Eur. Phys. J. C (2021) 81:707
https://doi.org/10.1140/epjc/s10052-021-09506-w

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Theoretical Physics

The origin of rest-mass energy

Fulvio Melia[a]

Department of Physics, The Applied Math Program, and Department of Astronomy, The University of Arizona, Tucson, AZ 85721, USA

Received: 27 June 2021 / Accepted: 28 July 2021

Abstract Today we have a solid, if incomplete, physical picture of how inertia is created in the standard model. We know that most of the visible baryonic 'mass' in the Universe is due to gluonic back-reaction on accelerated quarks, the latter of which attribute their own inertia to a coupling with the Higgs field – a process that elegantly and self-consistently also assigns inertia to several other particles. But we have never had a physically viable explanation for the origin of rest-mass energy, in spite of many attempts at understanding it towards the end of the nineteenth century, culminating with Einstein's own landmark contribution in his Annus Mirabilis. Here, we introduce to this discussion some of the insights we have garnered from the latest cosmological observations and theoretical modeling to calculate our gravitational binding energy with that portion of the Universe to which we are causally connected, and demonstrate that this energy is indeed equal to mc^2 when the inertia m is viewed as a surrogate for gravitational mass.

1 A brief history of $E = mc^2$

Today we take it for granted that a particle with *inertia*, m_i, carries an irreducible amount of energy – even when at rest with respect to the observer – given by Einstein's famous formula, $E = m_{\rm i}c^2$. Every object gains kinetic energy, K, under the accelerating influence of an external force, and it loses potential energy, Φ, when allowed to *fall* freely in a region where it experiences an attraction to something else. No matter how K and Φ change, however, the rest-mass energy $E = m_{\rm i}c^2$ is an immutable feature of $m_{\rm i}$. So why should inertia, which has no obvious connection to K and Φ, be associated with energy, and why is it possible for E to be converted back and forth into K and/or Φ when $m_{\rm i}$ is allowed to change, e.g., via the annihilation of a particle-antiparticle pair?

Contrary to conventional wisdom, Einstein was not the first to consider the possible conversion of 'mass energy' into other forms of energy, and actually did not formally prove their equivalence either. In 1881, the future Nobel laureate Thomson realized that – when viewed as a charged sphere – an electron moving through an 'aether' resists being accelerated more than a similarly uncharged object [1]. Much earlier, Stokes had drawn similar conclusions in the context of hydrodynamics, showing in 1844 that a body's inertia increases when moving through an incompressible perfect fluid [2]. Quite remarkably, both of these explanations for the origin of inertia would eventually constitute a historical echo of the Higgs mechanism (see Sect. 3), proposed more than a century later, though based on a surprisingly similar idea [3,4].

Thomson viewed this effect as arising from the electromagnetic field carried by the charge itself, so he assigned to it an effective momentum and an apparent *electromagnetic mass*. At least part of the mass of the electron could thus be viewed as arising from its electromagnetic self-energy – requiring some sort of equivalence between inertia and energy. Over the next two decades, this idea was fleshed out in considerable detail by Heaviside [5], Searle [6], Abraham [7] and Lorentz [8,9]. Its development proceeded to the point where the radiation reaction force, $\mathbf{F}_{\rm em}$, acting on a charged particle due to the momentum and energy carried away by the radiation it produces, could be formally incorporated into the Abraham-Lorentz equation [10].

It had been known since 1884, when Poynting published [11] his now famous theorem on the conservation of energy in an electromagnetic field, that Maxwell's equations contained the ingredients necessary to calculate both the energy and momentum density carried by a radiation field. The relationship between these two dynamical attributes, together with the Larmor equation yielding the rate of energy loss by an accelerated charge, could therefore be used to infer the

John Woodruff Simpson Fellow.

[a] e-mail: fmelia@email.arizona.edu (corresponding author)

Published online: 07 August 2021

particle's momentum loss rate, from which one could see that [10]

$$\mathbf{F}_{\rm em} \equiv \frac{2q^2}{3c^3}\ddot{\mathbf{v}}, \tag{1}$$

where q is the particle's electric charge. One could thus interpret from this that the field has an effective mass

$$m_{\rm em} \equiv \frac{2q^2}{3c^3}\frac{1}{\tau}, \tag{2}$$

yielding $\mathbf{F}_{\rm em} \equiv m_{\rm em}\dot{\mathbf{v}}$, if one identifies $\tau \equiv r_q/c$ as the light travel-time across the radius r_q of the charge – a reasonable estimate of the time associated with dynamical changes in $\dot{\mathbf{v}}$. And given that the electric self-energy of a charge q spread evenly across the surface of a sphere of radius r_q is

$$E_{\rm em} = \frac{1}{2}\frac{q^2}{r_q}, \tag{3}$$

one immediately infers the implied equivalence of the field's energy and effective mass:

$$E_{\rm em} \equiv \frac{3}{4}m_{\rm em}c^2. \tag{4}$$

Of course, this electromagnetic mass requires a particle to be charged, so it could not apply to everything. Nevertheless, one cannot but marvel at the strong similarity between Eq. (4) and Einstein's formula $E = m_{\rm i}c^2$. And this first formal attempt to find an equivalence between mass and energy preceded special relativity by several decades.

Following other developments in finding the 'correct' relationship between mass and energy, Hasenöhrl created in 1904 a thought experiment involving the heat (i.e., 'blackbody') energy inside a moving cavity [12,13]. As we shall see shortly, Einstein's own derivation of the relationship between inertia and energy was based on very similar physics. Hasenöhrl published several different versions of his argument, but one can appreciate the gist of his thought experiment by simply considering the first [12]. He imagined filling a perfectly reflecting cavity with 'heat,' i.e., blackbody radiation, emitted symmetrically at the two ends of a cylindrical container. Since identical radiation (or photons, in modern parlance) is emitted at each end according to an observer sitting inside, the external forces applied to counter the radiative reaction forces (analogous to Eq. 1) are equal and opposite.

But to an observer sitting in the laboratory, watching the same cavity moving past them at constant velocity, $\mathbf{v}$, the radiation emitted in the direction of $\mathbf{v}$ is Doppler blue-shifted, while that emitted in the opposite direction is red-shifted. And since blue-shifted photons carry more momentum than their red-shifted counterparts, the two external forces seen in the laboratory must now be different in order to maintain the cavity moving at constant velocity. Hasenöhrl applied the classical work-energy theorem, equating the net difference in work exerted by the external forces to the change in the cavity's kinetic energy, to show that the blackbody radiation has an equivalent mass $m_{\rm bb} = (4/3)E_{\rm bb}/c^2$. Actually, his first publication erroneously quoted this result as $m_{\rm bb} = (8/3)E_{\rm bb}/c^2$, but he corrected his algebraic mistake in a subsequent paper after receiving communication from M. Abraham.

The importance of this step was Hasenöhrl's extension of the result in Eq. (4) to non-charged particles. Indeed, as we shall see shortly, his thought experiment was very similar to that of Einstein, which was published the following year. One may thus wonder why his expression contained the factor $4/3$ instead of simply 1. As it turns out, this was not due to his use of classical physics, as one might suspect but, rather, to the fact that he incorrectly ignored the mass being lost by the cylinder's caps while they are emitting heat [14].

Such was the impact of Hasenöhrl's argument, however, that even as late as 1909, Max Planck [15] included in one of his lectures the statement "that the blackbody radiation possesses inertia was first pointed out by F. Hasenöhrl." But the correct answer, of course, was published by A. Einstein [16] in one of his four Annus Mirabilis papers of 1905. Couched in the language of special relativity, Einstein's argument was – in retrospect – remarkably simple though, in the end, he approximated away the relativistic parts anyway, so his answer is derivable purely from classical physics, based on the Doppler effect.

Einstein considered a single point particle, moving with velocity $\mathbf{v}$ in the laboratory frame, radiating away a quantity of energy $\Delta E'$ with front-back symmetry in its own rest frame. For simplicity, he assumed that $\Delta E'/2$ is radiated in a direction parallel to and anti-parallel to $\mathbf{v}$. According to the relativistic Doppler-shift formula, an observer in the laboratory sees the radiation carrying away an energy $(\Delta E'/2)\gamma(1 + \beta\cos\theta')$, where θ' is the angle between $\mathbf{v}$ and the direction of propagation of the radiation, and $\gamma \equiv 1/\sqrt{1-\beta^2}$ is the Lorentz factor in terms of $\beta \equiv |\mathbf{v}|/c$. Thus, the difference in kinetic energy of the particle between the laboratory and rest frames is simply

$$\Delta K - \Delta K' = \Delta E'(\gamma - 1). \tag{5}$$

In the low-velocity limit, where the relativistic parts are approximated away, this equation becomes

$$\Delta K - \Delta K' = \frac{1}{2}\frac{\Delta E'}{c^2}v^2. \tag{6}$$

He then argued that since the particle is giving away an amount of energy $\Delta E'$, its mass must have diminished by $\Delta m_{\rm i} = \Delta E'/c^2$.

It is important to note, however, that Einstein made several sweeping conclusions from this result, including (i) that it applies to all bodies and all forms of energy, and (ii) that it

remains true even at higher velocities (where relativity would indeed introduce corrections to the classical outcome). But he never actually proved any of these claims, even in the subsequent handful of papers he published on this topic over the next several decades. Today we know this result is correct because it has been verified experimentally to incredible accuracy. It has never been proven theoretically, however, and *the fundamental reason why inertia ought to be associated with energy has remained a complete mystery to this day*.

2 Inertia and gravitational mass

To properly address the question of why a 'rest mass' $m_{\rm i}$ represents an energy $m_{\rm i}c^2$, we first need to refine and clarify our concepts of inertia and gravitational charge, which we shall call $m_{\rm g}$ to properly distinguish it from $m_{\rm i}$. Newton viewed inertia to be a conserved and irreducible property of matter, and did not consider $m_{\rm i}$ and $m_{\rm g}$ to be distinct [17]. By 'inertia' we shall strictly refer to the proportionality constant between an applied force and an object's consequent acceleration, according to Newton's second law of motion,

$$\mathbf{F} = m_{\rm i}\mathbf{a}. \tag{7}$$

The quantity $m_{\rm i}$ retains this meaning in relativity, where it is considered to be the inertial mass in the object's *rest* frame. Since an observer in this frame can reduce their equation of motion to the classical limit shown in Eq. (7), they could with equal validity refer to $m_{\rm i}$ as either the object's inertia or its 'rest mass' m. No doubt, this is a very basic concept, but we need to be clear that 'inertial mass' strictly represents an object's resistance to acceleration when a force is applied to it in the classical limit.

Gravitational charge, on the other hand, arises in the context of Newton's universal law of gravitation,

$$\mathbf{F}_1 = -\frac{Gm_{\rm g1}m_{\rm g2}}{r^2}\hat{\mathbf{r}}, \tag{8}$$

expressing the force $\mathbf{F}_1$ experienced by particle 1 (with gravitational charge $m_{\rm g1}$) due to the gravitational influence exerted by particle 2 (with gravitational charge $m_{\rm g2}$). The radius vector $\mathbf{r} = r\hat{\mathbf{r}}$ points from 2 to 1, and we have explicitly included a negative sign in this equation, arising from the fact that gravity is *always* attractive – a feature that will shortly become highly relevant to our discussion concerning the relationship between $m_{\rm i}$ and $m_{\rm g}$. The quantity G is the 'gravitational' constant, whose numerical value and physical units depend on how we *choose* to define $m_{\rm g}$, say in terms of the (dimensionless) number of atoms in an object, or its inertial mass $m_{\rm i}$ in kilograms. The conventional value of G that we are all familiar with arises when we force the equality $m_{\rm i} = m_{\rm g}$.

The latter possibility – that $m_{\rm i}$ and $m_{\rm g}$ might be related, perhaps even equal – arises from the observation that they both represent the amount of 'something' in the object. Certainly, at the time of Newton, there weren't too many options to consider. If one were to double the *quantity of matter*, as Newton would have put it, one would reasonably expect from simple experimentation that its inertia would also double. Likewise, doubling the quantity of matter in object 1 would double the gravitational force in Eq. (8). Today we know much more and realize at a very fundamental level that these two 'quantities' need not be the same physically. For example, if we were to naively stick two identical objects together, we could double the attribute that gives rise to inertia, while also doubling the analogous (but different) attribute responsible for the gravitational charge. In both cases, $m_{\rm i} \to 2m_{\rm i}$ and $m_{\rm g} \to 2m_{\rm g}$, even though $m_{\rm i}$ and $m_{\rm g}$ might have nothing to do with each other. In the absence of any more definitive information, the best one could argue is therefore that $m_{\rm i} \propto m_{\rm g}$, certainly not that $m_{\rm i} = m_{\rm g}$. But even this statement is fraught with peril given what we now know about the 'equivalence' of mass and energy and the fact that, in general relativity, the spacetime curvature really responds to energy, not mass, as we shall discuss later in this paper. Nonlinear effects that increase the self- (or binding) energy of an object as its gravitational charge increases may therefore destroy the simple constancy of $m_{\rm i}/m_{\rm g}$ if inertia is unrelated to gravity [18].

But at least in this regard, experimentation does provide us with a very firm indication that $m_{\rm i}$ remains proportional to $m_{\rm g}$ over all the scales that have been tested thus far. Most of the experiments attempt to compare the acceleration of two laboratory-sized objects of different composition in the presence of an external gravitational field. Many high-precision Eötvös-type of measurements have been made, starting with the pendulum experiments of Newton and Bessel, to the classic torsion-balance version of Eötvös [19], Dicke [20] and others. In the latest version of these torsion-balance experiments, two objects of different composition are rested on a tray and suspended horizontally by a fine wire. For example, the 'Eöt-Wash' experiments have used such devices at the University of Washington to compare the accelerations of various materials toward movable laboratory masses, the Sun and the galaxy [21,22], reaching a relative precision [23] of 2×10^{-13}. (For a recent review, see Tino et al. [24].) Another way to say this is that, as far as we can tell, everything in an object that gives rise to inertia also contributes *proportionately* to its gravitational charge.

As is well known, this proportionality between $m_{\rm i}$ and $m_{\rm g}$ is the basis for Einstein's principle of equivalence. One can easily understand this from Eqs. (7) and (8), which show that particles ($j = 1...n$) – much closer to each other than the scale over which a gravitational field is changing – are all accelerated at an equal rate proportional to the constant $m_{{\rm g}j}/m_{{\rm i}j}$. An observer could therefore not distinguish this situation from an analogous one in which they were being

observed in a local, non-inertial frame accelerating uniformly in the opposite direction.

So why couldn't this equivalence apply to other forces as well? For example, why couldn't we argue that the amount of charge in an object is proportional to its matter content? Then the Coulomb force acting on it analogously to Eq. (8) would be proportional to its net charge, q_1. Doubling the quantity of matter would result in $q \to 2q$ and $m_\mathrm{i} \to 2m_\mathrm{i}$, so that the ratio q/m_i would always remain the same. In this case, we would see an equivalence between inertia and the electric charge, perhaps leading us to propose an alternative equivalence principle based on the notion that we could not distinguish between charges accelerated in an electromagnetic field and the analogous situation of charges being viewed in a non-inertial frame uniformly accelerated in the opposite direction.

The answer, of course, is that the other forces all lack the unique combination of properties that allow gravity to function in this way. Gravity has a single charge, unlike electromagnetism which has two, or quantum chromodynamics which has three (red, green and blue) and the corresponding antiquark colors. So gravity is always attractive, while the others can vary depending on the charge balance. In addition, gravitational charge cannot be annihilated, so that all forms of energy have an effective m_g that accumulates, as does inertia, while electric charge can be completely removed from an object. In other words, gravity is the only force for which the proportionality between its charge and m_i is guaranteed. And equally important, it is the only force for which one may reasonably expect its charge to extend its influence over a vast volume of space (i.e., the cosmos). In spite of the fact that the Coulomb force is itself an inverse-square law, it is energetically prohibitive to maintain a separation of charges over distances extending beyond the laboratory or, in the most extreme situation, beyond the magnetosphere of a pulsar, smaller than a typical city here on Earth. The Universe is therefore neutral on large scales – specifically because the electromagnetic force contains more than one charge. So the equivalence principle could only work for gravity, and we are led to the conclusion that inertial mass must therefore be proportional to the gravitational charge, which we shall henceforth sometimes call the 'gravitational mass.' And to simplify the discussion even further, we shall often 'choose' the relevant constants (such as G) to have values and units that allow us to set the inertial mass and gravitational charge equal to each other, thereby defining the *rest mass*, $m \equiv m_\mathrm{i} = m_\mathrm{g}$.

3 The Higgs and QCD inertia

Without unduly preempting our discussion in Sect. 5, the obvious question arising from the conclusion in the previous section centers on the issue of whether rest-mass energy, mc^2, can really be associated with the object's inertia, m_i, or whether it is in fact an energy due to a physical influence involving its gravitational charge, m_g. We would not be able to tell the difference since $m_\mathrm{i}/m_\mathrm{g} =$ constant, which permits inertia to act as a *surrogate* for m_g. In that case, it wouldn't even matter what the origin of inertia were, as long as we could identify the physics that generates an energy $m_\mathrm{g}c^2 \to mc^2$ (which we shall do in Sect. 5). Nevertheless, for the sake of clarity and completeness, we shall here first summarize the current situation concerning the origin of m_i.

In Newton's view of the world, inertia was an intrinsic property of matter, manifested by objects moving relative to an absolute space. But several early thinkers following Newton, notably Berkeley [25] and Mach [26], already questioned an independently defined absolute space, and instead proposed that inertial frames are those that are unaccelerated relative to the 'fixed stars' or, more accurately, relative to a carefully defined mean of all the matter in the Universe. Einstein called this 'Mach's principle' and considered it to be foundational in the development of his general relativity theory, but he eventually realized that these two are actually incompatible with each other [27,28]. Though the equivalence principle had suggested to him that inertia must be due to the gravitational influence of the whole Universe, Einstein eventually realized that this influence disappears completely for a particle in free-fall. While the particle experiences zero gravity in this frame, it nevertheless still exhibits inertial properties.

Mach himself never explicitly stated how or why his view of inertia ought to be formalized as some kind of new physical law, so he never provided a physical mechanism describing how the distant matter in the Universe affects the motion of a local particle. But Mach's principle has been invoked many times in the development of alternative gravity theories. For example, Dennis Sciama attempted in 1953 [29] to express Mach's principle in more quantitative terms by proposing the addition of an acceleration-dependent contribution to Newton's law of gravity (Eq. 8). Sciama called this effect an 'inertial induction.' Later, Brans and Dicke [30] incorporated Mach's principle into an alternative theory to general relativity, by setting up a framework in which the gravitational constant G is determined by the structure of the Universe. In their approach, the unit of inertial mass is taken to be the Planck mass (i.e., $m_\mathrm{P}^2 \equiv \hbar c/G$), so that a changing mass results from a changing G, which in turn can be viewed as the Machian consequence of a changing Universe.

But in spite of these attempts at physically interpreting inertia as an effect due to distant matter in the Universe, the situation today regarding Mach's principle is perhaps best summarized by Abraham Pais [31]: "It must be said that, as far as I can see, to this day, Mach's principle has not brought physics decisively farther. It must also be said that the origin of inertia is and remains the most obscure subject in the theory

of particles and fields." Quite remarkably, though, at least a partial answer appears to have been found in the intervening period.

In ordinary matter, ignoring for brevity and simplicity other possible issues associated with dark matter and dark energy in a cosmological context, inertia is overwhelmingly dominated by the nuclei, $m_{\rm i} \sim m_{\rm N}$, specifically, protons and neutrons. Electrons are far smaller ($m_{\rm e} < m_{\rm N}/1000$) and – if we take the liberty of borrowing the $E = mc^2$ result to convert the nuclear binding energy into an effective inertial mass – other contributions to the mass of the nucleus are but a small fraction of $m_{\rm N}$ (typically less than 1%). Thus, to understand the origin of atomic inertia and, by extension, most of the inertia of ordinary matter in the Universe, one must uncover the origin of proton and neutron masses and, to a lesser extent, the origin of electron mass.

Today, the standard model of particle physics is well established and experimentally confirmed. It encompasses electromagnetism, the weak force and strong interactions, and provides a self-consistent classification of all the known elementary particles. It is nevertheless still incomplete because it does not (i) include gravity, (ii) account for baryon asymmetry and dark matter and (iii) allow for the inclusion of dark energy, if the latter turns out to be something other than a cosmological constant, Λ. Some of the key steps in its development have been (i) the unification of the electromagnetic and weak interactions by Glashow [32], (ii) the incorporation by Weinberg and Salam of the Higgs mechanism to generate inertial masses for some of its particles [3,4,33,34] (more on this below), and (iii) the discovery of various new particles it predicted, such as the $W^{\pm}$, Z^0 and Higgs bosons (see, e.g., Oerter [35] for a detailed review).

Its structure contains six quarks (fermions that carry color charge), which are used in various combinations to form the meson and baryon hierarchy; six leptons (including electrons and neutrinos); twelve spin-1 gauge bosons that mediate the strong, weak, and electromagnetic interactions; and one spin-0 scalar boson, i.e., the recently discovered Higgs particle. The gauge bosons include the aforementioned $W^{\pm}$ and Z^0 carriers of the weak force, as well as the massless photon responsible for the electromagnetic interaction. The remaining eight gauge bosons are various color combinations of gluons that mediate the strong force inside mesons and baryons, such as the proton and neutron.

The quark, electron, $W^{\pm}$ and Z^0 inertial masses are generated via the Higgs mechanism that we shall discuss shortly. The proton and neutron masses, however, are much larger than the mere sum of their enclosed quark and gluon fields. As surprising as it may seem, it is actually possible to measure individual quark masses based on the reconstruction of jets they induce in high-energy collisions. This is the method used to measure the top quark mass, while the bottom and charm masses may also be inferred from the mass of meson resonances, such as bottomonium and charmonium, since these appear to be non-relativistic quark-antiquark bound states. The other three light quark masses (strange, down, up) may be inferred from the spectroscopy of low-lying pseudoscalar mesons, such as π, K, and η, whose inertial masses depend sensitively on the light-quark masses.

As noted earlier, however, this beautiful, self-consistent picture does not yet explain why the nucleon mass is ~ 20 times larger than the sum of the quark masses within it. Ironically, this is where the highly original development concerning the electromagnetic mass in the nineteenth century resurfaces (see Sect. 1 above), notably via arguments of the form expressed in Eqs. (1) and (2). That proposal was based on the idea that energy and momentum carried away by the electromagnetic field provided a back-reaction on the radiating particle being accelerated, thereby generating inertia. There are several fundamental reasons from quantum electrodynamics why this mechanism cannot work for the electron, in part because this mechanism produces infinite multiplicative factors representing the mass. Remarkably, however, a very similar approach does work in quantum chromodynamics. Detailed calculations from first principles have shown that most ($\sim 95\%$) of the nucleon's inertia is generated by the back-reaction of color gluon fields resisting the acceleration of quarks and (the similarly colored) gluons inside the baryons [36]. Actually, this process accounts very well for most of the inertia in the entire low-lying meson and baryon distribution.

Most of the inertial mass in ordinary matter can therefore be understood as arising from the back reaction of gluons on the quarks that radiate them in response to the acceleration they are subjected to by external forces. This is a rather profound statement because it tells us that inertia originates dynamically, principally to conserve momentum, rather than from some Newtonian definition of irreducible internal 'mass.' It should now become clearer why the statement made at the beginning of this section is so essential to this whole discussion. Attempting to assign 'rest energy' to inertia – when viewed as an emergent property – doesn't make much physical sense. Instead, interpreting inertia as a surrogate for how much '$m_{\rm g}$' a quark (say) possesses allows us to pursue a more physically meaningful investigation of how gravitational charge is involved in the generation of energy.

The story is not yet complete, however, because individual quarks and some of the leptons and gauge bosons also have inertial mass, which must be due to something else. Conventional wisdom today has it that this type of inertia, distinct from the one generated by the QCD interactions discussed above, is due to a coupling of these particles to a pervasive spin-0 scalar field [3,4] known as 'Higgs.' Much has been written about this mechanism [37], and the discovery of the Higgs boson itself appears to have cemented our basic understanding of how inertia is generated for particles in the

standard model that would otherwise have to remain massless in order to satisfy several required symmetries. The way this mechanism works is rather easy to explain, but it also contains an important caveat that will leave us wondering whether we have actually uncovered the whole truth.

All of the particles in the standard model (in the absence of a Higgs field, Φ) must have zero mass in order to comply with various (presumed) symmetries. The Lagrangian density representing gauge bosons, for example, cannot contain 'mass' terms, such as $m_w W_\mu^\pm W^{\pm\mu}$, which would violate gauge invariance. In physics, we measure distances and times, velocities and acceleration in order to infer the particle dynamics. But the latter results from 'forces,' not potentials from which the forces are derived. As long as one can shift the gauge of the potentials without affecting the forces, the description of the system should remain the same. But the mass term for the $W^\pm$ gauge bosons, for example, would not remain invariant if the gauge of $W_\mu^\pm$ were shifted, unless $m_w = 0$. Similarly, a mass term for fermions must necessarily mix left-handed and right-handed fermions, but these have different gauge quantum numbers, so a shift in gauge would not allow such a term in the Lagrangian density to remain invariant. The latter requirement is commonly referred to as chiral symmetry, meaning that the Dirac action ought to remain invariant under a chiral rotation.

The addition of a spin-0 scalar field to the standard model introduces an additional interaction for the fermions and gauge bosons, regulated by a unique coupling constant g_j for each particle species "j", chosen to produce consistency with the observed masses. The term associated with each particle-Higgs interaction appearing in the Lagrangian density is represented as a product $g_j\xi_j\Phi$, written in terms of g_j, the particle field ξ_j, and the Higgs field. But still nothing interesting would happen with this in terms of generating inertia if all the fields retained a zero expectation value in vacuum. This interaction term would then merely vary stochastically as the fields fluctuated about zero, and could in no way be linked to the highly stable masses we measure for the standard-model particles. To overcome this deficiency, the Higgs field is instead assigned a potential, $V(\Phi^\dagger\Phi)$, tuned to prevent its lowest-energy state from having $\Phi = 0$. This is done by postulating that

$$V(\Phi^\dagger\Phi) = -\mu^2\Phi^\dagger\Phi + \frac{1}{2}\lambda(\Phi^\dagger\Phi)^2, \tag{9}$$

with $\mu^2 > 0$. Does the Higgs field have some as yet unknown 'internal' property or 'structure' that produces such a potential? No one knows, but it is not difficult to see that, instead of being minimized at $\Phi = 0$, V attains its lowest value for the modulus

$$\Phi^\dagger\Phi = v^2 \equiv \frac{\mu^2}{\lambda}. \tag{10}$$

The quantity v is known at the Higgs *vacuum expectation value*. In other words, if we insist on vacuum corresponding to the lowest energy state for such a potential, Φ cannot be zero; it must have a vacuum expectation value consistent with Eq. (10).

This changes the nature of the interaction term completely, because now we may write $g_j\xi_j(v+\phi_1) = g_j\xi_j v + g_j\xi_j\phi_1$, in terms of the real part of Φ, given as $v+\phi_1$. Here, ϕ_1 represents a fluctuation of the Higgs field away from its otherwise constant vacuum expectation value v. This achieves the principal result because $g_j\xi_j v$ is a mass term for ξ_j, dependent only on g_j, μ and λ. We interpret this result to mean that a fermion or gauge boson (with $g_j \neq 0$) plowing through the pervasive Higgs field attracts Higgs bosons to itself, and its inertia increases in proportion to the mass carried by the latter [38].

But therein lies the crucial caveat. This mechanism is quite different from the QCD interaction we described earlier. Whereas the latter results from conservation of momentum and the back-reaction of gluons radiated by accelerated quarks, the Higgs interaction creates inertia for the standard-model particles by attracting them to massive Higgs bosons. To make this work, a potential of the form in Eq. (9) is essential, but we don't know where it comes from. With it, a nonzero Higgs field pervades all of space, very much like the aether proposed to mediate the propagation of electromagnetic waves back in the nineteenth century. More seriously, though, this ansatz for the Higgs potential includes a quantity μ with dimensions of mass. Indeed, the mass of the Higgs boson in this model is $m_{\rm H}^2 = 2\lambda v^2 = 2\mu^2$, and it appears as a *free parameter*. *There is no elucidation or explanation for where it comes from.* Yet clearly all of the standard-model masses derived with this mechanism are critically dependent on it.

To summarize, the Higgs mechanism endows standard-model particles with inertia, yet allows them to still satisfy all of the essential invariances arising from gauge and chiral symmetry. But to do so, the Higgs boson must itself already have inertial mass, and we have no idea where that comes from. And we should not forget that none of these features provide us with any elucidation of the complicated structure of quark and fermion masses and mixings. Why should the particles all have different couplings g_j to the Higgs field? And where do these values come from? It is fair to say that we have come a long way exploring the origin of inertia since the nineteenth century, but no one would claim that we fully understand it yet. And then there's the question of why inertia (or, more likely $m_{\rm g}$) ought to be associated with an energy $E = m_{\rm g}c^2$ $(= mc^2)$, which we shall address next.

4 The gravitational horizon in cosmology

If we believe the argument that rest-mass energy is more likely to be associated with m_g than some kind of emergent inertia, the next important factor to consider is the source of gravity that couples with the particle to produce this energy. Is it other nearby particles, the laboratory, galaxy or something even bigger? Certainly, no other force can be involved in this process because, as we have seen, the equivalence principle works only for gravity. And quite simply, no other force extends meaningfully to large enough distances to contribute non-negligibly to rest-mass energy. Thus, since the effects of gravity are cumulative, one should reasonably expect that all of the cosmic energy density in causal contact with the particle must be coupling gravitationally with it and contributing to its 'rest-mass' energy. But what fraction of the Universe should we include in this 'causally connected' region? Fortunately, recent work in cosmology provides us with several indispensable clues to answer this question, notably the role played by the so-called apparent (or gravitational) horizon in both the interpretation of observational measurements and their theoretical foundation [39].

To avoid any possible confusion, we should reiterate at this stage that the question of energy is entirely independent of how inertia arises. In Sect. 3 we described early attempts at explaining inertia based on the influence of distant matter in the Universe and found that Mach's principle has never been successfully incorporated into any working theory of gravity. Here, we are again invoking an interaction between local particles and the rest of the Universe, though it will now become clear that this interaction must be a gravitational one. And this gravitational influence is not at all responsible for creating inertia but, as we shall see shortly, it appears to be the origin of rest-mass energy.

Standard cosmology is based on the Friedmann–Lemaître–Robertson–Walker (FLRW) metric, describing a spatially homogeneous and isotropic three-dimensional space, expanding or contracting as a function of time:

$$ds^2 = c^2\,dt^2 - a^2(t)\left[\frac{dr^2}{(1-kr^2)} + r^2(d\theta^2 + \sin^2\theta\,d\phi^2)\right]. \tag{11}$$

This metric is written in terms of the cosmic time, t, and and comoving spherical coordinates (r, θ, ϕ), representing the perspective of a *free-falling* observer, analogous to their free-falling counterparts in the Schwarzschild and Kerr metrics. The expansion factor, $a(t)$, is independent of position, and the geometric constant k is $+1$ for a closed universe, 0 for a flat universe, and -1 for an open universe. The latest observations [40] are telling us that the Universe is flat (with $k = 0$), so we shall assume this condition throughout this paper.

It is also helpful to introduce the proper radius, $R(t) \equiv a(t)r$, which is often used to express changing (or 'physical') distances as the Universe expands. Sometimes, R is called the areal radius – the radius of two-spheres of symmetry – defined in a coordinate-independent way as $R \equiv \sqrt{A/4\pi}$, where A is the area of the two-sphere in the given geometry [41,42].

In a cosmology based on the FLRW metric, the term 'horizon' may refer to (i) the 'particle horizon,' characterizing the distance particles have traveled towards an observer since the big bang, (ii) the 'event horizon,' a membrane that separates causally connected spacetime events from those that are not, or (iii) several other constructs, each with its own customized application [43]. These all have their purpose, but as the measurements continue to improve, it is becoming quite clear that one particular definition is emerging as the most relevant for interpreting the observations – the (imaginary) surface separating all null geodesics receding from the observer from those that are approaching. This is how we formally define the apparent horizon, R_h, in general relativity. It turns out, however, that for an isotropic Universe (as described by Eq. 11), the apparent horizon coincides with the better known gravitational horizon [39,66] first identified in the Schwarzschild metric,

$$R_h = \frac{2GM_{MS}}{c^2}, \tag{12}$$

in terms of the Misner-Sharp mass [44],

$$M_{MS} \equiv \frac{4\pi}{3} R_h^3 \frac{\rho}{c^2}, \tag{13}$$

where ρ is the total energy density in the cosmic fluid.

We must be very clear about what this definition actually means, so let us take a moment to carefully dissect it. It follows the standard practice in general relativity of considering the source of gravity (or, more accurately, the spacetime 'curvature') to be the energy (in this case ρ). But this expression also redefines it in terms of a 'gravitational mass density' (ρ/c^2) by tacitly assuming the $E = mc^2$ relation. All the equations that follow then have this *ab initio* assumption built into them. One can see, however, that this conversion is merely one of convenience, for R_h can be re-written independently of ρ. Introducing the Friedmann equation,

$$H^2 = \frac{8\pi G}{3c^2}\rho, \tag{14}$$

obtained by putting $k = 0$, absorbing the cosmological constant Λ into ρ (if necessary), and inserting the FLRW metric coefficients into Einstein's equations [46], one can easily combine it with Eqs. (12) and (13) to show that $R_h = c/H$, the more familiar expression for the Hubble radius, written in terms of the Hubble parameter $H \equiv \dot{a}/a$. Yes, quite interestingly, the empirically derived Hubble radius in a cosmic setting turns out to be the apparent, or gravitational, radius.

The physical nature of $M_{\rm MS}$ first emerged from the pioneering work of Misner and Sharp [44] on spherical collapse problems in general relativity. It is sometimes also referred to as the Misner–Sharp–Hernandez mass, to include the subsequent contribution by Hernandez and Misner [45]. In the cosmic framework, however, $R_{\rm h}$ – and therefore $M_{\rm MS}$ – is not static. Unlike the situation with Schwarzschild, in which $R_{\rm h}$ is in fact the event horizon, $R_{\rm h}$ in cosmology continues to grow as the Universe expands, and may eventually turn into a cosmic event horizon, depending on the equation-of-state in the cosmic fluid, i.e., it depends on whether or not $H(t)$ eventually approaches a constant. In the next section, we shall demonstrate that a particle's rest-mass energy is none other than its *gravitational binding energy* to the Misner–Sharp 'mass' $M_{\rm MS}$. Though $M_{\rm MS}$ grows as the Universe expands, it is the ratio $M_{\rm MS}/R_{\rm h}$ (see Eq. 12) that sets the conversion factor from $m_{\rm g}$ to $m_{\rm g}c^2$ $(= mc^2)$, and this ratio remains constant as the Universe expands.

For the reader with a deeper understanding of general relativity, it may also be helpful to mention that the Misner-Sharp-Hernandez mass may not be the only definition one may use to specify a 'global' mass, though there are several good reasons for choosing it in the context of FLRW. First and foremost, it is not at all arbitrary, in the sense that only this definition is consistent with the g_{rr} metric coefficient. As a result, $M_{\rm MS}$ is the only mass that provides an apparent horizon allowing us to write the FLRW metric in terms of the proper radius, $R = a(t)r$, and the ratio $R/R_{\rm h}$, signalling how far the observer is from the gravitational horizon (see Eq. 18 below).

In general relativity, it is generally non-trivial to identify a 'physical mass-energy' in a non-asymptotically flat geometry [47]. But when the spacetime is spherically symmetric, as we have with FLRW, other possible definitions, such as the Hawking–Hayward quasilocal mass [48], reduce exactly to the Misner–Sharp–Hernandez construct. The same happens with another example, known as the Brown–York energy, which is defined as a two dimensional surface integral of the extrinsic curvature on the two-boundary of a spacelike hypersurface referenced to flat spacetime [49].

It is important to emphasize that our derivation of the radius $R_{\rm h}$ is fully self-consistent with the established understanding of apparent horizons in general relativity, which are generally defined – even for non-spherical spacetimes – by the subdivision of the congruences of outgoing and ingoing null geodesics relative to the observer. For the simpler case of a spherically-symmetric spacetime, these reduce to the outgoing and ingoing radial null geodesics from a two-sphere of symmetry [47,50–52]. Of course, the FLRW metric is always spherically symmetric, so the Misner–Sharp–Hernandez mass and apparent horizons are simply related via the Birkhoff theorem and its corollary. With spherical symmetry, the general definition of an apparent horizon thus always reduces exactly to Eq. (12) [47,51]. Another way to put this is that Birkhoff's theorem and its corollary allow us to define a 'gravitational horizon' in cosmology which, however, is simply identified as the 'apparent horizon' even in non-spherically-symmetric systems.

It is clear, therefore, that the apparent horizon $R_{\rm h}$ directly tells us which portion of the Universe is gravitationally coupled to the observer. Its observational and theoretical implications have been discussed extensively in both the primary [39] and secondary [46,47] literature, though there is still some confusion concerning its properties. The time-dependent gravitational horizon is not necessarily a null surface, but is sometimes confused with one. Some [53–56] have suggested that objects beyond $R_{\rm h}(t_0) \equiv c/H_0$ are observable today (at time t_0), which is not correct [57–59]. Almost certainly some of this discourse is due to a confusion between coordinate and proper speeds in general relativity. The former may exceed the speed of light c, but there is an absolute limit to the latter, whose value must be calculated using the curvature-dependent metric coefficients. A misunderstanding of this distinction can lead to claims of recessional speeds exceeding c, even within the observer's particle horizon [60].

An indication of $R_{\rm h}$'s relevance to our interpretation of the data is provided by the many cosmological observations [61] now pointing to what could only be called a very curious *coincidence*: the data are telling us that $\dot{R}_{\rm h}(t) = c$ [46]. Those familiar with the Schwarzschild horizon might at first find this similar to what they would see in free-fall towards a black hole as they cross its event horizon, which would also at that moment appear to be approaching them at speed c. But as we have pointed out, $R_{\rm h}$ in the cosmic context is not yet an event horizon (and may never turn into one), so it evolves in time at a rate dependent on the equation-of-state in the medium. Yet somehow, the observations are telling us that $R_{\rm h} = ct$ as a function of cosmic time t.

From a theoretical perspective, we know that the gravitational horizon in the cosmic setting expands linearly with time only if the cosmic fluid satisfies the zero active mass condition from general relativity, i.e., if its total energy density, ρ, and pressure, p, satisfy the constraint $\rho + 3p = 0$. One can easily understand this from the second Friedmann equation, more commonly referred to as the Raychaudhuri equation [62],

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}\,(\rho + 3p)\,, \tag{15}$$

from which one finds that $\ddot{a} = 0$ as long as $p = -\rho/3$.

A considerable amount of work has been expended over the past decade trying to understand why the Universe would evolve in this manner, and there are now clues – both observational and theoretical – pointing to some possible explanations [46,63]. Insofar as the topic of this paper is concerned, it is not essential for us to dwell on the details right

now, but it turns out that whether or not R_h equals ct is of utmost importance to the identification of rest-mass energy as a gravitational binding energy. As we shall see shortly, this interpretation works only if R_h is indeed expanding linearly with time.

5 Gravitational 'binding energy' and the origin of $E = mc^2$

The notion that an influence in cosmology ought to be restricted by a gravitational horizon is not easy to grasp because spatial flatness in the FLRW metric (Eq. 11) suggests the Universe is infinite. But we have to remember that the *relative* gravitational acceleration between two given spacetime points in the cosmic setting is due solely to the energy in the *intervening* medium. The Birkhoff theorem [64] and its corollary [65,66] help us to understand why every observer or particle – no matter where they are in the presumably infinite cosmos – is surrounded by a gravitational horizon a proper distance $R_\mathrm{h} = c/H$ away. Isotropy ensures that the rest of the Universe outside of a 'spherical shell' at R_h has zero influence on the interior. To be clear, this does not mean that the Universe possesses just a single spherical region bounded by R_h. There exists such a horizon centered on every observer, and every particle within the cosmic fluid. One should therefore expect such a restriction on the size of a causal region to have a significant impact on fundamental physics, especially the question concerning the origin of rest-mass energy. All of our discussion thus far points to the gravitational interaction between m_g and the gravitating energy lying within the particle's horizon R_h as the likely source of rest-mass energy. In this section, we prove this to be true so long as R_h is expanding linearly with time – which appears to be what the observations are telling us.

For reasons that will become clearer shortly, it will be helpful for us to complement the FLRW metric in Eq. (11) with its alternative form written in terms of the observer's 'physical' coordinates, which include the proper radius $R(t) = a(t)r$. The distinction between these two descriptions is that fixing the comoving radius r nevertheless still permits the proper distance $a(t)r$ to change, whereas the observer may choose to keep the physical distance fixed by setting R equal to a constant. It is not difficult to show that [39,67]

$$\begin{aligned} c^2\,dt^2 - a^2\,dr^2 &= \Phi\left[c^2\,dt^2 - \Phi^{-1}\,dR^2 \right.\\ &\quad \left. + 2c\,dt\left(\frac{R}{R_\mathrm{h}}\right)\Phi^{-1}\,dR\right]\\ &= \Phi\left[c\,dt + \left(\frac{R}{R_\mathrm{h}}\right)\Phi^{-1}\,dR\right]^2 - \Phi^{-1}dR^2 \end{aligned} \tag{16}$$

where, for convenience, we have introduced the function

$$\Phi \equiv 1 - \left(\frac{R}{R_\mathrm{h}}\right)^2, \tag{17}$$

which signals the dependence of the metric coefficients g_{tt} and g_{RR} on the proximity of R to the apparent horizon R_h.

If we now consider the worldlines of observers that have t as their proper time from one location to the next – essentially, the comoving observers – then we may introduce the proper speed $\dot{R} \equiv dR/dt$ in the line element and complete the square in Eq. (16). The FLRW metric thus becomes

$$\begin{aligned} ds^2 = \Phi\left[1 + \left(\frac{R}{R_\mathrm{h}}\right)\Phi^{-1}\frac{\dot{R}}{c}\right]^2 c^2dt^2\\ -\Phi^{-1}dR^2 - R^2d\Omega^2. \end{aligned} \tag{18}$$

The expert reader will see a similarity of this equation with that used to derive the Oppenheimer-Volkoff equations for the interior of a star [44,68]. The latter is static, however, whereas both $R(t)$ and $R_\mathrm{h}(t)$ vary with t in FLRW.

Written in this form, the FLRW metric allows us to see how its coefficients vary as a function of R, but even more importantly, in terms of the ratio R/R_h. In principle, we can use it to determine the variation of a particle's characteristics, such as its energy, with distance from the observer – all the way up to the gravitational horizon [69]. Let us define the 4-momentum of a particle

$$p^\mu \equiv (E/c,\, p^R,\, p^\theta,\, p^\phi), \tag{19}$$

written so that the quantity E has units of energy, and p^j ($j = 1, 2, 3$) represent the usual spatial components. We do not assume *a priori* the relationship between E and the vector $\mathbf{p}$, but insist on p^μ being a 4-vector. Then, the actual physical connection between E and $\mathbf{p}$ must be given by the invariance of the contraction $p^\mu p_\mu$ in the spacetime described by Eq. (18). For the metric coefficients in this line element, one has

$$\Phi\left[1 + \left(\frac{R}{R_\mathrm{h}}\right)\Phi^{-1}\frac{\dot{R}}{c}\right]^2\left(\frac{E}{c}\right)^2 - \Phi^{-1}\left(m\dot{R}\right)^2 = \kappa^2, \tag{20}$$

where the invariant contraction κ^2 is a scalar that we must now uncover. Notice that for simplicity and clarity, we have assumed in this expression that the particle's motion is restricted to the Hubble flow, i.e., that its velocity is purely radial, with $p^\theta = p^\phi = 0$ and

$$p^R = m\dot{R}, \tag{21}$$

in terms of the particle's *rest mass*, m.

One accustomed to the language of relativity might be tempted to include a time dilation factor in Eq. (21), which simply reduces to the Lorentz factor γ in Minkowski space, but that would be incorrect here, because the cosmic time t, used to infer the speed $\dot{R}$, also happens to be the local *proper*

time at every spacetime point in the medium. Eq. (20) therefore correctly yields the dependence of E on the particle's momentum $m\dot{R}$ – everywhere in the FLRW spacetime, starting at the origin ($R = 0$), where the observer is situated, all the way to the gravitational horizon at $R = R_\mathrm{h}$.

To bring out this physical connection between E and $\mathbf{p}$ more explicitly, let us re-write Eq. (20) in the form

$$E^2 = \frac{(c\kappa)^2\Phi + (mc)^2\dot{R}^2}{\left[\Phi + \left(\frac{R}{R_\mathrm{h}}\right)\frac{\dot{R}}{c}\right]^2}. \tag{22}$$

We interpret this expression to mean that the particle's energy, E, is a function of both its momentum, $m\dot{R}$, and its distance from the observer in the gravitating medium within R_h. We first consider what happens at the horizon, where $R = R_\mathrm{h}$ and $\dot{R} = c$, while $\Phi = 0$. Clearly,

$$E(R_\mathrm{h}) = mc^2. \tag{23}$$

We might find this hardly surprising, except for two critical facts. First of all, the particle's momentum at $R = R_\mathrm{h}$ is not zero, yet this expression appears to be giving us just the rest-mass energy. Second, notice that the value of E in Eq. (23) does not come from κ, which one would naively have assumed ab initio if we had set $p^\mu p_\mu = (mc)^2$. Instead, *this energy comes from the momentum p^R transitioning to its relativistic limit, $p^R \to mc$*, so that $E \to p^R c = (mc)c$ in Eq. (22). The contribution from κ itself actually gets redshifted away completely because $\Phi \to 0$ when $R \to R_\mathrm{h}$.

The limit $p^R \to mc$ when $R \to R_\mathrm{h}$ follows directly from the Hubble law, which says that the expansion velocity is $v = HR$, in terms of the Hubble parameter $H \equiv \dot{a}/a$ and proper distance R. Thus one may write $v = c(H/c)R$, which simply reduces to $v = cR/R_\mathrm{h}$, leading to the final result given in Eq. (24) with the definition $p^R \equiv mv$.

This remarkable result tells us that the observer sees the particle's energy approach what they can only interpret as an 'escape energy' upon reaching the gravitational horizon, and this quantity is exactly what they would normally consider to be its rest-mass energy mc^2. One must emphasize the phrase 'escape energy' in this conclusion, because this E is entirely due to the momentum $p^R = mc$ the particle needs to overcome its gravitational confinement within R_h. There is no contribution at all to E from κ at $R = R_\mathrm{h}$.

At any other radius $R < R_\mathrm{h}$, the particle's momentum may be written

$$m\dot{R} = mc\left(\frac{R}{R_\mathrm{h}}\right). \tag{24}$$

Equation (22) may thus be re-written as

$$E(R)^2 = (mc^2)^2\left[1 - \left(\frac{R}{R_\mathrm{h}}\right)^2\right]\left(\frac{\kappa}{mc}\right)^2 + (mc^2)^2\left(\frac{R}{R_\mathrm{h}}\right)^2. \tag{25}$$

For most FLRW cosmologies, R/R_h would be a function of time. Thus, E in Eq. (25) could not remain constant at any fixed radius R, regardless of what value κ has. Even so, this energy has the very interesting limit $E \to c\kappa$ when $R \to 0$, but gives no indication of what κ should be. Our argument relating rest-mass energy to the gravitational binding energy within R_h therefore does not appear to work very well for arbitrary FLRW metrics.

The situation changes dramatically for a gravitational horizon expanding at lightspeed, however, which is what the observations seem to be telling us today. In that case, both R and R_h scale linearly with t, and the righthand side of Eq. (25) is entirely independent of time. This is also true of the g_{tt} and g_{RR} coefficients in Eq. (18), which means that energy is conserved along the worldlines of these particular (comoving) observers [65,70]. An easy way to understand this is that a Universe with a linearly expanding R_h has zero active mass (see Sect. 4), so that everything within the gravitational horizon experiences zero net acceleration. The particle therefore cannot gain or lose energy from the background as the Universe expands. For this special case – and only this one – the energy E in Eq. (25) must thus be constant, which therefore means that $\kappa = mc$. Then we see that

$$E = mc^2 \tag{26}$$

everywhere and at all times.

This is a second remarkable result. It tells us that the particle's total energy E remains constant, independent of R, even though its momentum p^R transitions from zero at the origin to a maximum mc at R_h. According to the observer at the origin, the particle thus appears to have a gravitational binding energy mc^2 at their location, which gradually converts into kinetic energy as R increases, and E eventually becomes completely kinetic, equal to $(mc)c$, when $R \to R_\mathrm{h}$. No matter where the particle happens to be, however, its energy never deviates from the fixed value mc^2.

A particle with a peculiar velocity, i.e., a non-zero velocity relative to the Hubble flow, may have non-zero components p^θ and p^ϕ in Eq. (19), and its radial velocity – which we shall now call $\dot{R}_\mathrm{part}$ to distinguish it from the Hubble velocity $\dot{R}$ in the denominator – is not necessarily given by Eq. (24). It is easy to see that, in this more general case, Eq. (22) may instead be written

$$E^2 = \frac{(c\kappa)^2\Phi + (mc)^2\dot{R}_\mathrm{part}^2 + (cR)^2\Phi[p_\theta^2 + \sin^2\theta\, p_\phi^2]}{\left[\Phi + \left(\frac{R}{R_\mathrm{h}}\right)\frac{\dot{R}}{c}\right]^2}. \tag{27}$$

But again $\Phi \to 0$ and $\dot{R}_{\rm part} \to c$ as $R \to R_{\rm h}$, no matter the peculiar velocity, so that we recover the same limiting form of the 'escape' energy, $E \to (mc)c$ at the apparent (or gravitational) horizon.

Near the origin, however, $\Phi \to 1$ and Eq. (27) reduces to

$$E^2 \to (c\kappa)^2 + p^2c^2, \tag{28}$$

where $p^2 \to (m\dot{R}_{\rm part})^2 + R^2[p_\theta^2 + \sin^2\theta p_\phi^2]^2$. We already showed that $\kappa = mc$ leading up to Eq. (26), which must be preserved no matter the momentum, since the contraction $p^\mu p_\mu$ is invariant. And so we recover the well-known Lorentz invariant form of the energy-momentum equation,

$$E^2 = (mc^2)^2 + (pc)^2, \tag{29}$$

near the observer. The cosmological principle then ensures that this relation is the same for every observer throughout the FLRW spacetime.

6 Conclusion

It is important to emphasize the caveat raised above following Eq. (25), that the argument we are making in this paper for the origin of rest-mass energy works only if $R/R_{\rm h}$ has been independent of time throughout the Universe's history. That means that $\dot{R}_{\rm h}$ has been constant at the value c from the Big Bang to today. Among the strange coincidences in cosmology, the worst of them is the fact that the acceleration of the Universe, averaged over a Hubble time, is zero within the measurement error. Of course, this does not mean that $\dot{R}_{\rm h} = c$ from one moment to the next, but if this speed varied according to the prescription of the standard model without the zero active mass condition, the probability of seeing an average $\langle \dot{R}_{\rm h} \rangle = c$ today is 'astronomically' small, effectively zero. In addition, there is some evidence that the inclusion of zero active mass in ΛCDM may improve its consistency with the data [46].

Moreover adopting the zero active mass condition appears to eliminate all horizon problems [71,72], eliminate the standard model's initial entropy problem [73], and provide an explanation for how initial quantum fluctuations created in the early Universe might have classicalized to produce the large-scale structure we see today [74]. If the argument we are making here for the origin of rest-mass energy survives the test of time, perhaps it too may be used to argue in favour of zero active mass in the real Universe.

We are justified in calling mc^2 the particle's gravitational binding energy because the observer at the origin infers this to be the energy it needs to reach 'escape' velocity at $R_{\rm h}$ and free itself from its gravitational coupling to that portion of the Universe contained within this horizon. According to the Birkhoff theorem and its corollary, the rest of the Universe outside of $R_{\rm h}$ does not contribute to this interaction and is therefore not relevant to the question of rest-mass energy. Ironically, this interpretation suggests that all particles, those with inertia and those without, behave equivalently at $R \to R_{\rm h}$, in the sense that their energy there may be written as $E = p^R c$ in all cases. But whereas the momentum of massive particles drops to zero from its maximum value, mc, at the horizon, that of massless particles does not change. So while $E = pc$ always represents an energy associated purely with momentum for the latter, regardless of location, it gradually transitions to a 'rest' energy associated with $m_{\rm g}$ $(= m)$ for the former when viewed by the observer in their vicinity.

One may wonder how we reached this result without actually having 'calculated' the gravitational binding energy directly. This would be a non-trivial task to carry out, given that energy in general relativity is not an invariant quantity from one frame to the next, and would be very difficult to track non-locally. Instead, we have used the invariance of a contracted 4-vector to do this, which allowed us to measure the change in the particle's energy (as viewed from the origin) in terms of its momentum within the Hubble flow. The actual influence of gravity in this approach is represented by the factor $\Phi(R)$ in the metric. As we have seen, the redshift effect associated with $\Phi(R)$ accounts for the gravitational attraction the particle experiences to the rest of the cosmic fluid contained within $R_{\rm h}$.

A successful interpretation of rest-mass energy as a gravitational binding energy would lend some support to evidence emerging from cosmological observations that the equation-of-state in the cosmic fluid is apparently consistent with the zero active mass condition in general relativity. Significant effort is currently being expended addressing this issue, and the results of this investigation will be reported elsewhere.

Acknowledgements I am grateful to the anonymous referee for an excellent, thoughtful review of this manuscript, and for suggesting several key improvements to its presentation.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: This is a theoretical paper, dealing with a topic in fundamental physics, and does not require the use of any existing data, or newly generated data. As such, there are no data to be deposited in connection with this manuscript.]

Funded by SCOAP[3].

References

1. J.J. Thomson, Philos. Mag. **11**, 229 (1881)
2. G.G. Stokes, Trans. Camb. Philos. Soc. **8**, 105 (1844)
3. F. Englert, R. Brout, PRL **13**, 321 (1964)
4. P. Higgs, PRL **13**, 508 (1964)
5. O. Heaviside, Philos. Mag. **27**, 324 (1889)
6. G.F.C. Searle, Philosophical Magazine **44**, 329 (1897)
7. M. Abraham, Annalen der Physik **315**, 105 (1903)
8. H.A. Lorentz, Archives Néerlandaises des Sciences Exactes et Naturelles **25**, 363 (1892)
9. H.A. Lorentz, Proc. R. Neth. Acad. Arts Sci. **6**, 809 (1904)
10. F. Melia, *Electrodynamics* (University of Chicago Press, Chicago, 2001)
11. J.H. Poynting, Philos. Trans. R. Soc. Lond. **175**, 343 (1884)
12. F. Hasenöhrl, Annalen der Physik **320**, 344 (1904)
13. F. Hasenöhrl, Annalen der Physik **321**, 589 (1904)
14. S. Boughn, T. Rothman, (2011). arXiv:1108.2250
15. M. Planck, *General Dynamics. Principle of Relativity* (Columbia University Press, New York, 1909)
16. A. Einstein, Annalen der Physik **18**, 639 (1905)
17. I. Newton, *Philosophiae Naturalis Principia Mathematica* (1687)
18. T. Yarman, A. L. Kholmetskii, C. Marchal, O. Yarman & M. Arik, Journal of Physics **1251** (2019) id. 012051
19. R.V. Eötvös, V. Pekar, E. Fekete, Ann. Phys. (Leipzig) **68**, 11 (1922)
20. R.H. Dicke, *Memoirs of the American Philosophical Society. Jayne Lecture for 1969*, vol. 78 (American Philosophical Society, Philadelphia, 1970)
21. Y. Su et al., Phys. Rev. D **50**, 3614 (1994)
22. S. Baessler et al., Phys. Rev. Lett. **83**, 3585 (1999)
23. T.A. Wagner et al., Class. Quantum Gravity **29**, id. 184002 (2012)
24. G.M. Tino et al., Prog. Part. Nucl. Phys. **112**, id. 103772 (2020)
25. G. Berkeley, *The Principles of Human Understanding* (Jeremy Pepyat, Dublin, 1710)
26. E. Mach, *History and Root of the Principle of the Conservation of Energy* (English translation published by The Open Court Publishing Co., Chicago, 1911)
27. A. Einstein, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, pp. 142–152 (1917)
28. A. Einstein, *The Meaning of Relativity* (Methuen, London, 1946)
29. D.W. Sciama, MNRAS **113**, 34 (1953)
30. C. Brans, R.H. Dicke, Phys. Rev. **124**, 925 (1961)
31. A. Pais, *Subtle Is the Lord: The Science and the Life of Albert Einstein* (Oxford University Press, Oxford, 2005), pp. 287–288
32. S.L. Glashow, Nucl. Phys. **22**, 579 (1961)
33. S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967)
34. A. Salam, in *Elementary Particle Physics: Relativistic Groups and Analyticity. Eighth Nobel Symposium*, ed. by N. Svartholm (Almquvist and Wiksell, Stockholm, 1968)
35. R. Oerter *The Theory of Almost Everything: The Standard Model, the Unsung Triumph of Modern Physics* (Penguin Group, New York, 2006)
36. S. Dürr et al., Science **322**, 1224 (2008)
37. T.-P. Cheng, L.-F. Li, *Gauge Theory of Elementary Particle Physics* (Clarendon Press, Oxford, 1984)
38. D. Miller, (2008). http://www.scienceinschool.org/print/650
39. F. Melia, AJP **86**, 585 (2018)
40. Planck Collaboration et al., A&A **641**, A6, 67 pp (2020)
41. A.B. Nielsen, M. Visser, CQG **23**, 4637 (2006)
42. G. Abreu, M. Visser, Phys. Rev. D **82**, 044027, 10 pp (2010)
43. W. Rindler, MNRAS **116**, 662 (1956)
44. C.W. Misner, D.H. Sharp, Phys. Rev. **136**, 571 (1964)
45. W.C. Hernandez Jr., C.W. Misner, ApJ **143**, 452 (1966)
46. F. Melia, *The Cosmic Spacetime* (Taylor & Francis, Oxford, 2020)
47. V. Faraoni, *Cosmological and Black Hole Apparent Horizons* (Springer, New York, 2015)
48. A. Prain, V. Vitagliano, V. Faraoni, L. M. Lapierre-Léonard, CQG **33**, 145008, 13 pp (2016)
49. S. Chakraborty, N. Dadhich, J. High Energy Phys. **2015**, id.3, 19 pp (2015)
50. I. Ben-Dov, Phys. Rev. D **75**, 064007, 15 pp (2007)
51. V. Faraoni, Phys. Rev. D **84**, 024003, 15 pp (2011)
52. I. Bengtsson, J.M.M. Senovilla, Phys. Rev. D **83**, 044012, 30 pp (2011)
53. T.M. Davis, T.H. Lineweaver, PASA **21**, 97 (2004)
54. P. van Oirschot, J. Kwan, G.F. Lewis, MNRAS **404**, 1633 (2010)
55. G.F. Lewis, MNRAS **432**, 2324 (2013)
56. D.Y. Kim, A.N. Lasenby, M.P. Hobson, GRG **50**, id.29, 37 pp (2018)
57. O. Bikwa, F. Melia, A.S.H. Shevchuk, MNRAS **421**, 3356 (2012)
58. F. Melia, JCAP **09** (2012) 029, 10pp
59. F. Melia, CQG **30**, 155007, 14 pp (2013)
60. W.M. Stuckey, Am. J. Phys. **60**, 142 (1992)
61. F. Melia, MNRAS **481**, 4855 (2018)
62. A.K. Raychaudhuri, Phys. Rev. **90**, 1123 (1955)
63. F. Melia, Ann. Phys. **411**, id. 167997, 5 pp (2019)
64. G. Birkhoff, *Relativity and Modern Physics* (Harvard University Press, Cambridge, 1923)
65. S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972)
66. F. Melia, MNRAS **382**, 1917 (2007)
67. F. Melia, M. Abdelqader, IJMP-D **18**, 1889 (2009)
68. J.R. Oppenheimer, G.M. Volkoff, Phys. Rev. **55**, 374 (1939)
69. F. Melia, IJMP-A **34**, id. 1950055 (2019)
70. W. Killing, Journal für die reine und angewandte Mathematik **109**, 121 (1892)
71. F. Melia, A&A **553**, id. A76, 6 pp (2013)
72. F. Melia, EPJ-C Lett. **78**, 739 (2018)
73. F. Melia, EPJ-C **81**, 234 (2021)
74. F. Melia, PLB **818**, id. 136632, 14 pp (2021)