# Research Statement

[redacted]

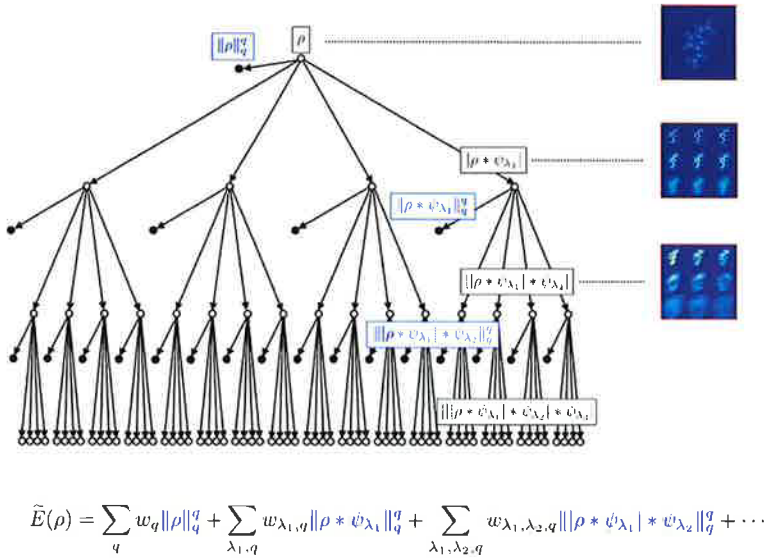
November 7, 2017

My research program is rooted in computational harmonic analysis, but spans a range of topics from mathematical foundations of data science to cutting edge research in data driven quantum chemistry and materials science. This program, in my view, is a microcosm for the Department of Computational Mathematics, Science & Engineering (CMSE), and serves as a bridge between the data science and scientific computation wings of the department. It has, thus far, been broadly appealing to students and postdocs as well, as my group (the CEDAR Team, for **C**ompl**E**x **D**ata **A**nalysis **R**esearch) currently has seven members: two postdocs supported by myself, three PhD students that I advise, one first year PhD student with no official advisor that I mentor, and another PhD student whom I do not advise, but with whom I work extensively. Additionally, the team is scientifically diverse, and includes one pure mathematician, two computational mathematicians, one computational statistician, one computational physicist, one computational bio-chemist, and an environmental economist.

Everyone in the CEDAR team, including myself, straddles the line between developing novel mathematical theory (primarily a mix of harmonic analysis, geometry, statistics) and state-of-the-art results in a particular scientific domain (applied computer science, biology, chemistry, materials science, economics). Each member is encouraged to find his or her own balance, but is required to excel in both regimes. By the time I come up for tenure, I expect that the research output of my team will mimic or even surpass my existing distribution of works, in which I have published articles in a variety of journals ranging from pure mathematics (*Linear Algebra and Its Applications*, *Mathematische Annalen*, *Revista Matemática Iberoamericana*, *Proceedings of the American Mathematical Society*), applied mathematics (*Applied and Computational Harmonic Analysis*, *Multiscale Modeling and Simulation*), electrical engineering (*Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*), computer science (*ICML Workshop on Computational Biology*, *Neural Information Processing Systems*), and papers soon to be submitted in biology (*Science*) and physics (*Journal of Chemical Physics*).

The core of my research is in developing mathematically provable machine learning algorithms to circumvent prohibitively expensive computations in scientific computing, thereby opening new avenues for scientific breakthroughs. Modern scientific advances in peta and exascale computing and high throughput technologies are enabling for the first time true scale bridging research in numerous computational fields such as quantum chemistry, materials science, computational biology, astrophysics, climate science, and fluid mechanics. These advances have led to massive amounts of high dimensional distributed data, and interpreting and analyzing this data is a fundamental problem facing science. Machine learning is one approach to extract information from these data sets, but new paradigms are needed that merge multiscale models with flexible learning architectures. In the sections below, I detail my efforts along three research avenues: multilayer learning and many body physics; geometric methods for biomedical data analysis; and the synthesis of interpolation theory and statistical learning.

**Multiscale, multilayer learning and many body physics** $N$-body and many body simulations are fundamental to numerous computational fields, including astrophysics, materials science, and computational chemistry. However, their scope is limited by the scale of the system (simulations of large swaths of the universe), the complexity of the interactions (modeling quantum effects), and the sheer number of systems (there are approximately $10^{60}$ small organic compounds). In order to push the boundary of what is capable in these simulations, my work develops new multiscale, physically based, machine learning algorithms that learn accurate physical models from known exemplars, and efficiently evaluate these models at new states.



$$\widetilde{E}(\rho) = \sum_q w_q \|\rho\|_q^q + \sum_{\lambda_1,q} w_{\lambda_1,q} \|\rho * \psi_{\lambda_1}\|_q^q + \sum_{\lambda_1,\lambda_2,q} w_{\lambda_1,\lambda_2,q} \||\rho * \psi_{\lambda_1}| * \psi_{\lambda_2}\|_q^q + \cdots$$

Figure 1: The scattering transform filtering a molecular density $\rho$, and recombining multiscale, multilayer features (in blue) to estimate the energy of the molecule.

[redacted] [*Multiscale Model. Simul., 2017*] introduced the wavelet scattering transform (Figure 1) for the regression of potential energies of $N$-body and many body systems. The scattering transform can be thought of as the machine learning version of a fast multipole method (FMM). Indeed, it reduces $O(N^\alpha)$ calculations ($\alpha \geq 2$) to $O(N \log N)$. Unlike FMMs, which analytically expand known physical kernels, the scattering transform learns the kernel from given physical states. In this work, as well as subsequent work with [redacted] [*NIPS, to appear*], I obtained state of the art numerical results on small organic molecules, achieving errors on the order of DFT algorithms at a fraction of the cost.

The scattering transform is one of the first machine learning algorithms to be fully adapted to the global physical properties of the potential energy: it is invariant to global isometries of the system, and Lipschitz stable to diffeomorphisms of the relative positions of the bodies. The method leverages an architecture similar to convolutional networks, but unlike such networks, each layer is a predefined wavelet modulus operator over the isometry group in $\mathbb{R}^3$. This structure ensures the learned model has the correct physical properties, and that the algorithm can learn an accurate model from few examples. Each path in the network encodes complex interactions that are coupled across a sequence of scales.

I proved (jointly with [redacted]) that pairwise potential energies of the form $U(\rho) = \iint \rho(x)\rho(y)V(x-y)\,dx\,dy$ can be regressed to $O(\epsilon)$ precision using $O(|\log \epsilon|^2)$ scattering paths, formalizing the earlier statement that scattering transforms learn physical kernels. However, in quantum chemistry, the potential energy also includes kinetic and exchange-correlation energies; the theoretical learning capability of scattering transforms for these energies will be investigated over the next three years. Relatedly, [redacted] and I are pursuing a conjecture that deeper scattering paths

encode information analogous to multipole moments. This result will facilitate local adaptability in the algorithm, enabling use of this work in large-scale astrophysics (e.g., by improving gravitational field evaluations).

Additionally, my group and I are pushing the envelope in materials science. Together with MSU Prof. [redacted] we aim to simulate the voltage and lithium transport in cathode materials with defects, which are currently beyond the capabilities of DFT. Balachandran and I will pursue work studying Perovskite structures, including the feasibility of learning properties related to complex defects, such as those arising from oxygen vacancies. This in turn could lead to the development of new computer chips with memory, that are able to resume computations in the event of power loss or intermittent power sources.

**Multiscale geometric methods for biological data analysis** In two separate papers ([*Appl. Comput. Harmon. Anal., 2014*] and [*Appl. Comput. Harmon. Anal., to appear*]), I developed with [redacted] diffusion based manifold learning (ML) for dynamic data sampled from a Riemannian manifold with a smooth family of metrics. Diffusion ML algorithms learn hierarchical organizations of data sampled from a manifold. We proved that one can learn from a finite data set a time inhomogeneous Markov chain that in the limit of infinite data converges to the heat kernel of $\partial_t u = \Delta_{g(t)} u$, thus extending seminal results on static manifolds (Belkin and Niyogi, [*Neural Comput.*]; [redacted], [*Appl. Comput. Harmon. Anal.*]).

Continuing theoretical work will incorporate: robust estimation of high order statistics in unstructured data, and geometric organization of non-manifold data (e.g., data sampled from metric spaces). Initial work in the latter direction considers how to learn metric tree structures from high dimensional sampled data. Regarding the former thrust, with [redacted] I am developing algorithms to learn data driven invariants, while extracting informative statistics, through unstructured versions of the scattering network described previously.

This research is driven by computational biology. In [*ICML WCB, 2016*], [redacted] and I leverage my work with [redacted] to devise a new diffusion based condensation process, which guides the Riemannian metric to find multiscale population structures in single cell data. With [redacted] and others, I am using the metric tree work to provide a low dimensional visualization of biological progressions in high-dimensional data (Figure 2), which is beyond the scope of existing nonlinear dimension reducing methods. The algorithm has been applied to a wide variety of big biological datasets including single-cell RNA sequencing and CyTOF data, where it reveals progression-forming variables (e.g., specific genes) and paths between developmental events in cellular state-space.

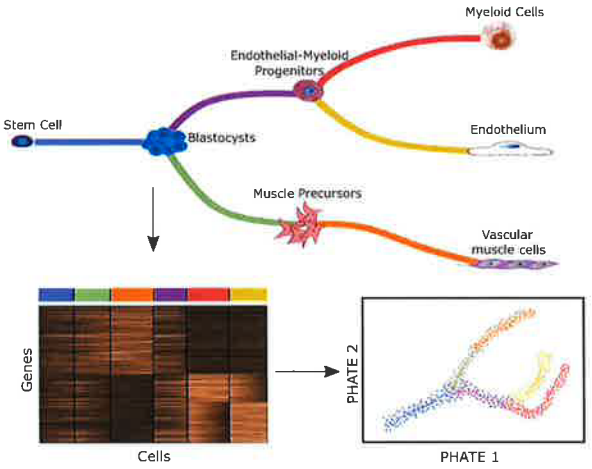



Figure 2: Geometric learning of single cell progressions.

Over the next three years, I will facilitate interpretation of high dimensional, high throughput biological data by continuing to develop tools and novel algorithms that allow biologists to extract meaningful and predictive information from the data. My collaborators and I aim to revolutionize

the analysis of single-cell data by revealing gene-gene interactions, differentiation trajectories, gene pathways, as well as progressions and multiscale cluster structures of cells and genes, all through a unified geometric framework.

**Smooth interpolation and statistical learning theory** Many learning algorithms, fundamentally, interpolate a function. I proved with [redacted] the existence of general Quasi Absolutely Minimal Lipschitz Extensions (quasi-AMLEs) [*Math. Ann., 2014*]. The study of AMLEs has a rich history and is linked to probability, PDEs, and computer vision. [redacted] and I developed an efficient algorithm for computing optimal interpolating functions (related to AMLEs) for the space $C^{1,1}(\mathbb{R}^d)$ [*Rev. Mat. Iberoam., 2017*]. Amongst Whitney type interpolation algorithms (e.g., Fefferman and Klartag, [*Ann. of Math.*]), this algorithm is the first to provably compute in $O(n \log n)$ time ($n$ is the number of interpolation points) the order of magnitude of the best Whitney constant to within a dimensionless factor. An interpolant with the same norm is computed by replacing Calderon-Zygmund decompositions with an intricate partition due to Wells [*J. Diff. Geom.*], which we prove can be reduced to the computation of a convex hull (Figure 3).

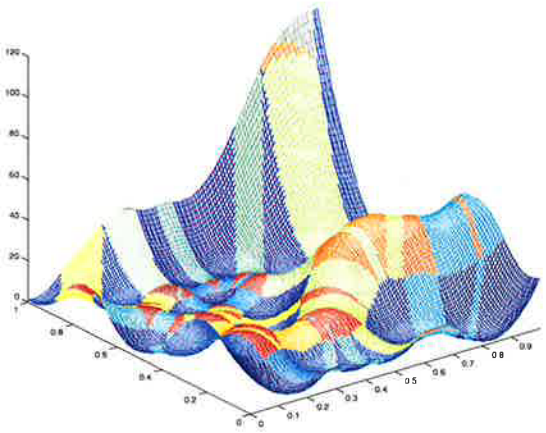


Figure 3: Computation of the optimal interpolant, colored by the intricate patchwork of local interpolants.

With [redacted] I am pushing this work into statistical learning theory. Optimizing the empirical risk over $C^{1,1}(\mathbb{R}^d)$, we obtain uniform bounds relating the empirical risk to the true risk, and show that the excessive risk is bounded by a given $\epsilon > 0$ with probability $1 - \delta$. This result, combined with the dimensionless factor result described previously, yields for the first time a practical means to incorporate smooth function classes into supervised machine learning. To continue pushing this synthesis of modern analysis and machine learning, I (with Le Gruyer) have partial results on the best Whitney constant for $C^{1,1}(\mathbb{R}^d, \mathbb{R}^m)$; I am also studying the same problem for $C^{2,1}(\mathbb{R}^d)$. Positive results along these lines, in addition to their theoretical impact, will yield further algorithmic developments. I also continue to study questions related to AMLEs, such as uniqueness and whether quasi-AMLEs can lead to AMLEs.

**Expansion of my research program** As one may infer from this research statement, I prefer a broad research program that ties together multiple fields. This approach forces one to uncover unifying principles when present, and to make novel connections otherwise. In this spirit, over the next reporting period I will augment my existing research program by pursuing three new directions, which leverage my existing strengths while significantly broadening its scope. These three directions will be: (1) multiscale learning for inverse problems, PDEs, and optimal transport; (2) mathematical foundations of geometric deep learning; and (3) multiscale geometric learning in neuroscience. In addition to maintaining funding in the mathematics of machine learning and machine learning for many body physics (NSF DMS, DARPA), I expect to supplement the latter with grants directed towards materials research (NSF DMR), while finding new sources of funding in the biomedical and/or neuroscience fields (NIH) and in data driven inverse problems/PDEs (ONR).

# Teaching Statement

[redacted]

November 7, 2017

During my tenure at MSU I have taught three courses, one at each level of instruction: (1) CMSE 201: Introduction to Computational Science (lower undergraduate); (2) Math 414: Linear Algebra II (upper undergraduate); (3) CMSE 820: Mathematical Foundations of Data Science (graduate qualifying course). In the spring of 2018, I will teach a topics level graduate course (Math 994) on computational harmonic analysis, further broadening my teaching experience. Of these courses, CMSE 201, CMSE 820, and Math 994 are all *new* courses, and thus I will have developed three new courses in my first three years at MSU. These course developments are significant endeavors, and have required an extensive amount of time beyond the teaching of an existing course.

While each course is unique and different levels of instruction require different specific approaches, I am a firm believer in pushing students to their limit, while coming to acknowledge that at certain junctures a light touch is required in order to avoid disenfranchisement. In Math 414, which was the first course I taught at MSU, I learned this lesson first hand. I taught the course in the standard lecture format, and assigned fairly difficult homework exercises so as to sharpen the students' skills and deepen their knowledge of the subject. Exams were in class, closed book tests, that required mastery of the rudiments of linear algebra. The result was a rich but difficult course, which in my estimation (gleaned from course grades and feedback), left too many students behind.

In subsequent courses I have strived to find a better balance, while incorporating more innovative teaching methods. CMSE 201, the next course I taught at MSU, was designed as a flipped class. The spring 2016 semester this was the first time this course was taught, and two sections ran concurrently, one taught by myself and one by Prof. [redacted] (CMSE/Physics). While Prof. [redacted] and a teaching specialist, Dr. [redacted] undertook the majority of course development, I was free to teach the course in my own fashion. As a flipped class, students do the majority of reading and standard lecture content before class (via videos produced by Prof. [redacted] and [redacted]), and worked in teams of two or four in class on various projects in scientific computing. In each class period I spent 5-10 minutes at the beginning of class reviewing the previous night's material, which could include particularly tricky mathematical derivations, or real time coding projected onto a screen at the front of class. The students then settled into their groups and worked on the proposed project for that class period, while I and the teaching assistant for the course ([redacted] went from group to group discussing the project and answering any questions. At various junctures in each class period I paused the groups to facilitate class wide discussion, and at the end of each class we had a final class dialogue going over

the project, trying to ascertain the key takeaways. One of the most important aspects of this course is the students' growth as team members, and the development of trust amongst the students not only within the groups, but across the whole class. In a flipped course such as CMSE 201, the course takes on a conversational tone, and thus for the course to fully realize its potential all students must feel confident, and safe, expressing their opinions in class. As such, in the first weeks of the course I helped lead the students in developing a type of "bylaws" for the course, in which they agreed upon acceptable behavior while in class, and pledged to maintain an open atmosphere.

The most recent course I taught at MSU was CMSE 820, which like CMSE 201, ran for the first time under my direction. CMSE 820 is a qualifying exam course, and focuses on the mathematical foundations of data science. This topic is currently of great interest, and in the emerging era of "big data" promises to be an important component of the core of computational science. Since the course serves as preparation for the qualifying exam, I returned to the lecture format but maintained the conversational component. As such, students in course volunteered numerous questions per class, poking and prodding at the various topics. Often these discussions would carry over beyond class; I estimate that on average I spent 30 minutes after each class conversing with a subset of the students, and hopefully enriching the course for them. Homeworks were a mixture of mathematical exercises (proofs of theory), and programming assignments illustrating the theory in practice. As part of my duties, I wrote and graded two qualifying exams for CMSE 820 (one in May, one in August), and organized a weekly summer study session for graduate students, in which I supplied them with example exams that I wrote each week. Students also undertook a course project in the later stages of the semester, linking ideas in data science to their own research. Overall, I view the course as a success, and many students have expressed a desire to further incorporate ideas from the course into their long term research plans.

As I continue to grow as an instructor, I will evolve my teaching style by adapting my practices to better meet the needs of my students, while reinforcing my strengths. In order to improve my teaching style, I plan on participating in the teaching essentials workshops at MSU. Additionally, to achieve better balance in my courses, I will request feedback at intermediate stages of the semester, not only at the end of the semester (this was a helpful component in CMSE 201). To facilitate in depth feedback, I will utilize the Academic Advancement Network (AAN) at MSU, which runs mid-semester feedback sessions. For every course I taught, I maintained an active webpage containing papers, book references, and typed notes which can serve as a lasting bank of knowledge for not only the students who took the course, but others interested in the material as well (for Math 414 and CMSE 820, this amounted to over 100 pages for each course). I maintain numerous office hours and invite students to stop by not only to discuss homework, but for general discussion regarding any aspect of the course and even beyond.

In addition to my teaching duties, for the upcoming academic year I have been appointed the chair of the undergraduate studies committee for CMSE. I look forward to this challenge and leadership opportunity, especially as CMSE continues to grow its undergraduate course offerings and develops the data science major.