arXiv:2203.02045v1 [math.NT] 3 Mar 2022

SERRE WEIGHTS, GALOIS DEFORMATION RINGS, AND LOCAL MODELS

DANIEL LE AND BAO V. LE HUNG

Abstract. We survey some recent progress on generalizations of conjectures of Serre concerning the cohomology of arithmetic groups, focusing primarily on the "weight" aspect. This is intimately related to (generalizations of) a conjecture of Breuil and Mézard relating the geometry of potentially semistable deformation rings to modular representation theory. Recently, B. Levin, S. Morra, and the authors established these conjectures in tame generic contexts by constructing projective varieties (local models) in mixed characteristic whose singularities model, in generic cases, those of tamely potentially crystalline Galois deformation rings for unramified extensions of $\mathbb{Q}_p$ with small regular Hodge–Tate weights.

1. Introduction

The study of congruences between automorphic forms has a long and rich tradition. A paradigm shift occurred when Deligne's construction of Galois representations attached to classical holomorphic Hecke eigenforms opened the door to the study of congruences of automorphic forms through congruences of Galois representations. In fact, conjectures of Fontaine–Mazur–Langlands and Serre suggest that these are really two sides of the same coin.

Let p be a prime. Recall that Serre's conjecture asserts that every continuous, odd, and irreducible Galois representation $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ of the absolute Galois group of $\mathbb{Q}$ is *modular*, i.e. arises from the reduction of a Galois representation attached to a classical modular form. It furthermore asserts a refinement which specifies the minimal weight and level at which one may find such a modular form in terms of local properties of $\overline{\rho}$. Turning the perspective around, if one begins with a modular $\overline{\rho}$, then the refinement predicts congruences between modular forms of many different levels and weights.

While the level part generalizes quite naturally, the weight part is subtler, because it turns out to be inextricably linked to integral p-adic Hodge theory. The goal of this survey is to describe the circle of ideas surrounding recent developments on the weight part of Serre's conjecture.

1.1. **Overview.** In §2, we articulate the main questions of interest in the context of cohomological automorphic forms, especially motivating the representation theoretic perspective on congruences.

In §3 and 4, we briefly discuss background on modular representation theory and Galois representations. This will be necessary to state the higher dimensional generalizations of the weight part of Serre's conjecture.

In §5, we state generalizations of Serre's conjecture due to Ash, Gee, Herzig, and Savitt among many others which provide conjectural answers to these questions in many cases. In §5.3, we narrow our focus to two cases, definite unitary groups in joint work with B. Levin and S. Morra and GL_n over CM fields, where we established some conjectures of [GHS18] when $\overline{\rho}$ is tame and sufficiently generic at p. In the proofs, the Kisin–Taylor–Wiles method plays a vital role in reducing a global problem to a local one. As the internal details of the Kisin–Taylor–Wiles method are orthogonal to our goals, we have chosen to axiomatize its essential output and explain how it is used.

Finally, §6 summarizes the key results on local models for local deformation rings which are the essential ingredients to prove the results on the weight part of Serre's conjecture.

1.2. **Acknowledgments.** This survey is based on two talks given by the authors at the International Colloquium on Arithmetic Geometry at TIFR in January 2020. We thank the organizers for the invitation to this wonderful mathematical and cultural event.

D.L. was supported by the National Science Foundation under agreement No. DMS-1703182. B.LH. acknowledges support from the National Science Foundation under grant Nos. DMS-1128155, DMS-1802037 and the Alfred P. Sloan Foundation.

2. Cohomology of arithmetic manifolds

Let $\mathbb{G}$ be a connected reductive group over $\mathbb{Q}$. Let A_∞° be the connected component of the group of $\mathbb{R}$-points of a maximal $\mathbb{Q}$-split torus in the center of $\mathbb{G}$, and let K_∞° be a maximal connected compact subgroup of $\mathbb{G}(\mathbb{R})$. For a compact open subgroup $K_f \subset \mathbb{G}(\mathbb{A})$, consider the adelic double quotient

$$Y(K_f) \stackrel{\text{def}}{=} \mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_f.$$

In various places, we require technical properties of K_f that can always be attained by passing to a finite index subgroup. Furthermore, these required properties are preserved under passing to a finite index subgroup (away from a finite set of places). We assume throughout that K_f is sufficiently small (cf. [LLHLM20a, §9.1]). In particular, K_f is neat so that $Y(K_f)$ is naturally a real manifold. Moreover, $Y(K_f)$ can be rewritten as a finite disjoint union of quotients of the symmetric space $\mathbb{G}(\mathbb{R})/A_\infty^\circ K_\infty^\circ$ by subgroups of $\mathbb{G}(\mathbb{Q})$ which are discrete and of finite covolume in $\mathbb{G}(\mathbb{R})$. Finally, $Y(K_f)$ is homotopy equivalent to its Borel–Serre compactification (cf. [BS73]), a compact real manifold with corners and, in particular, a finite CW complex.

2.1. **Rational cohomology.** Let V be an algebraic representation of $\mathbb{G}$ over $\mathbb{Q}$. Then let $\mathcal{V}_\mathbb{Q}$ be the $\mathbb{Q}$-local system

$$\mathbb{G}(\mathbb{Q})\Big\backslash(\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_f) \times V(\mathbb{Q}),$$

where $\mathbb{G}(\mathbb{Q})$ acts diagonally. Then the (finite-dimensional) sheaf cohomology groups $H^*(Y(K_f), \mathcal{V}_\mathbb{Q})$ have a convolution action by the double coset (Hecke) algebra $\mathbb{Q}[K_f\backslash\mathbb{G}(\mathbb{A}_f)/K_f]$. (We will consider sequences of cohomology groups as objects in appropriate bounded derived categories. In most instances, little is lost if $H^*(Y(K_f), \mathcal{V}_\mathbb{Q})$ is replaced by $\oplus_{i\in\mathbb{Z}} H^i(Y(K_f), \mathcal{V}_\mathbb{Q})$.) The significance of these cohomology groups stems from the fact that the Hecke module $H^*(Y(K_f), \mathcal{V}_\mathbb{Q}) \otimes_\mathbb{Q} \mathbb{C}$ can be computed by automorphic forms by work of Matsushima and Franke [Mat67, Fra98]. This is essentially Hodge theory for the locally symmetric manifolds $Y(K_f)$. Suppose that A_∞° acts on $V(\mathbb{C})$ through a character. If we let $\mathcal{A}(K_f)$ denote the space of automorphic forms on $Y(K_f)$, then

$$H^*(Y(K_f), \mathcal{V}_\mathbb{Q}) \otimes_\mathbb{Q} \mathbb{C} \cong H^*(\mathfrak{g}, \mathfrak{p}, (\mathcal{A}(K_f) \otimes_\mathbb{C} V(\mathbb{C}))^{A_\infty^\circ}), \tag{2.1}$$

where $\mathfrak{g}$ is the Lie algebra of the group of real points of the intersection of all kernels of rational characters of $\mathbb{G}$. Let $\mathcal{V}_\mathbb{C}$ be the local system

$$(\mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/K_f) \times V(\mathbb{C})\Big/A_\infty^\circ K_\infty^\circ,$$

where the right action of $A_\infty^\circ K_\infty^\circ$ is diagonal with the inverse of the left action on $V(\mathbb{C})$. Then the bijection

$$\mathbb{G}(\mathbb{A}) \times V(\mathbb{C}) \to \mathbb{G}(\mathbb{A}) \times V(\mathbb{C}) \tag{2.2}$$

$$(g, v) \mapsto (g, g_\infty^{-1} v) \tag{2.3}$$

induces an isomorphism $\mathcal{V}_\mathbb{Q} \otimes_\mathbb{Q} \mathbb{C} \xrightarrow{\sim} \mathcal{V}_\mathbb{C}$. This is the first step in establishing (2.1).

We will assume that we can write K_f as $K_\Sigma K^\Sigma$ where Σ is a finite set of finite places and $K^\Sigma = \prod_{\ell\notin\Sigma} K_\ell$ where for all $\ell \notin \Sigma$, K_ℓ is a hyperspecial subgroup of $\mathbb{G}(\mathbb{Q}_\ell)$ (in particular, we assume that $\mathbb{G}$ is unramified at places $\ell \notin \Sigma$). Then $T_{\mathbb{Q}}^\Sigma \stackrel{\text{def}}{=} \mathbb{Q}[K^\Sigma\backslash\mathbb{G}(\mathbb{A}^\Sigma)/K^\Sigma]$ is commutative (see §4.2). Since $H^*(Y(K_f),\mathcal{V}_\mathbb{Q})$ are finite dimensional $\mathbb{Q}$-vector spaces, the eigenvalues of the $T_{\mathbb{Q}}^\Sigma$-action on the part of $\mathcal{A}(K_f)$ which contributes to $H^*(Y(K_f),\mathcal{V}_\mathbb{C})$ (the *cohomological automorphic forms*) are algebraic numbers.

2.2. **Classical modular forms.** If $\mathbb{G} = \mathrm{GL}_2$, then $Y(K_f)$ is a modular curve and has the additional structure of a variety defined over $\mathbb{Q}$. Any irreducible algebraic representation of GL_2 is of the form $V(a,b) \stackrel{\text{def}}{=} \mathrm{Sym}^{a-b}(\mathbb{Q}^2)\otimes_\mathbb{Q}\det^b$. Let $T_\ell \in \mathbb{Q}[\mathrm{GL}_2(\mathbb{Z}_\ell)\backslash\mathrm{GL}_2(\mathbb{Q}_\ell)/\mathrm{GL}_2(\mathbb{Z}_\ell)]$ be the double coset operator

$$\mathrm{GL}_2(\mathbb{Z}_\ell)\begin{pmatrix}\ell & 0\\ 0 & 1\end{pmatrix}\mathrm{GL}_2(\mathbb{Z}_\ell).$$

A well-known incarnation of (2.1) is that there is a normalized Hecke eigenform

$$f(z) = \sum_{n=0}^\infty a_n q^n, \qquad \text{with } q = e^{2\pi i z}$$

(i.e. $a_1 = 1$) of weight $k \geq 2$ and level K_f (or $K_f \cap \mathrm{SL}_2(\mathbb{Z})$) if and only if there is a $T_{\mathbb{Q}}^\Sigma$-eigenvector in $H^1(Y(K_f),\mathcal{V}(b+k-2,b))$ such that T_ℓ acts by $\ell^b a_\ell$ for all $\ell\notin\Sigma$.

It is well-known that the space of modular forms has a basis with integral q-expansions whose $\mathbb{Z}$-span is Hecke stable. In particular, $(a_\ell)_\ell$ are not just algebraic numbers, but are in fact algebraic integers. This gives one way to make the notion of congruences between eigenforms precise: one asks for a congruence between the (integral) Fourier coefficients.

It turns out that there are a lot of congruences between q-expansions of integral Hecke eigenforms. A basic example comes from the Eisenstein series

$$G_k(z) \stackrel{\text{def}}{=} -\frac{B_k}{2k} + \sum_{n\geq 1}\sigma_{k-1}(n)q^n, \qquad k\geq 4 \text{ even},$$

where B_k is the k-th Bernoulli coefficient. Fixing a (rational) prime p, the mod p q-series $G_k \pmod p$ depends only on $k \pmod{p-1}$. Another well-known example is the congruence

$$\Delta(z) \stackrel{\text{def}}{=} q\prod_{m=1}^\infty (1-q^n)^{24} \equiv q\prod_{m=1}^\infty (1-q^n)^2(1-q^{11n})^2 \pmod{11} \tag{2.4}$$

between the unique normalized cuspforms of level $\Gamma(1)$ and weight 12 and of level $\Gamma_0(11)$ and weight 2, respectively.

The above notion of congruences between eigenforms is essentially equivalent to congruences between the system of Hecke eigenvalues on rational cohomology, and thus can also be detected by contemplating the action of (suitably integral) Hecke operators on cohomology with integral coefficients. It turns out that this shift of perspective from q-expansion to integral cohomology (initiated by Ash–Stevens) will give a systematic mechanism to explain congruences between automorphic forms via representation theory.

2.3. **Integral structure.** Fix a prime p and suppose that K_f factors as the product $K_f^p K_p$. We fix an algebraic closure $\overline{\mathbb{Q}}_p$ of $\mathbb{Q}_p$ and let E be a subfield of $\overline{\mathbb{Q}}_p$ of finite degree over $\mathbb{Q}_p$. By replacing

E if necessary, we will assume that E is sufficiently large. Let $\mathcal{O}$ be the ring of integers of E with uniformizer ϖ and $\mathbb{F}$ be the residue field. We define $\mathcal{V}_E$ to be the nonarchimedean analogue

$$(\mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_f^p) \times V(E)\Big/ K_p$$

of $\mathcal{V}_{\mathbb{C}}$ in §2.1, where K_p acts diagonally (using the natural right action on $\mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_f^p$ and the inverse of the natural left action on $V(E)$). Then as before, the map

$$\mathbb{G}(\mathbb{A}) \times V(E) \to \mathbb{G}(\mathbb{A}) \times V(E) \tag{2.5}$$

$$(g, v) \mapsto (g, g_p^{-1} v) \tag{2.6}$$

induces an isomorphism $\mathcal{V}_{\mathbb{Q}} \otimes_{\mathbb{Q}} E \xrightarrow{\sim} \mathcal{V}_E$. As K_p is a compact group, there exists a K_p-stable $\mathcal{O}$-lattice W in $V(E)$. If we let

$$\mathcal{W} \stackrel{\text{def}}{=} (\mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_f^p \times W)/K_p, \tag{2.7}$$

then the map

$$H^*(Y(K_f), \mathcal{W}) \to H^*(Y(K_f), \mathcal{V}_E) \cong H^*(Y(K_f), \mathcal{V}_{\mathbb{Q}}) \otimes_{\mathbb{Q}} E \tag{2.8}$$

gives a natural integral structure on $H^*(Y(K_f), \mathcal{V}_{\mathbb{Q}}) \otimes_{\mathbb{Q}} E$. In fact, the definition (2.7) makes sense for any $\mathcal{O}[\![K_p]\!]$-module and defines a functor from $\mathcal{O}[\![K_p]\!]$-modules to local systems on $Y(K_f)$. We caution that (2.8) may not be injective in any given degree. Indeed, $H^*(Y(K_f), \mathcal{W})$ may contain torsion, and in fact this torsion is expected to be abundant and to play an important role in connecting cohomological automorphic forms and Galois representations.

2.4. **Congruences between Hecke eigensystems.** Let $T_{\mathcal{O}}^\Sigma$ be the Hecke algebra $\mathcal{O}[K^\Sigma\backslash\mathbb{G}(\mathbb{A}^\Sigma)/K^\Sigma]$, which acts naturally on $H^*(Y(K_f), \mathcal{W})$. As $H^*(Y(K_f), \mathcal{W})$ is a finite $\mathcal{O}$-module, there are only finitely many maximal ideals $\mathfrak{m} \subset T_{\mathcal{O}}^\Sigma$ for which the localization $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ is nonzero. These localized modules record congruences between systems of Hecke eigenvalues: a Hecke eigenclass $H^*(Y(K_f), \mathcal{W})[\frac{1}{p}]$ survives in the localization $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}[\frac{1}{p}]$ if and only if its (automatically integral) system of Hecke eigenvalues lifts the mod p system given by $\mathfrak{m}$.

However, for a fixed $\mathfrak{m}$, there may be various local systems $\mathcal{W}$ for which $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ is nonzero. Indeed, we saw in §2.2 that if $\mathfrak{m}_{G_k}$ corresponds to (the system of Hecke eigenvalues of) the Eisenstein series $G_k \pmod p$ for $4 \le k \le p+1$ and $\mathcal{W}(a,b)$ corresponds to the lattice $W(a,b) \stackrel{\text{def}}{=} \mathrm{Sym}^{a-b} \mathbb{Z}_p^2 \otimes \det^b$ for $a \ge b$, then $H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})), \mathcal{W}(k'-2, 0))_{\mathfrak{m}_{G_k}}$ is nonzero for all $k' \equiv k \pmod{p-1}$. (Since $\mathrm{GL}_2(\widehat{\mathbb{Z}})$ is not neat, these cohomology groups should be interpreted as the cohomology groups of an orbifold.) Furthermore, if $\mathfrak{m}_\Delta$ corresponds to the Ramanujan Delta function mod 11, then both $H^*(Y(K_f), \mathbb{Z}_p)_{\mathfrak{m}_\Delta}$ and $H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})), \mathcal{W}(10,0))_{\mathfrak{m}_\Delta}$ are nonzero where K_f corresponds to the congruence subgroup $\Gamma_0(11)$.

The upshot of our discussion above is that congruences between eigenforms can be thought as the non-vanishing of localized cohomology for many different coefficient sheaves. Thus a complete classification of such congruences is equivalent to the following question:

Question 2.4.1. Given a mod p Hecke eigensystem $\mathfrak{m}$, for which $\mathcal{O}$-local systems $\mathcal{W}$ on $Y(K_f)$ is $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ nonzero?

Serre studied this question extensively in the case of GL_2 [Ser87]. This perspective of cohomology actually gives a natural explanation for the congruences for GL_2 in §2.2. We explain how it naturally leads to considerations in modular representation theory. From the short exact sequence $0 \to W \xrightarrow{\cdot p} W \to W \otimes_{\mathbb{Z}_p} \mathbb{F}_p \to 0$, we see that $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ is nonzero if and only if $H^*(Y(K_f), \mathcal{W} \otimes_{\mathbb{Z}_p} \mathbb{F}_p)_{\mathfrak{m}}$

is. While $W(a,b)\otimes_{\mathbb{Z}_p}\mathbb{Q}_p$ is an irreducible $\mathrm{GL}_2(\mathbb{Z}_p)$-module, $W(a,b)\otimes_{\mathbb{Z}_p}\mathbb{F}_p$ is irreducible if and only if $a-b\leq p-1$, in which case, we let $F(a,b)\stackrel{\mathrm{def}}{=}W(a,b)\otimes_{\mathbb{Z}_p}\mathbb{F}_p$. All (absolutely) irreducible $\mathrm{GL}_2(\mathbb{Z}_p)$-modules over $\mathbb{F}_p$ arise in this way, and $F(a,b)\cong F(c,d)$ if and only if $a-c=b-d\in(p-1)\mathbb{Z}$. Let $\mathcal{F}(a,b)$ be the corresponding local system.

Let us first revisit the congruences between Eisenstein series. For any $a'>b$, the submodule of $W(a',b)\otimes_{\mathbb{Z}_p}\mathbb{F}_p$ generated by a (nonzero) highest weight vector is isomorphic to $F(a,b)$ where a is the unique integer such that $0<a-b<p$ and $a\equiv a'\pmod{p-1}$. This gives a map

$$H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{W}(k-2,0)\otimes_{\mathbb{Z}_p}\mathbb{F}_p)_{\mathfrak{m}_{G_k}}\to H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{W}(k'-2,0)\otimes_{\mathbb{Z}_p}\mathbb{F}_p)_{\mathfrak{m}_{G_k}}$$

where $2<k<p+2$ and $k'>2$ with $k'\equiv k\pmod{p-1}$. It can be shown (for example by applying Hida's ordinary projector) that these maps are injective in each degree. This illustrates how modular representation theory can be used to produce infinite families of congruences between Hecke eigensystems.

The congruence (2.4) between cuspforms is simpler. Recall that here $p=11$ and that $\mathbb{G}=\mathrm{GL}_2$. Then the fact that $\mathfrak{m}_\Delta$ is *non-Eisenstein* implies that for all local systems $\mathcal{W}$, $H^*(Y(K_f),\mathcal{W})_{\mathfrak{m}_\Delta}$ is zero unless $*=1$. First, Shapiro's lemma now implies that $H^*(Y(K_f),\mathcal{W})_{\mathfrak{m}_\Delta}\cong H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{W}')_{\mathfrak{m}_\Delta}$, where $\mathcal{W}'$ corresponds to the principal series representation $\mathrm{Ind}_{B(\mathbb{F}_p)}^{\mathrm{GL}_2(\mathbb{F}_p)}W$. Second, the functor $W\mapsto H^1(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{W})_{\mathfrak{m}_\Delta}$ is an exact functor from the category of finite $\mathbb{Z}_{11}[\![\mathrm{GL}_2(\mathbb{Z}_{11})]\!]$-modules to the category of finite $\mathbb{Z}_{11}$-modules. Now $\mathrm{Ind}_{B(\mathbb{F}_{11})}^{\mathrm{GL}_2(\mathbb{F}_{11})}\mathbf{1}$ is naturally identified with the space of $\mathbb{F}_{11}$-valued functions on $\mathbb{P}^1(\mathbb{F}_{11})$ and decomposes as $\mathrm{Sym}^{10}\mathbb{F}_{11}^2\oplus\mathbf{1}$. Then the injection

$$H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{F}(10,0))_{\mathfrak{m}_\Delta}\hookrightarrow H^*(Y(K_f),\mathbb{F}_p)_{\mathfrak{m}_\Delta}$$

provides the desired congruence. This example illustrates the important phenomenon of how the weight and level of modular forms can interact mod p.

With these representation theoretic arguments in mind, Ash, Stevens, and others have suggested that one should narrow the focus of Question 2.4.1 to when K_p is a maximal compact open subgroup and $\mathcal{W}$ is an irreducible $\mathbb{F}$-local system.

Question 2.4.2 (The weight part of Serre's conjecture)**.** Suppose that K_p is a maximal compact open subgroup. Given a mod p Hecke eigensystem $\mathfrak{m}$, for which irreducible $\mathbb{F}$-local systems $\mathcal{W}$ on $Y(K_f)$ is $H^*(Y(K_f),\mathcal{W})_{\mathfrak{m}}$ nonzero?

While this is a substantial reduction since there are only finitely many such $\mathbb{F}$-local systems up to isomorphism, little is expected to be lost as we now explain. The following proposition is immediate.

Proposition 2.4.3. *If $H^*(Y(K_f),\mathcal{W})_{\mathfrak{m}}$ is nonzero, then $H^*(Y(K_f),\mathcal{F})_{\mathfrak{m}}$ is nonzero for some irreducible subquotient $\mathcal{F}$ of $\mathcal{W}\otimes_{\mathcal{O}}\mathbb{F}$.*

The converse to Proposition 2.4.3 holds in non-Eisenstein cases for $\mathbb{G}$ a Weil restriction of GL_n if expected vanishing conjectures hold (see §4.6). These vanishing conjectures generalize the vanishing outside of degree 1 for $\mathbb{G}=\mathrm{GL}_2$.

Question 2.4.2 turns out to be quite subtle. Nonisomorphic irreducible $\mathbb{F}$-local systems may contain the same Hecke eigensystems, i.e. not all congruences arise from modular representation theory. For example, with $p=23$, both $H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{F}(10,0))_{\mathfrak{m}_\Delta}$ and $H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})),\mathcal{F}(21,11))_{\mathfrak{m}_\Delta}$ are nonzero. If we write

$$\Delta(z)=\sum_{n=1}^{\infty}\tau(n)q^n,$$

then T_ℓ acts on $H^*(Y(\mathrm{GL}_2(\widehat{\mathbb{Z}})), \mathcal{F}(21,11))_{\mathfrak{m}_\Delta}$ by $\ell^{11}\tau(\ell)$ for all primes $\ell \neq 23$ by (2.1) (see §2.2). This implies that

$$\tau(n) \equiv \left(\tfrac{n}{23}\right)\tau(n) \pmod{23} \tag{2.9}$$

for all n coprime to 23. In other words, $23 \mid \tau(n)$ if $\left(\frac{n}{23}\right) = -1$.

3. An interlude on representation theory

3.1. **Serre weights.** In order to explore Question 2.4.2, the natural first step is to ask for a classification of simple $\mathbb{F}[\![K_p]\!]$-modules. If $K_p(1) \subset K_p$ is a normal finite index pro-p subgroup and W is a finite $\mathbb{F}[\![K_p]\!]$-module, then $W^{K_p(1)}$ is an $\mathbb{F}[\![K_p]\!]$-submodule of W which is *nonzero* (since the action of a p-group on a finite-dimensional $\mathbb{F}$-vector space must have a nonzero fixed vector). Then the action of K_p on a simple $\mathbb{F}[\![K_p]\!]$-module factors through the finite quotient $K_p/K_p(1)$, which can often be arranged to be a finite group of Lie type. In this section, we discuss the (modular) representation theory of these groups.

Let G be a connected reductive group over $\mathbb{F}_p$ which splits over $\mathbb{F}$. (We will eventually take G to be the mod p reduction of an integral model of $\mathbb{G}$.) An isomorphism class of a simple $\mathbb{F}[G(\mathbb{F}_p)]$-module is known as a *Serre weight for* $G(\mathbb{F}_p)$. Our goal now is to describe the (finite) set of Serre weights for $G(\mathbb{F}_p)$.

Let B be an $\mathbb{F}_p$-rational Borel subgroup in G with Levi subgroup T. We denote by W the Weyl group $N(T)/T$, which has a Bruhat partial order with a unique longest element w_0. We write $X(T)$ for the character group of T, which has an action of W and an induced action from the relative Frobenius F acting on T. This group has a subset

$$X_1(T) \stackrel{\text{def}}{=} \{\lambda \in X(T) : 0 \leq \langle \lambda, \alpha^\vee \rangle \leq p-1, \text{ for all simple } \alpha\}$$

which plays an important role in the modular representation theory of G. We let

$$X^0(T) \stackrel{\text{def}}{=} \{\lambda \in X(T) : \langle \lambda, \alpha^\vee \rangle = 0 \text{ for all roots } \alpha\}.$$

For any character $\lambda \in X(T)$, we can consider the algebraic induction $W(\lambda) \stackrel{\text{def}}{=} \mathrm{Ind}_B^G w_0\lambda$ (also known as the dual Weyl module), which is nonzero if and only if λ is dominant with respect to B. We let $L(\lambda)$ denote the socle of $W(\lambda)$, which is the simple submodule generated by a nonzero highest weight vector. Then we have the following result about Serre weights for $G(\mathbb{F}_p)$.

Theorem 3.1.1. *The map*

$$\frac{X_1(T)}{(F-1)X^0(T)} \to \{\textit{Serre weights for } G(\mathbb{F}_p)\} \tag{3.1}$$

$$\lambda \mapsto L(\lambda)(\mathbb{F})|_{G(\mathbb{F}_p)} \tag{3.2}$$

is a bijection.

We denote $L(\lambda)(\mathbb{F})|_{G(\mathbb{F}_p)}$ by $F(\lambda)$. (To avoid conflicts with the F-action on $X(T)$, we will write this action without parentheses.)

3.2. **Deligne–Lusztig representations.** While Question 2.4.2 only involves $\mathbb{F}$-local systems, we will see that it is inextricably linked to $\mathcal{O}$-torsion free local systems. It is then natural to ask for a classification of irreducible $G(\mathbb{F}_p)$-representations in characteristic 0. We now recall such a classification, provided by work of Deligne and Lusztig [DL76].

For an element $w \in W$, there exists $g_w \in G(\overline{\mathbb{F}}_p)$ such that $g_w^{-1}F(g_w) \in N(T)(\overline{\mathbb{F}}_p)$ represents w. Then we let T_w be the F-stable torus $g_w T g_w^{-1}$. Let $\widetilde{W}$ denote the extended affine Weyl group which

is the semidirect product $X(T) \rtimes W$. For an element $\widetilde{w} = (\mu, w) \in \widetilde{W}$, we define $\theta_{\widetilde{w}} : T_w(\mathbb{F}_p) \to E^\times$ to be the restriction of the character

$$\begin{aligned} T_w(\overline{\mathbb{F}}_p) &\to \overline{\mathbb{Q}}_p^\times \\ g &\mapsto [\mu](g_w^{-1} g g_w) \end{aligned}$$

to $T_w(\mathbb{F}_p)$ (here, $[\mu]$ denotes the Teichmüller lift of μ).

To a character $\theta_{\widetilde{w}}$ of a maximal rational torus $T_w(\mathbb{F}_p)$ of $G(\mathbb{F}_p)$, Deligne and Lusztig associate a virtual (Deligne–Lusztig) representation over E which they denote $\epsilon_G \epsilon_{T_w} R_{T_w}^{\theta_{\widetilde{w}}}$. We will instead denote this virtual representation by $R(\widetilde{w})$ and say that $\widetilde{w}$ is a presentation for $R(\widetilde{w})$. The map $\widetilde{w} \mapsto R(\widetilde{w})$ is not injective—two elements map to the same virtual representation if and only if they lie in the same orbit of the action of $\widetilde{W}$ on itself given by

$$(\nu, s) \cdot (\mu, w) = (s\mu + F\nu - swF(s)^{-1}(\nu), swF(s)^{-1}).$$

The simplest case of the above construction occurs when w is the identity. Then $T_w = T$, $\theta_{\widetilde{w}}$ is a character of $T(\mathbb{F}_p)$ and by inflation a character of $B(\mathbb{F}_p)$, and $R(\widetilde{w})$ is the principal series representation $\mathrm{Ind}_{B(\mathbb{F}_p)}^{G(\mathbb{F}_p)} \theta_{\widetilde{w}}$. Nonuniqueness of presentations can be seen from the existence of intertwiners between principal series representations.

The group $\widetilde{W}$ acts on $X(T)$ in the usual way—W acts on $X(T)$ by group automorphisms and $X(T)$ acts on itself by translation. Let m be a nonnegative integer and let $0 \in X(T)$ denote the trivial character. We say that $\widetilde{w} \in \widetilde{W}$ is (lowest alcove) m-generic if $\langle \widetilde{w}(0), \alpha^\vee \rangle > m$ for all simple roots α and $\langle \widetilde{w}(0), \alpha^\vee \rangle < p - m$ for all roots $\alpha^\vee$. We say that a Deligne–Lusztig representation R is m-generic if $R = R(\widetilde{w})$ for some m-generic $\widetilde{w}$. An m-generic $\widetilde{w}$ or R exists only if $mh < p$, where h denotes the Coxeter number of G. [DL76, Proposition 10.10] implies that if R is 0-generic, then R is in fact a genuine representation.

Let $\overline{R}$ denote the semisimplification of the reduction of any $G(\mathbb{F}_p)$-stable $\mathcal{O}$-lattice in a genuine $G(\mathbb{F}_p)$-representation R over E ($\overline{R}$ does not depend on the choice of lattice). In relation to Question 2.4.2, it is important to have an understanding of $\overline{R}$ for Deligne–Lusztig representations R. This is provided by Jantzen's formula for the reductions of Deligne–Lusztig representations in terms of virtual linear combinations of dual Weyl modules [Jan81]. If R is sufficiently generic, then the Jordan–Hölder factors of $\overline{R}$ admit the following description in terms of alcove geometry, which is in a sense independent of p. For convenience, we assume that G admits a *twisting element* $\eta \in X(T)$, defined up to $X^0(T)$, which by definition has the property that $\langle \eta, \alpha^\vee \rangle = 1$ for all simple roots α. The existence of an η can always be arranged by passing to a central extension of G by $\mathbb{G}_m$ (see [BG14, Proposition 5.3.1(a)]). We write $\cdot$ for the p-dot action so that $(\nu, w) \cdot \lambda = w(\lambda + \eta) - \eta + p\nu$. See [LLHLM20a, §2] for any unexplained notation below.

Proposition 3.2.1. [LLHLM20a, Proposition 2.3.6] *Let h be the Coxeter number of G. If $\widetilde{w} \in \widetilde{W}$ is $2h$-generic, then the Jordan–Hölder factors of $\overline{R}(\widetilde{w})$ are precisely the Serre weights of the form*

$$F(\pi^{-1}(\widetilde{w}_1) \cdot (\widetilde{w}\widetilde{w}_2^{-1}(0) - \eta))$$

with $\widetilde{w}_1 \in \widetilde{W}$ restricted and dominant, $\widetilde{w}_2 \in \widetilde{W}$ dominant, and $\widetilde{w}_1 \uparrow (\eta, w_0)\widetilde{w}_2$.

Remark 3.2.2. Of course, the description in Proposition 3.2.1 does not depend on the choice of twisting element η and could in fact be rephrased without any reference to η.

4. Relations to Galois representations

In order to address Question 2.4.2 (and to explain the congruence (2.9)), we introduce some conjectures and results concerning the relationship between cohomological automorphic forms and Galois representations. We follow the approach in [Gro99], which seems to be more standard when $\mathbb{G}$ is a general linear group or a unitary group. For a more canonical approach to conjectures concerning Galois representations attached to cohomological automorphic forms, see [BG14].

4.1. **Twisting element.** For a field F, let G_F denote the absolute Galois group $\mathrm{Gal}(F^{\mathrm{sep}}/F)$ where we fix some separable closure F^{sep}. Fix a maximal torus T and Borel subgroup B in $\mathbb{G}_{/\overline{\mathbb{Q}}}$. We assume now that $\mathbb{G}$ has a *twisting element* η which is by definition an element of $X(T)^{G_{\mathbb{Q}}}$ such that $\langle \eta, \alpha^\vee \rangle = 1$ for all simple roots α. If $\mathbb{G}$ is GL_n, we can take η to be $(n-1, n-2, \dots, 1, 0)$. As before, a twisting element always exists if we replace $\mathbb{G}$ by a central extension of $\mathbb{G}$ by $\mathbb{G}_m$. The effect of this on the constructions and questions in §2 is minimal. A twisting element is only unique up to $X^0(T)^{G_{\mathbb{Q}}}$.

4.2. **Satake parameters.** Fix a prime ℓ and suppose that there is a reductive model G over $\mathbb{Z}_\ell$ for $\mathbb{G}$ such that $G(\mathbb{Z}_\ell) = K_\ell$. Let $T \subset B \subset G$ be a maximal torus and Borel subgroup, respectively. We have the Satake isomorphism (see e.g. [Car79, §4.2])

$$\mathcal{S} : \mathbb{Z}[1/\ell][K_\ell \backslash \mathbb{G}(\mathbb{Q}_\ell)/K_\ell] \xrightarrow{\sim} \mathbb{Z}[1/\ell][T(\mathbb{Z}_\ell)\backslash T(\mathbb{Q}_\ell)/T(\mathbb{Z}_\ell)]^{W_s}$$

normalized using the choice of twisting element η as in [Gro98, Proposition 3.6], where W_s is the Weyl group of the maximal split torus in T.

If T is split, then $T(\mathbb{Z}_\ell)\backslash T(\mathbb{Q}_\ell)/T(\mathbb{Z}_\ell) \cong Y(T)$ (here $Y(T)$ denotes the cocharacter group of T) and E-valued characters of

$$\mathbb{Z}[1/\ell][K_\ell \backslash \mathbb{G}(\mathbb{Q}_\ell)/K_\ell] \cong \mathbb{Z}[1/\ell][Y(T)]^W \cong \mathcal{O}(\widehat{T}/\!\!/W)$$

are in bijection with semisimple conjugacy classes in the dual group $\widehat{G}(E)$ for any coefficient field E of characteristic not equal to ℓ. In general, E-valued characters χ of $\mathbb{Z}[1/\ell][K_\ell \backslash \mathbb{G}(\mathbb{Q}_\ell)/K_\ell]$ are in bijection with semisimple conjugacy classes C_χ of ${}^L G(E)$, where ${}^L G$ denotes the Langlands dual of G (see [Gro99, §16]).

4.3. **Conjectures on Galois representations associated to cohomological automorphic forms.** We fix a prime p and a sufficiently large subfield $E \subset \overline{\mathbb{Q}}_p$ of finite degree over $\mathbb{Q}_p$.

Conjecture 4.3.1. *Let $V(\lambda)_E$ denote the irreducible representation of $\mathbb{G}_{/E}$ of highest weight λ. Suppose that $\mathfrak{p} \subset T^\Sigma_{\mathbb{Q}} \otimes_{\mathbb{Q}} E$ is a maximal ideal such that the $\mathfrak{p}$-torsion $H^*(Y(K_f), \mathcal{V}(\lambda)_E)[\mathfrak{p}]$ is nonzero. Then there exists a continuous homomorphism*

$$\rho : G_{\mathbb{Q}} \to {}^L\mathbb{G}(E)$$

such that

1. *the composition of ρ with the projection ${}^L\mathbb{G}(E) \to G_{\mathbb{Q}}$ is the identity on $G_{\mathbb{Q}}$;*
2. *for $\ell \notin \Sigma$ and $\ell \neq p$, ρ is unramified at ℓ and $\rho(\mathrm{Frob}_\ell)$ is in C_χ where χ is the character*
$$\mathbb{Z}[1/\ell][K_\ell \backslash \mathbb{G}(\mathbb{Q}_\ell)/K_\ell] \subset T^\Sigma_{\mathbb{Q}} \otimes_{\mathbb{Q}} E \to (T^\Sigma_{\mathbb{Q}} \otimes_{\mathbb{Q}} E)/\mathfrak{p} \cong E$$
and Frob_ℓ is an arithmetic Frobenius element at ℓ;
3. *$\rho|_{G_{\mathbb{Q}_p}}$ is de Rham with Hodge–Tate cocharacter $\lambda + \eta$ (see* [BG14, §2.4]*; in our normalization the cyclotomic character corresponds to the cocharacter $\mathrm{id}_{\mathbb{G}_m}$) and is moreover crystalline if $p \notin \Sigma$; and*
4. *ρ is odd in the sense of* [Gro99, Conjecture 17.2(a)].

There is an analogous conjecture with torsion coefficients. As before, let $\mathbb{F}$ denote the residue field of E.

Conjecture 4.3.2. *If W is a finite $\mathbb{F}[\![K_p]\!]$-module, let $\mathcal{W}$ denote the $\mathbb{F}$-local system on $Y(K_f)$. Suppose that $\mathfrak{m} \subset T^{\Sigma}_{\mathcal{O}}$ is a maximal ideal such that $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ is nonzero. Then there exists a continuous homomorphism*

$$\overline{\rho} : G_{\mathbb{Q}} \to {}^L\mathbb{G}(\mathbb{F})$$

such that

(1) the composition of $\overline{\rho}$ with the projection ${}^L\mathbb{G}(\mathbb{F}) \to G_{\mathbb{Q}}$ is the identity on $G_{\mathbb{Q}}$;
(2) for $\ell \notin \Sigma$ and $\ell \neq p$, $\overline{\rho}$ is unramified at ℓ and $\overline{\rho}(\mathrm{Frob}_\ell)$ is in C_χ where χ is the character

$$\mathbb{Z}[1/\ell][K_\ell \backslash \mathbb{G}(\mathbb{Q}_\ell)/K_\ell] \to T^{\Sigma}_{\mathbb{Q}} \otimes_{\mathbb{Q}} \mathbb{F} \to T^{\Sigma}_{\mathbb{Q}} \otimes_{\mathbb{Q}} \mathbb{F}/\mathfrak{m} \cong \mathbb{F}$$

and Frob_ℓ is an arithmetic Frobenius element at ℓ; and
(3) $\overline{\rho}$ is odd in the sense of [Gro99, Conjecture 17.2(a)].

Remark 4.3.3. (1) One also expects that ρ satisfies a compatibility with the conjectural local Langlands correspondence at places in Σ.
(2) Comparing the two conjectures, observe that there is no property at p for $\overline{\rho}$. Such a property would be closely related to Questions 2.4.1 and 2.4.2.
(3) It is clear that ρ and $\overline{\rho}$ determine $\mathfrak{p}$ and $\mathfrak{m}$, respectively. On the other hand, the properties of ρ described in Conjectures 4.3.1 and 4.3.2 do not characterize ρ in general. When $\mathbb{G}$ is a Weil restriction of GL_n, the first two properties characterize the isomorphism class of the semisimplification of ρ by the Chebotarev density theorem and the Brauer–Nesbitt theorem. But even for tori, these properties do not characterize ρ (see [BG14, Remark 3.2.4]).
(4) The existence of torsion cohomology classes means that Conjecture 4.3.1 does not immediately imply Conjecture 4.3.2. A version of Conjecture 4.3.2 for all torsion coefficients implies Conjecture 4.3.1 (except for the third property) by taking a limit.

There are two cases, relevant to what follows, when both conjectures are known:

(1) the Weil restriction of a definite unitary group relative to a CM extension of a totally real field not equal to $\mathbb{Q}$ [Kot92, HT01, Lab99, Shi11, CH13], and

(2) the Weil restriction of GL_n over a CM field [Sch15].

In these cases, the attached Galois representations are determined up to semisimplification by Remark 4.3.3(3). We say that $\overline{\rho}$ (or $\mathfrak{m}$) is non-Eisenstein if $\overline{\rho}$ does not factor through a proper parabolic subgroup after any finite extension of $\mathbb{F}$.

4.4. **Modular Serre weights.** Fix p and E as before, and let $\mathbb{F}$ be the residue field of E. Suppose from now on that $\mathbb{G}_{/\mathbb{Q}_p}$ has an integral model $G_{/\mathbb{Z}_p}$. Having classified irreducible representations of $G(\mathbb{Z}_p)$ and introduced Galois representations, we now revisit Question 2.4.2 through that lens. We let K_p be $G(\mathbb{Z}_p)$. An irreducible $G(\mathbb{Z}_p)$-representation over $\mathbb{F}$ factors through the reduction map $G(\mathbb{Z}_p) \twoheadrightarrow G(\mathbb{F}_p)$ (whose kernel is pro-p), i.e. is the inflation of a Serre weight for $G(\mathbb{F}_p)$. For a Serre weight σ, let $\mathcal{F}_\sigma$ denote the corresponding $\mathbb{F}$-local system on $Y(K_f)$.

Fix an $\mathbb{F}$-valued Hecke eigensystem $\mathfrak{m}$. To understand Question 2.4.2 is to understand the following (finite) set.

Definition 4.4.1. Let $W(\mathfrak{m})$ be the set of isomorphism classes of Serre weights σ for $G(\mathbb{F}_p)$ for which $H^*(Y(K_f), \mathcal{F}_\sigma)_{\mathfrak{m}}$ is nonzero. If $\sigma \in W(\mathfrak{m})$, then we say that σ is a *modular Serre weight* (for $\mathfrak{m}$).

If there is a Galois representation $\overline{\rho} : G_{\mathbb{Q}} \to {}^L\mathbb{G}(\mathbb{F})$ satisfying the properties in Conjecture 4.3.2 for $\mathfrak{m}$, then we also write $W(\overline{\rho})$ for $W(\mathfrak{m})$ and say that σ is a modular Serre weight for $\overline{\rho}$ when $\sigma \in W(\overline{\rho})$.

4.5. **The case of GL_2.** Fix p and E as before. The first result on Conjecture 4.3.1 for nonabelian $\mathbb{G}$ was work of Deligne [Del71] in the case of $\mathbb{G} = \mathrm{GL}_2$ (building on work of Eichler and Shimura). In this case there is no torsion in cohomology, and so Conjecture 4.3.1 implies Conjecture 4.3.2. Deligne constructed ρ satisfying all the properties of Conjecture 4.3.1 except the third, which follows from subsequent work of Faltings on p-adic comparison theorems. (We set $\eta = (1, 0)$ here.)

Let $K_p = \mathrm{GL}_2(\mathbb{Z}_p)$. Fix a mod p Hecke eigensystem $\mathfrak{m}$. Assume that $\mathfrak{m}$ is non-Eisenstein, i.e. that the attached Galois representation $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F})$ is absolutely irreducible. (The data of $\mathfrak{m}$ is equivalent to that of the isomorphism class of $\overline{\rho}$ by Remark 4.3.3(3).) Since cohomology groups outside degree 1 do not admit non-Eisenstein mod p Hecke eigensystems, the functor $W \to H^1(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ is exact, as explained in §2.4. In particular, Question 2.4.1 reduces to Question 2.4.2, namely an investigation of $W(\overline{\rho})$.

If $H^1(Y(K_f), \mathcal{F}(a, b))_{\mathfrak{m}}$ is nonzero, then so is $H^1(Y(K_f), \mathcal{W}(a, b))_{\mathfrak{m}}$. A necessary condition for $F(a, b)$ to be in $W(\overline{\rho})$ is that $\overline{\rho}$ is the reduction of a representation $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(E)$ which is unramified outside of Σ and p and is crystalline at p of Hodge–Tate weights $a + 1$ and b. (Since $\overline{\rho}$ is irreducible, a $G_{\mathbb{Q}}$-invariant $\mathcal{O}$-lattice in ρ is unique up to scaling.) In particular, the restriction $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is the reduction of (a lattice in) a crystalline representation $\rho_p : G_{\mathbb{Q}_p} \to \mathrm{GL}_2(E)$ of Hodge–Tate weights $a + 1$ and b. The following result, known as the weight part of Serre's conjecture, is a local-global principle (i.e. $W(\overline{\rho})$ only depends on $\overline{\rho}|_{G_{\mathbb{Q}_p}}$) that asserts that the necessary condition is in fact sufficient.

Theorem 4.5.1 ([Gro99, Edi92, CV92])**.** *Suppose that $\mathfrak{m}$ is non-Eisenstein. Then $F(a, b) \in W(\overline{\rho})$ if and only if $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is the reduction of a crystalline representation of Hodge–Tate weights $a + 1$ and b.*

Remark 4.5.2.
(1) The above formulation is slightly different, albeit equivalent, from Serre's original formulation in [Ser87]. In *loc.cit.*, the recipe for $W(\overline{\rho})$ is completely explicit in terms of the "inertial weights" of $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ when $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is semisimple, with additional modfications in terms of ramification properties of an extension class in general. In particular, in the semisimple case, there is a simple combinatorial formula for a and b solely in terms of the inertial weights. Early generalizations of Serre's conjecture beyond $\mathrm{GL}_{2/\mathbb{Q}}$, e.g. [ADP02, Conjecture 3.1], involved similar formulas (see [GHS18, §7] for more recent formulas for a larger list of Serre weights). On the other hand, while [BDJ10] also contains formulas à la Serre, it emphasizes the above "crystalline lifts" perspective.
(2) Theorem 4.5.1 was generalized to (the Weil restriction of) the unit groups in quaternion algebras over totally real fields split at no more than one archimedean place and definite unitary groups over a totally real field when $p > 2$ (and under mild additional hypotheses) in a series of works by Gee, Newton, Kisin, Liu, and Savitt [New14, GK14, GLS14]. Some of these build on earlier work of Gee that introduced the Taylor–Wiles method to produce proofs rather different from the original proofs of Theorem 4.5.1.

We now revisit the mod 23 congruence for the Ramanujan Delta function. Let $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_{23})$ be the associated Galois representation. In this case, $\overline{\rho}|_{I_{\mathbb{Q}_{23}}} \cong \omega^{11} \oplus \mathbf{1}$, where ω denotes the reduction of the 23-adic cyclotomic character χ and $I_{\mathbb{Q}_{23}} \subset G_{\mathbb{Q}_{23}}$ denotes the inertial subgroup. Then $\overline{\rho}|_{G_{\mathbb{Q}_{23}}}$ is the reduction of both $\chi^{11} \oplus \mathbf{1}$ and $\chi^{11} \oplus \chi^{22}$ (up to unramified twists). Theorem 4.5.1 implies that $\{F(10, 0), F(21, 11)\} \subset W(\overline{\rho})$. In fact, this is an equality.

The behavior of the Ramanujan Delta function modulo 23 illustrates a rare phenomenon. For example, if we instead take $p = 19$ and $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_{19})$ the corresponding representation, then $\overline{\rho}|_{I_{\mathbb{Q}_{19}}}$ is a *nontrivial* extension of $\mathbf{1}$ by ω^{11}. Moreover, $\overline{\rho}|_{G_{\mathbb{Q}_{19}}}$ is the reduction of a crystalline representation which after restriction to inertia is a nontrivial extension of $\mathbf{1}$ by χ^{11}. However, any nontrivial extension of χ^{18} by any unramified twist of χ^{11} is *not* crystalline. In fact, we have $\{F(10,0)\} = W(\overline{\rho})$ in this case. One expects more generally that $W(\overline{\rho})$ is larger when $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is semisimple. Indeed, if $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is the reduction of an $\mathcal{O}$-lattice in $r : G_{\mathbb{Q}_p} \to \mathrm{GL}_2(E)$, then there is another $\mathcal{O}$-lattice (possibly after enlarging E) whose reduction is the semisimplification of $\overline{\rho}|_{G_{\mathbb{Q}_p}}$.

One approach to Theorem 4.5.1 is as follows. If $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ admits a local crystalline lift of Hodge–Tate weights $a+1$ and b, show that $\overline{\rho}$ admits a global lift ρ which is crystalline (at p) of Hodge–Tate weights $a+1$ and b. Then show that ρ comes from a modular form as predicted by the Fontaine–Mazur conjecture [FM95]. These steps can be executed using a combination of tools in the Taylor–Wiles method independently discovered by Khare–Wintenberger and Gee [KW09, Gee11].

4.6. **Vanishing conjectures for cohomology.** In the previous section, we were in the advantageous situation where $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ vanished outside of degree one for all $\mathcal{O}$-local systems $\mathcal{W}$ when $\mathfrak{m}$ is non-Eisenstein. While this is not true in general, this property does nevertheless admit a conjectural generalization. Let d_Y be the dimension of $Y(K_f)$. Define the integers $\ell_0 \stackrel{\mathrm{def}}{=} \mathrm{rk}\,\mathbb{G}(\mathbb{R}) - \mathrm{rk}\,A_\infty^\circ K_\infty^\circ$ and $q_0 = \frac{1}{2}(d_Y - \ell_0)$. (The group $\mathbb{G}$ admits discrete series if and only of $\ell_0 = 0$.)

Conjecture 4.6.1 ([CG18])**.** *Suppose that $\mathfrak{m}$ is non-Eisenstein. If $H^i(Y(K_f), \mathcal{W})_{\mathfrak{m}} \neq 0$, then $i \in [q_0, q_0 + \ell_0]$.*

[GN] shows that if Conjecture 4.6.1 holds, then so does the converse to Proposition 2.4.3 when $\mathbb{G}$ is a Weil restriction of GL_n, so that, as in our discussion above, it suffices to analyze Question 2.4.1 for irreducible mod p local systems (note that this is far from obvious in general, as it is *a priori* possible for the cohomology complex of irreducible local system to cancel each other out when they spread to several cohomological degrees). More seriously, Conjecture 4.6.1 plays a prominent role in the Taylor-Wiles patching method, which is the main tool to attack Question 2.4.1 in general, cf. Remark 5.1.2 below.

Unfortunately, there are few cases where Conjecture 4.6.1 is known. One trivial case is that of groups which are anisotropic modulo their center, e.g. definite unitary groups, when $Y(K_f)$ is a finite set of points. [CS17] have shown that for certain unitary groups ($\ell_0 = 0$) Conjecture 4.6.1 holds under some additional hypotheses. The case of the Weil restriction of GL_n (for $n > 1$) over a number field F (even CM fields) is open beyond the case $n = 2$ and F either totally real or imaginary quadratic.

5. Conjectures and results on the weight part of Serre's conjecture

5.1. **Taylor–Wiles patching.** Suppose for the moment that we are in a context where $\ell_0 = 0$ and Conjectures 4.3.1 and 4.6.1 hold (e.g., definite unitary groups). We fix p and E as before. We assume that a reductive integral model $G_{/\mathbb{Z}_p}$ of $\mathbb{G}_{/\mathbb{Q}_p}$ exists and continue to let $K_p = G(\mathbb{Z}_p)$. If $F(\lambda) \in W(\overline{\rho})$, then $H^*(Y(K_f), \mathcal{W}(\lambda))_{\mathfrak{m}}$ is nonzero (where $W(\lambda)$ is an $\mathcal{O}$-lattice in the irreducible algebraic representation $V(\lambda)$). In particular, $\overline{\rho}$ (attached to $\mathfrak{m}$) is the reduction of a crystalline representation of Hodge–Tate cocharacter $\lambda + \eta$. In light of Theorem 4.5.1, it is tempting to guess that the converse holds. However, counterexamples to this have been found for definite unitary groups in three variables [LLHLM18, Proposition 7.18]. The reason is that in contrast to the case

of GL_2, $W(\lambda)$ may include many Jordan–Hölder factors other than $F(\lambda)$. In fact, $F(\lambda)$ as a $\mathcal{O}[\![K_p]\!]$-module often does not lift to an $\mathcal{O}$-torsion-free module, which makes it more difficult to use the (expected) p-adic Hodge theoretic properties of Galois representations attached to automorphic forms. However, $F(\lambda)$ does lift *virtually*.

Suppose that $[F(\lambda)] = \sum_W c_W[W]$ in the Grothendieck group of $\mathbb{F}[\![K_p]\!]$-modules, where each W in the sum lifts to characteristic 0 (for example, one can take W running over the reductions of various $\mathcal{O}[G(\mathbb{F}_p)]$-modules using [Ser77, Theorem 33]). Then exactness gives us

$$[H^*(Y(K_f), \mathcal{F}(\lambda))_{\mathfrak{m}}] = \sum_W c_W [H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}].$$

Since the $\mathbb{F}$-vector spaces on the right hand side lift, their dimensions can in principle be computed in characteristic zero. However, they are of a global nature, and thus difficult to access. In contrast, we still expect that $W(\overline{\rho})$ depends only on $\overline{\rho}|_{G_{\mathbb{Q}_p}}$. The Taylor–Wiles method "patches" together cohomology functors (or rather, the total cohomology complex computing) $H^*(Y(K_f), \mathcal{F}_-)_{\mathfrak{m}}$ (for varying K_f) to obtain a functor $M_\infty(-)$ that plausibly depends (roughly speaking) only on $\overline{\rho}|_{G_{\mathbb{Q}_p}}$. Moreover, a control theorem guarantees that $M_\infty(\sigma)$ is nonzero if and only if $H^*(Y(K_f), \mathcal{F}_\sigma)_{\mathfrak{m}}$ is.

For a Galois representation $\overline{r} : G_{\mathbb{Q}_p} \to {}^L\mathbb{G}(\mathbb{F})$, let $R_{\overline{r}}$ denote the (framed) deformation ring parametrizing lifts $r : G_{\mathbb{Q}_p} \to {}^L\mathbb{G}(R)$ of $\overline{r}$ for complete local Noetherian $\mathcal{O}$-algebras R with residue field $\mathbb{F}$. Building on work of Kisin [Kis08] in the case when $\mathbb{G}_{/\mathbb{Q}_p}$ is a Weil restriction of GL_n, Balaji [Bal12] in particular defined a family of (reduced) semistable deformation rings $R_{\overline{r}}^{\lambda+\eta,\tau}$ whose $\overline{\mathbb{Q}}_p$-points correspond to potentially semistable Galois representations of Hodge–Tate cocharacter $\lambda+\eta$ and Galois type τ. In certain contexts where (enough of) an inertial local Langlands correspondence is known, one can define a finite dimensional locally algebraic $E[K_p]$-module $\sigma(\lambda,\tau) \overset{\mathrm{def}}{=} V(\lambda)\otimes_E \sigma(\tau)$. For example, when τ is tame, $\sigma(\tau)$ can be taken to be a certain combinatorially defined Deligne–Lusztig representation. For a ring A, let $A\text{-mod}^{\mathrm{fg}}$ denote the full subcategory of A-mod of finitely generated A-modules.

Axiom 5.1.1. There is an exact functor $M_\infty(-) : \mathcal{O}[\![K_p]\!]\text{-mod}^{\mathrm{fg}} \to R_{\overline{\rho}|_{G_{\mathbb{Q}_p}}}\text{-mod}^{\mathrm{fg}}$ with the following properties.

(1) For a Serre weight σ, $M_\infty(\sigma)$ is a maximal Cohen–Macaulay module on $R_{\overline{r}}^{\lambda+\eta,\tau} \otimes_{\mathcal{O}} \mathbb{F}$ for some λ and τ.
(2) $M_\infty(\sigma)$ is nonzero if and only if $H^*(Y(K_f), \mathcal{F}_\sigma)_{\mathfrak{m}}$ is nonzero.
(3) If $\sigma(\lambda,\tau)^\circ$ is an $\mathcal{O}$-lattice in $\sigma(\lambda,\tau)$, then $M_\infty(\sigma(\lambda,\tau)^\circ)$ is a maximal Cohen–Macaulay $R_{\overline{r}}^{\lambda+\eta,\tau}$-module of generic rank at most 1.
(4) If $\sigma(\lambda,\tau)^\circ$ is an $\mathcal{O}$-lattice in $\sigma(\lambda,\tau)$, then $M_\infty(\sigma(\lambda,\tau)^\circ)[1/p]$ is a generically free $R_{\overline{r}}^{\lambda+\eta,\tau}[1/p]$-module of rank 1.

Remark 5.1.2. (1) It may be impossible to arrange for the stringent rank conditions in Axiom 5.1.1, but the ranks can still be controlled and many arguments below can be successfully modified.
(2) If Conjecture 4.6.1 holds, then it is often possible to use the Taylor–Wiles method to construct a functor $M_\infty(-)$ satisfying the first three items of Axiom 5.1.1 with $\overline{r} \overset{\mathrm{def}}{=} \overline{\rho}|_{G_{\mathbb{Q}_p}}$ [CEG+16, GN], at least after adding formal variables to $R_{\overline{r}}^{\lambda+\eta,\tau}$ which we ignore. In general, the total (localized) cohomology complex may have non-vanishing cohomology groups in several degrees. The role of Conjecture 4.6.1 is to guarantee that after Taylor-Wiles patching, these complexes becomes concentrated into a single cohomological degree, i.e. they turn

into usual modules. This concentration effect relies on the "numerical coincidence" that powers the Taylor-Wiles method.

(3) It seems to be difficult to guarantee the last item (for all choices of (λ,τ)) except when $\mathbb{G}=\mathrm{GL}_2$ [Kis09]. Indeed, it is essentially equivalent to a modularity lifting result with very general p-adic Hodge theoretic hypotheses. However, there are some instances of specific (λ,τ) where the final item follows from the third: when $R^{\lambda+\eta,\tau}_{\overline{r}}$ is zero and when $R^{\lambda+\eta,\tau}_{\overline{r}}$ is a domain and $M_\infty(\sigma(\lambda,\tau)^\circ)$ is nonzero for some $\mathcal{O}$-lattice $\sigma(\lambda,\tau)^\circ\subset\sigma(\lambda,\tau)$.

Admitting Axiom 5.1.1 for the moment, there is the following strategy for determining $W(\overline{\rho})$. Let $d=\dim_E G_{/E}+\dim_E G_{/E}/B_{/E}$ where $B_{/E}\subset G_{/E}$ is a Borel subgroup. By a result of Kisin (see Theorem 6.1.1), d is the dimension of $R^{\lambda+\eta,\tau}_{\overline{r}}$ over $\mathcal{O}$ for any λ and τ. For each Serre weight σ, we write a presentation

$$[\sigma]=\sum_{(\lambda,\tau)} c_{\lambda,\tau}[\overline{\sigma(\lambda,\tau)}] \tag{5.1}$$

in the Grothendieck group of $\mathbb{F}[\![K_p]\!]$-modules. Then Axiom 5.1.1 guarantees that

$$Z(M_\infty(\sigma))=\sum_{(\lambda,\tau)} c_{\lambda,\tau}Z(R^{\lambda+\eta,\tau}_{\overline{r}}\otimes_{\mathcal{O}}\mathbb{F}), \tag{5.2}$$

where $Z(-)$ corresponds to taking the d-dimensional support cycle of the $R_{\overline{r}}$-module (note that all modules involved have support of dimension $\le d$). Since the right hand side of (5.2) only depends on $\overline{r}\overset{\mathrm{def}}{=}\overline{\rho}|_{G_{\mathbb{Q}_p}}$, so does the left hand side. In particular, this implies $W(\overline{\rho})$, being the set of σ with $Z(M_\infty(\sigma))\neq 0$ by the Cohen–Macaulay property, depends only on $\overline{r}$ as expected.

Another feature of the situation is that there are many possible choices of expressions (5.1), even if we restrict to the case when $c_{\lambda,\tau}=0$ for all $\lambda\neq 0$. Since the left hand side of (5.2) involves only σ, we get the surprising conclusion that the right hand side must be independent of the choice of expressions (5.1). In other words, there are many non-trivial relations between cycles of special fibers of potentially crystalline deformation rings. We summarize the above arguments as the following conjecture.

Conjecture 5.1.3. *(1)* [BM02, EG14] *The left hand side of* (5.2) *is independent of the presentation in* (5.1).

(2) [GHS18] *For any presentation as in* (5.1), *$\sigma\in W(\overline{\rho})$ if and only if the right hand side of* (5.2) *is nonzero.*

Remark 5.1.4. Conjecture 5.1.3(1) is purely local in the sense that it only involves $\mathbb{G}_{/\mathbb{Q}_p}$ and $\overline{r}$ while (2) is global since it involves $\overline{\rho}$. However, both follow from Axiom 5.1.1.

5.2. **Herzig's recipe.** The complexity of Conjecture 5.1.3 suggests that Question 2.4.2 does not admit a simple answer. Indeed, the case of GL_2 is perhaps misleading because of the simplicity of the geometry and representation theory involved. However, the class of semisimple representations $G_{\mathbb{Q}}\to {}^L\mathbb{G}(\mathbb{F})$ admits an essentially combinatorial classification, and so one could ask for a combinatorial description of $W(\overline{\rho})$ when $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is semisimple which generalizes Serre's recipe (see Remark 4.5.2). Herzig's recipe gives such a (conjectural) description.

We assume in this section that $\mathbb{G}_{/\mathbb{Q}_p}$ is unramified, with reductive integral model $G_{/\mathbb{Z}_p}$. We let K_p be $G(\mathbb{Z}_p)$. As before, we fix a twisting element η of $\mathbb{G}$. A *regular* Serre weight is a Serre weight $F(\lambda)$ with $0\le\langle\lambda,\alpha^\vee\rangle<p-1$ for all simple roots α. We define an involution $\mathcal{R}$ on the set of regular Serre weights $F(\lambda)$ by the formula

$$\mathcal{R}(F(\lambda))=F((-w_0(\eta),w_0)\cdot\lambda).$$

If $\overline{r} : G_{\mathbb{Q}_p} \to {}^L\mathbb{G}(\mathbb{F})$ is semisimple, then the restriction $\overline{r}|_{I_{\mathbb{Q}_p}}$ to the inertial subgroup factors through a torus. Its Teichmüller lift $[\overline{r}|_{I_{\mathbb{Q}_p}}] : G_{\mathbb{Q}_p} \to {}^L\mathbb{G}(E)$ is then a tame inertial type. Recall that to a tame inertial type τ (defined over E), one can attach through a tame inertial local Langlands a Deligne–Lusztig representation $\sigma(\tau)$ of $G(\mathbb{F}_p)$ (defined over E). Then K_p acts on $\sigma(\tau)$ by inflation, and we let $\overline{\sigma}(\tau)$ denote the semisimplification of the reduction of any K_p-stable $\mathcal{O}$-lattice in $\sigma(\tau)$. We have the following conjecture.

Conjecture 5.2.1 ([Her09, GHS18]). *The subset of regular Serre weights in $W(\overline{\rho})$ is $\mathcal{R}(\mathrm{JH}(\overline{\sigma}([\overline{\rho}|_{I_{\mathbb{Q}_p}}])))$.*

Conjectures 5.1.3 and 5.2.1 are of a rather different nature. For one thing, Conjecture 5.2.1 only applies to $\overline{\rho}$ which are semisimple locally at p, which is when one expects $W(\overline{\rho})$ is largest. However, Conjecture 5.2.1 is rather more explicit than Conjecture 5.1.3 when combined with Proposition 3.2.1.

5.3. **Results on the weight part of Serre's conjecture.** Conjecture 5.2.1 and a weakened version of Conjecture 5.1.3 is known when $\mathbb{G}$ is GL_2, the Weil restriction of the unit group in a quaternion algebra which is indefinite at no more than one archimedean place, or the Weil restriction of a definite unitary group in two variables under mild hypotheses (see Remark 4.5.2). Similar results [LLHLM18, LLHLM20b, LHLM22] are known for the Weil restrictions of definite unitary groups in three variables that are unramified at p under an additional *genericity* hypothesis.

Suppose now that $\mathbb{G}_{/E}$ is a product of GL_n over a finite set $\mathcal{J}$. For $\widetilde{w} \in \widetilde{W}$, $\widetilde{w}(0)$ is a tuple of elements in $\mathbb{Z}^n$ indexed by $\mathcal{J}$. We write $\widetilde{w}(0)_j \in \mathbb{Z}^n$ for the element corresponding to $j \in \mathcal{J}$. We say that a semisimple $\overline{r} : G_{\mathbb{Q}_p} \to {}^L\mathbb{G}(\mathbb{F})$ is sufficiently generic at p if $\sigma([\overline{r}|_{I_{\mathbb{Q}_p}}]) = R(\widetilde{w})$ where $0 \leq \langle \widetilde{w}(0), \alpha^\vee \rangle \leq p$ for all simple roots α and $p \nmid P(\widetilde{w}(0)_j)$ for an implicit polynomial $P \in \mathbb{Z}[X_1, \dots, X_n]$ which depends only on n (and not on p or j). If $\overline{r}$ is not semisimple, we say that it is sufficiently generic if its semisimplification is. In a precise sense, most $\overline{r}$'s are sufficiently generic as $p \to \infty$.

Theorem 5.3.1. *(1)* [LLHLM20a, Corollary 8.5.2] *Suppose that $\mathbb{G}_{/\mathbb{Q}_p}$ is an unramified Weil restriction of GL_n. Then Conjecture 5.1.3(1), restricted to presentations coming from Deligne–Lusztig representations, holds for sufficiently generic $\overline{r}$.*

(2) [LLHLM20a, Theorem 9.1.6] *Suppose that $\mathbb{G}$ is the Weil restriction of a definite unitary (over a nontrivial totally real extension of $\mathbb{Q}$) and that $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is sufficiently generic and semisimple. Then under mild Taylor–Wiles hypotheses, Conjectures 5.1.3(2) (for the restricted presentations in the previous item) and 5.2.1 hold.*

(3) [LLH] *Suppose that $\mathbb{G}$ is the Weil restriction of GL_n over a CM field and that $\overline{\rho}|_{G_{\mathbb{Q}_p}}$ is sufficiently generic and semisimple. Suppose moreover that $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}} \otimes_{\mathcal{O}} E$ is nonzero for some $\mathcal{W}$ corresponding to a Deligne–Lusztig representation. Then under mild Taylor–Wiles hypotheses, Conjectures 5.1.3(2) (for the restricted presentations in the previous item) and 5.2.1 hold.*

A critical tool to prove Theorem 5.3.1 is the following.

Theorem 5.3.2. *Suppose that $\overline{r} : G_{\mathbb{Q}_p} \to {}^L\mathbb{G}(\mathbb{F})$ is semisimple.*

(1) [LLHLM20a, Theorem 7.3.2(2)] *If τ is a sufficiently generic tame inertial type, then $R^{\eta,\tau}_{\overline{r}}$ is an integral domain (if it is nonzero).*

(2) [LLHLM20a, Proposition 6.2.7] *There exists a functor M_∞ (up to adding formal variables) satisfying Axiom 5.1.1(1)-(3), with $\mathbb{G}$ a Weil restriction of a definite unitary group, such that $M_\infty(\sigma(\tau)^\circ)$ is nonzero if $R^{\eta,\tau}_{\overline{r}}$ is nonzero for a sufficiently generic tame inertial type τ.*

Remark 5.3.3. (1) Theorem 5.3.2(2) combines the modularity of obvious weights [LLHL19] and the coherence conjecture for local models of Shimura varieties [Zhu14].
(2) In fact, Theorem 5.3.2 also holds for any $\lambda + \eta$ with λ dominant, though the implicit polynomial defining genericity depends on λ. See Theorem 6.1.3.

As alluded to in Remark 5.1.2(3), Theorem 5.3.2 implies the existence of a functor M_∞ satisfying Axiom 5.1.1(1)-(4) restricted to cases when $\lambda = 0$ and τ is a generic tame inertial type. The argument from §5.1 shows that Theorem 5.3.2 implies Theorem 5.3.1(1) for sufficiently generic *semisimple* $\overline{r}$. The nonsemisimple case follows from a simple argument using the global geometry of the Emerton–Gee stack relying on the fact that there is a semisimple $\overline{r}$ on every component (see Remark 6.2.3(1)). Moreover, Theorem 5.3.1(2) follows from Axiom 5.1.1(2).

Remark 5.3.4. The final part of Theorem 5.3.1 is a bit more subtle. When $\ell_0 = 0$, one expects $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}}$ to have little torsion. In contrast when $\ell_0 > 0$, one expects cohomology to be dominated by torsion and characteristic 0 classes to be rare. This means that the lifting hypothesis, i.e. that $H^*(Y(K_f), \mathcal{W})_{\mathfrak{m}} \otimes_{\mathcal{O}} E$ is nonzero in Theorem 5.3.1(3), is quite restrictive. This condition could be removed if one knew Conjecture 4.6.1 (see Remark 5.1.2). In lieu of this, the lifting hypothesis can be used to make an argument with Euler characteristics of the functor M_∞, whose image is a priori an object in $D^b(R_{\overline{\rho}|_{G_{\mathbb{Q}_p}}}\text{-mod}^{\mathrm{fg}})$, adopting Taylor's Ihara avoidance trick to this setting [ACC+18].

6. Local models for potentially crystalline deformation rings

The heart of the proof of Theorem 5.3.1 reduces, via the Taylor-Wiles method, to understanding the support of the patched modules $M_\infty(\sigma)$ in Axiom 5.1.1, and ultimately to geometric properties of the potentially crystalline deformation rings $R^{\lambda,\tau}_{\overline{r}}$. We achieve this by introducing and analyzing certain (finite type) group-theoretic moduli spaces which algebraize these deformation rings.

Fix a prime p. In this section, we will restrict to the case $\mathbb{G}_{/\mathbb{Q}_p} = \mathrm{Res}_{K/\mathbb{Q}_p}\mathrm{GL}_n$ for an unramified extension $K = \mathbb{Q}_{p^f}$ of $\mathbb{Q}_p$. In particular, we have a reductive integral model $G_{/\mathbb{Z}_p} = \mathrm{Res}_{\mathcal{O}_K/\mathbb{Z}_p}\mathrm{GL}_n$ of $\mathbb{G}$. Recall from §2.3 that E is a sufficiently large finite extension of $\mathbb{Q}_p$ with ring of integers $\mathcal{O}$, uniformizer ϖ, and residue field $\mathbb{F}$.

6.1. **Potentially crystalline deformation rings.** Let $\overline{r} : G_K \to \mathrm{GL}_n(\mathbb{F})$ be a mod p local Galois representation. Recall that we have $R_{\overline{r}}$ the (framed) deformation ring that classifies lifts of $\overline{r}$, and the $\overline{\mathbb{Q}}_p$-points of $R_{\overline{r}}$ correspond to p-adic Galois representations of G_K (lifting $\overline{r}$). Given a Hodge–Tate cocharacter λ and inertial type τ, one has the potentially crystalline deformation ring $R^{\lambda,\tau}_{\overline{r}}$ which is characterized as the unique reduced p-flat quotient of $R_{\overline{r}}$ whose $\overline{\mathbb{Q}}_p$-points correspond to lifts $r : G_K \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ which are potentially crystalline of type (λ, τ) (i.e. the Hodge–Tate weights of r are given by λ and $\mathrm{WD}(r)$ induces the inertial type τ). In the setting of GL_n, these rings were first constructed by Kisin, who also established their basic properties [Kis08].

Theorem 6.1.1 (Kisin)**.** *(1) $R^{\lambda,\tau}_{\overline{r}}[\frac{1}{p}]$ is regular.*
(2) $\dim_{\mathcal{O}} R^{\lambda,\tau}_{\overline{r}} = \dim_E G_{/E} + \dim_E G_{/E}/P_{\lambda/E}$, where P_λ is the parabolic subgroup corresponding to λ. In particular $\dim_{\mathcal{O}} R^{\lambda,\tau}$ is a constant d as λ varies over regular dominant cocharacters.

When λ is *regular dominant*, the rings $R^{\lambda,\tau}_{\overline{r}}$ play a pivotal role in the Taylor-Wiles method: they act on patched spaces of automorphic forms $M_\infty(\sigma(\lambda - \eta, \tau))$, which govern questions about modularity and congruences (cf. Axiom 5.1.1(3)). Even better, they are maximal Cohen–Macaulay modules, and hence must be supported on a union of irreducible components.

For global applications, it is essential to understand global properties of $R^{\lambda,\tau}_{\overline{r}}$ such as irreducibility. This turns out to be a notoriously difficult problem. There are roughly two reasons for this.

- Outside some special cases, $R^{\lambda,\tau}_{\overline{r}}$, being characterized by its $\overline{\mathbb{Q}}_p$-points, has no known moduli interpretation. This is related to the fact that integral p-adic Hodge theory is much less well understood than rational p-adic Hodge theory.
- The internal structure of $R^{\lambda,\tau}_{\overline{r}}$ is intrinsically complicated in general. Thus, one can not expect to have completely explicit descriptions for all λ and τ.

The second point is best illustrated by the Breuil–Mézard conjecture, which quantifies the complexity of the special fibers of $R^{\lambda+\eta,\tau}_{\overline{r}}$ as λ and τ vary in terms of the mod p representation theory of $\mathrm{GL}_n(\mathcal{O}_K)$ (we shift from λ to $\lambda+\eta$ for the rest of this subsection to be consistent with §5.1). We let $Z(R^{\lambda+\eta,\tau}_{\overline{r}}/\varpi)$ denote the d-dimensional cycle of $R^{\lambda+\eta,\tau}_{\overline{r}}/\varpi$, which counts the irreducible components with appropriate multiplicities. For a $\mathrm{GL}_n(\mathcal{O}_K)$-representation V over E, recall that $\overline{V}$ denotes the $\mathrm{GL}_n(\mathcal{O}_K)$-representation over $\mathbb{F}$ which is the semisimplification of the reduction modulo ϖ of any $\mathrm{GL}_n(\mathcal{O}_K)$-stable $\mathcal{O}$-lattice in V. The following is a reformulation of Conjecture 5.1.3(1).

Conjecture 6.1.2 (Breuil–Mézard, Emerton-Gee)**.** *There exist d-dimensional cycles $\mathcal{Z}_\sigma(\overline{r})$ in* $\mathrm{Spec}\, R_{\overline{r}}/\varpi$ *for each irreducible $\mathrm{GL}_n(\mathcal{O}_K)$-representation σ over $\mathbb{F}$ (i.e. a Serre weight for $G(\mathbb{F}_p)$) such that for all τ and λ,*

$$Z(R^{\lambda+\eta,\tau}_{\overline{r}}/\varpi) = \sum_\sigma m_{\lambda,\tau}(\sigma)\mathcal{Z}_\sigma(\overline{r}),$$

where $m_{\lambda,\tau}(\sigma)$ denotes the multiplicity of σ in $\overline{\sigma(\lambda,\tau)}$.

In other words, the special fibers $R^{\lambda+\eta,\tau}_{\overline{r}}/\varpi$ are built out of a *finite list* of basic cycles $\mathcal{Z}_\sigma(\overline{r})$, with multiplicities governed by the purely representation theoretic quantities $m_{\lambda,\tau}(\sigma)$. Conjecture 6.1.2 is known when $n = 2$ and $\lambda = 0$ by work of Gee and Kisin [GK14]. When τ is a generic tame type, $m_{0,\tau}(\sigma) = 1$ for 2^f Serre weights σ and is zero otherwise. In general, the quantities $m_{\lambda,\tau}$ are very complicated: if $\lambda = 0$ and τ is tame, $m_{\lambda,\tau}(\sigma)$ computes the multiplicities of a mod p Deligne–Lusztig representation, which for generic τ is given by periodic Kazhdan–Lusztig polynomials. In particular, as the rank of G grows, the special fibers $R^{\lambda+\eta,\tau}_{\overline{r}}/\varpi$ tend to be highly non-reduced.

As explained in §5.1, to prove Theorem 5.3.1, one needs to establish Axiom 5.1.1, particularly the main bottleneck (4). We do this by proving Theorem 5.3.2.

Theorem 6.1.3. [LLHLM20a, Theorem 7.3.2(2)] *Assume that $\overline{r}$ is* semisimple *and τ is a tame inertial type which is sufficiently generic relative to λ (in the sense of §5.3). Then $R^{\lambda+\eta,\tau}_{\overline{r}}$ is a domain (or zero).*

Remark 6.1.4. (1) Explicit computations that suggest that Theorem 6.1.3 is *false* without the tameness assumption on $\overline{r}$ when $n > 2$ unless $n = 3$ and $\lambda = 0$.

(2) If $R^{\lambda+\eta,\tau}_{\overline{r}} \neq 0$, then sufficient genericity of τ implies that of $\overline{r}$ and vice versa (generally with different choices of implicit polynomials). Because of this, the conclusion of Theorem 6.1.3 also holds if we let $\overline{r}$ be tame and sufficiently generic but impose no genericity hypothesis on τ.

6.2. **The Emerton-Gee stack.** In [EGa], Emerton–Gee constructed the moduli stack $\mathcal{X}_n$ over $\mathrm{Spf}\,\mathcal{O}$ of rank n étale (φ,Γ)-modules. By its construction, $\mathcal{X}_n$ interpolates framed deformation rings in the sense that the set $\mathcal{X}_n(\overline{\mathbb{F}}_p)$ is in bijection with the set of continuous representations $\overline{r} : G_K \to \mathrm{GL}_n(\overline{\mathbb{F}}_p)$, and framed deformation rings $R_{\overline{r}}$ are versal rings (in the sense of [EGb, Definition 2.2.9]) for $\mathcal{X}_n$. Furthermore, for a Hodge–Tate cocharacter λ and an inertial type τ, they

construct a p-flat p-adic formal algebraic closed substack $\mathcal{X}^{\lambda,\tau}$ which is characterized by the property that its points over any finite flat $\mathcal{O}$-algebra correspond to potentially crystalline representations r of type (λ,τ). Thus $\mathcal{X}^{\lambda,\tau}$ interpolates the potentially crystalline deformation rings $R^{\lambda,\tau}_{\overline{r}}$ as $\overline{r}$ varies.

The basic properties of these stacks are as follows:

Theorem 6.2.1 (Emerton–Gee)**.** *(1)* [EGa, Corollary 5.5.18] *$\mathcal{X}_n$ is a Noetherian formal algebraic stack.*

(2) [EGa, Theorem 4.8.12] *$\mathcal{X}^{\lambda,\tau}$ is a p-flat p-adic formal algebraic stack of dimension* $\dim G_{/E}/P_{\lambda/E}$.

(3) [EGa, Theorem 6.5.1] *The irreducible components of the underlying reduced stack $\mathcal{X}_{n,\mathrm{red}}$ are in bijection with the Serre weights of $G(\mathbb{F}_p)$.*

We let $\mathcal{C}_\sigma$ be the irreducible component labelled by σ. Let $\mathcal{Z}_{\lambda+\eta,\tau}$ denote the top dimensional cycle of $\mathcal{X}^{\lambda+\eta,\tau}/\varpi$, which has dimension independent of λ since $\lambda+\eta$ is regular dominant. One has the following interpolation of the Breuil–Mézard conjecture over $\mathcal{X}_n$:

Conjecture 6.2.2 (Conjecture 8.2.2 [EGa])**.** *For each Serre weight σ, there exists an effective top-dimensional cycle $\mathcal{Z}_\sigma$ on $\mathcal{X}_{n,\mathrm{red}}$ such that for all λ and inertial types τ, we have*

$$\mathcal{Z}_{\lambda+\eta,\tau}=\sum_\sigma m_{\lambda,\tau}(\sigma)\mathcal{Z}_\sigma.$$

Remark 6.2.3. (1) Conjecture 6.2.2 recovers Conjecture 6.1.2 by taking versal rings at $\overline{r}$. Conversely, knowledge of Conjecture 6.1.2 at *sufficiently many* $\overline{r}$ would imply Conjecture 6.2.2. This gives a mechanism to deduce Conjecture 6.1.2 for more general $\overline{r}$ from a few "basic $\overline{r}$". This allows us to reduce Theorem 5.3.1(1) to the case of semisimple $\overline{r}$.

(2) In [LLHLM20a, §8], using Taylor-Wiles patching, we constructed cycles $\mathcal{Z}_\sigma$ for sufficiently generic σ, which satisfies a (finite) subset of the equations postulated in Conjecture 6.2.2. As the conjectural cycles in Conjecture 6.2.2 is expected to be compatible with Taylor-Wiles patching, the cycles constructed in *loc. cit.* should be the "correct" ones.

(3) (cf. [LLHLM20a, Remark 1.4.11]) One expects $\mathcal{Z}_\sigma$ to contain the irreducible component $\mathcal{C}_\sigma$ with multiplicity one. That is, one should have a decomposition

$$\mathcal{Z}_\sigma=\sum_{\sigma'} b_{\sigma',\sigma}\mathcal{C}_{\sigma'}$$

with $b_{\sigma',\sigma}\geq 0$ and $b_{\sigma,\sigma}=1$. This is indeed true in the cases studied in [LLHLM20a, §8] and [GK14]. For example, in the setting of [GK14], the cycles $\mathcal{Z}_\sigma=\mathcal{C}_\sigma$, unless σ is a twist of the Steinberg weight (in particular such σ would be non-generic), in which case $\mathcal{Z}_\sigma$ is $\mathcal{C}_\sigma+\mathcal{C}_{\sigma'}$ for a suitable σ' (cf. [EGa, Theorem 8.6.2]). For $n>3$, it is quite difficult to compute $b_{\sigma',\sigma}$, and one does not expect $\mathcal{Z}_\sigma=\mathcal{C}_\sigma$ in general, even for generic σ. This is analogous to the situation of the locally analytic Breuil-Mézard conjecture studied in [BHS19].

6.3. **Local models and their geometric properties.** Let $L\mathcal{G}$ be the loop group, which is the ind-group scheme given by $L\mathcal{G}(R)=\mathrm{GL}_n(R(\!(v+p)\!))$ for any $\mathcal{O}$-algebra R. Consider the positive loop group scheme $L^+\mathcal{G}$ over $\mathcal{O}$ sending an $\mathcal{O}$-algebra R to the subgroup of $\mathrm{GL}_n(R[\![v+p]\!])$ consisting of matrices that are upper triangular mod v. Note that when p is invertible in R, $L^+\mathcal{G}(R)=\mathrm{GL}_n(R[\![v+p]\!])$ is the positive loop group for GL_n, whereas when $p=0$ in R, $L^+\mathcal{G}(R)=\mathcal{I}(R)$, the standard Iwahori group scheme.

The quotient $L^+\mathcal{G}\backslash L\mathcal{G}$ is represented by an ind-proper $\mathcal{O}$-ind-scheme $\mathrm{Gr}_{\mathcal{G}}$. This is a mixed characteristic version of the degeneration of affine Grassmannians introduced by Gaitsgory: indeed its generic fiber $\mathrm{Gr}_{\mathcal{G},E}$ is isomorphic to an affine Grassmannian, while the special fiber $\mathrm{Gr}_{\mathcal{G},\mathbb{F}}$ is isomorphic to the affine flag variety Fl.

The affine Grassmannian has the affine Schubert stratification $\mathrm{Gr}_{\mathcal{G},E} = \bigcup_\lambda L^+\mathcal{G}_E \backslash L^+\mathcal{G}_E(v+p)^\lambda L^+\mathcal{G}_E$, where λ runs over dominant coweights of GL_n. Similarly, the affine flag variety $\mathrm{Fl} = \bigcup_{\widetilde{w}} \mathcal{I}\backslash\mathcal{I}\widetilde{w}\mathcal{I}$, where $\widetilde{w}$ runs over the extended affine Weyl group $\widetilde{W}$.

For dominant λ, the Pappas–Zhu local model $M(\leq\lambda)$ is the Zariski closure of $L^+\mathcal{G}_E\backslash L^+\mathcal{G}_E(v+p)^\lambda L^+\mathcal{G}_E$ in $\mathrm{Gr}_{\mathcal{G}}$, cf. [PZ13].

Let $\mathbf{a} \in \mathcal{O}^n$. We now consider the condition

$$(\star) \qquad v\frac{dA}{dv}A^{-1} + A\mathrm{Diag}(\mathbf{a})A^{-1} \in \left(\frac{1}{v+p}\right) \mathrm{Lie}\, L^+\mathcal{G}$$

for $A \in L\mathcal{G}(R)$. This is an approximation to the monodromy condition coming from p-adic Hodge theory. This condition clearly descends to a closed condition on $\mathrm{Gr}_{\mathcal{G}}$.

Definition 6.3.1. The local model $M(\lambda, \nabla_{\mathbf{a}})$ is the Zariski closure in $M(\leq\lambda)$ of the locus cut out by $(\star)$ in $L^+\mathcal{G}_E\backslash L^+\mathcal{G}_E(v+p)^\lambda L^+\mathcal{G}_E$.

Note that right multiplication by the constant diagonal torus $T^\vee$ preserves $(\star)$. (Here, $T^\vee$ is a maximal torus in GL_n which is the dual group $G^\vee$ of the group $G = \mathrm{GL}_n$ which appeared in §3.1.) Thus, $M(\lambda, \nabla_{\mathbf{a}})$ inherits a $T^\vee$-action compatible with the $T^\vee$-action on $M(\leq\lambda)$.

By contemplating the interaction of condition $(\star)$ with the affine Schubert stratification, one observes:

- [LLHLM20a, Proposition 4.1.1] $M(\lambda, \nabla_{\mathbf{a}})_{/E}$ is isomorphic to $P_\lambda\backslash\mathrm{GL}_n$, hence is smooth and irreducible.
- [LLHLM20a, Theorem 4.2.4] Provided $\mathbf{a}$ mod p sufficiently regular, the locus cut out by $(\star)$ in each open Schubert cell $\mathcal{I}\backslash\mathcal{I}\widetilde{w}\mathcal{I} \subset \mathrm{Fl}$ is an affine space, with dimension combinatorially determined by $\widetilde{w}$.

Thus $M(\lambda, \nabla_{\mathbf{a}})$ is a degeneration of a partial flag variety, and one has control over its *reduced* special fiber.

Example 6.3.2. (1) For example, when $n = 2$ and $\lambda = (1, 0)$, condition $(\star)$ is empty, and $M(\lambda, \nabla_{\mathbf{a}}) = M(\leq\lambda)$ is a degeneration of $\mathbb{P}^1$ into a union of two $\mathbb{P}^1$ crossing transversely at a point. More generally, one has $M(\lambda, \nabla_{\mathbf{a}}) = M(\leq\lambda)$ if and only if λ is minuscule.

(2) Suppose $n = 3$ and $\lambda = (2, 1, 0)$, and $\mathbf{a}$ mod p is sufficiently regular. Then $\dim M(\leq\lambda) = 4$, whereas $\dim M(\lambda, \nabla_{\mathbf{a}}) = \dim B\backslash\mathrm{GL}_3 = 3$. The special fiber $M(\lambda, \nabla_{\mathbf{a}})_{\mathbb{F}}$ is reduced and has 9 irreducible components, six of which are isomorphic to the flag variety $B\backslash\mathrm{GL}_3$, while the remaining three are more complicated rational smooth varieties. Already in this case, the behavior of the intersections among the irreducible components is somewhat elaborate, cf. [LHLM22].

Remark 6.3.3. Around each point $\widetilde{z} \in \mathrm{Fl}$, one can write down an explicit open neighborhood $\mathcal{U}(\widetilde{z})$ of $\mathrm{Gr}_{\mathcal{G}}$ using the theory of the "big cell". This allows us to in principle give explicit coordinate charts for $M(\lambda, \nabla_{\mathbf{a}})$: the coordinate charts parametrize matrices A with polynomial entries whose degrees are bounded in terms of $\widetilde{z}$, and one then imposes elementary divisor conditions dictated by λ together with the explicit equation $(\star)$ and takes the p-saturation of the result. It is the p-saturation operation that makes this description rather difficult to work with.

In order to establish the connection between the above models to Galois deformation theory, we have to understand the behavior of $M(\lambda, \nabla_{\mathbf{a}})$ under completion. The essential difficulty is that an irreducible variety may break up into formal branches in some complicated way after completions: its singularities may not be *unibranch*. Unfortunately, $M(\lambda, \nabla_{\mathbf{a}})$ fails to be unibranch in general, and in such situations it is difficult to control the subset of the formal branches that are related

to Galois deformation theory. Fortunately, it turns out there is supply of special points where this difficulty does not manifest:

Theorem 6.3.4. ([LLHLM18, Theorem 3.7.1]) *There exists a nonzero polynomial $P \in \mathbb{Z}[X_1, \dots, X_n]$ such that if $P(\mathbf{a}) \neq 0 \mod p$, then for any T-fixed point $x \in M(\lambda, \nabla_{\mathbf{a}})(\overline{\mathbb{F}}_p)$, the completed local ring $\mathcal{O}^{\wedge}_{M(\lambda,\nabla_{\mathbf{a}}),x}$ is a domain (i.e., $M(\lambda, \nabla_{\mathbf{a}})$ is unibranch at its T-fixed points).*

This key result, whose proof we now sketch, underlies everything else. One first observes that the theorem holds (under a mild assumption on the characteristic) for the equal characteristic analogues of $M(\lambda, \nabla_{\mathbf{a}})$ where E is replaced by $\mathbb{F}(\!(t)\!)$. In this function field setting, there is an additional symmetry: there is an extra $\mathbb{G}_m$-action given by "loop rotation" which scales t. This implies that the T-fixed points look like cone points, i.e. the fixed point of an attracting torus action, and one observes that cone points are unibranch. We then deduce the mixed characteristic case by a spreading out argument. The essential point here is that unibranch can be phrased in terms of connectedness of fibers of the normalization map, and normalization is preserved by generic base change. This explains the occurence of the universal polynomial P: its vanishing locus is the obstruction to certain properties being preserved under base change.

6.4. **Local models and Emerton–Gee stacks.** Recall that we fixed a finite unramified extension $K/\mathbb{Q}_p$. Let k be the residue field of K. Let $\mathcal{J}$ be the set of embeddings $\mathrm{Hom}_{\mathbb{Q}_p}(K, \overline{\mathbb{Q}}_p)$ which we identify with $\mathrm{Hom}_{\mathbb{Q}_p}(K, E) = \mathrm{Hom}(k, \mathbb{F})$ using the inclusion $E \subset \overline{\mathbb{Q}}_p$.

To any *tame* inertial type τ for I_K, one can associate a collection $\mathbf{a}_\tau = (\mathbf{a}_{\tau,j})_{j\in\mathcal{J}}$, where $\mathbf{a}_{\tau,j} \in \mathcal{O}^n$ records the inertial weights of τ.

In the "lowest alcove" principal series case, $\mathbf{a}_\tau$ is defined so that τ is the direct sum

$$\bigoplus_{i=1}^{n} \prod_{j\in\mathcal{J}} j \circ \omega_K^{\mathbf{a}^{(i)}_{\tau,j}}$$

where $\omega_K : I_K \to k^\times$ is the reduction of the Lubin–Tate character $I_K \to \mathcal{O}_K^\times$. Set $\lambda = (\lambda_j)_{j\in\mathcal{J}} \in (\mathbb{Z}^n)^{\mathcal{J}}$, a Hodge–Tate cocharacter. Define

$$M_{\mathcal{J}}(\lambda, \nabla_{\mathbf{a}_\tau}) = \prod_{j\in\mathcal{J}} M(\lambda_j, \nabla_{\mathbf{a}_{\tau,j}})$$

where, for each $j \in \mathcal{J}$, the local models $M(\lambda_j, \nabla_{\mathbf{a}_{\tau,j}})$ are those appearing in Definition 6.3.1.

The relationship between the local models and the Emerton–Gee stacks is given by the following:

Theorem 6.4.1. ([LLHLM18, Theorem 7.3.2]) *If τ is sufficiently generic (with respect to λ), then there exist Zariski open covers $\bigcup_{\widetilde{z}} \mathcal{X}_{\mathrm{reg}}^{\leq\lambda,\tau}(\widetilde{z})$ and $\bigcup_{\widetilde{z}} U_{\mathrm{reg}}(\widetilde{z}, \leq\lambda, \nabla_{\mathbf{a}_\tau})^{\wedge_p}$ of $\bigcup_{\substack{\lambda'\leq\lambda \\ \lambda'\text{reg. dom.}}} \mathcal{X}^{\lambda',\tau}$ and $\bigcup_{\substack{\lambda'\leq\lambda \\ \lambda'\text{reg. dom.}}} M(\lambda', \nabla_{\mathbf{a}_\tau})^{\wedge_p}$, respectively, such that for each $\widetilde{z}$, there exists a local model diagram*

(6.1)

$$\begin{array}{ccc} & \widetilde{\mathcal{X}}_{\mathrm{reg}}^{\leq\lambda,\tau}(\widetilde{z}) & \\ \swarrow & & \searrow \\ \mathcal{X}_{\mathrm{reg}}^{\leq\lambda,\tau}(\widetilde{z}) & & U_{\mathrm{reg}}(\widetilde{z}, \leq\lambda, \nabla_{\mathbf{a}_\tau})^{\wedge_p} \end{array}$$

Remark 6.4.2. (1) In the above diagram, the T-fixed points of the local models have a simple Galois theoretic interpretation: they correspond to semisimple $\overline{r}$.

(2) When $\lambda = \eta$, one has $\bigcup_{\substack{\lambda' \leq \lambda \\ \lambda' \text{reg. dom.}}} \mathcal{X}^{\lambda',\tau} = \mathcal{X}^{\eta,\tau}$. Since potentially crystalline deformation rings of type (η, τ) are versal rings to $\mathcal{X}^{\eta,\tau}$, we see that they appear (up to smooth modifications) as the completion of local rings of $M(\eta, \nabla_{\mathbf{a}_\tau})$ at closed points. In particular, we deduce the irreducibility of the potentially crystalline deformation rings of Theorem 6.1.3 from the unibranch property of the local models (at the appropriate points). This completes the proof of Theorem 5.3.1(1) and the first half of Theorem 5.3.1(2) (see Remark 6.2.3(1)).

(3) Combining the theorem with Remark 6.3.3, we get an algorithm to write down explicit presentations of (unions of) $R^{\lambda,\tau}_{\overline{r}}$ (for regular λ).

We now give a slightly simplified outline of the proof of Theorem 6.4.1. The construction of the stacks $\mathcal{X}^{\lambda,\tau}$ comes in two steps:

- Using integral p-adic Hodge theory, one can attach Breuil–Kisin modules to lattices in (potentially) crystalline representations for G_K. Thus the first step is to construct a moduli stack of Breuil–Kisin modules $Y^{\leq\lambda,\tau}$ with tame descent data of type (λ, τ).
- As not all Breuil–Kisin modules come from lattices in (potentially) crystalline representations, one needs cut down $Y^{\lambda,\tau}$ by appropriate conditions to get $\mathcal{X}^{\lambda,\tau}$.

Accordingly, the proof is divided into two steps:

- In the first step, we show that $Y^{\leq\lambda,\tau}$ is locally modelled by the Pappas-Zhu model $M(\leq\lambda)$. This is not surprising, as Breuil–Kisin modules are a projective $\mathcal{O}_K[\![u]\!]$-modules with certain semi-linear structures, and thus are closely related to points of $\mathrm{Gr}_{\mathcal{G}}$. Using the open cover $\mathrm{Gr}_{\mathcal{G}} = \bigcup_{\widetilde{z}} \mathcal{U}(\widetilde{z})$ (cf. Remark 6.3.3), we get an analogue of the local model diagram (6.1) for $Y^{\leq\lambda,\tau}$ and induced open affine covers on every object in sight.
- After the first step, we get *two* closed substacks of $Y^{\leq\lambda,\tau}(\widetilde{z})$: the substack $\mathcal{X}^{\leq\lambda,\tau}(\widetilde{z})$ and the substack $\mathcal{X}^{\leq\lambda,\tau,\star}(\widetilde{z})$ induced by the p-adic completion of $\bigcup_{\lambda'\leq\lambda} M(\lambda', \nabla_{\mathbf{a}_\tau})$ along the local model diagram for $Y^{\leq\lambda,\tau}$. They are genuinely different substacks, because condition $(\star)$ is only the "first order term" of the condition cutting out $\mathcal{X}^{\leq\lambda,\tau}$ inside $Y^{\leq\lambda,\tau}$.

  However, the two substacks are p-adically close, and using the smoothness of the generic fiber of $M(\lambda, \nabla_{\mathbf{a}})$, one can produce a non-canonical embedding $\mathcal{X}^{\leq\lambda,\tau}(\widetilde{z}) \hookrightarrow \mathcal{X}^{\leq\lambda,\tau,\star}(\widetilde{z})$. Since both stacks turn out to have the same dimension, the maximal dimensional part $\mathcal{X}^{\leq\lambda,\tau}_{\mathrm{reg}}(\widetilde{z})$ of $\mathcal{X}^{\leq\lambda,\tau}(\widetilde{z})$ embeds into the maximal dimension part of $\mathcal{X}^{\leq\lambda,\tau,\star}(\widetilde{z})$. Now, using the results of [LLHL19] (which ultimately uses Taylor–Wiles patching, and hence automorphic forms), one obtains a lower bound on the number of irreducible components (of the spectrum of the structure sheaf) of the former, while Theorem 6.3.4 gives the same upper bound for the number of irreducible components (of the spectrum of the structure sheaf) of the latter. Thus the two maximal dimension parts are (non-canonically) isomorphic to each other, which concludes the proof.

As the above outline suggests, the arrows in the local model diagram are produced by Hensel-type lifting arguments, and thus are highly non-canonical. However, modulo p this issue disappears, and the local model diagrams on the open cover produced by Theorem 6.4.1 glue together. Consequently, the analysis of irreducible components of the special fibers of local models implies the following.

Theorem 6.4.3. *For τ sufficiently generic (with respect to λ):*

(1) $\mathcal{X}^{\lambda+\eta,\tau}_{\mathrm{red}} = \cup_\sigma \mathcal{C}_\sigma$, where the union runs over all Serre weights $\sigma \in \mathrm{JH}(\overline{\sigma(\lambda,\tau)})$.

(2) There is a natural bijection between the irreducible components of $\overline{M}(\lambda + \eta, \nabla_{\mathbf{a}_\tau})$ and the Jordan–Hölder factors of $\overline{\sigma(\lambda,\tau)}$.

(3) For each $\sigma \in \mathrm{JH}(\overline{\sigma(\lambda,\tau)})$, we have a mod p local model diagram:

(6.2)
$$\begin{array}{ccc} & \widetilde{\mathcal{C}}_\sigma & \\ \swarrow & & \searrow \\ \mathcal{C}_\sigma & & \overline{M}(\lambda+\eta,\nabla_{\mathbf{a}_\tau})_\sigma \end{array}$$

where $\overline{M}(\lambda+\eta,\nabla_{\mathbf{a}_\tau})_\sigma$ is the irreducible component of $\overline{M}(\lambda+\eta,\nabla_{\mathbf{a}_\tau})$ labelled by σ and both arrows are torsors for the torus $(T^\vee)^{\mathcal{J}}$ (with respect to different $(T^\vee)^{\mathcal{J}}$-actions).

Remark 6.4.4. (1) The proof of Theorem 6.4.3 does not go through Theorem 6.4.1. Because of that it holds under much milder genericity conditions compared to our other theorems: if $\sigma(\tau) = R$, we only require that R is m-generic for m sufficiently large depending on λ (larger than both $2\langle\lambda,\alpha^\vee\rangle + 2$ and $4n + \langle\lambda,\alpha\rangle$ for all roots α).

(2) Over $\mathbb{F}$, equation $(\star)$ becomes the equation cutting out a *deformed* affine Springer fiber in Fl, cf. [FZ10]. Thus the irreducible components $\overline{M}(\lambda+\eta,\nabla_{\mathbf{a}_\tau})_\sigma$ are irreducible components of a (product of) *deformed* affine Springer fiber(s). (It is immediate from the aforementioned [LLHLM20a, Theorem 4.2.4] that these irreducible components are rational varieties.) Theorem 6.4.3 then allows us to get a handle on the irreducible component $\mathcal{C}_\sigma$ of the reduced Emerton–Gee stack for generic σ. In particular, one can get a description of the semisimple points on $\mathcal{C}_\sigma$, and this is the critical ingredient for the verification of Herzig's recipe (Theorems 5.3.1(2) and 5.3.1(3)).

(3) The irreducible components $\overline{M}(\lambda+\eta,\nabla_{\mathbf{a}_\tau})_\sigma$ are fairly easy to implement on computer algebra systems such as Macaulay2. For any given n, this allows us to probe the structure of $\mathcal{C}_\sigma$ in a purely algorithmic manner.

Example 6.4.5. (1) (Fontaine-Laffaille components) For σ in the lowest alcove, the corresponding irreducible component of the deformed affine Springer fiber is isomorphic to a product of flag varieties $(B\backslash\mathrm{GL}_n)^{\mathcal{J}}$. We deduce from this that

$$\mathcal{C}_\sigma = [(N\backslash\mathrm{GL}_n)^{\mathcal{J}}/T^{\mathcal{J}}]$$

where N is the subgroup of unipotent upper triangular matrices, and $T^{\mathcal{J}}$ acts via *shifted conjugation*: $(t_j)\cdot(Ng_j) = (Nt_jg_jt_{j\circ\varphi}^{-1})$, where $\circ\varphi$ denotes pre-composition with Frobenius on K. In particular, for $n = 2$, all components $\mathcal{C}_\sigma$ for generic σ are of this form.

(2) When $n = 3$, there are two types of irreducible components at each factor $j \in \mathcal{J}$: the flag variety $B\backslash\mathrm{GL}_3$ or a more complicated rational and smooth variety (this can be extracted, for example, from the description of minimal primes in [LHLM22, Table 3]). In particular, there are 2^f types of $\mathcal{C}_\sigma$ (for generic σ) which correspond to the possible p-alcoves containing the highest weight of σ.

(3) For $n = 4$, there are generic σ for which $\mathcal{C}_\sigma$ is singular (e.g. for σ with highest weight in the highest p-restricted alcove). Thus the smoothness of $\mathcal{C}_\sigma$ appears to be a low rank coincidence.

References

[ACC+18] Patrick B. Allen, Frank Calegari, Ana Caraiani, Toby Gee, David Helm, Bao V. Le Hung, James Newton, Peter Scholze, Richard Taylor, and Jack A. Thorne, *Potential automorphy over CM fields*, 2018.

[ADP02] Avner Ash, Darrin Doud, and David Pollack, *Galois representations with conjectural connections to arithmetic cohomology*, Duke Math. J. **112** (2002), no. 3, 521–579. MR 1896473

[Bal12] Sundeep Balaji, *G-valued potentially semi-stable deformation rings*, ProQuest LLC, Ann Arbor, MI, 2012, Thesis (Ph.D.)–The University of Chicago. MR 3152673

[BDJ10] Kevin Buzzard, Fred Diamond, and Frazer Jarvis, *On Serre's conjecture for mod ℓ Galois representations over totally real fields*, Duke Math. J. **155** (2010), no. 1, 105–161. MR 2730374 (2012k:11067)

[BG14] Kevin Buzzard and Toby Gee, *The conjectural connections between automorphic representations and Galois representations*, Automorphic forms and Galois representations. Vol. 1, London Math. Soc. Lecture Note Ser., vol. 414, Cambridge Univ. Press, Cambridge, 2014, pp. 135–187. MR 3444225

[BHS19] Christophe Breuil, Eugen Hellmann, and Benjamin Schraen, *A local model for the trianguline variety and applications*, Publications mathematiques de l' IHES **130** (2019), 299–412.

[BM02] Christophe Breuil and Ariane Mézard, *Multiplicités modulaires et représentations de* $\mathrm{GL}_2(\mathbf{Z}_p)$ *et de* $\mathrm{Gal}(\overline{\mathbf{Q}}_p/\mathbf{Q}_p)$ *en* $l = p$, Duke Math. J. **115** (2002), no. 2, 205–310, With an appendix by Guy Henniart. MR 1944572 (2004i:11052)

[BS73] A. Borel and J.-P. Serre, *Corners and arithmetic groups*, Commentarii mathematici Helvetici **48** (1973), 436–483.

[Car79] P. Cartier, *Representations of p-adic groups: a survey*, Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, pp. 111–155. MR 546593

[CEG+16] Ana Caraiani, Matthew Emerton, Toby Gee, David Geraghty, Vytautas Paškūnas, and Sug Woo Shin, *Patching and the p-adic local Langlands correspondence*, Camb. J. Math. **4** (2016), no. 2, 197–287. MR 3529394

[CG18] Frank Calegari and David Geraghty, *Modularity lifting beyond the Taylor-Wiles method*, Invent. Math. **211** (2018), no. 1, 297–433. MR 3742760

[CH13] Gaëtan Chenevier and Michael Harris, *Construction of automorphic Galois representations, II*, Camb. J. Math. **1** (2013), no. 1, 53–73. MR 3272052

[CS17] Ana Caraiani and Peter Scholze, *On the generic part of the cohomology of compact unitary Shimura varieties*, Ann. of Math. (2) **186** (2017), no. 3, 649–766. MR 3702677

[CV92] Robert F. Coleman and José Felipe Voloch, *Companion forms and Kodaira-Spencer theory*, Invent. Math. **110** (1992), no. 2, 263–281. MR 1185584

[Del71] Pierre Deligne, *Formes modulaires et représentations l-adiques*, Séminaire Bourbaki. Vol. 1968/69: Exposés 347–363, Lecture Notes in Math., vol. 175, Springer, Berlin, 1971, pp. Exp. No. 355, 139–172. MR 3077124

[DL76] P. Deligne and G. Lusztig, *Representations of reductive groups over finite fields*, Ann. of Math. (2) **103** (1976), no. 1, 103–161. MR 0393266

[Edi92] Bas Edixhoven, *The weight in Serre's conjectures on modular forms*, Invent. Math. **109** (1992), no. 3, 563–594. MR 1176206

[EGa] Matthew Emerton and Toby Gee, *Moduli of étale (φ, Γ)-modules and the existence of crystalline lifts*, `http://arxiv.org/abs/1908.07185`, preprint (2019).

[EGb] ———, *'Scheme theoretic images' of morphisms of stacks*, Algebraic Geometry, to appear.

[EG14] ———, *A geometric perspective on the Breuil-Mézard conjecture*, J. Inst. Math. Jussieu **13** (2014), no. 1, 183–223. MR 3134019

[FM95] J-M. Fontaine and B. Mazur, *Geometric Galois representations*, Elliptic curves, modular forms, & Fermat's last theorem (Hong Kong, 1993), Ser. Number Theory, I, Int. Press, Cambridge, MA, 1995, pp. 41–78. MR 1363495 (96h:11049)

[Fra98] Jens Franke, *Harmonic analysis in weighted L_2-spaces*, Ann. Sci. École Norm. Sup. (4) **31** (1998), no. 2, 181–279. MR 1603257

[FZ10] Edward Frenkel and Xinwen Zhu, *Any flat bundle on a punctured disc has an oper structure*, Math. Res. Lett. **17** (2010), no. 1, 27–37. MR 2592725

[Gee11] Toby Gee, *Automorphic lifts of prescribed types*, Math. Ann. **350** (2011), no. 1, 107–144. MR 2785764 (2012c:11118)

[GHS18] Toby Gee, Florian Herzig, and David Savitt, *General Serre weight conjectures*, J. Eur. Math. Soc. (JEMS) **20** (2018), no. 12, 2859–2949. MR 3871496

[GK14] Toby Gee and Mark Kisin, *The Breuil-Mézard conjecture for potentially Barsotti-Tate representations*, Forum Math. Pi **2** (2014), e1, 56. MR 3292675

[GLS14] Toby Gee, Tong Liu, and David Savitt, *The Buzzard-Diamond-Jarvis conjecture for unitary groups*, J. Amer. Math. Soc. **27** (2014), no. 2, 389–435. MR 3164985

[GN] Toby Gee and James Newton, *Patching and the completed cohomology of locally symmetric spaces*, J. Inst. Math. Jussieu, to appear.

[Gro98] Benedict H. Gross, *On the Satake isomorphism*, Galois representations in arithmetic algebraic geometry (Durham, 1996), London Math. Soc. Lecture Note Ser., vol. 254, Cambridge Univ. Press, Cambridge, 1998, pp. 223–237. MR 1696481

[Gro99] ———, *Algebraic modular forms*, Israel J. Math. **113** (1999), 61–93. MR 1729443

[Her09] Florian Herzig, *The weight in a Serre-type conjecture for tame n-dimensional Galois representations*, Duke Math. J. **149** (2009), no. 1, 37–116. MR 2541127 (2010f:11083)

[HT01] Michael Harris and Richard Taylor, *The geometry and cohomology of some simple Shimura varieties*, Annals of Mathematics Studies, vol. 151, Princeton University Press, Princeton, NJ, 2001, With an appendix by Vladimir G. Berkovich. MR 1876802

[Jan81] Jens Carsten Jantzen, *Zur Reduktion modulo p der Charaktere von Deligne und Lusztig*, J. Algebra **70** (1981), no. 2, 452–474. MR 623819

[Kis08] Mark Kisin, *Potentially semi-stable deformation rings*, J. Amer. Math. Soc. **21** (2008), no. 2, 513–546. MR 2373358 (2009c:11194)

[Kis09] ———, *The Fontaine-Mazur conjecture for* GL_2, J. Amer. Math. Soc. **22** (2009), no. 3, 641–690. MR 2505297 (2010j:11084)

[Kot92] Robert E. Kottwitz, *On the λ-adic representations associated to some simple Shimura varieties*, Invent. Math. **108** (1992), no. 3, 653–665. MR 1163241

[KW09] Chandrashekhar Khare and Jean-Pierre Wintenberger, *Serre's modularity conjecture. II*, Invent. Math. **178** (2009), no. 3, 505–586. MR 2551764

[Lab99] Jean-Pierre Labesse, *Cohomologie, stabilisation et changement de base*, Astérisque (1999), no. 257, vi+161, Appendix A by Laurent Clozel and Labesse, and Appendix B by Lawrence Breen. MR 1695940

[LHLM22] Daniel Le, Bao Viet Le Hung, Brandon Levin, and Stefano Morra, *Serre weights for three-dimensional wildly ramified Galois representations*, arXiv:2202.03303 (2022).

[LLH] Daniel Le and Bao Viet Le Hung, *Serre weights for* GL_n *over CM fields*, in preparation.

[LLHL19] Daniel Le, Bao V. Le Hung, and Brandon Levin, *Weight elimination in Serre-type conjectures*, Duke Math. J. **168** (2019), no. 13, 2433–2506. MR 4007598

[LLHLM18] Daniel Le, Bao V. Le Hung, Brandon Levin, and Stefano Morra, *Potentially crystalline deformation rings and Serre weight conjectures: shapes and shadows*, Invent. Math. **212** (2018), no. 1, 1–107. MR 3773788

[LLHLM20a] ———, *Local models for Galois deformation rings and applications*, arXiv:2007.05398 (2020).

[LLHLM20b] ———, *Serre weights and Breuil's lattice conjectures in dimension three*, Forum Math. Pi **8** (2020), e5, 135. MR 4079756

[Mat67] Yozô Matsushima, *A formula for the Betti numbers of compact locally symmetric Riemannian manifolds*, J. Differential Geometry **1** (1967), 99–109. MR 222908

[New14] James Newton, *Serre weights and Shimura curves*, Proc. Lond. Math. Soc. (3) **108** (2014), no. 6, 1471–1500. MR 3218316

[PZ13] George Pappas and Xinwen Zhu, *Local models of Shimura varieties and a conjecture of Kottwitz*, Invent. Math. **194** (2013), no. 1, 147–254. MR 3103258

[Sch15] Peter Scholze, *On torsion in the cohomology of locally symmetric varieties*, Ann. of Math. (2) **182** (2015), no. 3, 945–1066. MR 3418533

[Ser77] Jean-Pierre Serre, *Linear representations of finite groups*, Springer-Verlag, New York-Heidelberg, 1977, Translated from the second French edition by Leonard L. Scott, Graduate Texts in Mathematics, Vol. 42. MR 0450380

[Ser87] ———, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Duke Math. J. **54** (1987), no. 1, 179–230. MR 885783 (88g:11022)

[Shi11] Sug Woo Shin, *Galois representations arising from some compact Shimura varieties*, Ann. of Math. (2) **173** (2011), no. 3, 1645–1741. MR 2800722

[Zhu14] Xinwen Zhu, *On the coherence conjecture of Pappas and Rapoport*, Ann. of Math. (2) **180** (2014), no. 1, 1–85. MR 3194811

Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067

Email address: ledt@purdue.edu

Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, Illinois 60208, USA

Email address: lhvietbao@googlemail.com