonance Imaging*, 2020. 1, 2
- [48] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. *arXiv preprint arXiv:2103.15916*, 2021. 1
- [49] Kévin Moulin, Ilya A Verzhbinsky, Nyasha G Maforo, Luigi E Perotti, and Daniel B Ennis. Probing cardiomyocyte mobility with multi-phase cardiac diffusion tensor MRI. *PLoS one*, 15(11):e0241996, 2020. 6
- [50] Dmitry S Novikov, Valerij G Kiselev, and Sune N Jespersen. On modeling. *Magnetic resonance in medicine*, 79(6):3172–3193, 2018. 1, 6
- [51] Dmitry S. Novikov, Valerij G. Kiselev, and Sune N. Jespersen. On modeling. *Magnetic Resonance in Medicine*, 79(6):3172–3193, 2018. 1
- [52] Marco Pizzolato, Erick Jorge Canales-Rodríguez, Mariam Andersson, and Tim B Dyrby. Axial and radial axonal diffusivities and radii from single encoding strongly diffusion-weighted mri. *Medical Image Analysis*, 86:102767, 2023. 1
- [53] J. M. Portegies, R. H. J. Fick, G. R. Sanguinetti, S. P. L. Meesters, G. Girard, and R. Duits. Improving fiber alignment in hardi by combining contextual pde flow with constrained spherical deconvolution. *Plos One*, 10(10), 2015. 8
- [54] David A Porter, Fernando Calamante, David G Gadian, and Alan Connelly. The effect of residual nyquist ghost in quantitative echo-planar diffusion imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(2):385–392, 1999. 6
- [55] Alexis Reymbaut and Maxime Descoteaux. Advanced encoding methods in diffusion MRI. *arXiv preprint arXiv:1908.04177*, 2019. 1

- [56] Monica Roascio, Gabriele Arnulfo, Domenico Tortora, Mariasavina Severino, Andrea Rossi, Antonio Napolitano, Marco M Fato, et al. Influence of adaptive denoising on diffusion kurtosis imaging at 3t and 7t. *Computer Methods and Programs in Biomedicine*, 234:107508, 2023. [1](#)
- [57] Ariel Rokem. Stanford hardi surfaces, 2016. [6](#), [8](#)
- [58] Vladimir Rokhlin and Mark Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008. [4](#)
- [59] Kurt G Schilling, Shreyas Fadnavis, Joshua Batson, Mereze Visagie, Anna JE Combes, Colin D McKnight, Francesca Bagnato, Eleftherios Garyfallidis, Bennett A Landman, Seth A Smith, et al. Denoising of diffusion mri in the cervical spinal cord—effects of denoising strategy and acquisition on intra-cord contrast, signal modeling, and feature conspicuity. *Neuroimage*, 266:119826, 2023. [1](#)
- [60] J-Donald Tournier, Fernando Calamante, and Alan Connelly. Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage*, 35(4):1459–1472, 2007. [8](#)
- [61] Rosella Denise Trò, Monica Roascio, Domenico Tortora, Mariasavina Severino, Andrea Rossi, Eleftherios Garyfallidis, Gabriele Arnulfo, Marco Massimo Fato, and Shreyas Fadnavis. Data-driven characterization of preterm birth through intramodal diffusion mri. *bioRxiv*, pages 2023–01, 2023. [1](#)
- [62] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012. [1](#), [4](#), [6](#)
- [63] Jelle Veraart, Dmitry S. Novikov, Daan Christiaens, Benjamin Ades-Aron, Jan Sijbers, and Els Fieremans. Denoising of diffusion MRI using random matrix theory. *NeuroImage*, 142:394–406, 2016. [4](#), [6](#)
- [64] Shusen Wang, Alex Gittens, and Michael W Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *International Conference on Machine Learning*, pages 3608–3616. PMLR, 2017. [7](#)
- [65] Van J Wedeen, Robert M Weisskoff, and Brigitte P Poncelet. MRI signal void due to in-plane motion is all-or-none. *Magnetic resonance in medicine*, 32(1):116–120, 1994. [6](#)
- [66] Qiuting Wen, Mark Graham, Sourajit Mustafi, Ivana Drobnyak, Hui Zhang, and Yu-Chien Wu. Comparing the LPCA and MPPCA denoising approaches for diffusion MRI using simulated human data. *International Society for Magnetic Resonance in Medicine*, 2017. [8](#)
- [67] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2), 2014. [4](#), [6](#)
- [68] Tiange Xiang, Mahmut Yurt, Ali B Syed, Kawin Setsompop, and Akshay Chaudhari. ddm^2 : Self-supervised diffusion mri denoising with generative diffusion models. *arXiv preprint arXiv:2302.03018*, 2023. [1](#)
- [69] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. [1](#)

Patch2Self2: Self-supervised Denoising on Coresets via Matrix Sketching [Supplement]

Shreyas Fadnavis*
Johnson and Johnson R&D
Cambridge, MA
sfadnavi@its.jnj.com

Agniva Chowdhury
Oak Ridge National Laboratory
Oak Ridge, TN
chowdhurya@ornl.gov

Joshua Batson
Anthropic
San Francisco, CA
joshb@anthropic.com

Petros Drineas
Purdue University
West Lafayette, IN
pdrineas@purdue.edu

Eleftherios Garyfallidis
Indiana University Bloomington
Bloomington, IN
elef@iu.edu

Acknowledgments

SF and EG were supported by the National Institute of Biomedical Imaging And Bioengineering (NIBIB) of the National Institutes of Health (NIH) under Award Numbers R01EB027585 and R01EB017230. PD and AC were partially supported by NSF grants CCF-2209509, CCF-1814041, DMS-1760353, and DOE grant DE-SC0022085. AC was also partially supported by the U.S. Department of Energy, through the Office of Advanced Scientific Computing Research’s “Data-Driven Decision Control for Complex Systems (DnC2S)” project, FWP ERKJ368. Oak Ridge National Laboratory is operated by UT-Battelle LLC for the U.S. Department of Energy under contract number DEAC05-00OR22725. This manuscript has been authored in part by UT-Battelle, LLC. The US government retains, and the publisher, by accepting the article for publication, acknowledges that the US government retains, a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to the results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

1. Model Fitting Comparisons

In order to compare the model fitting gains of P2S2 against P2S, an analysis similar to P2S was performed where the goodness of fit for the diffusion tensor model [1] and the diffusion kurtosis [4] model was computed. To do this a k -fold cross-validation approach was used where signal data corresponding to the k^{th} fold were held-out and predicted using

the held-in data from the rest of the folds. An R^2 metric was computed to see changes in the goodness of fit by the procedure described in [6] and implemented in DIPY [3]. This goodness of fit was computed for each voxel of 80^{th} slice in the HCP 7T [7] data after masking the data (17433 voxels were used). The improvement in the R^2 scores over noisy data was shown using strip-plots after subtracting the R^2 values of the noisy data from the R^2 values of both P2S and P2S2. As one can note from the strip-plots shown in

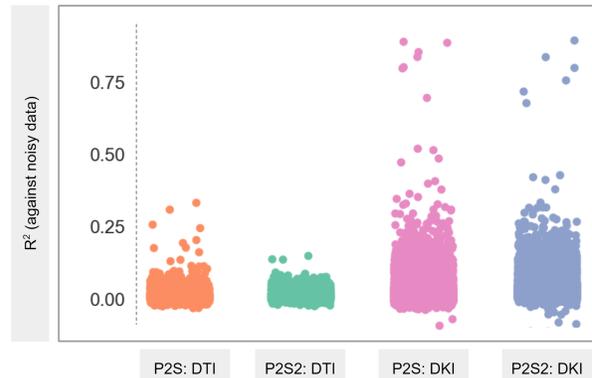


Figure S1: Depicts improvement in the R^2 scores to show a voxel-wise improvement in the goodness of fit via DTI and DKI models both via P2S and P2S2 for each model. The improvement is a subtraction of the R^2 scores obtained from fitting the model to the noisy data from the R^2 scores after fitting to P2S and P2S2 denoised data.

Fig. S1, the model fitting performance improved approximately the same after denoising with P2S2 in comparison with P2S. Some variability in the strip-plots shows is expected but the overall tendency remains the same both for

DTI and DKI model representations. This explains why the DKI metrics shown in the main text result in the same denoising performance and similar noise suppression leading to reduced fitting degeneracies (number of black voxels in the Radial Kurtosis and Mean Kurtosis maps).

2. Residual Maps Comparison

To show that the noise suppressed by P2S and P2S2 are very similar to one another, we show the residual maps for a random slice from a random diffusion weighted volume of the HCP 7T denoised data. Here volume #139 and slice #77 has been used for qualitative comparison. The same data was denoised both by P2S and P2S2 and the residuals were compared by computing root mean squared error for each individual voxel in the data shown in Fig. S2. Note

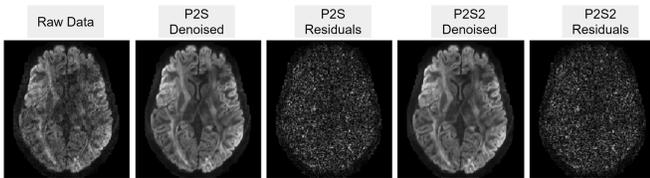


Figure S2: Shows the residual maps computed for after denoising the same subject via P2S and P2S2. Note that both show very similar noise suppression resulting in approximately the same denoised data and residual maps.

that both P2S and P2S2 show very similar performance and do not result in smoothing or suppression of signal relating to anatomy. This can be seen by visually inspecting the residual maps shown in Fig. S2 for structure. If the data was smoothed, anatomical structure would start becoming visible in the residual maps. It is also important to note that P2S2 was denoised by training the denoiser only on 50K of the 61M available training samples, i.e. only 0.083% of the data, and yielded exactly the same denoising performance.

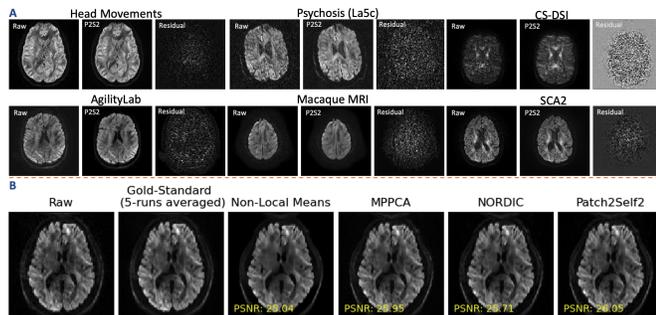


Figure S3: Benchmarking results showing Peak Signal-to-Noise Ratio (PSNR) comparisons across different datasets, illustrating the superior performance of P2S2 over SOTA algorithms NORDIC and MPPCA.

3. Comparison against other baseline methods

To benchmark against unsupervised methods, we used Peak Signal-to-Noise Ratio (PSNR) for comparison, employing the EDDEN dataset. PSNR calculations were based on an average of 5 runs, serving as a standard for clean data. Notably, P2S2 outperformed even in this benchmark, tailored for SOTA algorithms like NORDIC [5] and MPPCA [8]. Additionally, we included 6 datasets (denoised via P2S2 in ≈ 20 seconds) from OpenNeuro, where SOTA algorithms either fail or are inapplicable. Names of the datasets are included in the figure.

4. Code

All comparisons in the paper that have been done against P2S are based on the DIPY implementation [3] [2]. The instructions in the tutorial in the documentation [here: https://dipy.org/documentation/1.4.1./examples_built/denoise_patch2self/#example-denoise-patch2self] were followed to use P2S optimally.

P2S2 has been implemented keeping a similar API to P2S for broad adoption by the community. The anonymous code implemented for matrix sketching [File Path: /Patch2Self2/matrix_sketching] and consequently the Patch2Self2 module [File Path: /Patch2Self2/models/patch2self2.py] are available in the link here: <https://figshare.com/s/078c34809ce3e39a236e>.

To reproduce the results in the paper, jupyter notebooks [File Path: /Patch2Self2/notebooks] have been added to the codebase with the data available in the anonymous link below. A README file has been made available along with the code to explain how to reproduce the results, details of the jupyter notebooks and explanations of the different modules.

NOTE: All experiments in the paper and supplement were performed on a standard i7 CPU with 16 GB RAM Laptop.

5. Data

All the data that is used in the main text and the supplement has been made available in an anonymous link here: <https://figshare.com/s/87f6ffee972510bfda76>. The data do not have any license on usage and is openly available. The names of the data used in the experiments match exactly the examples in the jupyter notebook for reproducibility. The porcine ex-vivo data used was obtained from the link here: https://med.stanford.edu/cmrgroup/data/ex_vivo_dt_mri.html.

6. Licenses

DIPY: The Patch2Self code used for comparison and as the base code for implementing Patch2Self2 has a BSD-3 License available here: <https://github.com/dipy/dipy/blob/master/LICENSE>. This code implicitly also relies on Scikit-learn, Numpy and Scipy for different internal operations. Their licenses are as below:

Scikit-learn (BSD-3): <https://github.com/scikit-learn/scikit-learn/blob/main/COPYING>

Numpy (BSD-3): <https://github.com/numpy/numpy/blob/main/LICENSE.txt>

Scipy (BSD-3): <https://github.com/scipy/scipy/blob/master/LICENSE.txt>

The matrix sketching code used for implementing Patch2Self2 has been adopted and modified from PyRLA which has MIT License available here: <https://github.com/wangshusen/PyRLA/blob/master/LICENSE>.

References

- [1] P.j. Basser, J. Mattiello, and D. LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1): 259–267, 1994. [1](#)
- [2] Shreyas Fadnavis, Joshua Batson, and Eleftherios Garyfallidis. Patch2self: Denoising diffusion MRI with self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 16293–16303. Curran Associates, Inc., 2020. [2](#)
- [3] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian and Nimmo-Smith. Dipy, a library for the analysis of diffusion MRI data. *Frontiers in Neuroinformatics*, 8, 2014. [1](#), [2](#)
- [4] Jens H. Jensen and Joseph A. Helpert. MRI quantification of non-gaussian water diffusion by kurtosis analysis. *NMR in Biomedicine*, 23(7):698–710, 2010. [1](#)
- [5] Steen Moeller, Pramod Pisharady Kumar, Jesper Andersson, Mehmet Akcakaya, Noam Harel, Ruoyun Ma, Xiaoping Wu, Essa Yacoub, Christophe Lenglet, and Kamil Ugurbil. Diffusion imaging in the post hcp era. *Journal of Magnetic Resonance Imaging*, 2020. [2](#)
- [6] Ariel Rokem, Jason D. Yeatman, Franco Pestilli, Kendrick N. Kay, Aviv Mezer, Stefan van der Walt, and Brian A. Wandell. Evaluating the accuracy of diffusion MRI models in white matter. *PLOS ONE*, 10(4):e0123272, 2015. [1](#)
- [7] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012. [1](#)
- [8] Jelle Veraart, Dmitry S. Novikov, Daan Christiaens, Benjamin Ades-Aron, Jan Sijbers, and Els Fieremans. Denoising of dif-

fusion MRI using random matrix theory. *NeuroImage*, 142: 394–406, 2016. [2](#)