Research Statement

In the 21st century, mathematics will increasingly be done by artificial intelligence, even though the prevalence right now is low. The current methods we have for generating AI models require the consumption of huge amounts of data. Large data sets are thus helpful for conjecture-formation by both humans and computers. Over the past decade I have worked on one of the largest existing mathematical databases, the L-functions and modular forms database (LMFDB), which contains over 2TB of contents, over 2 billion objects, and required roughly a millennium of computation time to create.

The LMFDB focuses on the Langlands program, which brings together many different areas of mathematics, from number theory and arithmetic geometry to representation theory to complex, p-adic and harmonic analysis. The level of interest in this data can be measured by the engagement of the global mathematical community: the LMFDB has almost 1000 citations, as well as roughly 60,000 users and 500,000 page views per year from 176 countries and all 50 US states.

In addition to my work on mathematical databases, I am also pursuing several projects in p-adic computation. This topic has seen much less attention than numerical computation over the real numbers, and has applications to counting points on varieties over finite fields, finding rational points on curves, computing L-functions, and studying p-adic analytic spaces associated to varieties.

1 Mathematical Databases

There are two kinds of work that go into building an online mathematical database. First, the underlying data must be computed, often using computer algebra systems such as Sage [Sage] or Magma [BCP97]. This process starts with an initial schema describing how the mathematical objects of interest will be represented using basic types such as integers, strings and floating point values. Code for producing this data needs to be structured to run at large scale and to include verifications supporting the reliability and reproducibility of the computations. As an ancillary benefit, running over huge numbers of examples often reveals bugs in the underlying systems which can be reported and fixed, benefiting other users. Moreover, building such datasets often yields interesting mathematical problems of independent interest.

Second, the data must be made available to the user through a website. This process requires enough mathematical understanding to build a helpful presentation and to anticipate how the user may want to search through the data, enough web development background to create a usable interface, and enough facility with databases to connect the webpages to the underlying data. Within the LMFDB, I have worked to streamline this process for others, and I intend to use this experience to facilitate the creation of other mathematical databases outside the LMFDB.

In the sections below, I describe my roles in creating several sections of the LMFDB: finite groups, classical modular forms, modular curves, and abelian varieties over finite fields.

1.1 Finite groups

The database of finite groups in the LMFDB [CJP+23] provides a searchable database of over 500,000 groups [LMF24b] that are small in some way: either with small cardinality (from the SmallGroup database in GAP and Magma), with a small permutation representation (abstract isomorphism classes of groups from the transitive group database), arising as a matrix group in small dimension over a ring such as $\mathbb{Z}$, $\mathbb{F}_q$ or $\mathbb{Z}/N\mathbb{Z}$, or with a short composition series (simple, perfect and almost simple groups). For each group, the subgroup lattice, character table, and other properties are stored, to the extent that these computations are feasible. One of the main improvements that it offers over previous group databases is that it includes subgroup and quotient relationships rather than just information on how to construct each group.

I have been highly involved in this effort over the last three years, working on writing Magma code and structuring the computations to gather group theoretic invariants as completely as possible. I am excited about this database because it offers a qualitative difference with the rest of the LMFDB in terms of its connections with areas of mathematics outside number theory. Because finite groups arise so ubiquitously, there is room for cooperation with researchers studying graph theory, error correcting codes, and quantum computation.

1.2 Classical modular forms

Modular forms have played a central role in number theory over the last several decades, from the proof of Fermat's last theorem and subsequent applications to Diophantine equations to connections with quadratic forms and automorphic representations. For a positive integer k and finite index subgroup $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$, the space of modular forms of weight k and level Γ is finite dimensional. Each has a basis whose Fourier coefficients provide a connection to arithmetic geometry and L-functions. In 2019, I helped build the most extensive existing database of modular forms [BBB+21]. It currently contains 281,965 newforms [LMF24d], corresponding to 14,417,694 modular forms over $\mathbb{C}$. In addition to covering a broader range of weights and levels than previous databases, it made great strides in enumerating weight 1 modular forms, where modular symbol algorithms are not applicable. As part of our effort to ensure the reliability of the data, I designed and implemented a test suite for the database that ran extensive internal and external consistency checks.

1.3 Modular curves

The modular forms database played a key role in the creation of a new database of modular curves [LMF24c]. I helped organize three workshops at MIT [BCE+22a; BCE+22b; BCE+24] to kick off the creation of this database, and then worked to put together contributions from the 46 participants in those workshops and incorporate it into the LMFDB.

Modular curves parameterize elliptic curves together with some extra structure. Specifically, given any open subgroup $H \subseteq \mathrm{GL}_2(\hat{\mathbb{Z}})$, the points on the modular curve X_H parameterize elliptic curves E so that the image of the adelic Galois representation ρ_E is contained within H, up to conjugacy. Each such X_H has three basic invariants:

1. the level N, so that H is equal to the full preimage of its reduction modulo N,
2. the index i of H inside $\mathrm{GL}_2(\hat{\mathbb{Z}})$,
3. the genus g of X_H as a curve.

The modular curve database in the LMFDB contains all curves with level up to 70, models (explicit equations) for many curves of genus up to 24, and rational points coming from the LMFDB's databases of elliptic curves (the Galois images were computed using an algorithm of Zywina [Zyw22] as part of this project).

Work on this database is still ongoing. We aim to include modular curves of higher level, to search for points using the explicit models (hopefully finding new examples of elliptic curves with interesting Galois representations), to add quotient curves under Atkin-Lehner operators, and to add automorphism groups and exceptional isomorphisms. In addition, we hope to connect this database to the extensive literature on modular curves so that it can serve as a dynamic and growing resource for the community.

1.4 Abelian varieties over finite fields

While many interesting questions about elliptic curves remain open, a lot of focus in computational number theory has shifted in recent years to higher genus curves (elliptic curves are genus 1 curves equipped with a base point). One of the central objects of study for a curve of genus g is its Jacobian, defined in terms of formal sums of points on the curve. Jacobians are examples of abelian varieties, higher dimensional analogues of elliptic curves that are simultaneously projective varieties and abelian groups. For $g > 2$, not all abelian varieties are Jacobians; the characterization of Jacobians among all abelian varieties is known as the Schottky problem.

Abelian varieties were originally studied over the complex numbers, but arithmetic geometers prefer to work over number fields and finite fields. Over a finite field $\mathbb{F}_q$, the Honda-Tate theorem [WM71] provides a powerful tool for studying abelian varieties. It classifies them up to isogeny, which is an equivalence relation determined by the presence of a homomorphism with finite kernel, and gives a bijection[1] between isogeny classes over $\mathbb{F}_q$ of dimension g and integer polynomials of degree $2g$ whose roots in $\mathbb{C}$ all have absolute value $\sqrt{q}$. Using this correspondence, Taylor Dupuy, Kiran Kedlaya, Christelle Vincent and I built a database of abelian varieties [DKRV21], including quantities like point counts, endomorphism algebras,

[1] the image needs to be adjusted slightly by requiring that certain irreducible factors occur with multiplicity

twists, primitivity, and angle ranks that are isogeny invariant and can be determined from the corresponding polynomial.

The current database [LMF24a] contains almost 3 million isogeny classes, of dimension up to 6. I have been working in two directions to enhance it. First, Stefano Marseglia and Mckenzie West have joined us and we are making progress on dividing isogeny classes up into isomorphism classes. This process proceeds in two steps, working first with unpolarized abelian varieties and then computing polarizations (a polarization of A is an isomorphism from A to its dual abelian variety). Given an isogeny class corresponding to a polynomial $f(x)$, the polynomial defines an order[2] and, under certain constraints on f, unpolarized isomorphism classes correspond to ideal classes in this order. Polarizations and explicit isogenies can be computed in terms of the same data. We have completed a draft of the resulting dataset, and aim to include it in the LMFDB in the next several months.

The second direction relates to the Schottky problem. When $g > 2$, there is no known method for determining whether a single isogeny class (specified by a polynomial) contains a Jacobian. Instead, various authors have enumerated all curves of genus g over $\mathbb{F}_q$ for g up to 5 and small values of q. I am working with Kiran Kedlaya on using descriptions of genus 6 and 7 curves [Muk95] to carry out a similar project in higher genus. Longer term, enough of this form may make it possible to use machine learning algorithms to search for a pattern in which polynomials correspond to isogeny classes containing a Jacobian.

2 p-adic computation

My interest in p-adic computation began as an undergraduate, with a project that used p-adic cohomology to count points on surfaces over finite fields [AKR09] and a project that studied the Coleman-Mazur eigencurve [CM98] in the 3-adic setting [Roe14]. My contributions developed in a more theoretical direction with my thesis on the local Langlands correspondence for tame unitary groups [Roe11], two papers with Clifton Cunningham on a function-sheaf dictionary for characters of group schemes [CR16; CR21], and a paper with Moshe Adrian on rectifiers in the local Langlands correspondence [AR16].

From a computational point of view, there are two main challenges in implementing p-adic arithmetic: tracking precision through a computation, and handling algebraic extensions, which are far more complicated than for real numbers. I have led the development of p-adic arithmetic in Sage [Sage] for the last 15 years, creating an open source platform that other mathematicians have used in work on p-adic L-functions [PS11], quadratic Chabauty [BD19], Gröbner bases over Tate algebras [CVV19], and many other applications. The implementation includes many methods for tracking precision (one of which is discussed in the next section), as well as the ability to compute in extension fields.

2.1 Tracking precision

Both real and p-adic arithmetic can only be implemented to finite precision on a computer, but error tracking is easier in the p-adic context because p-adic fields are non-Archimedean. For example, arbitrarily many inexact values can be added and the result will have the same precision as the least precise input. Yet once addition and multiplication are mixed, precision loss occurs both in theory and practice. Partly as a consequence of these non-Archimedean advantages, far less work had been done in the p-adic setting than over the real numbers, where numerical methods are commonly studied. In a sequence of papers with Xavier Caruso and Tristan Vaccon [CVR14; CVR15; CVR16; CVR17; CVR18], we explored a new method for propagating precision bounds through a computation, based on the following foundational lemma.

Let V and W be vector spaces over $\mathbb{Q}_p$ of dimension m and n. The p-adic ball $B_V(r)$ of radius $r = p^{-a}$ around 0 in V is a lattice, $p^a\mathbb{Z}_p^m$. More general ellipsoids can be modeled as arbitrary lattices $H \subset V$, and an imprecise element of V can be modeled as a coset $v + H$.

Lemma ([CVR14, Lem. 3.4])**.** *Suppose that $f : V \to W$ is differentiable at $v \in V$ and that the differential $f'(v)$ is surjective. Then, for all $\rho \in (0, 1]$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for all $r \in (0, \delta)$ and all lattices H with $B_V(\rho r) \subset H \subset B_V(r)$ one has*

$$f(v + H) = f(v) + f'(v)(H).$$

[2] a subring of full rank inside the number field defined by the same polynomial

The unique feature of this lemma is the *equality* of $f(v+H)$ and $f(v)+f'(v)(H)$, which implies that tracking precision through lattices and differentials is *optimal*, since the image of $v+H$ is also given by a lattice. Accompanying the final paper in this series, we implemented lattice precision within Sage, making it broadly and easily usable by other mathematicians.

The main downside to the method is that the complexity grows dramatically with the dimension, when compared to simpler methods. There is a lot of room for student projects that use the lattice approach to model precision loss theoretically for specific problems, while designing numerically stable algorithms for computing approximations without precision tracking.

2.2 Extensions and Galois groups

Unlike $\mathbb{R}$, which has a unique nontrivial algebraic extension (the complex numbers), any p-adic field $\mathbb{Q}_p$ has infinitely many extensions (see [Ser79] for example). To any such extension $K/\mathbb{Q}_p$ we may associate its Galois group $G = \mathrm{Gal}(K/\mathbb{Q}_p)$, a permutation group that controls how K is situated among other extensions $L/\mathbb{Q}_p$. For example, for each odd prime p there is a unique quartic extension with Galois group C_2^2, and it contains all the quadratic extensions of $\mathbb{Q}_p$. Moreover, there is a natural filtration on G (the ramification filtration), and corresponding sequences of subfields of K and of the normal closure of K. In addition to the standard problem of computing G and its filtration given K, there are two computational problems naturally arising in this context:

1. Given G, find all K with $\mathrm{Gal}(K/\mathbb{Q}_p) \cong G$ (there are finitely many),
2. Build a database of extensions $K/\mathbb{Q}_p$, giving a canonical defining polynomial for each K.

Toward the first question, I gave an algorithm [Roe19] for counting the number of K with a given Galois group, as long as $p \neq 2$. There are some natural necessary conditions on G for such a K to exist, but no known, easy-to-state sufficient conditions. Moreover, there is no efficient algorithm known for enumerating such K.

I am currently working with Jordi Guardia, John Jones, Kevin Keating, Sebastian Pauli, and David Roberts on a collaboration (through the SQuaRE program at AIM) to build on work of Monge [Mon14] in understanding how Galois groups vary in families of Eisenstein polynomials, as well as improving the Jones-Roberts database [JR06] within the LMFDB. We have found a method for choosing a canonical polynomial defining each K; the analogous function for number fields, `polredabs` in Pari [BBB+85], has proven very useful since it enables looking up an unknown number field from a list of existing fields.

2.3 Hypergeometric L-functions

Hypergeometric motives are a class of motives that provide an avenue to constructing a wide variety of L-functions that are not easily accessible via direct computation from an explicit algebraic variety. They are defined in terms of very simple data: a rational function $f(T)/g(T)$ so that f and g have equal degree and can be expressed as products of $T^m - 1$ for varying m, together with a specialization parameter $t \in \mathbb{Q}$. A full definition is given in the survey [RR22], and the Dirichlet coefficient a_p of the associated L-function can be expressed [Wat15] as an explicit sum with p summands. Computing the L-function to precision N using this formula directly thus requires $O(N^2)$ operations. Together with Edgar Costa and Kiran Kedlaya [CKR20; CKR23], we have designed an implemented an algorithm that uses methods originally due to Costa, Gerbicz and Harvey [CGH14] to compute the L-function in $\widetilde{O}(N)$ operations.

The performance impact is dramatic in practice, opening up the possibility of computing many hypergeometric L-functions. Such a source of L-functions has two immediate applications: examples are important in pinning down Euler factors at wild primes in order to get a full conjecture on what the conductor of a hypergeometric L-function should be, and a large body of high degree L-functions will be valuable in determining how widely the new murmurations phenomenon [HLOP22] applies.

References

[AKR09] Tim Abbott, Kiran S. Kedlaya, and David Roe. "Bounding Picard numbers of surfaces using p-adic cohomology". In: *Arithmetic, Geometry and Coding Theory (AGCT 2005), Séminaires et Congrès **21**, Societé Mathématique de France*. 2009, 125–159 (cit. on p. 3).

[AR16] Moshe Adrian and David Roe. "Rectifiers and the local Langlands correspondence: the unramified case". In: *Math. Res. Letters* 23.3 (2016), 593–619 (cit. on p. 3).

[BCE+22a] Jennifer Balakrishnan, Edgar Costa, Noam Elkies, David Roe, Andrew Sutherland, and John Voight. *Modular curves workshop 1*. `https://math.mit.edu/~edgarc/MCW.html`. Mar. 2022 (cit. on p. 2).

[BCE+22b] Jennifer Balakrishnan, Edgar Costa, Noam Elkies, David Roe, Andrew Sutherland, and John Voight. *Modular curves workshop 2*. `https://math.mit.edu/~edgarc/MCW2.html`. Nov. 2022 (cit. on p. 2).

[BCE+24] Jennifer Balakrishnan, Edgar Costa, Noam Elkies, David Roe, Andrew Sutherland, and John Voight. *Modular curves workshop 3*. `https://math.mit.edu/~edgarc/MCW3.html`. Mar. 2024 (cit. on p. 2).

[BD19] Jennifer S. Balakrishnan and Netan Dogra. "Code for Quadratic Chabuaty and rational points I: p-adic heights". `https://github.com/jbalakrishnan/QCI`. 2019 (cit. on p. 3).

[BBB+85] Christian Batut, Karim Belabas, Dominique Benardi, Henri Cohen, and Michel Olivier. *User's guide to PARI-GP*. 1985-2013 (cit. on p. 4).

[BBB+21] Alex Best, Jonathan Bober, Andrew Booker, Edgar Costa, John Cremona, Maarten Derickx, Min Lee, David Lowry-Duda, David Roe, Andrew Sutherland, and John Voight. "Computing classical modular forms". In: *Arithmetic Geometry, Number Theory, and Computation*. Simons Symp. Switzerland: Springer, 2021, 131–213 (cit. on p. 2).

[BCP97] Wieb Bosma, John Cannon, and Catherine Payoust. "The Magma algebra system. I. The user language." In: *J. Symbolic Comput.* 24.3-4 (1997), pp. 235–265 (cit. on p. 1).

[CVR14] Xavier Caruso, Tristan Vaccon, and David Roe. "Tracking p-adic precision". In: *LMS Journal of Computation and Mathematics* 17 (Special issue A) (2014), pp. 274–294 (cit. on p. 3).

[CVR15] Xavier Caruso, Tristan Vaccon, and David Roe. "p-adic stability in linear algebra". In: *Proceedings of the 2015 ACM on International Symposium on Symbolic and Algebraic Computation*. New York: ACM, 2015, 101–108 (cit. on p. 3).

[CVR16] Xavier Caruso, Tristan Vaccon, and David Roe. "Division and Slope Factorization of p-Adic Polynomials". In: *Proceedings of the 2016 ACM on International Symposium on Symbolic and Algebraic Computation*. New York: ACM, 2016, 159–166 (cit. on p. 3).

[CVR17] Xavier Caruso, Tristan Vaccon, and David Roe. "Characteristic polynomials of p-adic matrices". In: *Proceedings of the 2017 ACM on International Symposium on Symbolic and Algebraic Computation*. New York: ACM, 2017, pp. 389–396 (cit. on p. 3).

[CVR18] Xavier Caruso, Tristan Vaccon, and David Roe. "`ZpL`: a p-adic precision package". In: *Proceedings of the 2018 ACM on International Symposium on Symbolic and Algebraic Computation*. New York: ACM, 2018, pp. 119–126 (cit. on p. 3).

[CVV19] Xavier Caruso, Tristan Vaccon, and Thibaut Verron. "Gröbner Bases Over Tate Algebras". In: *Proceedings of the 2019 ACM on International Symposium on Symbolic and Algebraic Computation*. Beijing: ACM, 2019, pp. 74–81 (cit. on p. 3).

[CM98] Robert Coleman and Barry Mazur. "The eigencurve". In: *Galois representations in algebraic geometry*. London Math Soc. Lecture Note Ser., no. 254. Cambridge UP, 1998, pp. 1–113 (cit. on p. 3).

[CJP+23] Lewis Combes, John Jones, Jennifer Paulhus, David Roe, Manami Roy, and Sam Schiavone. "Creating a Dynamic Database of Finite Groups". Available at `math.mit.edu/~roed`. 2023 (cit. on p. 1).

[CGH14] Edgar Costa, Robert Gerbicz, and David Harvey. "A search for Wilson primes". In: *Math. Comp.* 83 (2014), 3071–3091 (cit. on p. 4).

[CKR20] Edgar Costa, Kiran S. Kedlaya, and David Roe. "Hypergeometric L-functions in average polynomial time". In: *Proceedings of the Fourteenth Algorithmic Number Theory Symposium (ANTS-XIV)*. Vol. 2. Open Book Series. Berkeley: Math. Sci. Pub., 2020, 143–159 (cit. on p. 4).

[CKR23] Edgar Costa, Kiran S. Kedlaya, and David Roe. "Hypergeometric L-functions in average polynomial time II". arXiv:2310.06971. 2023 (cit. on p. 4).

[CR16] Clifton Cunningham and David Roe. "From the function-sheaf dictionary to quasicharacters of p-adic tori". In: *J. Inst. Math. Jussieu* 17.1 (2016), 1–37 (cit. on p. 3).

[CR21] Clifton Cunningham and David Roe. "Commutative character sheaves and geometric types for supercuspidal representations". In: *Ann. Henri Lebesgue* 4 (2021), 1389–1420 (cit. on p. 3).

[DKRV21] Taylor Dupuy, Kiran Kedlaya, David Roe, and Christelle Vincent. "Isogeny classes of abelian varieties over finite fields in the LMFDB". In: *Arithmetic Geometry, Number Theory, and Computation*. Simons Symp. Switzerland: Springer, 2021, 375–448 (cit. on p. 2).

[HLOP22] Yang-Hui He, Kyu-Hwan Lee, Thomas Oliver, and Alexey Pozdnyakov. "Murmurations of elliptic curves". arXiv:2204.10140. 2022 (cit. on p. 4).

[JR06] John Jones and David Roberts. "A Database of Local Fields". In: *J. Symbolic Comput.* 41 (2006), pp. 80–97 (cit. on p. 4).

[LMF24a] The LMFDB Collaboration. *L-functions and modular forms database — Abelian varities over $\mathbb{F}_q$*. July 2024 (cit. on p. 3).

[LMF24b] The LMFDB Collaboration. *L-functions and modular forms database — Finite groups*. July 2024 (cit. on p. 1).

[LMF24c] The LMFDB Collaboration. *L-functions and modular forms database — Modular curves*. July 2024 (cit. on p. 2).

[LMF24d] The LMFDB Collaboration. *L-functions and modular forms database — Modular forms*. July 2024 (cit. on p. 2).

[Mon14] Maurizio Monge. "A family of Eisenstein polynomials generating totally ramified extensions, identification of extensions and construction of class fields". In: *Int. J. Number Theory* 10.7 (2014), 1699–1727 (cit. on p. 4).

[Muk95] Shigeru Mukai. "Curves and symmetric spaces I". In: *Amer. J. Math.* 117.6 (1995), 1627–1644 (cit. on p. 3).

[PS11] Rob Pollack and Glenn Stevens. "Overconvergent modular symbols and p-adic L-functions". In: *Annales scientifiques de l'ENS* 44.1 (2011), pp. 1–42 (cit. on p. 3).

[RR22] David Roberts and Fernando Rodriguez Villegas. "Hypergeometric motives". In: *Notices Amer. Math. Soc.* 69.6 (2022), 914–929 (cit. on p. 4).

[Roe11] David Roe. "The local Langlands correspondence for tamely ramified groups". PhD thesis. Harvard University, 2011 (cit. on p. 3).

[Roe14] David Roe. "The 3-adic eigencurve at the boundary of weight space". In: *Int. J. Number Theory* 10.7 (2014), 1791–1806 (cit. on p. 3).

[Roe19] David Roe. "The inverse Galois problem for p-adic fields". In: *Proceedings of the Thirteenth Algorithmic Number Theory Symposium (ANTS-XIII), Open Book Series*. Vol. 2. Berkeley: Math. Sci. Pub., 2019, 393–409 (cit. on p. 4).

[Ser79] Jean-Pierre Serre. *Local Fields*. New York: Springer-Verlag, 1979 (cit. on p. 4).

[Sage] The Sage Development Team. *SageMath, the Sage Mathematics Software System*. 2005-2024 (cit. on pp. 1, 3).

[WM71] William C. Waterhouse and James S. Milne. "Abelian varieties over finite fields". In: *Proc. Sympos. Pure Math.* 1971, 53–64 (cit. on p. 2).

[Wat15] Mark Watkins. "Hypergeometric motives over $\mathbb{Q}$ and their L-functions". `https://magma.maths.usyd.edu.au/~watkins/papers/known.pdf`, accessed 2023. 2015 (cit. on p. 4).

[Zyw22] David Zywina. "Explicit open images for elliptic curves over $\mathbb{Q}$". arXiv:2206.14959. 2022 (cit. on p. 2).