Article

Gravity as a Quantum Field Theory

Roberto Percacci $^{1,2}$

$^1$ Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, 34134 Trieste, Italy; percacci@sissa.it
$^2$ INFN, Sezione di Trieste, 34127 Trieste, Italy

Abstract: Classical gravity is understood as the geometry of spacetime, and it seems very different from the other known interactions. In this review, I will instead stress the analogies: Like strong interactions, the low energy effective field theory of gravity is related to a nonlinearly realized symmetry, and like electroweak interactions, it is a gauge theory in Higgs phase, with a massive connection. I will also discuss the possibility of finding a UV complete quantum field theoretic description of all interactions.

Keywords: quantum gravity; gauge theories; unification

1. Introduction

There is a wide gap between the current descriptions of gravity and those of the other interactions, which is partly due to objective, and partly to historical reasons. It is widely expected that this gap will eventually be filled, but there is no consensus on the way in which this will happen.

There are two known long range interactions, gravity and electromagnetism, that admit a purely classical description, and two other fundamental interactions, the weak and strong nuclear forces, that act only at very short distance scales and that manifest themselves as purely quantum phenomena.

For many applications, classical gravity is well-described by Newton’s theory, whose fundamental equations are almost identical to those of classical electrostatics. Maxwell’s theory is richer, since it also describes magnetism, and requires a vector potential in addition to the electrostatic scalar potential. For strong gravitational fields, Newton’s theory fails and is superseded by General Relativity (GR), which is even richer. In fact, in certain situations, GR has a close similarity to electromagnetism $^{[1,2]}$, but a full description of all its features requires a tensor field. Aside from containing more variables, conceptually, GR differs from all previous theories, insofar as space and time stop being the fixed arena in which particles and fields evolve, and themselves become dynamical.

Classical electromagnetism evolved in a completely different direction. In order to account for the structures of atoms and the properties of black body radiation, the electromagnetic field had to be treated according to the laws of quantum mechanics and came to be seen as being composed of elementary particles called photons. Since Maxwell’s theory is Lorentz invariant, a proper description of photons and their interactions required the development of special relativistic quantum mechanics. Thus, Quantum Electrodynamics (QED) came into existence. It was the first example of a Quantum Field Theory (QFT), and leads to stunningly precise predictions. Through a complicated history $^{[3]}$, it was understood that the weak and nuclear interactions could be described by (quantum) Yang–Mills (YM) theories, which are non-abelian generalizations of QED. For the electroweak interactions, the non-abelian group is $SU(2) \times U(1)$, which is “spontaneously broken” to the electromagnetic $U(1)$, at the Fermi scale. For Quantum Chromodynamics (QCD), the theory of the strong interactions, the group is $SU(3)$; the main features of this theory are asymptotic freedom, confinement, chiral symmetry breaking, and the dynamical generation of a low energy characteristic scale.
So now we have these extremely successful theories, which are based on widely different conceptual foundations and mathematical tools. They describe phenomena occurring at widely different scales: sub-atomic distances for QFTs, and macroscopic to cosmic distances for GR. As long as we stay away from certain currently unreachable situations, these theories are perfectly adequate. However, it seems unlikely that the status quo is the ultimate truth. Rather, one would expect a more uniform description of all interactions to be possible, which in all likelihood means that gravity also has to be subjected to the laws of quantum mechanics. Hence, the search for Quantum Gravity (QG). Even more ambitiously, one may look for a unified theory: besides the use of similar conceptual foundations and mathematical tools, one could strive for a description where all forces are manifestations of a single fundamental force, their differences arising from some kind of symmetry breaking, perhaps akin to what is seen in the electroweak theory, or perhaps due to some dimensional reduction or still some other mechanism.

Leaving aside the issue of unification, this paper deals mainly with the search of uniformity. In this respect, it may be useful to think for a moment of possible different historical developments that could have led to very different perspectives. YM theories have been thought of from the start as QFTs, but they can also be treated as classical field theories. In fact, around the 1970s, it emerged that classical YM fields can play important roles in QFT, in the form of solitons and instantons. This led to a very fruitful interaction between physicists and mathematicians, as it became clear that the latter had developed a formalism, the theory of connections in fiber bundles, which is the right mathematical framework to describe YM fields, much in the same way as Riemannian geometry, which was found to be the right mathematical formalism to describe gravity. Thus, we could have first understood YM theories as classical field theories of connections, as the mathematicians did, without knowing about the strong and weak interactions, and we could have regarded them as being essentially geometrical, though this geometry concerns internal spaces instead of space and time. The discovery of the $W$ and $Z$ particles, and of gluons, would then have been seen as a proof that “geometry is quantized”.

The opposite could have happened with gravity. In his lectures on gravitation [4], Feynman imagines a world in which QED had been developed before GR. Presented with the motion of a massive body whose trajectory is bent due to the presence of another massive body, we would have attributed the bending to the exchange of some bosonic particle (“force mediator”), analogous to the photon. The general properties of gravity lead us to identify such mediators as spin-two particles, that we would call “gravitons”. The propagator of the gravitons must have a special form (that was found by Fierz and Pauli), exhibiting a gauge symmetry. This gauge symmetry also severely constrains the form of the self-interactions between gravitons. Feynman worked out the three-point vertex, which is sufficient to calculate the precession of perihelia to a reasonable accuracy, but there are infinitely many higher vertices that cannot be found in practice in this way. The way to reconstruct all the vertices was found later by Deser [5], and interestingly, it requires treating the connection as an independent variable. Thus, one can arrive at GR “from the bottom up”, by studying the interactions of spin-two particles. This theory is perturbatively non-renormalizable, but it can be consistently treated as an Effective (Quantum) Field Theory (EFT). Using the Pauli-Fierz propagator in tree level diagrams, one can immediately reconstruct the Newtonian form of the scattering potential. One loop diagrams reconstruct general relativistic corrections to the potential, as well as genuine quantum corrections. It is remarkable that in spite of non-renormalizability, these corrections are entirely unambiguous [6].

This contradicts the often made statement that “we still do not have a quantum theory of gravity”, and reveals a “split brain syndrome” [7]. If we call gravity the force that makes apples fall, then certainly, Feynman’s theory of gravitons in flat space, and its modern extensions, qualifies as a quantum theory of gravity. The reason for the split brain is that, having learned from Einstein that classical gravity can be interpreted as the geometry of spacetime, we are reluctant to accept a theory based on flat space (or
any other fixed background) as a quantum version of GR. What is commonly meant by "quantum gravity" is something much more ambitious, for which the alternative name of "quantum theory of spacetime" would perhaps be more appropriate. However, we should not necessarily expect such a theory to bear much resemblance to GR. (The attitudes of particle physicists to the role of geometry have changed in the course of time. Throughout the 1970s, the prevailing attitude was similar to Feynman’s. For example, Weinberg’s book [8] famously contains several comments that tend to minimize the importance of the "geometric analogy". On the other hand, geometry plays an important role in string theory.)

One possible UV completion of the EFT of gravity is Asymptotic Safety (AS) [9,10]. It is a rather conservative approach based on standard continuum QFT methods, but it is nonperturbative in two ways: it does not rely on an expansion around flat space, and it does not rely on an expansion for small coupling. In practice, one can still expand around a background field, but one tries to keep the background arbitrary. There is then a sense in which "all backgrounds is no background", thus addressing some of the issues mentioned above. AS is unfamiliar in particle physics. Examples of AS theories in four dimensions containing gauge fields, fermions, and scalars interacting via the usual couplings have only been found quite recently [11], but the techniques to discover them already existed in the 1970s, so this is another historical accident.

The most important common feature of gravity and YM theories is the connection: in both cases, it is necessary to describe parallel transport and to define covariant derivatives, but unlike in YM theories, in GR, the connection is not dynamical. Theories of gravity with independent metric and connection are called Metric-Affine Gravity (MAG) theories. As I will discuss in Section 3, in a generic MAG, the connection acquires a large mass due to a kind of symmetry breaking mechanism that is very similar to the one that operates in electroweak theory, or in superconductivity. This property of MAGs is not widely appreciated, for a variety of reasons. First of all, Einstein’s original formulation of GR adequately describes all observations, and the presence of the connection gives rise to complications that are unnecessary for most purposes. Then, the most familiar way of introducing a dynamical connection is via the Palatini action, which produces vacuum field equations that are equivalent to those of GR. Furthermore, every MAG can be reformulated as a metric theory of gravity coupled to some specific tensor fields, that represent either torsion, or nonmetricity, or both. One can justifiably view these as special cases of metric gravity coupled to matter. ("Torsion is just another tensor", see e.g., the discussion between S. Weinberg and F. Hehl in [12].) Finally, in Poincaré gauge theories of gravity, the tetrad is viewed as a translational gauge field, and torsion as its curvature, see e.g., [13]. This is an attractive interpretation from one point of view, but it hides the nonlinear nature of the frame field and its role as an order parameter, which I want to emphasize here.

Viewing gravity as a theory with a massive connection leads one to suspect the existence of an alternative “unbroken” phase with dramatically different properties. It is quite possible that spacetime is “emergent” in a sense to be properly defined, and that the more fundamental description of nature is based on entirely different concepts. How much of the structure of GR will have to be given up is one of the central questions for the quantum theory of spacetime. Conservative approaches such as AS have the advantage that many familiar tools remain available, and the relation to the low energy EFT is easier to understand, at least in principle. More radical, “top down” approaches typically have greater difficulty connecting to low energy physics.

The plan of this paper is as follows. In Section 2, I will start by reviewing the formulation of gravity as an EFT, which provides a close parallel with strong interactions at low energy, and is by now completely well established. In Section 3, I will discuss gravity as a gauge theory. This highlights analogies to electroweak interactions and suggests the natural extensions of the EFT. The rest of the paper is of a more speculative nature. In Section 4, I will discuss AS in gravity, and in Section 5, I will mention some of the numerous questions that remain open.
2. Gravity as an Effective Field Theory

There was a time when only renormalizable theories were considered worthy of attention. In non-renormalizable theories, every possible monomial in the Lagrangian has a divergent coefficient at some order in perturbation theory, so that all the coefficients of local terms in the Lagrangian have to be determined by an experiment, undermining the predictivity of the theory. This logic has played a central role in establishing the SM. It has emerged later that non-renormalizable theories can also be predictive and useful, as long as one only considers processes that happen at scales E below a characteristic “cutoff” scale M. (Here, the cutoff is to be understood as a physical scale setting the upper limit of the domain of validity of the theory; not as a mathematical device that is used to regularize ill-defined integrals.) The point is that, at a given energy E ≪ M, and for measurements of some given finite precision, only a finite number of Feynman diagrams contribute.

2.1. The Chiral Models

Typical examples of EFTs are theories carrying a nonlinear realization of some symmetry group, where the Lagrangian is non-polynomial in the field. Probably the most important case is the chiral model describing the dynamics of pions, viewed as (pseudo-)Goldstone bosons of the spontaneously broken SU(2) chiral symmetry of QCD with two massless quarks. The Goldstone bosons are nonlinear fields having values in the coset space SU(2)\_L × SU(2)\_R / SU(2)\_V, which is diffeomorphic to SU(2). Thus, we can describe the elements of the coset by a group-valued field

\[ U(x) = \exp \left( i \frac{\pi^a(x) \sigma_a}{2 F_\pi} \right), \]

where \( \pi^a \) are the pion fields, \( \sigma_a \) are the Pauli matrices, and \( F_\pi = 92 \text{ MeV} \) is the pion decay constant. The most general SU(2)\_L × SU(2)\_R—\textit{invariant} action can be written as sums of traces of powers of the Lie algebra-valued field \( U^{-1} \partial_\mu U \). The first terms in an expansion in the number of derivatives are \( \mathcal{L} = \mathcal{L}_2 + \mathcal{L}_4 + O(\partial^6) \), with

\[ \mathcal{L}_2 = -\frac{F_\pi^2}{4} \text{tr}(U^{-1} \partial_\mu U)^2 \]

(2)

\[ \mathcal{L}_4 = \ell_1 \text{tr}(((U^{-1} \partial_\mu U)^2)^2) + \ell_2 (\text{tr}(U^{-1} \partial_\mu U)^2)^2. \]

(3)

Let us concentrate for a moment on the first term, containing two derivatives. If we expand the exponential, it gives rise to a canonical pion kinetic term plus infinitely many interaction terms, all involving two derivatives and increasing powers of \( g = 1/F_\pi \):

\[ L_2 = \frac{1}{2} \left[ (\partial_\mu \pi^a)^2 - \frac{1}{12} \delta^2 \left[ (\pi^a \pi^a) (\partial_\mu \pi^a)^2 - (\pi^a \partial_\mu \pi^a)^2 \right] \right. \]

\[ \left. + \frac{1}{360} \delta^4 \left[ (\pi^a \pi^a) (\partial_\mu \pi^a)^2 - (\pi^a \partial_\mu \pi^a)^2 \right] + O(\pi^8) \right]. \]

(4)

We see that the perturbative coupling in this action is \( g \). It has dimensions of length, and so via power counting, this theory must be non-renormalizable. Furthermore, tree level scattering cross-sections grow with momentum and would violate unitarity at a scale of \( M = 16 \pi F_\pi \), which is of the order of the GeV. This is the effective cutoff for these models.

A systematic analysis [14] shows that at order \( n \) in \( (p/M)^2 \), one must take into account diagrams with \( n - 1 \) loops constructed from \( L_2 \) and \( n - 2 \) loops constructed from \( L_4 \), down to tree diagrams from \( L_n \). In practice for low-energy meson physics, one needs \( F_\pi, \ell_1, \ell_2 \), and a bunch of other parameters that are related to the quark masses. Calculations at one loop in \( F_\pi \) and at tree level in \( \ell_1, \ell_2 \) successfully describe a rich phenomenology [15]. The lesson we learn from this example is that non-renormalizable theories can be used to systematically calculate processes by expanding in powers of \( p/M \).
We shall see in Section 3.2 that the same model can be reinterpreted as the low energy EFT of Goldstone bosons in the electroweak theory.

2.2. Gravity

The dynamical variable of General Relativity is the metric tensor. Viewing it as a tensor is, however, misleading: the metric is subject to the constraint of being non-degenerate, with the eigenvalues of specific signs. This implies that at each spacetime point, the metric is a field with values in the coset space $GL(4)/O(1,3)$. It therefore carries a nonlinear realization of the linear group, and is kinematically very similar to nonlinear sigma models [16]. In this sense, one can refer to the metric as a Goldstone boson. One can expand the metric analogously to (1)

$$g_{\mu\nu} = \bar{g}_{\mu\rho}(e^{h/m_P})^\rho_{\mu}$$

where $h_{\mu\nu} = \bar{g}_{\mu\rho}h^\rho_{\mu}$ is the graviton field and $m_P$ is the Planck mass. A general diffeomorphism-invariant Lagrangian for a metric $g$ is a combination of scalars constructed with the Riemann tensor $R_{\mu\nu\rho\sigma}$, its contractions $R_{\mu\nu} = R_{\lambda\mu}^\lambda_{\nu}$, $R = R^\mu_{\mu}$, and the covariant derivatives thereof. Expanding in derivatives, the first few terms read $L = L_0 + L_2 + L_4 + O(\partial^6)$, where

$$L_0 = m_P^2 \Lambda$$

$$L_2 = -\frac{1}{2}m_P^2R$$

$$L_4 = \alpha R^2 + \beta R_{\mu\nu}R^{\mu\nu} + \gamma R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma}$$.

If we recall that the Christoffel symbols have the structure $\Gamma \sim g^{-1}\partial g$, and that the curvature tensors contains terms of the form $\Gamma^2 \sim (g^{-1}\partial g)^2$, this is very similar to the chiral action (2), with $m_P$ in place of $F_{\pi}$. (The curvature scalar also contains terms $\partial^2 \Gamma$ with the second derivatives of the metric, but these terms can be eliminated with some integrations by parts.) In particular, both actions are nonpolynomial and describe massless particles that have derivative interactions.

The most striking difference is the appearance of the overall factor $\sqrt{g}$, which is necessary to ensure the diffeomorphism invariance of the measure. Another, related, difference is that a constant term in the Lagrangian can be dropped in the chiral model but not in GR, where it gives rise to non-derivative interactions. Both differences disappear in the unimodular version of GR, which is the theory where $\text{det} g = 1$. Indeed, in this case, the “cosmological term” becomes a field-independent constant in the action, and as such, can be safely ignored. We also note that unimodular gravity has the smallest possible gauge group that is compatible with locality: the three-parameter volume-preserving diffeomorphisms, instead of the four-parameter diffeomorphism group [17].

The power counting in the EFT of gravity works as in the chiral model, with $E/m_P$ playing the role of the expansion parameter. Thus, tree level diagrams with vertices from $L_2$ give contributions of order $(E/m_P)^2$ while those with vertices from $L_4$ are of order $(E/m_P)^4$, one-loop diagrams with vertices from $L_2$ give contributions of order $(E/m_P)^4$, and so on. In this way, using this expansion in $E/M$, it is in principle possible to compute the quantum effects for energies that are below the Planck scale.

Note that the coefficients of local terms in $L$ cannot be calculated: they must be determined by some experiment. However, the calculations of cross-sections involve non-local terms that do not suffer from such renormalization ambiguities, as we shall see in the following example.

2.3. The Leading Corrections to the Newtonian Potential

The gravitational potential can be studied by scattering two massive, non-relativistic particles. It has the familiar Newtonian term, plus corrections that can be systematically
calculated with Feynman diagrams [18]. On dimensional grounds, the potential must have the form

\[ V(r) = -\frac{G m_1 m_2}{r} \left[ 1 + a \frac{G (m_1 + m_2)}{r c^2} + b \frac{G \bar{h}}{r^2 c^4} + \ldots \right] \]  

(9)

The first term is the Newtonian potential. It can be reproduced in perturbation theory by considering the tree level diagram where two almost static particles exchange a graviton. The second term in the bracket does not contain factors of \( \bar{h} \) and is a classical effect. It represents a general-relativistic correction. The third term is linear in \( \bar{h} \) and represents the leading quantum correction to the potential. One would like to calculate the coefficients \( a \) and \( b \). In perturbation theory at one loop, one has to evaluate several diagrams. These are all divergent, and one may worry about the renormalization ambiguities. As a matter of fact, the terms that enter in the evaluation of the coefficients \( a \) and \( b \) are completely immune from these ambiguities. The Fourier transforms of the corrections to the potential in (9) behave as \( \frac{1}{|p|} \) and \( \log \left( \frac{p^2}{\mu^2} \right) \). These momentum-space amplitudes are non-analytic in \( p \), and are clearly distinct from the contributions of local counterterms, which give analytic corrections to the amplitude. The higher-derivative, UV-divergent terms appear as contact interactions, whereas the terms that we are interested in originate from the low energy part of the momentum integrations. It is therefore possible to neatly disentangle the UV effects, which are subject to renormalization ambiguities, from the IR effects, which are not.

The actual calculation yields the following values for the coefficients [19–21]:

\[ a = 3, \quad b = \frac{41}{10 \pi} \]  

(10)

Several comments are in order at this point. In order to avoid misunderstandings, it should be emphasized that (9) is a scattering potential: it is defined operationally as the Fourier transform of the scattering amplitude, and cannot be measured by a Cavendish experiment, for example. The radius \( r \) is defined by the background flat metric. Different operational definitions of the potential may require different calculations, see e.g., [22].

The second term in the square bracket in (9) is classical. It had been computed originally in [23]. Such modifications reproduce the leading terms of the expansion of the Schwarzschild metric, written in the harmonic gauge [24]. Similar diagrams involving photons generate the leading terms of the Reissner–Nordstrom metric [25]. The fact that loop diagrams reproduce the classical general relativistic correction to the Newtonian potential provides a counterexample to the general belief that the expansion in loops is also an expansion in powers of \( \bar{h} \). Another counterexample, and a detailed explanation of the reasons behind the failure of the standard argument, has been given in [26].

The third term in the square bracket in (9) is of truly quantum origin. The coefficient \( b \) is unaffected by higher loop corrections, which give contributions that fall off even faster with distance. It will be different in the presence of other massless particles, in particular, the photon [27,28]. The smallness of the quantum effects is the result of the wide separation between the scales where observations are made, and the “scale of new physics”, which in this case is the Planck scale. In no other EFT is the separation of scales so large, and the expansion parameter so small. Somewhat paradoxically, this EFT of gravity is the most perturbative QFT. It also has a very broad range of applicability.

This calculation provides an explicit proof of the fact that the theory of gravitons in flat space is a consistent framework to compute observable quantum gravity effects at low energy. There is no reason to doubt that these calculations are not only mathematically, but also physically correct. Of course, it remains true that this EFT does not solve any of the problems that are usually given as motivations for research in quantum gravity: UV problems (the EFT methods break down at the Planck scale), IR problems (this EFT does not seem to say anything about the cosmological constant), and strong field problems (gravitational singularities). Additionally, it is certainly not a quantum theory of spacetime.
3. Gravity as a Gauge Theory in Higgs Phase

It is instructive to think of gravity as a gauge theory, because in this way, the close similarities to other phenomena emerge more clearly. To this end, we will begin by recalling the main features of some “spontaneously broken gauge theories”, or more precisely, “gauge theories in Higgs phase”, at very low energy.

3.1. Superconductors

In the context of condensed matter physics, the basic properties of superconductivity can be deduced from the statement that in the bulk of a superconductor the electromagnetic $U(1)$ gauge invariance is spontaneously broken [29]. In a superconductor, the current is carried by Cooper pairs, which are loosely bound states of two electrons. The microscopic BCS theory explains how these pairs form, but it is otherwise not necessary to understand the large scale behavior of the material. For this purpose, we can use the simpler and older Ginzburg–Landau theory, which is just an effective low-energy theory that had been introduced in the study of phase transitions. The Ginzburg–Landau theory for superconductivity is basically scalar electrodynamics, with the scalar representing the density of Cooper pairs. We can decompose the scalar field in its radial and angular parts

$$\phi(x) = \rho(x) \exp(i\varphi(x)).$$

The radial mode is the order parameter. If its expectation value is zero, which happens to be the case above the critical temperature, we are in the normal phase. Below the critical temperature $\rho$ develops a VEV $\rho_0$ and we are in the Higgs phase, which happens to correspond to the superconducting state. The second derivative of the potential around the VEV is the mass of the radial mode. If we consider the material at scales that are much larger than the inverse of this mass, we can assume that $\rho$ is frozen and the only remaining degree of freedom is the phase $\varphi$, which acts as a Goldstone boson. In this phase, the electromagnetic $U(1)$ group is non-linearly realized, because the Goldstone boson will enter the Lagrangian either through its covariant derivative $D_\mu \varphi = \partial_\mu \varphi - A_\mu$ or through other gauge-invariant combinations involving other degrees of freedom. In the relativistic generalization of the Ginzburg–Landau theory, the kinetic term of the Goldstone boson is

$$-\frac{1}{2} \rho_0^2 (D\varphi)^2.$$

One can choose the “unitary” gauge $\varphi = \text{const}$, and then this kinetic term becomes a mass terms for the electromagnetic field, $\rho_0^2 A_\mu A^\mu$. This is the essence of the Higgs phenomenon. Furthermore, in a static situation, the energy is minimized by

$$D_\varphi \varphi = 0,$$

which implies that the magnetic field $B_\varphi = 0$. This is known as the Meissner effect: the superconductor expels the magnetic field lines. An external magnetic field will decay exponentially from the surface inward, with a characteristic length that is inversely proportional to the mass acquired by the electromagnetic field. Note that the electromagnetic field strength can be interpreted geometrically as curvature, so that the Meissner effect means that the superconductor is “stiff”.

3.2. Electroweak Theory

The Standard Model (SM) offers another prominent example. In this case, the group $U(1)$ is replaced by the electroweak group $SU(2)_L \times U(1)$, and the complex scalar $\phi$ is replaced by the complex Higgs doublet $H$. The VEV $\rho_0$ is replaced by $v = 246$ GeV. The
Higgs potential is invariant under $SO(4) \sim SU(2)_L \times SU(2)_R$, and the angular modes of the Higgs field (the Goldstone bosons) parametrize a three-sphere. At energies that are much lower than the mass of the Higgs (radial) mode, we can assume that the latter is frozen, and the dynamics reduce just to the Goldstone bosons coupled to the gauge fields [30,31]. The kinetic term of the Goldstone bosons is

$$L_g = -\frac{\upsilon^2}{2} D_\mu \varphi^\alpha D^\mu \varphi^\beta h_{\alpha \beta} \quad \text{where} \quad D_\mu \varphi^\alpha = \partial_\mu \varphi^\alpha + g A^a_\mu K^a_\alpha (\varphi),$$

where $h_{\alpha \beta}$ is an invariant metric on the sphere and $K^a_\alpha (\varphi)$ are the Killing vectors generating the action of the gauge group. In this regime, the electroweak group is non-linearly realized. We note that in the absence of gauge fields, this Lagrangian is actually identical to (2), except for the replacement of $F_\pi$ with $\upsilon$. At energies much below $\upsilon$, the analog of (11) will hold:

$$D_\mu \varphi^\alpha = 0,$$

which is just a gauge invariant way of saying that of the four gauge fields, only the electromagnetic $U(1)$ can be excited, while the $W$ and $Z$ (and their curvatures) must vanish.

The core of the Higgs phenomenon is the fact that the kinetic term of the Goldstone bosons, when written in the unitary gauge, becomes a mass term for the $W^{\pm}_\mu$ and $Z_\mu$. To say it differently, the Goldstone bosons are pure gauge variables, whose presence is necessary in order to write the mass term of the gauge fields in a gauge-invariant way.

The electroweak theory also contains a radial mode, but so far, we cannot be sure whether it is a fundamental field or a composite state. Thus, in spite of the striking success of the SM, we still do not have a deep understanding of electroweak symmetry breaking, and the SM itself may be just an effective field theory, akin to the Ginzburg–Landau model.

### 3.3. Gravity in Arbitrary Frames

Before we allow the connection to become an independent variable, let us discuss briefly the question: if gravity is a gauge theory, what is its gauge group? This question could be answered in many different ways and it is not a physically meaningful one. The reason for this is that only global symmetry transformations have a physical meaning; by definition, local gauge transformations map field configurations to mathematically different but physically indistinguishable field configurations, and therefore, they are merely convenient redundancies in our description of nature. The convenience lies therein, that by accepting gauge redundancy, we are able to describe nature with local fields and local interactions. When one tries to discover the true physical degrees of freedom in a gauge theory, these are found to be non-local. (For example, at a perturbative level, the photon is described using a four component vector, but its longitudinal component is a gauge degree of freedom. If we try to describe the photon just by the transverse components of the vector, we have to use the inverse Dalambertian, which is a non-local operator. More generally, a gauge field configuration could be completely described by giving the “Wilson loops” (or “holonomies”), which are gauge-invariant but non-local.)

The answer that one chooses is thus a matter of convenience: it depends on what variables one wants to work with. As already mentioned, the smallest gauge group for gravity that still allows us to maintain locality is the group of volume-preserving diffeomorphisms [17]. On the other hand, there is no upper limit on the size of the gauge group: one can always artificially introduce new degrees of freedom, and at the same time make them unphysical by choosing a Lagrangian that does not depend on them. It is sometimes useful to do so, as demonstrated by Stückelberg’s treatment of massive QED [32]. In practice, in addition to diffeomorphisms, the largest gauge group that has some reasonable justification is the group of local $GL(4)$ transformations. (This gauge group can be characterized globally as the automorphisms of the bundle of frames. Sometimes, one also considers the affine group instead of the linear group. We will refrain from doing so, because it leads to an interpretation of the frame field as a gauge potential, whereas
here we view it as a “Goldstone” field.) This choice makes sense, for two reasons. First, the local $GL(4)$ freedom amounts just to the freedom of choosing an independent frame at each spacetime point, and therefore, it has a clear and natural geometrical meaning. We also understand intuitively that physical quantities cannot depend on the choice of frame that one chooses to describe them in. Second, the holonomies of a spacetime connection belong, in general, to $GL(4)$. (Or to the Lorentz subgroup, if one assumes a priori that the connection is metric-compatible.)

Let us therefore start from a reformulation of gravity that has manifest $GL(4)$ invariance [33,34]. To this end, we shall work with generic linear bases, or “frames”, $\{\theta_a\}$, where $\theta_a = \theta^\mu_a \partial_\mu$, or equivalently, the dual co-frames (a.k.a. “soldering form”) $\theta^a = \theta^a_\mu \, dx^\mu$. The frame (or co-frame) field will be a dynamical variable, generalizing the tetrad. In addition we have a dynamical metric, whose components in the frame $\{\theta_a\}$ are $g_{ab}$.

We have already mentioned that the metric is a nonlinear field, having values in the coset $GL(4)/O(1,3)$, and that the same is true of the frame, which is constrained to be nondegenerate, and therefore locally has values in $GL(4)$. For this reason, I will refer to $\theta^a_\mu$ and $g_{ab}$ as Goldstone bosons.

The components of the metric and connection in a coordinate basis are given by

$$g_{\mu\nu} = \theta^a_\mu \theta^b_\nu \, g_{ab},$$

$$\Gamma^a_{\mu\nu} = \theta^{-1} a_\mu \Gamma^a_{\lambda \nu} \theta^\lambda + \theta^{-1} a_\nu \partial_\lambda \theta^\mu.$$

They are invariant under the local $GL(4)$ transformations, which act on the fields as follows:

$$\theta'^a_\mu = \Lambda^{-1} a_\mu \theta^c_\nu \Lambda^d_b + \Lambda^{-1} a_\mu \partial_\lambda \Lambda^c_b,$$

$$\theta'^a_\mu = \Lambda^{-1} a_\mu \theta^c_\nu,$$

$$g'_{ab} = g_{cd} \Lambda^c_a \Lambda^d_b.$$

Choosing $\Lambda = \theta$, we see that we can enforce the gauge condition $\theta'^a_\mu = \delta^a_\mu$. This brings us to a coordinate basis and completely breaks $GL(4)$. In this gauge, (14) shows that there is no difference between Latin and Greek indices. This corresponds to the standard formulation of gravity in terms of a metric. This will be called the metric gauge.

Alternatively, one can choose the gauge so that $g'_{ab} = \eta_{ab} = \text{diag}(-1,1,1,1)$, leaving an unbroken $O(1,3)$ gauge group. This means that we are using an orthonormal frame. Equation (14) are the usual relations holding in the tetrad formalism, relating the components of the metric and connection in a coordinate frame to those in an orthonormal frame. This will be called the tetrad gauge.

It is crucial that there is not enough gauge freedom to fix both gauges simultaneously. Therefore, unlike the theories of Sections 3.1 and 3.2, where the Goldstone bosons could be completely gauged away, one of the two Goldstone bosons survives at low energy as a physical dynamical variable. Its dynamics is then very similar to that of the chiral models, as we have seen.

The components of the Levi-Civita Connection in a general frame are

$$\Gamma_{abc} = \frac{1}{2} (E_{cab} + E_{abc} - E_{bac}) + \frac{1}{2} (C_{abc} + C_{bac} - C_{cab}),$$

where $E_{cab} = \theta^{-1} c_\lambda \partial_\lambda \delta_{ab}$ and $C_{abc} = g_{ad} \theta^d_\lambda (\theta^{-1} b_\mu \partial_\mu \theta^{-1} c_\lambda - \theta^{-1} c_\mu \partial_\mu \theta^{-1} b_\lambda)$. In coordinate frames, $C_{abc} = 0$, and this formula reduces to the Christoffel symbols, while in orthonormal frames, $E_{abc} = 0$.

3.4. Metric-Affine Gravity

In GR, an important role is played by the connection and its curvature, but the connection is assumed to be metric-compatible and torsionfree, and this determines it uniquely as a function of the metric and/or frame field. We now consider Metric-Affine theories
of Gravity (MAGs), where parallel transport and covariant derivatives are constructed with an independent, dynamical connection, whose components in a general frame are $A_{\mu}^a = \theta^a_{\mu b} A^b$. It transforms under change of frame in the same way as $\Gamma$ in (15).

The new ingredients in MAG are torsion and non-metricity, which are the covariant derivatives of the Goldstone bosons with respect to the independent connection:

$$T^a_{\mu v} = \partial_\mu \theta^a_v - \partial_v \theta^a_\mu + A^a_{\mu b} \theta^b_v - A^a_{v b} \theta^b_\mu ,$$

$$Q_{\lambda ab} = -\partial_\lambda g_{ab} + A_{\lambda}^c a g_{cb} + A_{\lambda b} g_{ac} .$$

Note that this statement is obscured if one works in coordinate or orthonormal frames. Indeed, in the metric gauge (coordinate frames), torsion appears to be just an algebraic combination of connection components

$$T^a_{\mu v} = A^a_{\mu v} - A^a_{v \mu} ,$$

whereas in the tetrad gauge (orthonormal frames), it is the non-metricity that appears to be a purely algebraic combination of connection components:

$$Q_{\lambda ab} = A_{\lambda ab} + A_{\lambda ba} .$$

In order to recognize that these tensors play very similar roles, it is necessary to work with generic frames.

The Lagrangian of MAG is a scalar constructed with the tensors $T$, $Q$, the curvature

$$F_{\mu \nu}^a = \partial_\mu A^a_{\nu b} - \partial_\nu A^a_{\mu b} + A^a_{\mu c} A^c_{\nu b} - A^a_{\nu c} A^c_{\mu b}$$

and their covariant derivatives. If we restrict ourselves to terms of dimension four or less, there are more than 900 independent terms of this type [35]. However, at low energy, the dynamics are dominated by the dimension-two terms, of which there are only 12. (I will disregard the dimension-zero cosmological term.) The most general dimension-two Lagrangian is

$$L_{(2)} = L_p + L_{TQ} .$$

Here, $L_p = -m^2_p F_{ab}^{\mu \nu}$ is the Palatini Lagrangian, and

$$L_{TQ} = T^{\mu \nu} (a_1 T_{\mu \nu} + a_2 T_{\mu \nu}^2) + \frac{1}{3} T^{\mu \nu} T_{\mu} + Q^{\mu \nu} (a_4 Q_{\mu \nu} + a_5 Q_{\nu \nu})$$

$$+ a_6 Q^a_{\mu} Q_{\mu} + a_7 Q^a_{\mu} \tilde{Q}_{\mu} + a_8 Q^{a b} Q_{a b} + a_9 T^{\mu \nu} Q_{\mu \nu} + a_{10} T^{\mu \nu} Q_{\mu} + a_{11} T^{\mu \nu} \tilde{Q}_{\mu} ,$$

where $T_{\mu} = T^a_{\lambda \mu a}$, $Q_{\mu} = Q_{\mu \lambda a} \tilde{Q}_{\mu} = Q_{a \lambda \mu}$. We recognize that $L_{TQ}$ are the kinetic terms of the Goldstone bosons, whereas the Palatini term has no analog in the Yang–Mills theories.

Let us assume that the vacuum state is flat Minkowski space: $F_{abcd} = 0$, $T_{abcd} = 0$, $Q_{\mu \nu \rho \sigma} = 0$. In a suitable gauge, it can be represented as $A_{\mu}^a = 0$, $g_{ab} = \eta_{ab}$, $\theta^a_\mu = \delta^a_\mu$. Denoting $a_{\mu}^a$ the fluctuation of the connection, we see that

$$T^a_{\mu v} = a^a_{\mu v} - a^a_{v \mu} ,$$

$$Q_{\mu ab} = a_{\mu ab} + a_{\mu ba} .$$

Then, $L_{TQ}$ just becomes a mass term for the connection. The curvature also contains a term quadratic in the gauge field, so that the Palatini lagrangian also contributes to the mass term $-m^2_p (a_{\mu a} a_{\nu b} - a_{\mu b} a_{\nu a})$. For the generic values of the coefficients, the mass matrix will be non-degenerate, and all components of the connection will become massive. This is a gravitational version of the Higgs phenomenon [34,36,37].

There is an alternative and more general way of arriving at the same conclusion, which is based on a different set of dynamical variables. Any connection $A$ can be split uniquely
as $A = \Gamma + \Phi$, where $\Phi$ is called the distortion tensor. Then, in the action, we may replace $A$ by $\Phi$:

$$S(A^a_{\mu b}, \theta^a_{\mu}, g_{ab}) = S(\Gamma^a_{\mu b} + \Phi^a_{\mu b}, \theta^a_{\mu}, g_{ab}) = S'(\Phi^a_{\mu b}, \theta^a_{\mu}, g_{ab}).$$

With these variables,

$$T^a_{\mu v} = \Phi^a_{\mu v} - \Phi^a_{v \mu}, \quad Q_{\mu ab} = \Phi_{\mu ab} + \Phi_{\mu ba}$$

are independent variables, unrelated to the Goldstone bosons. Thus, without making any assumptions about the nature of the vacuum and the choice of gauge, $\mathcal{L}_{TQ}$ is seen to be a mass term for $T$, $Q$, or equivalently for the distortion. We assume that the mass matrix is generic, and therefore, nondegenerate. This implies that, for any $\theta_{\mu}^a$ and $g_{ab}$, the deviation of $A$ from the Levi–Civita connection is massive. At low energy, it is therefore a very good approximation to assume that $\Phi = 0$, or in other words, $T = 0$ and $Q = 0$. In this way, we understand that these conditions, which in GR are simply postulated, are natural properties of any MAG at low energy. Furthermore, we recognize that these conditions have exactly the same origin and the same status as condition (11) for superconductivity, or condition (13) for the low energy electroweak theory.

The Higgs phenomenon removes the connection degrees of freedom from the low energy spectrum, and one of the two fields $\theta_{\mu b}^a, g_{ab}$ can be gauged away. The leftover Goldstone boson has no analog in the case of superconductivity or of the electroweak theory, so whereas in these two cases the connection at low energy can only be flat, in gravity, there can still be curvature, but it is strictly related to the derivatives of the metric. The low energy limit of MAG then agrees with GR as an EFT.

Taken at face value, a MAG is a theory with new physical degrees of freedom, but if we regard it as an EFT below the Planck scale, it is not granted that new particles can be seen in the domain of validity of the EFT. In fact, there are two possibilities. The first and perhaps most natural possibility is that all eigenvalues of the mass matrix are of the order of the Planck mass. This is suggested by the fact that the Palatini term already generates some mass terms that are of that order of magnitude. In this case, all the components of the connection have masses at or above the cutoff of the theory, and cannot be seen as new particles. From this point of view, MAG as an EFT would be essentially indistinguishable from GR as an EFT. Still, MAG has a somewhat greater explanatory power than GR because it gives a dynamical explanation for the conditions $T = 0, Q = 0$, that otherwise have to be postulated. (It is well known that if one assumes that torsion (or nonmetricity) vanishes, and uses the Palatini Lagrangian, then also nonmetricity (or torsion) vanishes as a result of the equations of motion, but one cannot obtain both conditions simultaneously. The terms $T^2$ and $Q^2$ are essential to obtain both these conditions from the dynamics.)

The second and much more interesting possibility is that some eigenvalues of the mass matrix are much smaller than the Planck mass. In this case, also switching on the dimension-four terms in the Lagrangian, there could be new propagating particles in the domain of the validity of the EFT, leading to genuinely new physics. What states propagate depends on the detailed values of the coefficients in the Lagrangian. See [38] for an interesting example. Unfortunately, among these new degrees of freedom, there are generally ghosts or tachyons, at least in a naive perturbative treatment. There are several known subclasses of ghost- and tachyon-free MAGs [35,39–43], but the radiative stability of these theories is doubtful. We shall discuss this a bit further in the next section.

There is an interesting alternative (but physically equivalent) view, the so-called teleparallel theories. In this case, we assume that in a vacuum, $F = 0$ (a condition reminiscent of the Meissner effect), but $\phi$ can be nonzero. In this case, the Hilbert Lagrangian is replaced by a specific combination of terms that are quadratic in torsion and/or nonmetricity. More generally, one can show that any Lagrangian for a metric formulation of gravity has a teleparallel equivalent [35].
4. UV Completion

The EFT treatment of gravity breaks down at a scale of the order of the Planck mass. An internal symptom for this is that the tree level cross-section for gravitational scattering violates the unitarity bound. This must be the case, because for dimensional reasons, the scattering amplitude must be proportional to \( G p^2 \), where \( p \) is the external momentum. What replaces this EFT at higher energies (assuming that this concept makes sense at all!) is a wide open question. Here, we shall discuss the possibility of a UV completion in the context of QFT.

Given the similarity of the low energy gravitational EFT to chiral models of strong interactions, one can look at that case for inspiration. The strong interactions are described, at a fundamental level, by QCD, whose main feature is asymptotic freedom: the perturbative expansion becomes better as the energy increases. As we have already discussed, at sufficiently low energy, perturbation theory is again applicable to strong interactions, but in the form of the chiral models. There is no overlap between these two perturbative domains: they are separated by a region of strong coupling that is hard to describe in any way other than by brute force computer simulations. It is tempting to take this scenario as a template for a possible UV completion of gravity. It has been known for a long time that higher derivative gravity is renormalizable [44], and for certain choices of signs of the couplings, asymptotically free [45]. For decreasing energy, the coupling becomes strong, and it is conceivable that the Planck scale arises, like the strong scale, from a process of dimensional transmutation. At low energy, we would then recover the EFT of gravity that is discussed in Section 2 [46]. The main difficulty that this scenario must overcome is the presence of new unwanted (ghostly or tachyonic) states in the perturbative definition of the theory. Arguments to get around these issues go way back [47], but there has been much more work recently [48–55]. There are also independent arguments from noncommutative geometry, suggesting that bosons do not propagate at ultra-Planckian scales [56].

Perhaps an even closer analog to QCD would be a MAG. The quantum properties of MAGs have not been explored much, but viewing MAG as a metric theory of gravity (with curvature squared terms) coupled to a peculiar form of matter, suggests that they may not be too dissimilar from those of quadratic curvature gravity, and it is quite possible that MAG is asymptotically free for some choices of parameters. However, as already mentioned, ghosts or tachyons often appear in these theories, and it is very unlikely that there are any parameter choices that give unitarity and renormalizability at the same time. Thus, also in this case, one would have to appeal to some unfamiliar mechanism to explain the ghosts away.

A different possibility is that once the EFT becomes strongly coupled, non-perturbative effects somehow manage to stop the growth of the amplitude, and one would have a UV complete QFT, possibly with the same field content as the EFT. This scenario is known as Asymptotic Safety (AS), and in its simplest form, it amounts to saying that \( G \) (as well as all other gravitational couplings) has to be treated as a running coupling and the dimensionless combination \( \tilde{G} = G p^2 \) has to reach a finite limit (a fixed point) at high energy.

The general idea goes back to the 1970s [57,58], but most of the literature has appeared beginning from the 1990s, and is based on the use of non-perturbative methods. There are two main tools. The first is discretization (“lattice”), of which Regge calculus [59] and Causal Dynamical Triangulations [60] are two variants. The second method, which I will concentrate on, is the Functional (or Exact) RG, a version of Wilson’s nonperturbative RG, adapted to the generator of 1PI Green functions. The general idea is as follows. One artificially introduces in the functional integral a scale \( k \), such that the propagation of modes with momentum below \( k \) is suppressed. This is achieved by modifying the inverse propagator by the addition of a momentum-dependent term \( R_k(q^2) \) that goes rapidly to zero for \( q^2 > k^2 \). Thus, \( k \) acts as a momentum-dependent IR cutoff. The effective action computed in the presence of this cutoff is called \( \Gamma_k \). Note that the functional integral still has all the usual UV divergences. However, when one computes the difference of \( \Gamma_k \) for
different values of $k$, the UV divergences cancel. In particular, the $k$-derivative of $\Gamma_k$ turns out to be presented using the very simple formula [61,62]:

$$k\partial_k \Gamma_k = \frac{1}{2} \text{Tr} \left( \Gamma^{(2)}_k + R_k \right)^{-1} k\partial_k R_k,$$

(26)

where $\Gamma^{(2)}_k$ is the (full) inverse propagator, and the r.h.s. is both IR and UV finite. At this point, one can “kick the ladder” and take the mathematically well-defined expression for $k\partial_k \Gamma_k$ as the defining equation for the theory. Fixing an initial condition for $\Gamma_k$ at some initial scale $k_0$, one can compute the effective action at any other value of $k$ by integrating the RG equation, which is a first order functional differential equation. In particular, integrating the RG equation down to $k = 0$ yields the usual effective action, which contains all of the information about the QFT. If the theory is to be treated as an EFT, one just fixes the initial condition at the UV cutoff. Otherwise, one can explore the UV behavior by integrating the RG equation in the direction of increasing $k$. If the RG trajectory reaches a fixed point, the theory is expected to be well behaved in the UV and is said to be AS.

This scenario opens up trans-Planckian physics to exploration via QFT methods. In principle, an AS theory is UV complete, but the real bonus is not so much the possibility of taking the UV limit in itself, which is anyway a mathematical abstraction, but rather, the increased predictivity. It is generally expected that a fixed point will have only a finite number of directions along which it can be reached in the UV. This is then the number of free parameters in the theory. (In the case of asymptotic freedom, the UV-attractive directions correspond precisely to the couplings with positive (and possibly zero) mass dimension. Thus, the only free parameters are the coefficients of renormalizable or super-renormalizable interactions.) Integrating the flow towards the IR, the whole low energy effective action is completely determined, except for the dependence on these parameters.

The FRGE has been tested in many different contexts and gives not only quantitatively but also qualitatively good results, e.g., for critical phenomena [63], and it can also account for previously unexplained phenomena, e.g., some regimes in turbulence [64]. The application of the FRGE to gauge theories, and gravity in particular, involves further approximations and complications, such as the use of the background field, gauge fixing etc., but these can be managed, as demonstrated in the case of low energy QCD, see e.g., [65]. Its first application to gravity [66] involved the so-called Einstein–Hilbert truncation, where the (Euclidean) effective action is assumed to have the form

$$\Gamma_k = \frac{1}{16\pi G_k} \int dx \sqrt{g} (2\Lambda_k - R).$$

Here, $G_k$ and $\Lambda_k$ are the running Newton coupling and cosmological constant. One then finds a nontrivial FP that is now known as the Reuter FP. Subsequent applications involving many more terms in the action have confirmed the existence of this FP. I refer to [9,10] for extended introductory reviews.

In order to be able to calculate, one has to truncate $\Gamma_k$. There are two main systematic approaches. The first is the derivative expansion, the same that is used in EFT. Unlike EFT, however, here one can prove that the theory extends beyond the Planck scale, and the expansion seems to closely reflect the order of relevance of the perturbations around the nontrivial FP. This fact can be seen very clearly for example when one assumes that the Lagrangian is a polynomial in $R$ [67], and it is taken as a sign that the FP is only mildly non-perturbative. It is remarkable that one can find a nontrivial FP also if the Lagrangian is a general function $f(R)$. See [68] for a review. It is in this context that the functional RG comes most powerfully into play. In the derivative expansion, one truncates the dependence on momenta, but keeps the full field dependence. The other systematic approach, the vertex expansion, consists of keeping the full momentum dependence, but expanding in powers of the field. This is best suited for the expansion around flat space, and is more similar to the particle physics applications of QFT [69].
The two approaches give a broadly consistent picture, but there remain a number of issues, some of which are rooted in the method (the FRGE and the necessity of making approximations), while others are peculiar to gravity. See [70] for a discussion and critical review. Here we briefly mention some of the main problems.

It has been pointed out [71,72] that the notion of a running Newton coupling cannot have the same usefulness as that of a running gauge coupling, for example. Basically, this is due to the fact that the RG only gives the dependence of a coupling on a single energy/momentum scale, and this is not sufficient to describe particle physics observables, which can depend in different ways on several different energy/momentum scales. This observation highlights that the notion of running coupling can be given a direct physical interpretation only in specific cases where the physics depends on a single scale. The behavior of the theory in all possible kinematical configurations is given by studying the \( n \)-point functions, and preliminary case studies in special symmetric momentum configurations confirm that the momentum-dependence is closely related to the dependence on the RG scale \( k \) [73]. (There are also general arguments to this effect [57].)

Besides this, and irrespective of its physical interpretation, a running Newton coupling is necessary to achieve scale invariance, which is the hallmark of a fixed point [74]. To appreciate this point, let us recall that in the case of perturbatively renormalizable theories, classical scale invariance only requires zero mass, but scale invariance is broken in the quantum theory because of the presence of the renormalization scale \( \mu \). For example, the scalar self-interaction \( \frac{\lambda}{4!} \phi^4 \) is scale-invariant under the transformation

\[
\delta \epsilon x^\mu = \epsilon x^\mu , \quad \delta \epsilon \phi = -\frac{d - 2}{2} \epsilon \phi , \quad \delta \epsilon \lambda = 0 ,
\]

but the effective action for this theory is not: the Ward identity of scale transformations is

\[
\delta \epsilon \Gamma = -A(\epsilon) , \quad \text{where} \quad A(\epsilon) = \epsilon \beta \int d^4x \frac{1}{4!} \phi^4 , \quad \text{with} \quad \beta \equiv \mu \partial_\mu \lambda = \frac{3 \lambda^2}{16 \pi^2} . \tag{27}
\]

is the anomaly.

In the case of a general non-renormalizable theory, considering the scale-dependent effective action \( \Gamma_k \), and assuming that the bare action in the functional integral \( S \) is scale-invariant, the anomalous Ward identity receives an additional contribution from the IR cutoff \( k \) [75]:

\[
\delta \epsilon \Gamma_k = -A(\epsilon) + \epsilon k \partial_k \Gamma_k . \tag{28}
\]

This equation defines the anomaly \( A \). For example, if we write

\[
\Gamma_k = \sum_i \lambda_i(k) \hat{O}_i , \tag{29}
\]

we obtain

\[
A(\epsilon) = \epsilon \sum_i \tilde{\beta}_i \hat{\mathcal{O}}_i , \tag{30}
\]

where \( \tilde{\lambda}_i = \lambda_i k^{-d_i} , \tilde{\beta}_i = k \partial_k \tilde{\lambda}_i , \hat{\mathcal{O}}_i = \mathcal{O}_i k^{d_i} , d_i \) being the dimension of \( \lambda_i \). Note that this agrees with the earlier definition for renormalizable couplings. Furthermore, we see that the anomaly vanishes at an FP. However, \( \Gamma_k \) at an FP is not scale-invariant, according to the original definition of \( \delta \epsilon \). For example, for (29), we have

\[
\delta \epsilon \Gamma_k = \epsilon \sum_i d_i \lambda_i \hat{\mathcal{O}}_i ' . \tag{31}
\]

This vanishes only if the couplings \( \lambda_i \) are dimensionless.
In general, the way in which scale invariance is achieved at an FP is more subtle: we have to define a different transformation $\delta_{\epsilon} k = -\epsilon k$, and agrees with $\delta_{\epsilon}$ on all other quantities. Then,

$$\delta_{\epsilon} \Gamma_k = \delta_{\epsilon} \Gamma_k - c k \partial_k \Gamma_k = -A(\epsilon),$$

where we used the Ward identity (28). We see that the \textit{anomaly is the Wilsonian RG}, and that scale invariance is realized at a fixed point in the sense of the transformation $\delta_{\epsilon}$. In particular, (30) implies that in order to have scale invariance at a fixed point for gravity, we must have $kd_k \bar{G} = 0$.

The central importance of scale invariance is that via a general argument given in [57,58], it gives a good high energy behavior for observables; for example, the cross-sections must decrease with the square of energy.

So far, we have ignored one complication: the fact that not all couplings need to reach a fixed point. Those that can be altered by redefining the fields cannot enter in the $S$-matrix, and their behaviors are unconstrained. These couplings are called "redundant" or "inessential", in contrast to the "essential" ones, which are required to reach a fixed point. In many calculations so far, this has not been necessary, because all couplings seem to reach a fixed point. However, eliminating the inessential ones significantly simplifies the calculations [76], and it seems likely that this point is going to become important in future work.

The fixed point of gravity, if it exists, will contain higher derivative terms. As in the discussion of quadratic curvature gravity and MAG, the question of unitarity is therefore present. In the context of AS, given that the effective action will contain terms with arbitrarily high order in derivatives, the ghost states may very well just be artefacts of truncating at some finite order in derivatives. Another possible explanation comes from the previously mentioned restriction to essential couplings. It is well-known that in perturbation theory, the curvature-squared terms can be eliminated by redefining the metric. Then, in an expansion around flat spacetime, all of the terms with three and more curvatures only contribute to the vertices, and not to the propagator. In this way, the propagator remains the same as in GR and is free of ghosts [77].

Another technical issue is the (near-) necessity of working in Euclidean space. This arises from the requirement that the condition $p^2 < k^2$ defines a bounded domain of momenta. (In Minkowski space, this condition is compatible with arbitrarily large space momenta). The alternative is to impose separate conditions on space momenta and on energies, which leads to the breaking of local Lorentz symmetry. In contrast, in flat space, it is often convenient to work in Euclidean signature and to perform a Wick rotation in the end; in the context of curved spacetimes, this procedure is fraught by ambiguities and no unique prescription clearly stands out [78,79]. However, one may expect that in the low energy perturbative regime where spacetime is approximately flat, one could use the standard Wick rotation to turn the predictions of AS into Lorentzian observables.

One peculiarity of quantum calculations in gravity is the nearly unavoidable use of the background field method. As a result, the effective action becomes a functional of two fields. When we decompose the metric into a background and fluctuation, as in $g_{\mu\nu} = \bar{g}_{\mu\nu} + h_{\mu\nu}$, the action is automatically invariant under the "split symmetry"

$$\delta \bar{g}_{\mu\nu} = \epsilon_{\mu\nu}, \quad \delta h_{\mu\nu} = -\epsilon_{\mu\nu}.$$  

On the other hand, in order for the cutoff term to be quadratic in the quantum field, it must have the form

$$\int d^4x \sqrt{\bar{g}} h_{\mu\nu} R^{\mu\nu\rho\sigma} R_k (-\nabla^2) h_{\rho\sigma},$$
where $\tilde{K}$ is some combination of background metrics, and $\nabla$ is the covariant derivative of the background metric. This is a separate function of the background and fluctuation field that breaks the split symmetry. As a result, the whole effective action is also not split symmetric, i.e., it is a function $\Gamma_k(h; g)$ separately.

In the context of scalar field theory, the use of the background field method gives rise to various pathologies, but the effect can be controlled by using the anomalous Ward identity of split symmetry, and the correct physics can be reconstructed [80]. In the case of gravity, this is harder but still possible in some simplified situations [81–83]. Another proposal to minimize this effect is to use “vanishing regulators” [84,85]. It has been shown, at least in the case of scalar field theories treated with the background field method, that all pathological features disappear when the regulator becomes sufficiently small, in a suitable sense. There are also other ways to address this problem [86].

The nature of observables is a thorny issue in most approaches to quantum gravity. The preceding discussion of AS has been phrased mostly with particle physics observables in mind. For example, on an asymptotically flat background, we can try to calculate the gravitational scattering of SM particles. (In this way, we avoid the additional difficulties associated with the notion of scattering gravitons.) However, Planckian cross-sections are the least likely way in which quantum gravity effects will be detected. It is more likely that some quantum effect can be seen in early universe cosmology, and in this context, it is likely that more “primitive” observables will play an important role.

The FRGE is a tool that allows us to compute beta functions. However, even in a perturbative, renormalizable theory, the beta functions are not observable: higher loops are generally renormalization scheme-dependent. In the case of non-renormalizable interactions, scheme dependence appears already at one loop level. In order to arrive at some physical observable, one has to look for a scheme-independent quantity. In condensed matter applications, these are usually related to the scaling exponents in the vicinity of the FP. Due to the use of various approximations, in practice, these still exhibit some degree of scheme dependence.

Finally, let me stress that asymptotic safety of gravity is not enough: all interactions must have this property for the program to be successful. Thus, unlike Hamiltonian approaches but like string theory, a UV-complete QFT including gravity must be a “theory of everything”. Note that the unification of forces is a desirable but not a necessary condition for this. The presence of matter undoubtedly generates additional complications over and above those that are already present for gravity, but it also presents opportunities.

Indeed, while the properties of the theory at Planckian energies will remain inaccessible for some time, the predictions of an FP propagate to low energy, and there are many particle physics observables that may be restricted in this way. The presence of matter alters the gravitational beta functions, and too many fields of certain types may destroy the Reuter FP [87]. When quantitatively precise estimates of these effects become available, we will have new ways to rule out particle physics models, or AS. Conversely, gravity affects the matter beta functions, and a nontrivial FP for gravity necessarily turns on certain types of self-interactions of matter [88].

There is a hope that the enhanced predictivity of the AS scenario may fix some of the arbitrary parameters of the SM. The prediction of the mass of the Higgs particle was a striking success [89]. This also seems to work for the mass of the top quark in simplified models with a single family of quarks [90]. Studies of semi-realistic, SM-like theories including three families of quarks and have revealed a large number of FPs, but so far, no additional predictions have emerged [91].

5. Deeper Questions

Beyond predicting its own breakdown near the Planck scale, the EFT of gravity outlined in Section 2 does not contain any hint as to what replaces it at higher energies. One possibility is that gravity as we know it is an “emergent” phenomenon. This may mean different things, depending on what structures one is willing to give up. For example,
in “induced gravity”, the whole kinematical machinery of GR is present from the start, but the action is assumed to be generated by quantum fluctuations of matter. At the other extreme, one can give up the whole notion of continuous manifold and assume that the physical degrees of freedom that make up the universe somehow spontaneously organize themselves in an extended finite dimensional structure [92–94]. This may very well be the case, but it has the drawback in that one would have to give up much of the powerful machinery of QFT. The already mentioned numerical simulations show that the emergence of a continuous manifold is not an easy task to achieve.

In this connection, it is worthwhile to stress that the use of continuum QFT tools does not imply that continuous fields have a direct physical meaning. In practice, in an EFT with UV cutoff at \( \Lambda \), there is no physical meaning to lengths that are shorter than \( 1/\Lambda \). The notion of asymptotic safety implies that we can take the limit \( \Lambda \to \infty \), and this is normally thought to give a physical meaning to the continuum limit. However, even in this case, the operational meaning of measuring arbitrarily short distances and high energies is not obvious. To see this, note that every measurement of a dimensionful quantity is really the measurement of the dimensionless ratio between that quantity and a suitable unit. The standard units in daily use are based on atomic spectroscopy, and the energy levels of the atoms are ultimately set by the Higgs VEV, \( \nu \). Thus, in the early universe before the electroweak phase transition, such units certainly do not exist. The units of last resort seem to be the Planck units. However, in the asymptotic safety scenario, even these units become problematic. For example, if we want to measure a temperature in Planck units, we may use the dimensionless ratio \( T\sqrt{G_0} \), where \( G_0 \) is the low energy value of \( G \), or \( T\sqrt{G_T} \), where the running \( G \) is taken at the scale \( k = T \). Which one of these is actually relevant will depend on the way in which the measurement is performed. The two definitions agree for all presently accessible systems, but the second one becomes of order one near the Planck scale, and then, if asymptotic safety holds, it does not grow further. In some sense, there would then be an upper bound on temperature. Similar arguments may apply to momenta, lengths, etc. [95,96].

Keeping in mind these points, the discussion in Sect.3 suggests a “mild” form of emergence that does not require us to give up the continuum language. We have seen that metric compatibility and torsionlessness are emergent, low energy properties that can be naturally interpreted as the result of the theory being in Higgs phase. In this view the metric plays the role of order parameter, its VEV discriminating between different phases. This point of view gives a new meaning to the Planck scale, which is seen to be analogous to the QCD scale (i.e. the cutoff for the low energy theory of Goldstone bosons), but also to the Fermi scale (i.e. the mass scale of the gauge fields), and it shifts the focus of the questions that one would like to be answered by the quantum theory of spacetime. Just as the central issue of the electroweak theory is the origin and real nature of the Higgs VEV, the central issue for the quantum theory of spacetime is the origin and real nature of the metric [97]. Furthermore, we recall that the order parameter in superconductivity and the pions in QCD are not fundamental fields, but rather composites of fermions. (In electroweak theory it is possible to think of the Higgs as a fundamental field, but compositeness is not ruled out.). Thus, it is possible that also the metric is not to be regarded as a fundamental field. From this point of view the connection may be more fundamental than the metric itself. Even though we shall not discuss this here, this also opens the door to a possible unification of forces [34,98–100].

The question of the nature of the metric cannot be properly addressed within the normal formulation of gravity, which is based on nonlinear fields. A tentative answer is to use linear fields. In superconductivity, a more fundamental description is obtained in terms of the linearly transforming complex field \( \phi \), and going from \( \phi \) to \( \phi \) one increases the number of dynamical fields by one. Similarly in the electroweak case, the sub-Fermi scale description in terms of Goldstone bosons is extended by embedding them in the Higgs doublet, which has one more degree of freedom. In this regard, gravity is different, because the constraints on the metric are formulated as inequalities rather than equalities.
To go from the nonlinearly realized theory to the linear one, it is enough to relax those constraints. Then, the metric really becomes a tensor. Unfortunately, when one relaxes these constraints, the metric may become degenerate or even zero, and it is very difficult to formulate dynamics under these circumstances. Some attempts at a self-consistent bi-metric dynamics were made in [101,102].

Funding: This research received no external funding.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References


42. Marzo, C. Ghost and Tachyon Free Propagation up to spin-3 in Lorentz Invariant Field Theories. *Phys. Rev. D* **2022**, *105*, 065017. [CrossRef]


84. Baldazzi, A.; Percacci, R.; Zambelli, L. Functional renormalization and the \( \overline{\text{MS}} \) scheme. *Phys. Rev. D* 2021, 103, 076012. [CrossRef]
95. Floreanini, R.; Percacci, R. Chirality in unified theories of gravity. *Phys. Rev. D* 2010, 81, 025010. [CrossRef]