# Robert J. Webber

*Research Statement*

## Executive summary

In the 21st century, we have witnessed increases in data availability and computational resources, yet algorithms for data analysis remain painfully slow. Chemists struggle to find clusters in data sets with $N \geq 10^5$ molecular configurations, because spectral clustering would require factorizing an $N \times N$ matrix that is too large to even store on a computer. Meanwhile, geophysicists struggle to simulate Category 5 hurricanes reaching Boston, because generating high-resolution simulations would require thousands of computers operating for years. Motivated by such seemingly impossible computations, my role as an applied mathematician is to build new *randomized* algorithms that offer speed-up factors of at least $10\times$–$100\times$ and support these algorithms by proving rigorous guarantees.

Randomized algorithms accelerate computations by creating a balance between *exploration* of possible solutions and *exploitation* of information regarding the optimal solution. Randomization has achieved success in numerical linear algebra [MT20] and optimization [CBS14]. Yet, many problems are still solved using deterministic algorithms, which can only exploit information without any exploration. The time is ripe to deploy randomness to speed up computations further.

As my first research program, I am using randomized algorithms to compute low-rank approximations of kernel matrices which arise in machine learning. Traditionally, kernel machine learning has been challenging for data sets with $N \geq 10^5$ points because the methods require $\mathcal{O}(N^3)$ floating point operations. However, I am investigating a randomized strategy that reduces the cost from $\mathcal{O}(N^3)$ to $\mathcal{O}(k^2N)$ operations where $k$ is the kernel approximation rank, which can be set orders of magnitude smaller than $N$ [Che+23; Día+23]. Before it was common to run spectral clustering codes for over a day on high-performance workstations [Roh+11]. In contrast, my randomized low-rank approach accurately clusters $N = 2.5 \times 10^5$ data points **in just 10 seconds on a laptop** [Che+23]. To support this approach, I have proved error bounds that quantify the relationship between approximation accuracy and kernel matrix eigenvalue decay [TW23; Che+23].

Second, I am developing a rare event sampling algorithm inspired by the survival of the fittest. The algorithm uses just a small ensemble of simulations. To assess the probability of a rare event, the algorithm duplicates the "fittest" simulations (deemed likely to lead to the rare event) and randomly eliminates the "least fit" simulations (deemed unlikely to lead to the rare event). The algorithm leads to a **7× speed-up** when simulating intense hurricanes near Boston [Web+19] and a **30× speed-up** when simulating Mercury's close encounters with Venus [Abb+21]. I have proved the algorithm produces unbiased estimates [Web+19] and am working to establish optimality. This research has been featured in Forbes.com, the California Business Journal, and SIAM News.

Below, I describe my research programs in more detail and state my overall research vision.

## Program 1: Accelerate kernel machine learning

Kernel methods are a popular set of algorithms for prediction and clustering. Kernel methods perform *interpretable AI*, since they use an explicit kernel function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to quantify the similarity between points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. The strong kernel hypothesis posits that kernel

*Robert J. Webber*
*rwebber@caltech.edu* • *rwebber.people.caltech.edu*

methods can compete with machine learning algorithms based on neural networks [Rad+22; Lee+20; Aro+20], but it is currently impossible to certify or refute this hypothesis because kernel methods are too costly to be applied to the largest data sets. My vision is to resolve the strong kernel hypothesis by building the computational tools needed to deploy kernel methods at scale.

The computational bottleneck in a kernel method is a linear system or eigenvalue problem involving the kernel matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ with entries $k_{ij} = \phi(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Storing the kernel matrix is difficult with $N \geq 10^5$ data points, and traditional matrix inversion or eigendecomposition requires an exorbitant $\mathcal{O}(N^3)$ perations. As a cheaper alternative, I have designed algorithms that replace the matrix $\boldsymbol{K}$ with a randomized rank-$k$ approximation $\hat{\boldsymbol{K}}$, reducing the cost to $\mathcal{O}(k^2 N)$ operations. Figure 1 shows that a rank $k = 150$ leads to nearly perfect kernel spectral clustering when applied to a biochemistry data set with $N = 2.5 \times 10^5$ points. For this example, the difference between $N^3$ and $Nk^2$ is **a factor of** $3$ **million**, and the method runs in ten seconds on a laptop [Che+23].

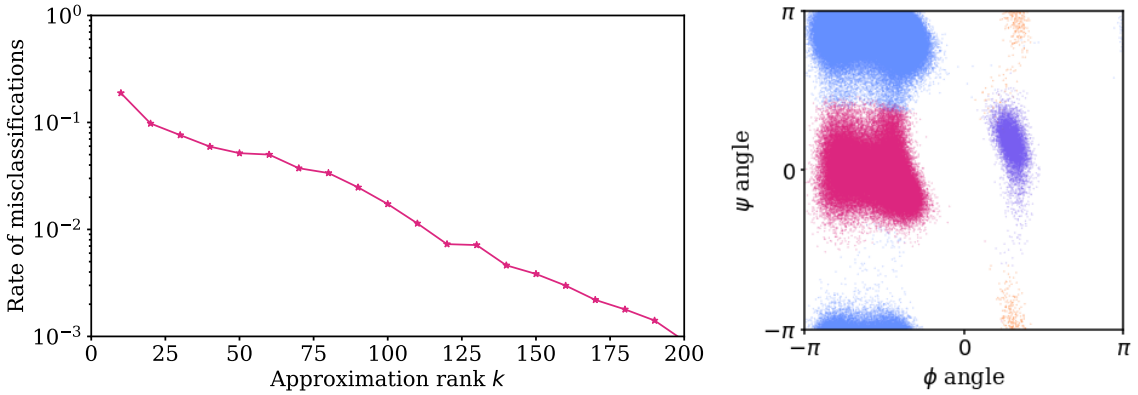


Figure 1: Randomized kernel matrix approximation leads to near-perfect clustering of chemistry data (left), revealing 2-dimensional clusters in a 30-dimensional data space (right). See [Che+23].

To support the randomized low-rank approach, I have proved error bounds for "randomized block Krylov iteration", which produces an accurate low-rank approximation at a modest cost, as well as "randomly pivoted Cholesky", which produces a very low-cost approximation that is accurate if the kernel eigenvalues decay rapidly. Let $\lfloor \boldsymbol{K} \rfloor_r$ denote the theoretically optimal rank-$r$ approximation of a positive semidefinite matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$, from an $r$-truncated eigendecomposition. Randomized block Krylov iteration generates a low-rank approximation $\hat{\boldsymbol{K}}$ that satisfies

$$\mathbb{E}\,\|\boldsymbol{K} - \hat{\boldsymbol{K}}\| \leq (1+\varepsilon)\,\|\boldsymbol{K} - \lfloor \boldsymbol{K} \rfloor_r\|, \quad \text{for} \quad r \geq 1, \quad \varepsilon > 0,$$

using an approximation rank $k = \tilde{\mathcal{O}}(r\varepsilon^{-1/2})$, where the $\tilde{\mathcal{O}}$ notation suppresses logarithmic factors [TW23, Thm. 7.1]. Similarly, randomly pivoted Cholesky satisfies

$$\mathbb{E}\,\mathrm{tr}(\boldsymbol{K} - \hat{\boldsymbol{K}}) \leq (1+\varepsilon)\,\mathrm{tr}(\boldsymbol{K} - \lfloor \boldsymbol{K} \rfloor_r), \quad \text{for} \quad r \geq 1, \quad \varepsilon > 0,$$

using a larger approximation rank $k = \tilde{\mathcal{O}}(r\varepsilon^{-1})$ [Che+23, Thm. 3.1]. Bounds for randomized block Krylov iteration were known before my work [MM15], but I improved these bounds to be numerically accurate and contain explicit constants. My bounds for randomly pivoted Cholesky are novel.

**Next steps.** The efficiency of kernel machine learning can potentially be improved even further, to just $\mathcal{O}(k^3 + kN)$ operations and $\mathcal{O}(k^2 + N)$ storage. Such a low cost is possible because randomized low-rank approximation can be implemented in as few as $\mathcal{O}(k^3 + kN)$ operations with rejection sampling [EM23] and memory requirements can be minimized with stochastic gradient descent [MB17; MB19; ABP23]. Over the next two years, I will combine these ingredients to form an accelerated strategy with error bounds building on my earlier techniques [Che+23; Día+23].

*Robert J. Webber*
*rwebber@caltech.edu* • *rwebber.people.caltech.edu*

## Program 2: Accelerate rare event probability calculation

Extreme weather events, such as hurricanes, floods, and heat waves, are traditionally studied using repeated simulations from a high-resolution model. Yet, a rare event with probability $p$ occurs just once per $p^{-1}$ simulations. Generating enough samples of the rare event might require a thousand or ten thousand simulations, and each high-resolution simulation can require weeks of computing.

To prevent these exploding computing costs, I introduced a rare event sampling algorithm that makes replicas of the "fittest" simulations that are leading to a rare event, while randomly eliminating many of the "least fit" simulations [Web+19]. I proved the algorithm yields unbiased estimates and established a variance formula to enable *a posteriori* error estimation [Abb+21]. Figure 2 shows an example in which the algorithm (right) produces several samples of Category 5 hurricanes near Boston (probability $p = 0.002$), while direct simulation (left) produces none.

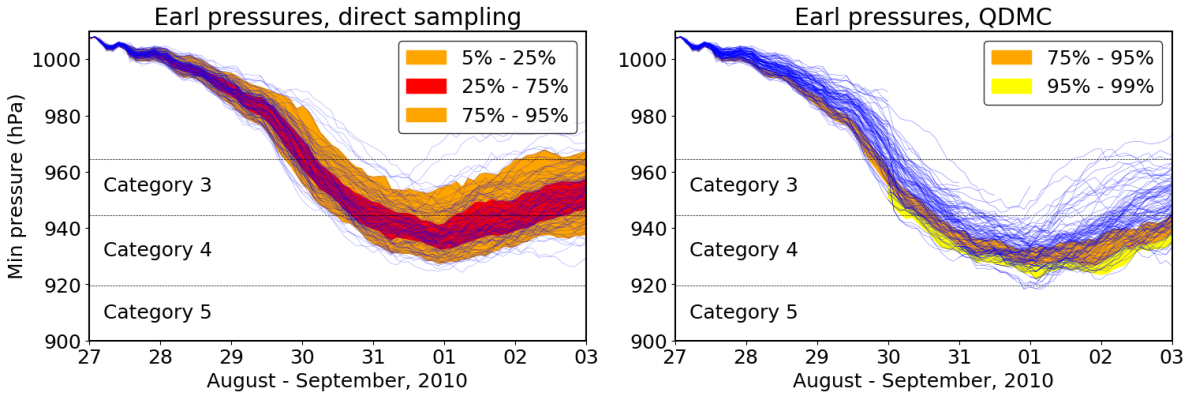


Figure 2: Survival-of-the-fittest algorithm (right) produces several Category 5 storms; direct sampling (left) produces none. See [Web+19].

**Next steps**. I am advising a team of planetary scientists that is using the survival-of-the-fittest algorithm to evaluate Mercury's probability of experiencing a close encounter with Venus [Abb+21; Abb+23a; Abb+23b]. Meanwhile, I am developing a theory to explain the optimal variance that any Monte Carlo rare event sampling scheme can achieve [WAS22; Ari+23], and I plan to extend this theory to establish an optimality guarantee for the survival-of-the-fittest algorithm.

## Additional future research directions

As an example of a future research direction, I will use *iterative random sparsification* to accelerate large-scale fixed-point calculations in physics and chemistry, such as the power method for computing ground states of molecules [SO17]. In this approach, I replace a convergent but computationally expensive fixed-point iteration $\boldsymbol{x}_{t+1} = \boldsymbol{F}(\boldsymbol{x}_t)$ with a cheaper iteration $\boldsymbol{x}_{t+1} = \boldsymbol{F}(\boldsymbol{y}_t)$, where $\boldsymbol{y}_t = \text{sparse}(\boldsymbol{x}_t)$ is a sparse randomized approximation. Related techniques have been used to solve eigenvalue problems as large as $10^{108} \times 10^{108}$ in quantum chemistry [She+12], as well as *infinite-dimensional* linear systems in PDE analysis [CDD01], but without guaranteed accuracy. By studying these techniques mathematically, I will identify opportunities to apply randomness in a principled way, leading to faster and more reliable calculations.

I am also excited to investigate *active learning* algorithms, which identify a small representative subset $\boldsymbol{x}_{s_1}, \ldots, \boldsymbol{x}_{s_k}$ of a large data set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. The selected data points serve as "sensors" for interpolating functions across the complete data set. The most powerful active learning algorithms [Mus+22] currently require expensive pre-processing steps, involving a complete pass

*Robert J. Webber*
*rwebber@caltech.edu* • *rwebber.people.caltech.edu*

over the data. Yet I perceive the possibility of a high-performing *online active learning* algorithm, where the user is presented with random data points and accepts or rejects each data point with probability depending on the previous selections. An early version of this approach was suggested in [EM23], but I see the potential to take many fewer samples while achieving higher accuracy.

## Conclusion: My research vision

In the 21st century, we are transitioning from deterministic algorithms to modern randomized algorithms, which combine exploration and exploitation to optimize computational efficiency. During this transition, new insights are needed to ensure the accuracy of the randomized algorithms. My research has already shown how randomized algorithms exploit low-rank structure [TW23; Che+23] and sparse structure [WW23] to provide fast and accurate solvers. My vision is to expand this library of exploitable structures further, making randomized methods so well-understood and effective that they are taught in every graduate math and science curriculum. In support of my vision, I will develop new randomized algorithms which address the needs of scientists struggling with large-scale computations. Meanwhile, I will benefit from interacting with scientists, by using their computational challenges to prompt new developments in linear algebra and probability theory.

## References


[Abb+23a] Dorian S. Abbot, David M. Hernandez, Sam Hadden, Robert J. Webber, Georgios P. Afentakis, and Jonathan Weare. "Simple physics and integrators accurately reproduce Mercury instability statistics". In: *The Astrophysical Journal* 944.2 (2023), p. 190. DOI: 10.3847/1538-4357/acb6ff.

[Abb+21] Dorian S. Abbot, Robert J. Webber, Sam Hadden, Darryl Seligman, and Jonathan Weare. "Rare event sampling improves Mercury instability statistics". In: *The Astrophysical Journal* 923.2 (2021), p. 236. DOI: 10.3847/1538-4357/ac2fa8.

[Abb+23b] Dorian S. Abbot, Robert J. Webber, David M. Hernandez, Sam Hadden, and Jonathan Weare. *Mercury's chaotic secular evolution as a subdiffusive process*. 2023. arXiv: 2306.11870 [astro-ph.EP].

[ABP23] Amirhesam Abedsoltan, Mikhail Belkin, and Parthe Pandit. "Toward Large Kernel Models". In: *Proceedings of the 40th International Conference on Machine Learning*. 2023. URL: https://proceedings.mlr.press/v202/abedsoltan23a.html.

[Ari+23] D. Aristoff, J. Copperman, G. Simpson, R. J. Webber, and D. M. Zuckerman. "Weighted ensemble: Recent mathematical developments". In: *The Journal of Chemical Physics* 158.1 (2023), p. 014108. DOI: 10.1063/5.0110873.

[Aro+20] Sanjeev Arora, Simon S. Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. "Harnessing the power of infinitely wide deep nets on small-data tasks". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rkl8sJBYvH.

[CBS14] Volkan Cevher, Stephen Becker, and Mark Schmidt. "Convex Optimization for Big Data: Scalable, randomized, and parallel algorithms for big data analytics". In: *IEEE Signal Processing Magazine* 31.5 (2014), pp. 32–43. DOI: 10.1109/MSP.2014.2329397.

[Che+23] Yifan Chen, Ethan N. Epperly, Joel A. Tropp, and Robert J. Webber. *Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations*. 2023. arXiv: 2207.06503 [math.NA].

[CDD01] Albert Cohen, Wolfgang Dahmen, and Ronald Devore. "Adaptive Wavelet Methods for Elliptic Operator Equations: Convergence Rates". In: *Mathematics of Computation* 70.233 (2001), pp. 27–75. URL: http://www.jstor.org/stable/2698924.



*Robert J. Webber*

*rwebber@caltech.edu* • *rwebber.people.caltech.edu*

[Día+23] Mateo Díaz, Ethan N. Epperly, Zachary Frangella, Joel A. Tropp, and Robert J. Webber. *Robust, randomized preconditioning for kernel ridge regression*. 2023. arXiv: 2304.12465 [math.NA].

[EM23] Ethan N. Epperly and Elvira Moreno. *Kernel quadrature with randomly pivoted Cholesky*. 2023. arXiv: 2306.03955 [math.NA].

[Lee+20] Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. "Finite versus infinite neural networks: An empirical study". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020. URL: https://dl.acm.org/doi/10.5555/3495724.3496995.

[MB17] Siyuan Ma and Mikhail Belkin. "Diving into the shallows: A computational perspective on large-scale shallow learning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017. URL: https://dl.acm.org/doi/10.5555/3294996.3295135.

[MB19] Siyuan Ma and Mikhail Belkin. *Kernel machines that adapt to GPUs for effective large batch training*. 2019. arXiv: 1806.06144 [stat.ML].

[MT20] Per-Gunnar Martinsson and Joel A. Tropp. "Randomized numerical linear algebra: Foundations and algorithms". In: *Acta Numerica* 29 (2020), pp. 403–572. DOI: 10.1017/S0962492920000021.

[MM15] Cameron Musco and Christopher Musco. "Randomized block Krylov methods for stronger and faster approximate singular value decomposition". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 2015. URL: https://dl.acm.org/doi/10.5555/2969239.2969395.

[Mus+22] Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. "Active Linear Regression for $\ell_p$ Norms and Beyond". In: *IEEE 63rd Annual Symposium on Foundations of Computer Science*. 2022. DOI: 10.1109/FOCS54457.2022.00076.

[Rad+22] Adityanarayanan Radhakrishnan, George Stefanakis, Mikhail Belkin, and Caroline Uhler. "Simple, fast, and flexible framework for matrix completion with infinite width neural networks". In: *Proceedings of the National Academy of Sciences* 119.16 (2022), e2115064119. DOI: 10.1073/pnas.2115064119.

[Roh+11] Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. "Determination of reaction coordinates via locally scaled diffusion map". In: *The Journal of Chemical Physics* 134.12 (2011), p. 124116. DOI: 10.1063/1.3569857.

[She+12] James J. Shepherd, George Booth, Andreas Grüneis, and Ali Alavi. "Full configuration interaction perspective on the homogeneous electron gas". In: *Physical Review B* 85 (8 2012), p. 081103. DOI: 10.1103/PhysRevB.85.081103.

[SO17] Attila Szabo and Neil S Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mineola, NY: Dover Publications, 2017.

[TW23] Joel A. Tropp and Robert J. Webber. *Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications*. 2023. arXiv: 2306.12418 [math.NA].

[WW23] Jonathan Weare and Robert J. Webber. *Randomly sparsified Richardson iteration is really fast*. 2023. arXiv: 2309.17270 [math.NA].

[WAS22] Robert J. Webber, David Aristoff, and Gideon Simpson. *A splitting method to reduce MCMC variance*. 2022. arXiv: 2011.13899 [math.NA].

[Web+19] Robert J. Webber, David A. Plotkin, Morgan E O'Neill, Dorian S. Abbot, and Jonathan Weare. "Practical rare event sampling for extreme mesoscale weather". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.5 (2019). DOI: 10.1063/1.5081461.

*Robert J. Webber*
*rwebber@caltech.edu* • *rwebber.people.caltech.edu*