**PAPER • OPEN ACCESS**

# Categorical representation learning and RG flow operators for algorithmic classifiers

To cite this article: Artan Sheshmani *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 015012

View the article online for updates and enhancements.

## MACHINE LEARNING
### Science and Technology

**PAPER**

# Categorical representation learning and RG flow operators for algorithmic classifiers

Artan Sheshmani[1,2,3,4], Yi-Zhuang You[5,*] , Wenbo Fu[6] and Ahmadreza Azizi[6]

[1] Department of Mathematics, and Harvard University Physics department, Center for Mathematical Sciences and Applications, Harvard University, Jefferson Laboratory, 17 Oxford St, Cambridge, MA, 02138, United States of America
[2] Institute of the Mathematical Sciences of the Americas, University of Miami, 1365 Memorial Drive Ungar 515, Coral Gables, FL, 33146, United States of America
[3] National Research University Higher School of Economics, Russian Federation, Laboratory of Mirror Symmetry, NRU HSE, 6 Usacheva str., Moscow, 119048, Russia
[4] NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[5] Department of Physics, Condensed matter group, UC San Diego, University of California San Diego, 9500 Gilman Dr, La Jolla, CA, 92093, United States of America
[6] QGNai INC. (Quantum Geometric networks for artificial intelligence), 83 Cambridge Parkway, Unit W806, Cambridge, MA, 02142, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: yzyou@physics.ucsd.edu

## Abstract

Following the earlier formalism of the categorical representation learning, we discuss the construction of the 'RG-flow-based categorifier'. Borrowing ideas from the theory of renormalization group (RG) flows in quantum field theory, holographic duality, and hyperbolic geometry and combining them with neural ordinary differential equation techniques, we construct a new algorithmic natural language processing architecture, called the RG-flow categorifier or for short the RG categorifier, which is capable of data classification and generation in all layers. We apply our algorithmic platform to biomedical data sets and show its performance in the field of sequence-to-function mapping. In particular, we apply the RG categorifier to particular genomic sequences of flu viruses and show how our technology is capable of extracting the information from given genomic sequences, finding their hidden symmetries and dominant features, classifying them, and using the trained data to make a stochastic prediction of new plausible generated sequences associated with a new set of viruses which could avoid the human immune system.

## 1. Introduction

The renormalization group (RG) [2] is a powerful and useful set of methods developed in statistical physics and quantum field theory to deal with many-body problems. It helps physicists to establish the connection between the microscopic laws of physics and the macroscopic collective behaviors of the system. It starts with a many-body system at the microscopic scale and then performs the coarse-graining iteratively to group the fundamental building blocks together into larger and larger clusters. Meanwhile, it constructs effective descriptions of the clusters at each different scale and extracts the effective interaction among them. In the end, the many-body system can be reduced to a few-body system at the highest scales, which enables the understanding of complex systems and their collective behaviors at a large scale.

This idea can be particularly useful for representation learning and classification tasks in machine learning. There are many examples of many-body systems in machine learning tasks. For instance, an image can be viewed as a system of many pixels, and a sequence can be viewed as a system of many tokens. It is desired to see whether the idea of RGs can also be applied to extract the overall representations of images and sequences from their microscopic representations.

In terms of mathematics, the existence of profound connections between quantum field theory and geometry/topology has been a source of many exciting research activities. One of them, as an example, is the

interesting connection between the theory of RG flows for quantum field theories in physics, and the geometric theory of Ricci flows in mathematics. The theory of Ricci flows was developed by Richard Hamilton in the 80's [3–7]. Given a smooth manifold, $M$, a Riemannian metric, $g$, on $M$ defines a bilinear positive-definite product on tangent space, $T_pM$, for each point $p \in M$. This bilinear form is a 2-tensor locally in an open neighborhood $U \subset M$ of $p$, which will have a matrix representation. One can then investigate whether infinitesimal deformations of the metric on $M$ would provide interesting information about its geometry or topology. For instance, given a 1-parameter family, $g_t, t \in (a,b)$ of metrics on $M$, one can study the variation of $g$ with respect to the parameter $t$. The derivative $\dfrac{\partial g_t}{\partial t}$ will then provide for every fixed choice of $t$ and fixed point $p$ a bilinear inner product form (i.e. a 2-tensor) on $T_pM$. It turns out that variation of metric in a 1-parameter family provides one with a differential equation

$$\frac{\partial g_t}{\partial t} = -2\mathrm{Ric}^{g_t},$$

where the term on the right-hand side is the Ricci curvature tensor, $\mathrm{Ric}^{g_t}$, named after Gregorio Ricci-Curbastro, measuring that, how for each fixed choice of $g_t$, the geometry of space is curved as one moves along the geodesics on the manifold $M$.

The connection between RG flows for nonlinear sigma models in physics and the Ricci flow for Riemannian manifolds in mathematics has been quite known for a while since the earlier work of Daniel Freidan [8], Zamalodchikov [9], Tseytlin [10], as well as groundbreaking work of Gregory Perelman [11] in proof of Poincare conjecture, and more recently Carfora [12]. At the end of this article, in the appendix section, we briefly provide an expository account of RG flows in the context of Ricci geometry following the work of Carfora [12]. It must be noted that our focus in the current article is to implement RG flows for developing algorithmic architectures in mathematical artificial intelligence; therefore, later, we quickly diverge from its connection to Ricci geometry and focus solely on RG networks. We encourage interested readers to study the resources provided above to gain a deeper understanding of the connections between the two frameworks in physics and mathematics.

When it comes to implementing RG flow theory in machine learning, the key challenge lies in the difficulty of constructing the coarse-graining transformation at each RG step. In physics, the RG rules are usually specified by humans, such as the majority vote in real-space RG or the momentum-shell integration in field-theoretic RG. These intuitions may not be immediately applicable to the realistic dataset of images and sequences, as the underlying coarse-graining rules may be much more complicated compared to physics systems. This calls for machine learning methods to enable the algorithm to design and optimize the RG transformation in adaptation to the given dataset. One important idea is borrowed from the holographic duality in physics, which states that the RG transformation can be viewed as a holographic mapping of a field configuration from a flat (boundary) space to a hyperbolic (bulk) space with one-higher dimension, such that the long-range correlation in the original field configuration can be equivalently represented as short-range correlation in the bulk space. So the optimal RG transformation can be defined as a bijective holographic mapping that disentangles the features at different hierarchies as much as possible. This allows us to embed the bijective holographic map in the flow-based generative model and use the unsupervised machine learning technique to train the optimal RG transformation. This idea is first proposed in [13] and further developed in later works [14, 15]. The current article further develops the machine-learning RG method by combining the RG-flow model with neural ordinary differential equation (ODE) techniques [16], and explores its application to representation learning of sequential data.

## 2. From conventional RG to machine-learning RG

The similarity between the RG and deep learning has been noticed in the literature [17–27]. Recently, it is realized that the conventional RG flow can be viewed as an optimal transport of the correlated data distribution towards uncorrelated noise distribution [28]. See appendix for the mathematical formulation for the conventional RG. However, there are several aspects that should be upgraded before the idea of RG can find useful applications in machine learning. The main differences between the conventional RG and the machine-learning RG are summarized in table 1, and discussed as follows.

### 2.1. Continuous v.s. discrete
The conventional RG in the quantum field theory typically assumes that the field is defined on a smooth base manifold. However, this assumption is typically not the case for machine learning applications. For example, images are defined on discrete pixels, and texts are defined on discrete words. The discrete nature of most datasets in machine learning requires us to generalize the base manifold from continuous space to discrete lattice.

**Table 1.** Differences between conventional RG and machine-learning RG proposed in [13, 14].

|  | Conventional RG | Machine-learning RG |
|---|---|---|
| Base space | Smooth manifold | Discrete lattice |
| RG flow | Continuous | Discrete |
| RG equation | Differential equation | Recurrence equation |
| RG fixed point | Conformal | General |
| RG scheme | Human-specified (fixed) | Machine-designed (learnable) |
| Data driven | No | Yes |
| Algebraic structure | Semigroup | Group |
| Invertibility | Non-invertible | Invertible |
| Holographic bulk | Not available | Available |
| Hyperbolic geometry | Not defined | Emergent |

The discretization of the base manifold also forces the RG flow to be discrete because it is no longer possible to perform infinitesimal dilation on a discrete lattice. Therefore, instead of writing down a differential equation to describe the continuous RG flow, the discrete RG flow should be described by a *recurrent* equation. However, in the continuum limit (when the lattice spacing approaches zero), the recurrent equation should converge to the differential equation, which will be shown in section 3.11.

### 2.2. Semigroup v.s. group

The conventional RG keeps decimating information in each step of coarse-graining. As a result, the conventional RG is not invertible and only forms a *semigroup* instead of a *group*, despite of its inaccurate name of renormalization 'group'. Recent development in physics [29] reveals that the RG flow can actually be viewed as a holographic mapping, which is invertible. This not only makes a profound connection from RG to quantum gravity but also promotes RG to a group.

The conventional RG studies how perturbations of the action (or deformations of the field configuration) get renormalized at larger and larger scales. The invertible RG has a completely different mindset: it aims to answer how the correlated field configurations on the holographic boundary can be disentangled to uncorrelated noises in the holographic bulk or how the strongly-coupled quantum field theory on the holographic boundary can be reformulated as the weakly-coupled dual gravitational theory in the holographic bulk. By establishing the holographic mapping, any deformation of the field configuration on the holographic boundary can be translated into an excitation in the holographic bulk and analyzed more conveniently. Therefore invertible RG is a more powerful paradigm of RG. Nevertheless, it can always fall back to the conventional RG by a forgetful map that forgets about the holographic bulk degrees of freedom.

### 2.3. Human v.s. machine

The conventional RG scheme is designed by humans. Due to the limitation of human intelligence, the conventional RG typically assumes that the action must take a fixed form with specific types of terms, and the RG flow only changes the coefficients of these terms, such that the action can only flow within a predefined moduli space. Although the moduli space allows us to parameterize the action conveniently, it also restricts our imagination. A more general RG flow can go beyond the moduli space, as new terms can be generated under RG, and even the field content can change under RG (microscopic and macroscopic descriptions of a system can be fundamentally different, as advocated by the emergence principle). However, such a general RG scheme is not analytically tractable by humans. It is not even clear how to design the RG scheme if the form of the action and the field content are all unknown. Thus it becomes desirable to introduce artificial intelligence to learn the optimal RG scheme automatically from the big data of field configurations generated by a field theory. By learning to generate similar field configurations from independent random noise in the holographic bulk, the machine will create the optimal holographic mapping, which also specifies the optimal (invertible) RG scheme.

### 2.4. Conformal v.s. general fixed point

Conventional RG typically assume a conformal fixed point to start with. Given the conformal symmetry at the fixed point, the RG transformation is always taken to be the dilation operator in the conformal group, which corresponds to the rescaling of spacetime and fields together. Given the RG transformation, one can study how a perturbation (or deformation) of the field evolves under dilation. If the perturbation grows stronger/weaker at larger scales, then the perturbation is said to be *relevant/irrelevant* (with respect to the conformal fixed point). More quantitatively, the conformal dimensions can be defined as the eigenvalues of the dilation generator, such that relevant/irrelevant fields are simply distinguished by their positive/negative

conformal dimensions. Intuitively, relevant fields are low-energy/slow-varying modes to be kept under coarse-graining, and irrelevant fields are high-energy/fast-varying modes to be decimated (or integrated out).

However, the more general machine-learning RG does not assume a conformal fixed point because the real-world data (like images or texts) may not be scale-invariant and hence not respect the conformal symmetry. Therefore, the dilation operator is not well-defined, and one can not prescribe an explicit RG scheme from the beginning. The RG scheme has to be learned from data using a data-driven approach. In fact, the real-world data is more likely to be closer to Gaussian fixed points. So even if one learns the RG scheme, it is not immediately clear whether the RG transformation can be used to infer the conformal dimension, as the data could be far from any conformal fixed point.

### 2.5. Relevant v.s. irrelevant
Therefore, the traditional idea of calculating scaling dimensions as eigenvalues of the dilation generator no longer works in more general RG approaches. We need a different way to define what is relevant and what is irrelevant. [14] proposes an elegant and universal definition of irrelevant degrees of freedom using holographic duality and information theory. The key idea is that irrelevant fields are those degrees of freedom that should be decimated under coarse-graining, so they should appear to us as random noise (i.e. independent/uncorrelated random variables). Since the irrelevant fields are actually the holographic bulk field under the holographic duality, the above idea can also be rephrased to a statement that holographic bulk fields are almost uncorrelated. The goal of machine-learning RG is to learn the RG transformation that automatically identifies and separates such irrelevant degrees of freedom in a field theory. We will explain this approach in more detail in section 3.5, after introducing the concrete construction of the machine-learning RG algorithm.

For now, we would like to comment that the information-theoretic definition of the irrelevant field is consistent with the conformal dimension definition in the conformal limit. Because a *negative* conformal dimension in the conformal field theory (CFT) indicates that the field correlation will *decay* exponentially in the dual anti-de Sitter holographic bulk, which is equivalent to the statement that the holographic bulk field is *short-range correlated*, which look like independent random noises beyond a finite correlation length, and are therefore *irrelevant* in the information-theoretic sense.

## 3. Machine-learning RG via flow-based generative models

### 3.1. Sequential data and quantum field on one-dimensional lattice
The idea of RG can be used to construct novel generative models for unsupervised learning. The discussion will mainly focus on sequential data, although generalizations to images and graphs are possible. A sequence is an ordered set of objects $a = (a_1, a_2, \ldots)$, where each object $a_i \in A$ is taken from an object set $A$ (also known as the vocabulary). In machine learning, each object $a_i$ is usually embedded as a vector $\phi_i$ in a finite-dimensional vector space $\mathbb{R}^n$ (assuming the dimension to be $n$). Denote the embedding map as $E : A \to \mathbb{R}^n$, the sequence can be represented as an ordered set of vectors $\phi = (\phi_1, \phi_2, \ldots)$, where $\phi_i = E(a_i) \in \mathbb{R}^n$.

One can also view $\phi_i$ as a quantum field on a one-dimensional discrete lattice, as described by the mapping $\phi : I \to \mathbb{R}^n$, where $I \subset \mathbb{N}$ denotes the index set (equipped with an ordering). Each index $i \in I$ labels an object (or its vector embedding) in the sequence and the set $I$ describes the one-dimensional lattice. The size (cardinality) $|I|$ of the index set corresponds to the length of the sequence. Let $\mathrm{Map}(I, \mathbb{R}^n) := \{\phi : I \to \mathbb{R}^n\}$ be the associated space of all maps from the index set $I$ to the vector space $\mathbb{R}^n$. The objective of unsupervised machine learning is to model the probability measure $p(\phi)\mathcal{D}\phi$ given the dataset of sequences.

### 3.2. Conventional renormalization forms a semigroup
The conventional notion of RG transformation $\mathcal{R} : \mathrm{Map}(I, \mathbb{R}^n) \to \mathrm{Map}(I', \mathbb{R}^n)$ corresponds to a coarse-graining map that extracts the relevant (coarse-grained) field $\phi' = \mathcal{R}(\phi)$ from the original (fine-grained) field $\phi$ and discards the remaining (irrelevant) field degrees of freedom. The renormalization transformation always reduces the degrees of freedom. Therefore the index set will become smaller $|I'| \leqslant |I|$ under the renormalization transformation. Because of the information loss, it is no longer possible to recover the original field configuration $\phi$ from the coarse-grained configuration $\phi'$. Therefore the renormalization transformation $\mathcal{R}$ is not invertible and only forms a *semigroup*.

### 3.3. Invertible renormalization forms a group
The key idea to make the renormalization transformation invertible is to keep the irrelevant field $\zeta'$ together with the relevant field $\phi'$ as the joint output of the renormalization transformation. Intuitively, the relevant/irrelevant fields are the low-/high-energy modes in the field configuration. What the

renormalization transformation does is to separate the irrelevant field $\zeta'$ and the relevant $\phi'$ field given the original field $\phi$ as input. The criterion to separate irrelevant fields will be elaborated in section 3.5.

Invertible renormalization was first proposed under the name of exact holographic mapping [29], which further leads to applications in flow-based generative models for unsupervised machine learning [13–15]. An invertible renormalization transformation is a bijective map $\hat{\mathcal{R}} : \mathsf{Map}(I, \mathbb{R}^n) \to \mathsf{Map}(I', \mathbb{R}^n) \otimes \mathsf{Map}(J', \mathbb{R}^n)$, under which the original field $\phi$ splits to the relevant field $\phi'$ and the irrelevant field $\zeta'$

$$(\phi', \zeta') = \hat{\mathcal{R}}(\phi), \tag{1}$$

where $\phi' = (\dots, \phi'_{i'}, \dots)_{i' \in I'} \in \mathsf{Map}(I', \mathbb{R}^n)$ and $\zeta' = (\dots, \zeta'_{j'}, \dots)_{j' \in J'} \in \mathsf{Map}(J', \mathbb{R}^n)$. The bijectivity requires $|I'| + |J'| = |I|$, i.e. the numbers of relevant and irrelevant features must add up to the total number of features in the original field. The inverse renormalization transformation $\hat{\mathcal{R}}^{-1}$ will also be called the generation transformation $\hat{\mathcal{G}}$, denoted as

$$\phi = \hat{\mathcal{G}}(\phi', \zeta') := \hat{\mathcal{R}}^{-1}(\phi', \zeta'). \tag{2}$$

As the transformation is invertible, the RG is promoted from a semigroup to a *group*.

### 3.4. RG flow

The invertible renormalization transformation enables us to define invertible RG flow on both the field configuration level and the probability measure (or the action) level.

*3.4.1. RG flow on the field level*

Repeating the invertible renormalization transformation, an RG flow can be defined (on the field configuration level) via the following iteration

$$\left( \phi^{(k)}, \zeta^{(k)} \right) = \hat{\mathcal{R}}^{(k)} \left( \phi^{(k-1)} \right), \tag{3}$$

where $\phi^{(k)} \in \mathsf{Map}(I^{(k)}, \mathbb{R}^n)$, $\zeta^{(k)} \in \mathsf{Map}(J^{(k)}, \mathbb{R}^n)$ are the relevant and irrelevant fields, and $\hat{\mathcal{R}}^{(k)} : \mathsf{Map}(I^{(k-1)}, \mathbb{R}^n) \to \mathsf{Map}(I^{(k)}, \mathbb{R}^n) \otimes \mathsf{Map}(J^{(k)}, \mathbb{R}^n)$ is the (bijective) renormalization transformation at the $k$th step. The condition $|I^{(k)}| + |J^{(k)}| = |I^{(k-1)}|$ is always satisfied as a necessary condition for the bijectivity. The iteration defines a flow of quantum fields, called the *renormalization flow* ($\mathcal{R}$-flow):

$$\phi \equiv \phi^{(0)} \xrightarrow[\hat{\mathcal{R}}^{(1)}]{} \phi^{(1)} \xrightarrow[\hat{\mathcal{R}}^{(2)}]{} \phi^{(2)} \xrightarrow[\hat{\mathcal{R}}^{(3)}]{} \cdots \atop \quad\searrow\qquad\quad\searrow\qquad\quad\searrow \atop \qquad \zeta^{(1)} \qquad\quad \zeta^{(2)} \qquad\quad \cdots \tag{4}$$

Along the $\mathcal{R}$-flow, the field configuration will be coarse-grained progressively $\phi^{(0)} \to \phi^{(1)} \to \phi^{(2)} \to \dots$, and the relevant degrees of freedom will be reduced (as $|I^{(0)}| \geqslant |I^{(1)}| \geqslant |I^{(2)}| \geqslant \dots$). Through this process, a sequence of irrelevant fields $\zeta^{(1)}, \zeta^{(2)}, \dots$ is also produced, which was discarded in the conventional renormalization approach but kept in the invertible renormalization approach. Suppose all the relevant degrees of freedom are eliminated after $K$ steps of the renormalization transformation (i.e. $|I^{(K)}| = 0$), the entire $\mathcal{R}$-flow corresponds to a map that encodes the original field $\phi \equiv \phi^{(0)}$ to the collection of irrelevant fields $\zeta \equiv \{\zeta^{(k)}\}_{k=1:K}$, denoted as $\zeta = \hat{\mathcal{R}}(\phi)$.

Retaining these irrelevant fields allows the RG flow to be inverted. The inverse flow is also called the *generation flow* ($\mathcal{G}$-flow) that reconstructs the original field, as defined by the following inverse iteration

$$\phi^{(k-1)} = \hat{\mathcal{G}}^{(k)} \left( \phi^{(k)}, \zeta^{(k)} \right) := \left( \hat{\mathcal{R}}^{(k)} \right)^{-1} \left( \phi^{(k)}, \zeta^{(k)} \right), \tag{5}$$

or given by the dual diagram of equation (4)

$$\cdots \xrightarrow[\hat{\mathcal{G}}^{(3)}]{} \phi^{(2)} \xrightarrow[\hat{\mathcal{G}}^{(2)}]{} \phi^{(1)} \xrightarrow[\hat{\mathcal{G}}^{(1)}]{} \phi^{(0)} \equiv \phi \atop \quad\nearrow\qquad\qquad\nearrow\qquad\qquad\nearrow \atop \cdots \qquad \zeta^{(2)} \qquad\qquad \zeta^{(1)} \tag{6}$$

The entire $\mathcal{G}$-flow corresponds to a map that decodes the irrelevant fields $\zeta$ to the original field $\phi$, denoted as $\phi = \hat{\mathcal{G}}(\zeta)$.

### 3.4.2. RG flow on the probability measure (action) level

The RG flow of field $\phi \to \zeta$ induces a flow of the associated probability distribution over $\mathrm{Map}(I, \mathbb{R}^n)$. Under the bijective map between the original field $\phi$ and the irrelevant field $\zeta$, the probability measure must remain invariant

$$p_{\Phi}(\phi)\mathcal{D}\phi = p_Z(\zeta)\mathcal{D}\zeta. \tag{7}$$

Given $\zeta = \hat{\mathcal{R}}(\phi)$ and $\phi = \hat{\mathcal{G}}(\zeta)$, equation (7) implies that the probability distributions are related by

$$p_Z(\zeta) = p_{\Phi}(\phi)\left| \det \frac{\partial \hat{\mathcal{R}}(\phi)}{\partial \phi} \right|^{-1}, \quad p_{\Phi}(\phi) = p_Z(\zeta)\left| \det \frac{\partial \hat{\mathcal{G}}(\zeta)}{\partial \zeta} \right|^{-1}, \tag{8}$$

where $\left| \det \partial_{\phi} \hat{\mathcal{R}}(\phi) \right|$ denotes the absolute value of the Jacobian determinant of the transformation $\hat{\mathcal{R}}$, and similarly for $\left| \det \partial_{\zeta} \hat{\mathcal{G}}(\zeta) \right|$. More specifically, in each step of the transformation, the probability measure is deformed by (along the $\mathcal{G}$-flow)

$$p_{\Phi}^{(k-1)}\left(\phi^{(k-1)}\right) = p_{\Phi}^{(k)}\left(\phi^{(k)}\right) p_Z^{(k)}\left(\zeta^{(k)}\right) \left| \det \frac{\partial \hat{\mathcal{G}}^{(k)}\left(\phi^{(k)}, \zeta^{(k)}\right)}{\partial \left(\phi^{(k)}, \zeta^{(k)}\right)} \right|^{-1}. \tag{9}$$

In quantum field theory, the field action is defined as the negative log-likelihood of the field configuration, i.e. $S_{\Phi}^{(k)} = -\log p_{\Phi}^{(k)}$ and $S_Z^{(k)} = -\log p_Z^{(k)}$. In terms of the field action, the transformation relates

$$S_{\Phi}^{(k-1)}\left(\phi^{(k-1)}\right) = S_{\Phi}^{(k)}\left(\phi^{(k)}\right) + S_Z^{(k)}\left(\zeta^{(k)}\right) + S_{\Phi Z}^{(k)}\left(\phi^{(k)}, \zeta^{(k)}\right), \tag{10}$$

where the coupling action $S_{\Phi Z}^{(k)}(\phi^{(k)}, \zeta^{(k)})$ is defined to be the log Jacobian determinant of the $\hat{\mathcal{G}}^{(k)}$ transformation,

$$S_{\Phi Z}^{(k)}\left(\phi^{(k)}, \zeta^{(k)}\right) := \log \left| \det \frac{\partial \hat{\mathcal{G}}^{(k)}\left(\phi^{(k)}, \zeta^{(k)}\right)}{\partial \left(\phi^{(k)}, \zeta^{(k)}\right)} \right|. \tag{11}$$

Therefore the renormalization transformation $\hat{\mathcal{R}}$ of the relevant field $\phi^{(k)}$ induces the deformation $\bar{\mathcal{G}}$ of the relevant field action $S_{\Phi}^{(k)}(\phi^{(k)})$ along the generative direction

$$
\begin{array}{ccccccccc}
\phi = \phi^{(0)} & \longrightarrow & \cdots & \longrightarrow & \phi^{(k-1)} & \xrightarrow{\hat{\mathcal{R}}^{(k)}} & \phi^{(k)} & \longrightarrow & \cdots & \longrightarrow & \phi^{(K)} = \emptyset \\
\downarrow & & & & \downarrow & & \downarrow & & & & \downarrow \\
S_{\Phi} = S_{\Phi}^{(0)} & \longleftarrow & \cdots & \longleftarrow & S_{\Phi}^{(k-1)} & \xleftarrow{\bar{\mathcal{G}}^{(k)}} & S_{\Phi}^{(k)} & \longleftarrow & \cdots & \longleftarrow & S_{\Phi}^{(K)} = 0
\end{array}. \tag{12}
$$

In this way, the renormalization flow of the action $\bar{\mathcal{R}} := \bar{\mathcal{G}}^{-1}$ is defined as the pullback of the renormalization flow $\hat{\mathcal{R}}$ of the field. The RG transformation is invertible on both the field and the action level, making the RG literally a group.

### 3.5. Criterion to separate irrelevant fields

What has not been explained so far is the criterion to separate relevant fields from irrelevant fields. [14] argues that the irrelevant field should look like independent random variables (or random maps), because the irrelevant fields are supposed to be discarded under the conventional RG flow, meaning that (in the ideal limit) they do not contain information and should appear like random noise. Guided by this intuition, [14] further proposes the minimal bulk mutual information (minBMI) principle as the designing principle of renormalization flow, that the optimal renormalization transformations $\{\hat{\mathcal{R}}^{(k)}\}_{k=1:K}$ should be defined as the maps that minimize the mutual information among all irrelevant fields

$$\sum_{k,k',j,j'} I\left(\zeta_j^{(k)} : \zeta_{j'}^{(k')}\right) = \sum_{k,k',j,j'} \int \mathcal{D}\zeta \, p_Z(\zeta) \log \frac{p_Z\left(\zeta_j^{(k)}, \zeta_{j'}^{(k')}\right)}{p_Z\left(\zeta_j^{(k)}\right) p_Z\left(\zeta_{j'}^{(k')}\right)}. \tag{13}$$

The minimum is achieved when the irrelevant fields are statistically independent, i.e.

$$p_Z(\zeta) = \prod_k \prod_{j \in J^{(k)}} p_Z\left(\zeta_j^{(k)}\right), \tag{14}$$

such that all mutual information vanishes.

The optimal solution of $\hat{\mathcal{R}}$ that converges to this limit can be found using machine learning approaches by constructing a trainable bijective map $\hat{\mathcal{G}} := \hat{\mathcal{R}}^{-1}$ (as the composition of smaller bijective maps $\hat{\mathcal{G}}^{(k)}$ at each RG step) to reproduce the data distribution $p_\Phi(\phi)$ starting from the independent prior distribution $p_Z(\zeta)$ in equation (14). The related methods were developed in [13–15] under the name of neural-RG. A conventional choice is to take each $p_Z(\zeta_j^{(k)}) = (2\pi)^{-n/2} \exp(-\frac{1}{2}\|\zeta_j^{(k)}\|^2)$ to be the standard normal distribution (Gaussian with zero mean and unit variance), such that

$$S_Z^{(k)}(\zeta^{(k)}) = \frac{1}{2} \sum_{j \in J^{(k)}} \|\zeta_j^{(k)}\|^2 + \text{const.}. \tag{15}$$

This action describes that the irrelevant field fluctuation is massive in the holographic bulk, which is compatible with the idea of holographic duality.

### 3.6. Hierarchical structure and hyperbolic space

As the renormalization transformation reduces the relevant degrees of freedom, the size of the relevant index set gradually reduces $|I^{(k)}| \leqslant |I^{(k-1)}|$. To be more concrete, we can restrict our discussion to the case where the degrees of freedom is reduced by half under each renormalization transformation, i.e. $|I^{(k)}| = |I^{(k-1)}|/2$, such that

$$|I^{(k)}| = 2^{-k}|I^{(0)}|. \tag{16}$$

Then the condition $|I^{(k)}| + |J^{(k)}| = |I^{(k-1)}|$ implies $|J^{(k)}| = 2^{-k}|I^{(0)}|$. The RG flow will stop when $|I^{(K)}| < 1$, which sets the total number $K$ of RG steps to be

$$K = \log_2|I^{(0)}|. \tag{17}$$

As illustrated in figure 1, the hierarchical structure of the RG flow generates a ordered collection of index sets $\{J^{(k)}\}_{k=1:K}$, which can be union into a hyperbolic lattice (a discrete hyperbolic space), described by

$$J = \bigcup_{k=1:K} J^{(k)}. \tag{18}$$

Instead of thinking the irrelevant fields as separate mappings $\zeta^{(k)} \in \mathsf{Map}(J^{(k)}, \mathbb{R}^n)$, we can treat them jointly as a field $\zeta \in \mathsf{Map}(J, \mathbb{R}^n)$ defined on the hyperbolic lattice $J$. Therefore, the $\mathcal{R}$-flow $\zeta = \hat{\mathcal{R}}(\phi)$ and the $\mathcal{G}$-flow $\phi = \hat{\mathcal{G}}(\zeta)$ respectively define the encoding and decoding maps that connect the field $\phi$ in one-dimensional flat space to the field $\zeta$ in two-dimensional hyperbolic space, which explicitly realize the holographic duality in quantum gravity.

### 3.7. Realization of bijective transformation

To optimize the renormalization transformation $\hat{\mathcal{R}}$, one relies on the construction of a trainable bijective map to model $\hat{\mathcal{R}}$. Machine learning community has provided several realizations of trainable bijective maps, including real-valued non-volume preserving (real NVP) [30] and neural ordinarty differential equation (ODE) [16]. In the following, we will focus on the neural ODE realization, as it can capture multi-modular features better than real NVP, which is more suitable for processing sequences of discrete objects.

#### 3.7.1. Neural ODE
Each single-step renormalization transforms $(\phi', \zeta') = \hat{\mathcal{R}}(\phi)$ can be realized by an ODE. Starting from $\phi(0) = \phi$, first evolve $\phi(t)$ from $t = 0$ to $t = 1$ following

$$\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = f_\theta(\phi(t), t), \tag{19}$$

where $f_\theta$ is a trainable function (realized as a neural network) parameterized by neural network parameters $\theta$. Then split the result as $\phi(1) = (\phi', \zeta')$ to obtain $\phi'$ and $\zeta'$. $t$ is considered as an auxiliary time. The inverse transformation is given by the time-reversal evolution, therefore, the mapping is indeed bijective as desired.
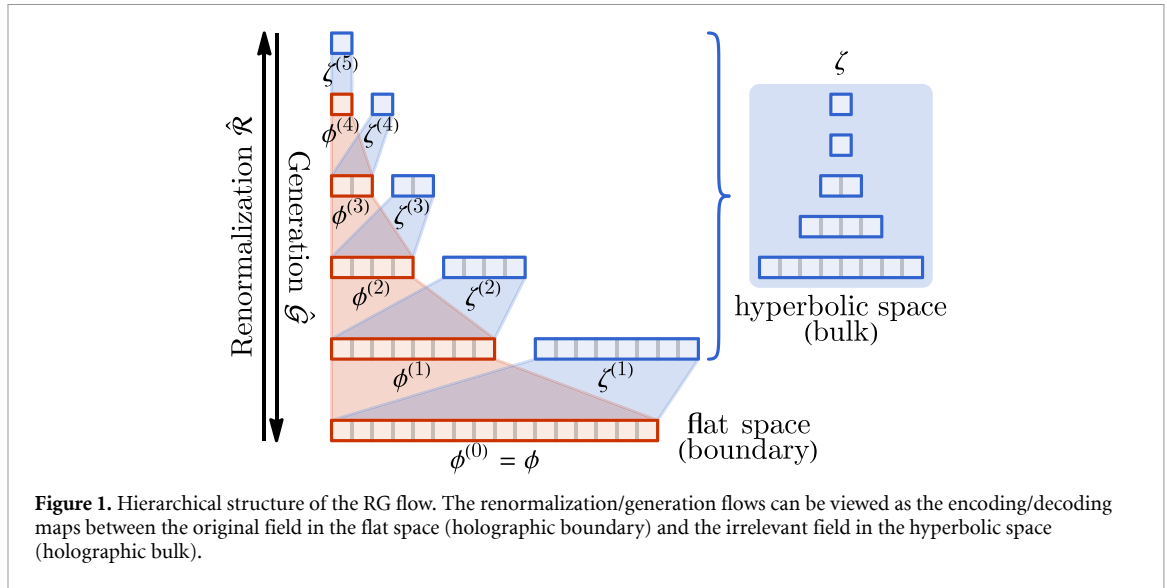
**Figure 1.** Hierarchical structure of the RG flow. The renormalization/generation flows can be viewed as the encoding/decoding maps between the original field in the flat space (holographic boundary) and the irrelevant field in the hyperbolic space (holographic bulk).

Apart from the transformation, the log Jacobian determinant of $\hat{\mathcal{R}}$ can also be evaluated. Based on the ODE in equation (19), one have

$$\log \left| \det \frac{\partial \phi (t + \mathrm{d}t)}{\partial \phi (t)} \right| = \mathrm{Tr} \left( \frac{\partial f_\theta (\phi (t), t)}{\partial \phi (t)} \right) \mathrm{d}t, \tag{20}$$

which can be integrated to

$$\log \left| \det \frac{\partial \hat{\mathcal{R}} (\phi)}{\partial \phi} \right| = \int_0^1 \mathrm{Tr} \left( \frac{\partial f_\theta (\phi (t), t)}{\partial \phi (t)} \right) \mathrm{d}t. \tag{21}$$

Given that $\hat{\mathcal{G}} := \hat{\mathcal{R}}^{-1}$, its log Jacobian determinant is simply given by a negation,

$$\log \left| \det \frac{\partial \hat{\mathcal{G}} (\phi', \zeta')}{\partial (\phi', \zeta')} \right| = -\int_0^1 \mathrm{Tr} \left( \frac{\partial f_\theta (\phi (t), t)}{\partial \phi (t)} \right) \mathrm{d}t, \tag{22}$$

which will be useful for the evaluation of the coupling action in equation (11).

*3.7.2. Locality and translational symmetry*
It is possible to design the ODE function $f_\theta$ to respect locality and translational symmetry. The idea is to realize $f_\theta$ using layers of convolutional neural networks with finite kernel followed by element-wise activations.

**3.8. Objective function**
The objective is to train the generative model, such that the model distribution $p_\Phi (\phi)$ matches the data distribution $p_{\mathrm{dat}} (\phi)$ as much as possible. The objective can be achieved by minimizing the Kullback–Leibler (KL) divergence

$$\mathcal{L} = D_{\mathrm{KL}} (p_{\mathrm{dat}} \| p_\Phi) = \int \mathcal{D}\phi \, p_{\mathrm{dat}} (\phi) \log \left( \frac{p_{\mathrm{dat}} (\phi)}{p_\Phi (\phi)} \right) = \mathop{\mathbb{E}}_{\phi \sim p_{\mathrm{dat}}} S_\Phi (\phi) - H (p_{\mathrm{dat}}), \tag{23}$$

where $S_\Phi (\phi) = -\log p_\Phi (\phi)$ is the model action (as the negative log-likelihood), and $H(p_{\mathrm{dat}}) = -\int \mathcal{D}\phi \, p_{\mathrm{dat}} (\phi) \log p_{\mathrm{dat}} (\phi)$ is the data entropy. As the data entropy $H(p_{\mathrm{dat}})$ is independent of the model parameter, it can be dropped from the loss function $\mathcal{L}$. Therefore the loss function is essentially the ensemble average of the model action on the dataset. By minimizing the average action, the ODE function $f_\theta$ in each RG transformation will get trained. Upon convergence, the algorithm will find the optimal invertible RG flow that maps the (presumably) strongly coupled original field $\phi$ on the holographic boundary to the weakly coupled irrelevant field $\zeta$ in the holographic bulk.

### 3.9. Summary of the algorithm

Given a set of sequences from the data, the learning algorithm goes as follows.

(a) For each given sequence $a = (a_1, a_2, \ldots)$, represent each object $a_i$ in the sequence as a vector $\phi_i = E(a_i) \in \mathbb{R}^n$. Denote the sequence of vectors as a vector field $\phi = (\phi_1, \phi_2, \ldots) \in \mathsf{Map}(I, \mathbb{R}^n)$.

(b) Starting with $\phi^{(0)} = \phi$, apply the renormalization transformation iteratively,

$$\left( \phi^{(k)}, \zeta^{(k)} \right) = \hat{\mathcal{R}}^{(k)} \left( \phi^{(k-1)} \right),$$

for $K = \log_2 |I|$ steps (until all relevant fields are eliminated).

1. Each step of the transformation is implemented by solving an ODE

$$\frac{\mathrm{d}\phi^{(k-1)}(t)}{\mathrm{d}t} = f_\theta^{(k)} \left( \phi^{(k-1)}(t), t \right),$$

starting from the initial condition $\phi^{(k-1)}(0) = \phi^{(k-1)}$, integrating from $t = 0$ to $t = 1$, and then splitting the final result into $\phi^{(k-1)}(1) = (\phi^{(k)}, \zeta^{(k)})$.

2. While solving the ODE, simultaneously integrate along the time evolution to obtain the coupling action

$$S_{\Phi Z}^{(k)} \left( \phi^{(k)}, \zeta^{(k)} \right) = -\int_0^1 \mathrm{Tr} \left( \frac{\partial f_\theta^{(k)} \left( \phi^{(k-1)}(t), t \right)}{\partial \phi^{(k-1)}(t)} \right) \mathrm{d}t.$$

(c) Starting from the initial condition $S_\Phi^{(K)} = 0$, collect the action in the reverse order (along the generation flow)

$$S_\Phi^{(k-1)} \left( \phi^{(k-1)} \right) = S_\Phi^{(k)} \left( \phi^{(k)} \right) + S_Z^{(k)} \left( \zeta^{(k)} \right) + S_{\Phi Z}^{(k)} \left( \phi^{(k)}, \zeta^{(k)} \right),$$

where $S_Z^{(k)}(\zeta^{(k)})$ is given by

$$S_Z^{(k)} \left( \zeta^{(k)} \right) = \frac{1}{2} \sum_{j \in J^{(k)}} \|\zeta_j^{(k)}\|^2.$$

The resulting total action will be denoted as $S_\Phi(\phi) := S_\Phi^{(0)}(\phi^{(0)})$.

(d) Train the model to minimize the loss function

$$\mathcal{L} = \mathbb{E}_{\phi \sim p_{\mathrm{dat}}} S_\Phi(\phi).$$

### 3.10. Potential applications and advantages after training, the model could potentially be used for the following tasks

- Inference of hierarchical latent representation. Using $\zeta = \hat{\mathcal{R}}(\phi)$, one can infer the hierarchical latent representation $\zeta$ of any sequence encoding $\phi$. The high-level representations ($\zeta^{(k)}$ with a large $k$) can be viewed as the encoding of the entire sequence, which can be used in downstream tasks like classification and translation.

- Likelihood estimation. Using $S_\Phi(\phi)$, one can estimate the probability density $p_\Phi(\phi) \propto \exp(-S_\Phi(\phi))$ for any field configuration $\phi$. This will be useful for anomaly detection.

- Sample generation. As a generative model, new samples can be generated by first sampling $\zeta$ in the hyperbolic space and then transforming to $\phi = \hat{\mathcal{G}}(\zeta)$ using the generation flow, which may find applications in completing missing objects in a sequence.

The proposed algorithm is advantageous in the following aspects.

- Disentangled features in scales. The optimal RG flow distills features at different scales, allowing the model to capture the long-range and multi-scale correlation in the sequential data. The features are automatically arranged in a hyperbolic space, making it easy to access/control.

- Efficient inference/generation. The hierarchical and iterative approach enables the model to infer latent fields or generate original fields in $\Theta(N)$ complexity (given the sequence length $N$), which is superior compared to the $\Theta(N^2)$ complexity of transformer-based approaches, especially when the sequence is long.

- Ability to process hierarchical structure. The renormalization transformation can progressively extract coarse-grained features from fine-grained features, making it capable of capturing global features (such as the parity of bit strings). In comparison, as shown in [31], self-attention-based models can not efficiently model hierarchical structures unless the number of layers/heads increases with sequence length.

### 3.11. Recovering conventional RG by integrating out irrelevant fields

Finally, we would like to comment that the invertible renormalization can fall back to the conventional renormalization by integrating out irrelevant fields. Recall equation (3) that in each step of the invertible renormalization transformation, the original (fine-grained) field $\phi^{(k-1)}$ is separated into the relevant $\phi^{(k)}$ and irrelevant $\zeta^{(k)}$ fields by $(\phi^{(k)}, \zeta^{(k)}) = \hat{\mathcal{R}}^{(k)}(\phi^{(k-1)})$. The invertible renormalization $\hat{\mathcal{R}}^{(k)}$ can be downgraded to a non-invertible renormalization $\mathcal{R}^{(k)}$ by a forgetful map which forgets about the irrelevant field $\zeta^{(k)}$, such that $\phi^{(k)} = \mathcal{R}^{(k)}(\phi^{(k-1)})$ only transforms the fine-grained field $\phi^{(k-1)}$ to the coarse-grained field $\phi^{(k)}$.

According to equation (10), the actions are related by

$$S_{\Phi}^{(k-1)}\left(\phi^{(k-1)}\right) = S_{\Phi}^{(k)}\left(\phi^{(k)}\right) + S_{Z}^{(k)}\left(\zeta^{(k)}\right) + S_{\Phi Z}^{(k)}\left(\phi^{(k)}, \zeta^{(k)}\right), \tag{24}$$

where the irrelevant field $\zeta^{(k)}$ is massive, and is described by the Gaussian action $S_{Z}^{(k)}(\zeta^{(k)}) = \frac{1}{2}\|\zeta^{(k)}\|^2$ as in equation (15). Because $\zeta^{(k)}$ represents the high-energy modes that should be integrated out under renormalization, one can argue that the fluctuation of $\zeta^{(k)}$ can be treated perturbatively due to its large mass, which justifies the expansion of the action around $\zeta^{(k)} \to 0$,

$$\begin{aligned} S_{\Phi}^{(k-1)}\left(\phi^{(k-1)}\right) &\simeq S_{\Phi}^{(k)}\left(\phi^{(k)}\right) + S_{\Phi Z}^{(k)}\left(\phi^{(k)}, 0\right) + \zeta^{(k)} \cdot \partial_{\zeta^{(k)}} S_{\Phi Z}^{(k)}\left(\phi^{(k)}, 0\right) \\ &+ \frac{1}{2}\|\zeta^{(k)}\|^2 + \frac{1}{2}\zeta^{(k)} \cdot \partial_{\zeta^{(k)}} \partial_{\zeta^{(k)}} S_{\Phi Z}^{(k)}\left(\phi^{(k)}, 0\right) \cdot \zeta^{(k)} + \dots. \end{aligned} \tag{25}$$

As the approximate action is quadratic in $\zeta^{(k)}$, one can perform a Gaussian integration for $\zeta^{(k)}$, under which the action becomes

$$S_{\Phi}^{(k-1)}\left(\phi^{(k-1)}\right) = S_{\Phi}^{(k)}\left(\phi^{(k)}\right) + S_{\Phi Z}^{(k)}\left(\phi^{(k)}, 0\right) + \frac{1}{2}\left(\partial_{\zeta^{(k)}}^2 S_{\Phi Z}^{(k)}\left(\phi^{(k)}, 0\right) - \left(\partial_{\zeta^{(k)}} S_{\Phi Z}^{(k)}\left(\phi^{(k)}, 0\right)\right)^2\right). \tag{26}$$

Therefore one can define the renormalization transformation $\bar{\mathcal{R}}^{(k)}$ on the action via $S_{\Phi}^{(k)} = \bar{\mathcal{R}}^{(k)}(S_{\Phi}^{(k-1)})$, in correspondence to the field renormalization $\phi^{(k)} = \mathcal{R}^{(k)}(\phi^{(k-1)})$. Based on equation (26), the explicit form of the renormalization operator $\bar{\mathcal{R}}$ can be given

$$\bar{\mathcal{R}}\left(S_{\Phi}\right) = S_{\Phi} - S_{\Phi Z} - \frac{1}{2}\left(\partial_{\zeta}^2 S_{\Phi Z} - (\partial_{\zeta} S_{\Phi Z})^2\right), \tag{27}$$

such that

$$\bar{\mathcal{R}}(S_{\Phi})(\phi) = S_{\Phi}(\mathcal{R}(\phi)), \tag{28}$$

which reproduces the pullback construction of the action renormalization. If one further define the infinitesimal generator of $\bar{\mathcal{R}}$ as $\bar{\mathfrak{r}} = \log \bar{\mathcal{R}}$, the renormalization flow can be expressed as a differential equation [2, 32] $\partial_{\ell} S_{\Phi} = \bar{\mathfrak{r}} S_{\Phi} = -S_{\Phi Z} - \frac{1}{2}(\partial_{\zeta}^2 S_{\Phi Z} - (\partial_{\zeta} S_{\Phi Z})^2)$.

## 4. Experiments on genomic sequences

### 4.1. Problem overview

Extracting the hidden information of genomic sequences has been a critical subject in biological research, with relevance to epidemiology, immunology, protein design and many other subfields. With its great similarity to natural language processing problems, there are numerous studies on applying machine learning techniques to extract information from genomic sequences. Various machine learning architectures include word2vec [33, 34], bidirectional long short-term memory [35], transformer [36] etc. However while the existing algorithms can provide a single-gene level embedding, they do not provide a canonical sequence embedding and the hierarchical information is not clear from the natural language models. Thus we apply the RG idea from the previous sections to the genomic sequence representation problem, where the hierarchical structure provides the biological information at different energy levels, i.e. the deeper layer can capture the longer correlation in the sequence and thus provide a canonical embedding of the sequence with

```
Seq('MKTIIALSYILCLVFAQKLPGNDNSTATLCLGHHAVPNGTIVKTITNDQIEVTN...ICI')
Seq('MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHN...ICI')
Seq('MEAKLFVLFCAFTALKADTICIGYHANNSTDTVDTILEKNVTVTHSVNLLENNH...ICI')
Seq('MKAILVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKH...ICI')
Seq('MKAILVVLLYTFTTANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLENRH...ICI')
Seq('MKTILVVLLYTFATADADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKH...ICI')
Seq('MKTIIALSYILCLVFAQKIPGNDNSTATLCLGHHAVPNGTIVKTITNDRIEVTN...ICI')
Seq('MEKIVLLLAIIGLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHN...ICI')
```

**Figure 2.** Sample Influenza HA amino acid sequences.

the deepest layer. We take the Influenza HA amino acid sequences as an example[7], where the sequences are regarded as one-dimensional lattices as described in section 3.1. Below are samples of the sequence data; there are clear global features embedded, as one can see the similarities between different sequences.

## 4.2. Single amino acid distribution learning

Before proceeding with the sequence into the RG scheme, we need to verify that local features are efficiently learned. Thus we first look at the single amino acid distribution learning. As shown in figure 2, at a fixed location $i$ among the sequences, there is a discrete distribution labeled by amino acid, we pick up that amino acid from each sequence. Then each sample is labeled by $a = (a_i)$, where $a_i \in A$ represents the single amino acid. We apply the pre-trained single-amino-acid level embedding $E : A \to \mathbb{R}^n$ from [35] where $n = 20$. After embedding, the boundary field is $\phi = (\phi_i)$, where $\phi_i = E(a_i) \in \mathbb{R}^n$. To remove the difficulty in transforming the boundary discrete distribution to the bulk uncorrelated continuous Gaussian distribution, we add small randomness on the boundary field, i.e. $\overline{\phi_i} = \phi_i + \epsilon$, where $p_Z(\epsilon)$ takes the standard normal distribution with small variance. For simplicity, in the following, we still use $\phi_i$ to denote the fields with small randomness.

With this setup, we train a neural ODE model to realize the bijective transformation between the data distribution to a Gaussian distribution as described in section 3.7.1. To further speed up the training process, we have added the Jacobian and Kinetic regularization to find an optimal bijective map as in [37]. The input data is the 4th single amino acid from 1000 Influenza HA sequences.

The ODE transformation $f_\theta(x, t)$ is constructed by a feed-forward network with four sequential hidden layers as shown in figure 3. The hidden layers are the concatsquash (CS) layers defined in [38]:

$$f_{\text{CS}}(x, t) = W_0 x \cdot \sigma(W_1 t) + W_2 t \tag{29}$$

where $W_0, W_1, W_2$ are parameter matrices with shape $(d_o, d_i), (d_o, 1), (d_o, 1)$ respectively, $d_i$ and $d_o$ are input and output dimensions. We can consider $x$ as a $d_i$-dimensional vector, and concate the time variable, then $f_{\text{CS}} : \mathbb{R}^{d_i+1} \to \mathbb{R}^{d_o}$. The hyperbolic tangent activation functions are applied after the first three CS layers. Then the model is trained such that $f_\theta(x, t)$ describes the flow from the data to a Gaussian variable.

As shown in figure 4, a two-dimensional feature space can be obtained by applying the t-distributed stochastic neighbor embedding (t-SNE) algorithm [39] to the original data and the flow generated data embedding vectors. The flow-generated data are obtained by taking the inverse transformation $\hat{\mathcal{R}}^{-1}$ from vectors with the Gaussian distribution. The original data distribution with the multi-modular feature can be perfectly captured after training.

## 4.3. Amino acid sequence distribution learning

To train on sequences, we adapt the hierarchical RG scheme described in section 3.6. As discussed in the previous section, the input sequence is represented as labels $a = (a_1, a_2, \ldots, a_I)$ with the cardinality $I$ denoting the length of the sequence. With the pretrained embedding $\phi_i = E(a_i) \in \mathbb{R}^n$, the boundary fields are represented as $\phi = (\phi_1, \phi_2, \ldots, \phi_I)$. Thus we have the initial boundary fields $\phi^{(0)} = \phi$, and we can run the renormalization flow using equation (4). On the other hand, the generation flow equation (6) reconstructs the original field. Following the notation of MERA networks, each renormalization transformation layer consists of a disentangler layer and a decimator layer:

$$\hat{\mathcal{R}}^{(k)} = \hat{\mathcal{R}}^{\text{dec},(k)} \circ \hat{\mathcal{R}}^{\text{dis},(k)} \tag{30}$$

where a disentangler layer disentangles the local correlations, and a decimator layer separates the decimated fields out as the bulk fields. In figure 5, we illustrate the model structure with green blocks as disentanglers

---

[7] Can be downloaded from the 'Protein Sequence Search' section of https://www.fludb.org.

**Figure 3.** Neural ODE structure. The input is the single amino acid vector representation $\phi \in \mathcal{R}^n$ with $n = 20$. The hidden layers are concatsquash (CS) layers. The latent variable has dimension 64, and the output has dimension $n = 20$.



(A) Initial distribution          (B) Distribution after training

**Figure 4.** Training result on a single amino acid with t-SNE representation. (A): The original data distribution (red) has a multi-modular structure, while the initial flow model's distribution (blue) is standard normal. (B): After training with neural ODE, two distributions coincide.
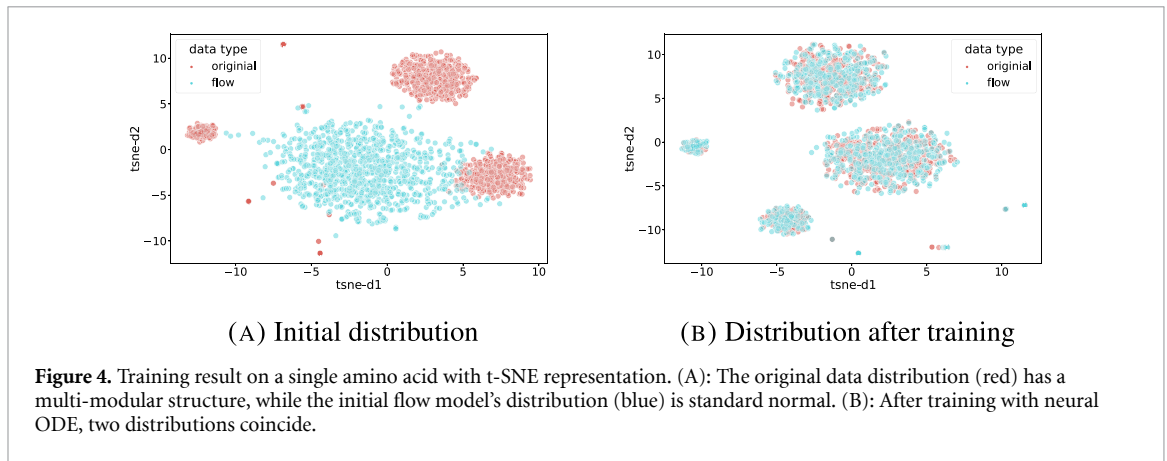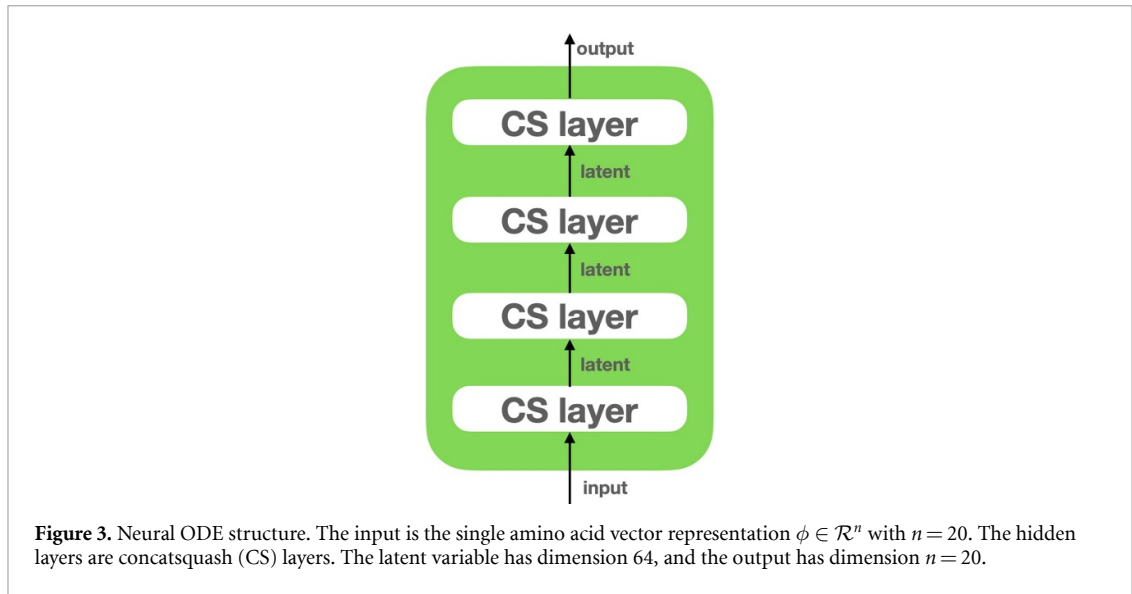
**Table 2.** Average normalized logarithmic probability from original data and generated data with sequence length 4 and 16.
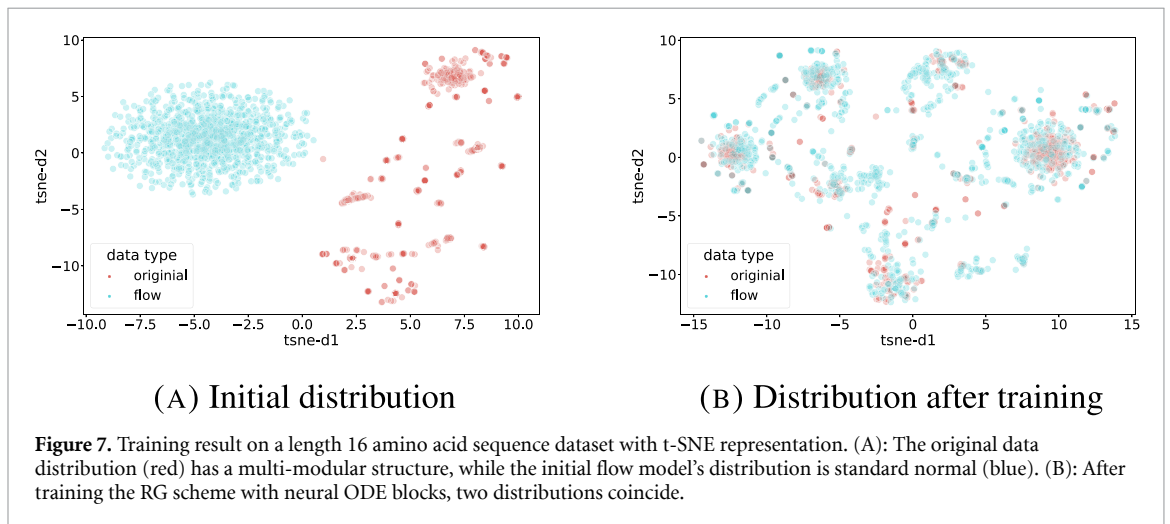
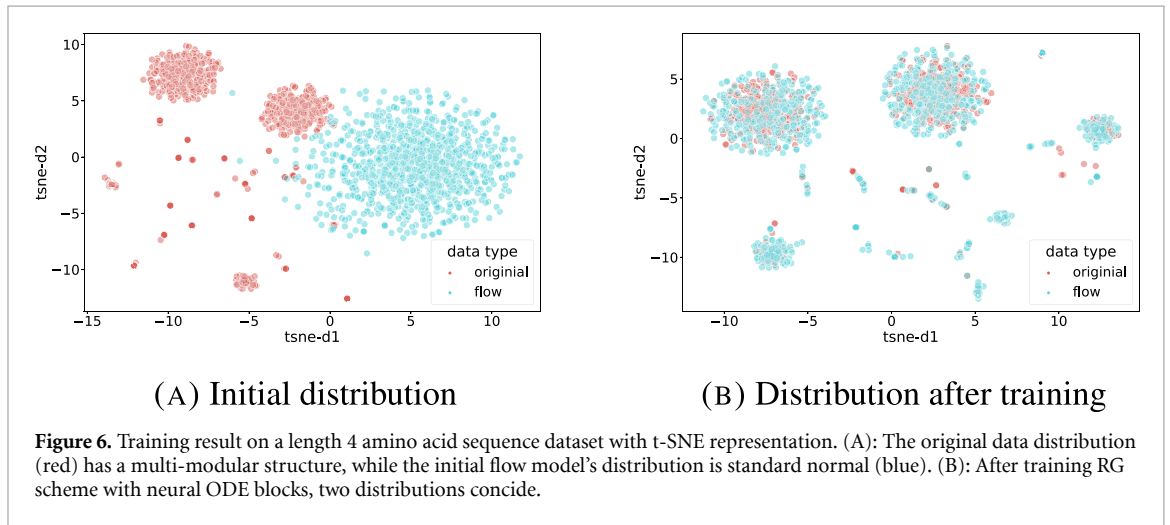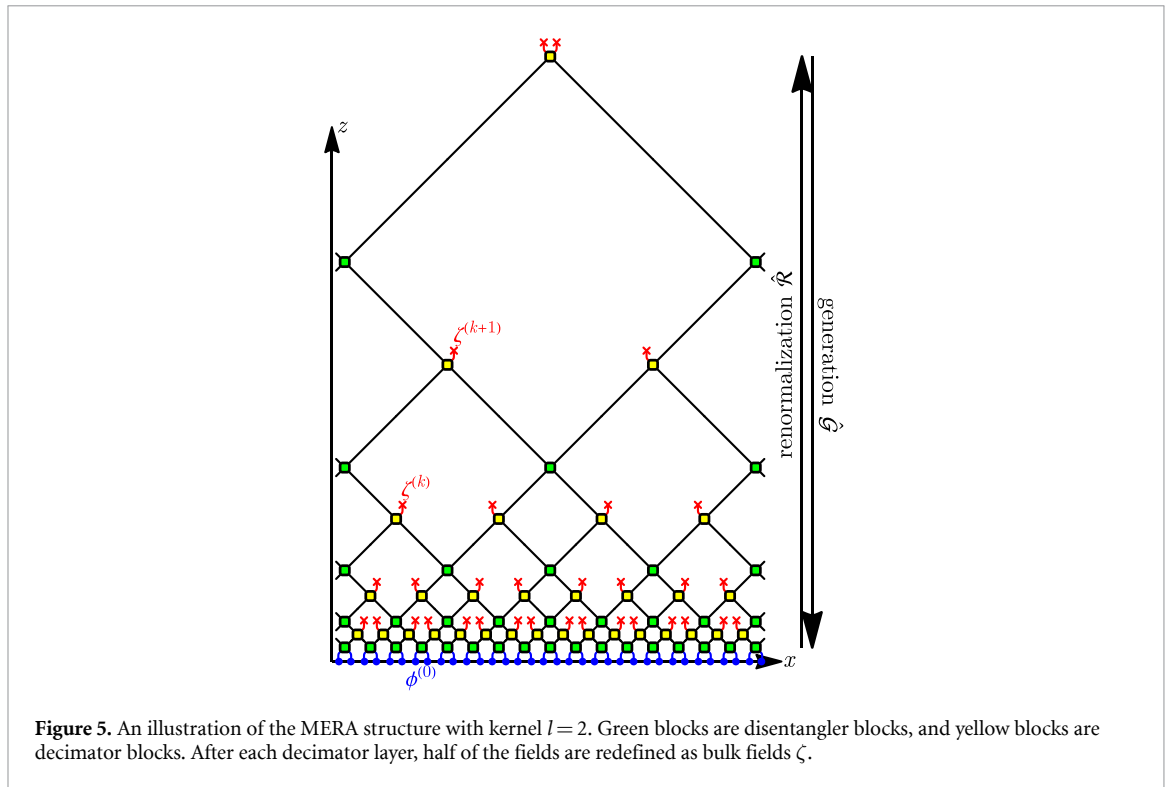|  | Normalized log prob | |
|---|---|---|
|  | Seq length $= 4$ | Seq length $= 16$ |
| Original data | 1.52 (0.45) | 1.52 (0.25) |
| Generated data | 1.50 (0.19) | 1.50 (0.19) |

and yellow blocks as decimators. Each block is a bijective transformation with the neural ODE structure as in figure 3. We can further explicitly write down the transformation equations: the covering length of a disentangler or a decimator is defined as the kernel length *l*. Then there are $\frac{I}{2^k l}$ blocks in the *k*th layer. For the *m*th block in the *k*th layer, where $m \in \{0, \ldots, \frac{I}{2^k l} - 1\}$, the transformation is given by

$$
\begin{aligned}
\left(\psi^{(k)}_{2^k(ml+a)}\right)_{a \in \{1,\ldots,l\}} &= \hat{\mathcal{R}}^{\mathrm{dis},(k)}_m \left(\left(\phi^{(k)}_{2^k(ml+a)}\right)_{a \in \{1,\ldots,l\}}\right) \\
\left(\phi^{(k+1)}_{2^k(ml+l/2+a)}\right)_{a \in \{1,\ldots,l\}} &= \hat{\mathcal{R}}^{\mathrm{dec},(k)}_m \left(\left(\psi^{(k)}_{2^k(ml+l/2+a)}\right)_{a \in \{1,\ldots,l\}}\right).
\end{aligned}
\tag{31}
$$

Half of the resulting fields are redefined as the bulk fields: $\zeta^{(k+1)}_{2^{k+1}(ml+a)} := \phi^{(k+1)}_{2^{k+1}(ml+a)}$. Here we have chosen a scheme that after each layer, every other existing field is redefined as a new bulk field. Since there are position-dependent features among the sequences, to respect the local features of the sequence, we take independent block transformations as they are labeled by both layer index *k* and block index *m*. With this setup, we train on the objective $\mathcal{L} = \mathbb{E}_{\phi \sim p_{\mathrm{dat}}} S_\Phi(\phi)$ as described in section 3.9.

In figures 6 and 7, we show the result when $I = 4, l = 2$ and when $I = 16, l = 4$ with the same set of data in the previous section. To compare the joint distribution, we concatenate the vector embeddings of the 4 and 16 amino acids for each sequence and train the t-SNE algorithm with these concatenated vectors. We also

**Figure 5.** An illustration of the MERA structure with kernel $l = 2$. Green blocks are disentangler blocks, and yellow blocks are decimator blocks. After each decimator layer, half of the fields are redefined as bulk fields $\zeta$.



(A) Initial distribution       (B) Distribution after training

**Figure 6.** Training result on a length 4 amino acid sequence dataset with t-SNE representation. (A): The original data distribution (red) has a multi-modular structure, while the initial flow model's distribution is standard normal (blue). (B): After training RG scheme with neural ODE blocks, two distributions concide.



(A) Initial distribution       (B) Distribution after training

**Figure 7.** Training result on a length 16 amino acid sequence dataset with t-SNE representation. (A): The original data distribution (red) has a multi-modular structure, while the initial flow model's distribution is standard normal (blue). (B): After training the RG scheme with neural ODE blocks, two distributions coincide.

**Figure 8.** Sample of possible mutations in the test sequence. The first column (word) represents the mutant amino acid, and the second column (position) indicates the mutation position in the sequence.
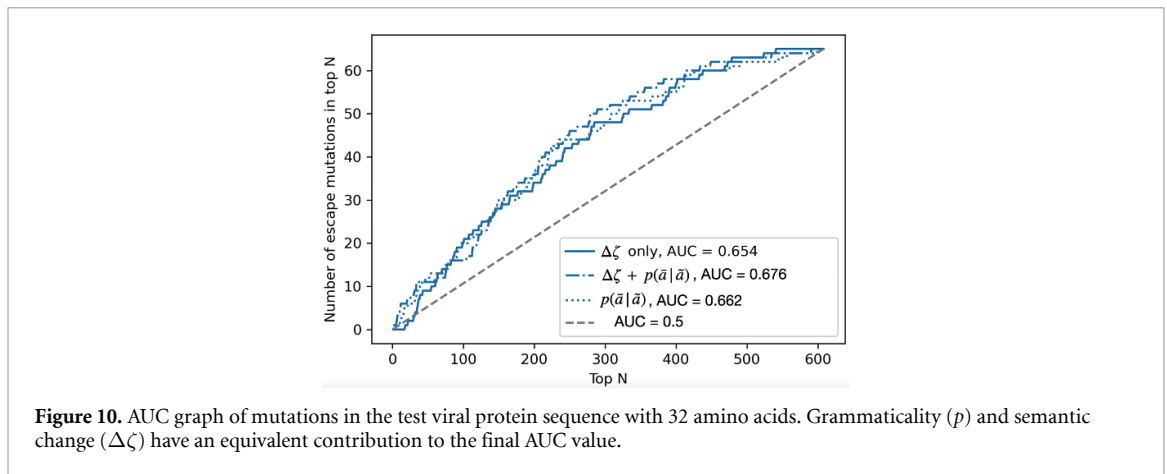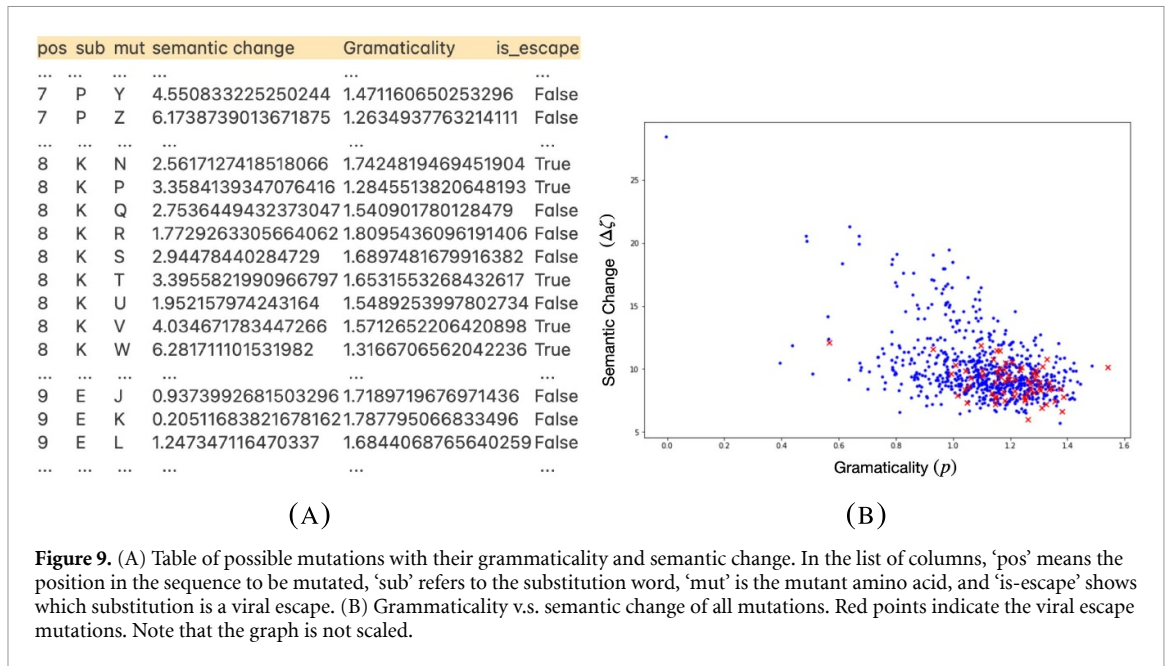
computed the normalized logarithmic probability defined as $\log_n p = \log p/(nI)$ with $n$ the embedding dimension and $I$ the sequence length, as shown in table 2. The numbers in the parentheses are normalized logarithmic probability before training. Both results show that the original data joint distribution can be captured using the RG scheme with local neural ODE blocks.

With training on the full sequence, one can have hierarchical information from each layer. This may give a natural way for the search for the escaping virus. The shallow layers mainly capture the local information, while the deeper layers hold the global information. The escaping virus should have good local fitness (i.e. the sequence must be grammatically correct locally; otherwise, it does not represent a functioning virus) and have significant mutations from the known virus in the dataset to escape the immune system. Then one can use this separation of information levels to design rules for escaping viruses or train on a downstream classification task.

### 4.4. Learning viral escape mutation

We conclude our investigation on learning protein sequence distribution by studying the predictive performance of viral escapes with our model. Viral escapes are those mutations in viral protein sequences that make them unrecognizable from the human immune system. In other words, although they are still effective on the human body and cause infection, the immune system does not flag the mutated sequence as a threat to the body. Such mutations can be single or multiple; that is, only one or few amino acids can be instantly mutated; hence, identifying underlying patterns in viral escape mutations will be essential for viral vaccine development. As described in [35], in terms of language models, a viral sequence can be regarded as textual data, and a viral escape mutation is seen as a word change in a sentence that does change the semantics of the sentence; however, the sentence is still grammatically meaningful. With this analogy, a viral escape is the one capable of making the immune system falsely flag the mutant as a harmless sequence (change in sentiment), while the mutant preserves the virus evolutionary structure (grammatically correct). Therefore, among all possible mutations in a viral sequence, we search for viral escapes, which result in a high semantic change and high grammaticality in our model. Figure 8 depicts an example of all possible mutations in the test sequence.

Following this idea (constrained semantic change search (CSCS) [35]), we first train our model on a corpus of viral sequences in an unsupervised fashion, then take a given viral protein sequence with its known viral escape mutations and rank the mutations based on their grammaticality and semantic change. In our construction, the semantic change caused by a mutation is regarded as the change in the internal representation of the deepest layer in our construction before and after the mutation happens.

In other words, given the test sequence $a = (a_1, a_2, \ldots, a_i, \ldots, a_I)$ and its mutant counterpart $\bar{a} = (a_1, a_2, \ldots, \bar{a}_i, \ldots, a_I)$, a semantic change will be noted as $\Delta\zeta = \left| \zeta^{(K)}(a) - \zeta^{(K)}(\bar{a}) \right|$ where $K$ indicates the deepest layer in the hierarchical structure of our model.

| pos | sub | mut | semantic change | Gramaticality | is_escape |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| 7 | P | Y | 4.550833225250244 | 1.471160650253296 | False |
| 7 | P | Z | 6.1738739013671875 | 1.2634937763214111 | False |
| ... | ... | ... | ... | ... | ... |
| 8 | K | N | 2.5617127418518066 | 1.7424819469451904 | True |
| 8 | K | P | 3.3584139347076416 | 1.2845513820648193 | True |
| 8 | K | Q | 2.7536449432373047 | 1.540901780128479 | False |
| 8 | K | R | 1.7729263305664062 | 1.8095436096191406 | False |
| 8 | K | S | 2.94478440284729 | 1.6897481679916382 | False |
| 8 | K | T | 3.3955821990966797 | 1.6531553268432617 | True |
| 8 | K | U | 1.952157974243164 | 1.5489253997802734 | False |
| 8 | K | V | 4.034671783447266 | 1.5712652206420898 | True |
| 8 | K | W | 6.281711101531982 | 1.3166705562042236 | True |
| ... | ... | ... | ... | ... | ... |
| 9 | E | J | 0.9373992681503296 | 1.7189719676971436 | False |
| 9 | E | K | 0.20511683821678162 | 1.787795066833496 | False |
| 9 | E | L | 1.247347116470337 | 1.6844068765640259 | False |
| ... | ... | ... | ... | ... | ... |

(A)                    (B)

**Figure 9.** (A) Table of possible mutations with their grammaticality and semantic change. In the list of columns, 'pos' means the position in the sequence to be mutated, 'sub' refers to the substitution word, 'mut' is the mutant amino acid, and 'is-escape' shows which substitution is a viral escape. (B) Grammaticality v.s. semantic change of all mutations. Red points indicate the viral escape mutations. Note that the graph is not scaled.



**Figure 10.** AUC graph of mutations in the test viral protein sequence with 32 amino acids. Grammaticality ($p$) and semantic change ($\Delta\zeta$) have an equivalent contribution to the final AUC value.

According to the CSCS objective, grammaticality can be defined as how probable a mutation is in $a$, i.e. the probability value that the model evaluates for a mutation. With this definition, one natural definition of grammaticality is the conditional probability $p(\bar{a}_i|a)$ on the mutation $\bar{a}_i$ in the test protein [35]. In our model, however, the joint probability $p_\Phi(a)$ is the optimization objective that is used to evaluate the input grammaticality. Therefore the final score for each mutation is defined as:

$$\text{score} := \Delta\zeta + p(\bar{a}_i|a). \tag{32}$$

Note that throughout evaluating our model on viral escape mutations, we only consider a single mutation in test data. We also keep the size of our samples to 32 amino acids, and with only 25 different amino acids as the building blocks of sequences, there will be 768 ($24 \times 32$) possible mutations. Among those, a small subset will be viral escape mutations already given to us. For this experiment, we used the escape mutations dataset in [40] that indicates 65 out of those 768 mutations are viral escapes. After calculating both sentiment change and grammaticaly of mutations, we ranked each mutation based on their ranking score in equation (32). Mutations with the highest ranking value will be considered as predicted viral escapes, and consequently, lower rankings indicate less probability for a mutation to be a viral escape. figures 9(A) and (B) illustrate grammaticality and semantic change of all mutations, including the viral escapes (red points), which clearly shows that viral escapes tend to show a high grammaticality.

We also calculated the area-under-curve (AUC) of the ranking scores of mutations in figure 10. Our results clearly indicate that both grammaticality and semantic change quantities have a similar impact on the AUC value.

## 5. Conclusion

We proposed a hierarchical flow-based generative model motivated by the RG in physics. The key idea of RG is to separate relevant and irrelevant features at each scale recursively. The conventional RG throws away irrelevant features and is information lossy. For this reason, the conventional RG is irreversible and only forms a semigroup. We promote the RG to a group by keeping the irrelevant features and formulating RG as a bijective holographic map. The renormalization map encodes the original data into (almost) uncorrelated irrelevant features. The renormalization map can be learned from the data as the optimal transport that transform the data distribution toward the uncorrelated normal distribution, which can be trained within the framework of a flow-based generative model. The generative map (as the reverse of the renormalization map) can then decode randomly sampled irrelevant features into data to achieve the generation task. We apply the construction to bio-sequence modeling. After training, the RG-flow generative model can not only generate new sequences but also estimate the log-likelihood of any given sequence. Using the log-likelihood estimation, we can evaluate whether a sequence is grammatically valid. This enables us to predict viral escape by looking for grammatically valid sequences with large enough semantic changes. Our approach can achieve similar performance as the current state of the art.

As for the limitations, the current machine-learning RG algorithm still relies on an explicit hierarchical structure in the neural network, which has less flexibility in designing the RG scheme (for example, the sequence must be reduced by half at every RG step). It is desired to go beyond this constraint and formulate a more flexible RG scheme. One possible future direction is to incorporate the idea of RG with diffusion-based generative models, which formulates RG flow as a diffusion process. The mathematical foundation has been formulated in appendix, which shows that the conventional RG flow can be understood as a Ricci flow in response to the continuous deformation of the base manifold. This can be naturally combined with diffusion-based generative models. We will explore this possibility in future works.

## Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

## Acknowledgment

## Appendix. Analytic construction of RG flow operator on moduli space of smooth maps

**A special example of our construction, using Laplacians and curvature form**

A large part of current section is based on work of Carfora [12] in relating RG flows for a specific set of QFT's and the Ricci flow construction for Riemannian manifolds. Moreover, the content in this section owes its existence to another highly recommended source, specially for a working mathematician, that is the outstanding work of Kevin Costello [41] in his mathematical formulation of perturbative quantum field theory.

For the time being, we use the introduction to a geometric construction of the RG flow, outlined below, as it is suitably intuitive, pleasantly elegant, and mainly since it will provide us later with the shortest pathways to generalize our constructions in several ways, for instance: by deviating from the classical setup, via altering (and generalizing) our action integrals made with Laplacians and curvature forms to more general actions, or by altering the base geometrical spaces from smooth manifolds to non-smooth algebraic varieties, or discrete lattices.

Let $C, X$ denote respectively a compact oriented Riemann surface and a compact oriented smooth manifold of dimension at least 2, both equipped with a Riemannian metric, and defined over a base number field $\mathbb{K}$. Let $\mathrm{Map}(C, X) := \{f : C \to X\}$ be the associated space of all continuous maps from domain $C$ to $X$. The construction of RG flow is based on considering a family of Lagrangians $\mathcal{L}(f, \phi_i, i = 1, \dots, n)$ associated to this space, defined as a morphism

$$\mathcal{L}(f, \phi_i, i = 1, \dots, n) := \mathrm{Map}(C, X) \times H^*(X, \mathbb{K})^{\otimes n} \to C^\infty(C, X)$$

taking a tuple of fields $(f, \phi_1, \ldots, \phi_n)$ on $X$ to the space of smooth integrable functions on $X$. Note that here the notation $H^*(X, \mathbb{K})^{\otimes n}$ means that the fields $\phi_i, i = 1, \ldots, n$ are realized as sections of a sheaf of differentially graded algebras over $X$, sitting in appropriate cohomological degrees on $X$. Moreover, we require the Lagrangians to be invariant under the action of diffeomorphism groups, $\mathcal{D}iff(C), \mathcal{D}iff(X)$ on $C$ and $X$ respectively.

Integrating the Lagrangian over the associated domain Riemann surface induces the Lagrangian action integral

$$\int_C \mathcal{L}(f, \phi_i, i = 1, \ldots, n).$$

Let the metric tensors on $C$ and $X$ be respectively denoted by $\mu_{mn}, m, n = 1, 2$ and $g_{ij}, i, j = 1, \ldots n$. Suppose that the local coordinates on $C$ are given by $x$ (that is $x := (x_1, x_2)$). Then a typical form of such Lagrangian action integral as defined above is given as

$$\int_C \mathcal{L}(f, \phi_i, i = 1, \ldots, n) = \int_C \lambda^{-1} \left[ \mu_{mn}(x) \partial_m f_i(x) \partial_n f_j(x) g_{ij}(f(x)) + \lambda \rho(f) \mathcal{K} \right] d\nu_C$$
$$m, n = 1, 2 \text{ and } i, j = 1, \ldots, n, \tag{33}$$

where $\lambda$ is a coupling parameter, $\nu_C$ is a measure on $C$, $\rho : X \to \mathbb{K} \in C^\infty(X)$ is a smooth function on $X$, and $\mathcal{K}$ is the Gaussian curvature on $C$ with respect to the metric $\mu$. Here the fields associated to the Lagrangian action integral are given as

$$\phi = \lambda^{-1}(g, \lambda \rho).$$

**Remark 1.** By this notation we mean that the coupling constant $\lambda$ has dependence on parameters $g, \rho$.

**Remark 2.** By writing the action in terms of the Laplacian + curvature form in equation (33), we have assumed to study the RG flow of this particular CFT. However, RG flow can be more generally defined for any field theory with any action to start with, not necessarily near a conformal fixed point. See section 2.4 for more discussions of RG flow around general fixed points.

**Deformation family of Lagrangian action integrals**

Let us denote

$$\mathcal{S}(f, \phi_0) := \int_C \mathcal{L}(f, \phi_i, i = 1, \ldots, n) = \int_C \lambda^{-1} \mu_{mn}(x) \partial_m f_i(x) \partial_n f_j(x) g_0^{ij}(f(x)),$$

where $\phi_0 := \lambda^{-1}(g_0, 0)$ is a field associated to the fixed choice of $g_0$. One interesting case of study is to identify the moduli space (the geometric space representing the family) of smooth maps $f : C \to X$ which minimizes the action integral $\mathcal{S}(f, \phi_0)$ for a fixed choice of metric $g_0$ over $X$. These are often identified with vacuum states of the underlying governing physical theory for our system of particles. A rather more interesting question is whether the vacuum states of the underlying theory are stable with respect to infinitesimal deformations of the geometry of $C$ and $X$, respectively, especially in quantum physics, where fields and geometry of space undergo algebraic or analytic fluctuations. This question could be rigorously studied via inducing deformations of the fields involved in our physical theory, that is

$$\phi_0 \to \phi_0 + \partial \phi_0 = \lambda^{-1}(g_0 + h, 0 + \lambda \rho),$$

where the function $h \in C^\infty(X, T^\vee X^{\otimes 2})$ is a symmetric bilinear smooth differential form on $X$ and $\rho \in C^\infty(X, \mathbb{K})$ is a smooth function on $X$. Introducing these deformation parameters, one can study the set of extremizing maps $f : C \to X$ of the action integral $\mathcal{S}(f, \phi)$, that is smooth harmonic maps minimizing $\mathcal{S}(f, \phi)$, where $\mathcal{S}(f, \phi)$ is obtained as a local deformation around $\mathcal{S}(f, \phi_0)$ induced by deforming the geometry of $C, X$. Let us consider a generalized deformed Lagrangian action

$$\mathcal{S}(f, \phi) = \mathcal{S}(f, \phi_0) + \int_C \lambda^{-1} \mu_{mn}(x) \partial_m f_i(x) \partial_n f_j(x) g_{ij}(f(x)) d\nu_C$$
$$+ \lambda^{-1} \int_C \lambda \rho(f) \mathcal{K} d\nu_C + \lambda^{-1} \int_C \Gamma(f) d\nu_C + \lambda^{-1} \int_C f^* \omega d\nu_C \tag{34}$$

where as before $h \in C^\infty(X, T^\vee X^{\otimes 2})$, $\Gamma \in C^\infty(X, \mathbb{K})$, and $\omega \in C^\infty(X, \wedge^2 T^\vee X)$ an antisymmetric bilinear form are all regarded as infinitesimal induced deformation parameters. Note that here the deformation parameters

$\phi_1 := \lambda^{-1}h$, $\phi_2 = \lambda^{-1}(\lambda\rho)$, $\phi_3 := \lambda^{-1}U$ and $\phi_4 := \lambda^{-1}\omega$ may, roughly speaking, be regarded as local coordinates in the space of deformations of $\mathcal{S}(f,\phi_0)$. Hence we can rewrite one such deformation in terms of the other as an extension

$$\mathcal{S}(f,\phi) = \mathcal{S}(f,\phi_0) + \sum_{i\geqslant 1}\int_C O_i(f,\phi_i). \tag{35}$$

Moreover, it must be noted that depending on the underlying physical theory, one may consider situations where $\mathcal{S}(f,\phi_0)$ is required to be invariant under conformal transformations $(C,\mu_{mn}) \to (C,e^{-\psi}\mu_{mn})$, in which case, shall one be interested to preserve the conformal invariance of the deformed Lagrangian action $\mathcal{S}(f,\phi)$, one requires that the deformation fields $\rho$ and $U$ vanish, as they break the conformal symmetry, however the deformations $h,\omega$ can be non-vanishing, as their associated integrals are preserved under conformal group action on $C$.

**Moduli functors associated to deforming fields and maps simultaneously**

We mimic the approach of algebraic geometers for constructing our moduli spaces. Consider the following situation. Let $\mathcal{T} \to \mathrm{Spec}\mathbb{K}$ be a finite type parametrizing scheme. The notation means that $\mathcal{T}$ is a space (known as parametrizing scheme in algebraic geometry terms) constructed over field of numbers $\mathbb{K}$ that is topologically compact. Let $\mathfrak{Map}(C,X) : \mathcal{S}\mathrm{ch}/\mathbb{K} \to \mathcal{A}\mathrm{b}/\mathbb{K}$ be defined as a two-category (i.e. a category which contains objects, their morphisms, and their morphisms of morphisms , also known as 2-morphisms), such that the category is fibered over a base category of finite type (parametrizing) schemes over $\mathbb{K}$. The objective of such functor is to produce families of maps from $C$ to $X$ parametrized by schemes such as $\mathcal{T}$. To state the latter functionality of $\mathfrak{Map}(C,X)$ in more mathematical formal terms, we say that the groupoid sections of $\mathfrak{Map}(C,X)$ over any $\mathcal{T}$ are given by the sheaf of Abelian groups of $\mathcal{T}$-families of smooth maps from $C$ to $X$, that is the groupoid sections of our moduli functor are given by families of maps

$$\mathfrak{Map}(C,X)(\mathcal{T}) \cong \{\tilde{f} : C_{\mathcal{T}} := \mathcal{T} \times_{\mathbb{K}} C \to X\} \tag{36}$$

such that for any $t \in \mathcal{T}$ the $t$-fibers of the family $\tilde{f}|_t \cong \{f_t : C \to X\}$ are given by smooth continuous maps from domain Riemann surface $C$ to $X$. Roughly speaking, the functor $\mathfrak{Map}(C,X)$ provides us with a platform to parametrize the smooth maps from $C$ to $X$ in a systematic way over any chosen parametrizing scheme. For instance, given any $\mathcal{T} := \mathrm{Spec}(\mathbb{K})$, geometric reduced point, the groupoid sections of $\mathfrak{Map}(C,X)(\mathcal{T})$ are given by single maps $f : C \to X$. Similarly, the fibers of $\mathfrak{Map}(C,X)$ over a line, $L$ (which as a geometric scheme belongs to our category, $\mathcal{S}\mathrm{ch}/\mathbb{K}$, of schemes of $\mathbb{K}$) provides a one dimensional family of maps $f_L : C_L \to X$, and the fibers of $\mathfrak{Map}(C,X)$ over a surface provides a two dimensional family of maps, etc.

Now as the geometric structure of $C,X$, and hence $f$ undergo deformations in our theory, similar to Feynman path integration formalism, we compute the vacuum states of the theory, by taking a stochastic average over all admissible weighted morphisms $f : C \to X$ which satisfy smoothness property. In doing so, we further allow certain induced correlation fields, defined in our theory, induced by evaluating the map $f$ at a finite number of smooth distinct marked points $p_1,\ldots,p_l \in C$. Moreover, we use the Lagrangian action integral constructed in previous section as a weight function associated to each single map $f : C \to X$. Doing so, we obtain an integral over the space parametrizing tuples $(f : C \to X, p_1,\ldots,p_l)$, where $p_i, i = 1,\ldots,l$ are distinct smooth marked points on $C$

$$Z[C,X,p_1,\ldots,p_n,\phi] := \frac{1}{Z_0}\int_{\mathfrak{Map}(C,p_1,\ldots,p_l,X)} D_\phi[f](f(p_1,\ldots,f(p_l)))e^{-\mathcal{S}(f,\phi)}. \tag{37}$$

Here $D_\phi(f)$ is a measure over $\mathfrak{Map}(C,p_1,\ldots,p_l,X)$. Note that by construction $\mathcal{S}(f,\phi)$ is regarded as a deformation of $\mathcal{S}(f,\phi_0)$, hence following the construction in (35), one may rewrite correlation function (37) in terms of $\mathcal{S}(f,\phi_0)$ as follows

$$Z[C,X,p_1,\ldots,p_n,\phi] = \frac{1}{Z_0}\int_{\mathfrak{Map}(C,p_1,\ldots,p_l,X)} D_\phi[f](f(p_1,\ldots,f(p_l)))e^{-\mathcal{S}(f,\phi_0)}\prod_{i\geqslant 1}\int_C O_i(f,\phi_i). \tag{38}$$

**Renormalization semigroup flow**

The construction of the renormalization semigroup flow is based on the fact that, in order to make the above integrals well-defined, one may merely consider certain controllable deformation regimes for the fields $\phi_i$, that is; one would like to consider a family of the fields $\phi_i(\mathcal{T})$, where the scheme $\mathcal{T}$ is the parametrizing scheme, used in (36) governing the geometric deformations of maps $f_{\mathcal{T}} : C_{\mathcal{T}} \to X$ induced by perturbation of geometric structures of $C$ and $X$. The idea is to consider an infinitesimal deformation flow, called

renormalization semigroup flow (as it turns out that our construction in this example only provides a semigroup rather than a group), over the moduli space of maps and field deformations, that is, to consider a morphism

$$
\mathcal{RG}_{\mathcal{T}} := \mathfrak{Map}(C,X)(\mathcal{T}) \times H^*(X,\mathbb{K})^{\otimes n} \to \mathfrak{Map}(C,X)(\mathcal{T}) \times H^*(X,\mathbb{K})^{\otimes n}
$$
$$
\times (f, \phi_1, \ldots, \phi_n) \to (f_{\mathcal{T}}, \phi_{1,\mathcal{T}}, \ldots, \phi_{n,\mathcal{T}}) \tag{39}
$$

which has a lift to a morphism on moduli space of action integrals

$$
\overline{\mathcal{RG}}_{\mathcal{T}} := \mathcal{A}ct(C,X) \cong \left(\mathfrak{Map}(C,X)(\mathcal{T}) \times H^*(X,\mathbb{K})^{\otimes n}\right)^{\vee} \to \mathcal{A}ct(C,X) \cong \left(\mathfrak{Map}(C,X)(\mathcal{T}) \times H^*(X,\mathbb{K})^{\otimes n}\right)^{\vee} \tag{40}
$$

taking $\mathcal{S}(f,\phi_0)$ to $\mathcal{S}(f_{\mathcal{T}},\phi_{\mathcal{T}})$, which satisfies the semigroup property.

**Remark 3.** We remark again that we are considering, generally speaking, our fields $\phi_i, i=1,\ldots,n$ as living in our field algebra, that is the vector space $H^*(X,\mathbb{K})^{\otimes n}$ generated by differentially graded forms on $X$. Moreover, the action integrals are regarded as morphisms from $\mathfrak{Map}(C,X)(\mathcal{T}) \times H^*(X,\mathbb{K})$ to the underlying ground field $\mathbb{K}$, and hence, realized as the dual space $(\mathfrak{Map}(C,X)(\mathcal{T}) \times H^*(X,\mathbb{K})^{\otimes n})^{\vee}$.

We now elaborate further on renormalization flow. In order to define it we need to formulate a deformation process, applied to geometry of $C, X$, then compute the induced deformations of associated fields $\phi_i$ and $f$ with support on deformed $X$ as shown in equation (39). Note that the functorial construction of the moduli space of maps allows us to perform this task in a rigorous algebraic manner. Take a scheme $\mathcal{T}$ (naively speaking, schemes have as their skeleton the geometrical spaces however, they come further equipped with extra topological or algebraic properties). As we noted above, the fibers of the moduli functor $\mathfrak{Map}(C,X)$ over $\mathcal{T}$ (i.e. $\mathfrak{Map}(C,X)(\mathcal{T})$) provide us with a $\mathcal{T}$-family of maps from $C \to X$ as in (36). Now choose an algebraic deformation (a perturbation) of $\mathcal{T}$ and denote it by $\mathcal{T}'$. Then the fibers $\mathfrak{Map}(C,X)(\mathcal{T}')$ provide a $\mathcal{T}'$-family

$$
\{\tilde{f}' : C_{\mathcal{T}'} := \mathcal{T}' \times_{\mathbb{K}} C \to X\},
$$

realized as a deformation of the former $\mathcal{T}$-family maps from $C$ to $X$.

One way of constructing such algebraic deformation is to construct $\mathcal{T}'$ as a *nilpotent thickening* of $\mathcal{T}$. We elaborate on this notion, using the language of ideals over the ring of polynomial functions.

Take the polynomial ring $\mathbb{C}[x_1,\ldots,x_n]$. In classical algebraic geometry, the set of prime ideals generated by different expressions involving the variables $x_1,\ldots,x_n$ makes a space, isomorphic to the '*affine*' space $\mathbb{C}^n$. Now in order to obtain more interesting spaces, one may consider an ideal, say as an example $\mathcal{I} = (x_1 x_2 - x_3^2)$, and consider the quotient ring $\mathbb{C}[x_1,\ldots,x_n]/\mathcal{I}$. This expression means that all polynomials generated by the expression $x_1 x_2 - x_3^2$ vanish on this quotient ring. Now the set of prime ideals $p \subset \mathbb{C}[x_1,\ldots,x_n]/\mathcal{I}$ provides us with set of geometric points of the algebraic space (algebraic variety) given as the solution set to the polynomial equation $x_1 x_2 - x_3^2 = 0$. Let us denote this algebraic variety as $\mathcal{T}$. In order to obtain a nilpotent thickening of $\mathcal{T}$ one can simply construct the quotient ring $\mathbb{C}[x_1,\ldots,x_n]/\mathcal{I}^l$ for some $l$. The set of prime ideals in the latter provides one with the set of geometric points of the variety obtained as the solution set to $(x_1 x_2 - x_3^2)^l = 0$, call the latter space $\mathcal{T}'$. Due to the natural inclusion of ideals $\mathcal{I}^l \subset \mathcal{I}$, one can immediately obtain a natural inclusion of $\mathcal{T} \hookrightarrow \mathcal{T}'$. This deformation is called a nilpotent extension of $\mathcal{T}$ of order $l$. Given a deformation as such nilpotent extension, $\iota_{TT'} : T \hookrightarrow T'$, as we elaborated earlier, the renormalization flow must satisfy the property that

$$
\iota_{TT'}^* \overline{\mathcal{RG}}_{\mathcal{T}} \left(\mathcal{S}(f,\phi_0)\right) = \mathcal{S}(\iota_{TT'}^* \mathcal{RG}_{\mathcal{T}'}(f,\phi_0)).
$$

Since the action of the RG flow is realized as a pullback in our construction, one is able to define its induced action on the correlation function defined in (38) as follows

$$
\overline{\mathcal{RG}}_{\mathcal{T}}(Z[C,X,p_1,\ldots,p_n,\phi]) = \frac{1}{Z_0} \int_{\mathcal{RG}_{\mathcal{T}}(\mathfrak{Map}(C,p_1,\ldots,p_l,X))} D_{\phi}[f](f(p_1),\ldots,f(p_l)))e^{-\mathcal{S}(f,\phi)}
$$
$$
= \frac{1}{Z_0} \int_{\mathfrak{Map}(C,p_1,\ldots,p_l,X)} \mathcal{RG}_{\mathcal{T}}^*(D_{\phi}[f](f(p_1),\ldots,f(p_l)))e^{-\overline{\mathcal{RG}}_{\mathcal{T}}^*(\mathcal{S}(f,\phi))}. \tag{41}
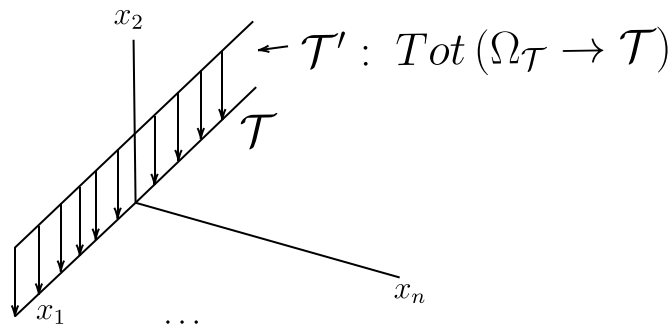$$

Let us work out a concrete example.

**Example 4.** For simplicity, let us assume that $\mathbb{K}$ is given as a field of characteristic zero, such as $\mathbb{C}$, the field of complex numbers. Consider the case where $\mathcal{T} := \mathrm{Spec}(\mathbb{K}[x_1, x_2, \ldots, x_n]/(x_2, \ldots, x_n)) \cong \mathbb{A}^1$ is given by taking the Zariski spectrum of the affine line in the direction $x_1$, given by ideal $\mathcal{I} = (x_2, \ldots, x_n)$ over $\mathbb{K}$. Locally, after choosing a coordinate chart $(x_1, \ldots, x_n)$, the set geometric points in $\mathcal{T}$ is the set of points on the $x_1$ axis in $\mathbb{C}^n$. Now we introduce an infinitesimal deformation of $\mathcal{T} \hookrightarrow \mathcal{T}'$, induced by a nilpotent extension of order 2, by taking $\mathcal{T}' := \mathrm{Spec}(\mathbb{K}[x_1, \ldots, x_n]/\mathcal{I}^2)$. There exists a canonical short exact sequence

$$0 \to \mathcal{I}/\mathcal{I}^2 \to \mathbb{K}[x_1, \ldots, x_n]/\mathcal{I}^2 \to \mathbb{K}[x_1, \ldots, x_n]/\mathcal{I} \to 0$$

whose kernel is governed by the conormal sheaf (which in here is identified by sheaf of differential one forms on $\mathcal{T}$, that is $\Omega_{\mathcal{T}}$). This roughly speaking realizes the second order nilpotent thickening of $\mathcal{T}$, as the cotangent bundle, $\Omega_{\mathcal{T}}$, of $\mathcal{T}$. We would like to deform the correlation function (38) in the direction of fibers of $\Omega_{\mathcal{T}}$. This amounts to setting $\mathcal{RG}_{\mathcal{T}}$ as the differential operator which deforms the fields in direction of fibers of cotangent bundle of $\mathcal{T}$, that is, RG flow acts on the fields as a map $\phi \to \phi + d\phi$ and hence its induced action on the action integral is given by



$$\overline{\mathcal{RG}}_{\mathcal{T}}(Z[C, X, p_1, \ldots, p_n, \phi]) = \frac{1}{Z_0} \int_{\mathcal{RG}_{\mathcal{T}}(\mathfrak{Map}(C, p_1, \ldots, p_l, X))} D_\phi[f](f(p_1), \ldots, f(p_l)))e^{-\mathcal{S}(f, \phi)}$$

$$= \frac{1}{Z_0} \int_{\mathfrak{Map}(C, p_1, \ldots, p_l, X)} \mathcal{RG}_{\mathcal{T}}^*(D_\phi[f](f(p_1), \ldots, f(p_l))))e^{-\overline{\mathcal{RG}}_{\mathcal{T}}^*(\mathcal{S}(f, \phi))}$$

$$\frac{1}{Z_0} \int_{\mathfrak{Map}(C, p_1, \ldots, p_l, X)} (D_{\phi+d\phi}[f](f(p_1), \ldots, f(p_l))))e^{-(\mathcal{S}(f, \phi+d\phi))} = -\frac{1}{Z_0} \int_{\mathfrak{Map}(C, p_1, \ldots, p_l, X)} \frac{\partial \phi(x_n)}{\partial x_n} \frac{\partial}{\partial \phi}$$

$$\times \left( D_\phi[f](f(p_1), \ldots, f(p_l)))e^{-\mathcal{S}(f, \phi)} \right). \tag{42}$$

Therefore, viewing the RG flow as a differential operator acting on the action integral, $Z$, and rewriting the variation of $Z$, induced by nilpotent deformation of $\mathcal{T}$, in terms of $Z$ itself, we obtain a differential equation governing the change in $Z$, that is

$$\left[ \overline{\mathcal{RG}}_{\mathcal{T}} + \frac{\partial \phi(x_n)}{\partial x_n} \frac{\partial}{\partial \phi} \right] Z[C, X, p_1, \ldots, p_n, \phi] = 0. \tag{43}$$

## ORCID iD

Yi-Zhuang You ● https://orcid.org/0000-0003-4080-5340

## References

[1] Sheshmani A and Yizhuang Y 2021 Categorical representation learning: morphism is all you need *Mach. Learn.: Sci. Technol.* **3** 015016
[2] Polchinski J 1984 Renormalization and effective Lagrangians *Nucl. Phys.* B **231** 269–95
[3] Hamilton R S 1982 Three-manifolds with positive Ricci curvature *J. Differ. Geom.* **17** 255–306
[4] Hamilton R S 1999 Non-singular solutions of the Ricci flow on three-manifolds *Commun. Anal. Geom.* **7** 695–729
[5] Hamilton R S 1993 Eternal solutions to the Ricci flow *J. Differ. Geom.* **38** 1–11
[6] Gage M and Hamilton R S 1986 The heat equation shrinking convex plane curves *J. Differ. Geom.* **23** 69–96
[7] Hamilton R S 1995 The formation of singularities in the Ricci flow *Surveys in Differential Geometry, Vol II (Cambridge, MA, 1993)* (Cambridge, MA: Int. Press) pp 7–136
[8] Freidan D 1980 Nonlinear models in $2 + \varepsilon$ dimensions *Phys.Rev.Lett.* **45** 1057
[9] Zomolodchikov A B 1986 Irreversibility of the flux of the renormalization group in a 2D field theory *Pisma Zh. Eksp. Teor. Fiz.* **43** 565–7
[10] Tseytlin A A 1986 Conformal anomaly in a two-dimensional sigma model on a curved background and strings *Phys. Lett.* B **178** 34–40

[11] Perelman G 2003 Ricci flow with surgery on three-manifolds (arXiv:math/0303109)
[12] Carfora M 2010 Renormalization group and the Ricci flow *Milan J. Math.* **78** 319–53
[13] Li S-H and Wang L 2018 Neural network renormalization group *Phys. Rev. Lett.* **121** 260601
[14] Hu H-Y, Li S-H, Wang L and You Y-Z 2020 Machine learning holographic mapping by neural network renormalization group *Phys. Rev. Res.* **2** 023369
[15] Hu H-Y, Wu D, You Y-Z, Olshausen B and Chen Y 2020 RG-Flow: a hierarchical and explainable flow model based on renormalization group and sparse prior (arXiv:2010.00029)
[16] Chen R T Q, Rubanova Y, Bettencourt J and Duvenaud D 2018 Neural ordinary differential equations (arXiv:1806.07366)
[17] Bény C 2013 Deep learning and the renormalization group (arXiv:1301.3124)
[18] Mehta P and Schwab D J 2014 An exact mapping between the variational renormalization group and deep learning (arXiv:1410.3831)
[19] Bény C and Osborne T J 2015 The renormalization group via statistical inference *New J. Phys.* **17** 083005
[20] Oprisa D and Toth P 2017 Criticality & deep learning I: generally weighted nets (arXiv:1702.08039)
[21] Oprisa D and Toth P 2017 Criticality & deep learning II: momentum renormalisation group (arXiv:1705.11023)
[22] Lin H W, Tegmark M and Rolnick D 2017 Why does deep and cheap learning work so well? *J. Stat. Phys.* **168** 1223–47
[23] Koch-Janusz M and Ringel Z 2018 Mutual information, neural networks and the renormalization group *Nat. Phys.* **14** 578–82
[24] Lenggenhager P M, Ringel Z, Huber S D and Koch-Janusz M 2018 Optimal renormalization group transformation from information theory (arXiv:1809.09632)
[25] Li S-H and Wang L 2018 Neural network renormalization group *Phys. Rev. Lett.* **121** 260601
[26] Hu H-Y, Li S-H, Wang L and You Y-Z 2020 Machine learning holographic mapping by neural network renormalization group *Phys. Rev. Res.* **2** 023369
[27] Chung J-H and Kao Y-J 2021 Neural Monte Carlo renormalization group *Phys. Rev. Res.* **3** 023230
[28] Cotler J and Rezchikov S 2022 (arXiv:2202.11737)
[29] Qi X-L 2013 Exact holographic mapping and emergent space-time geometry (arXiv:1309.6282)
[30] Dinh L, Sohl-Dickstein Js and Bengio S 2016 Density estimation using real NVP (arXiv:1605.08803)
[31] Hahn M 2019 Theoretical limitations of self-attention in neural sequence models (arXiv:1906.06755)
[32] Ma H and Lee S-S, 2020 Constraints on beta functions in field theories (arXiv:2009.11880)
[33] Du J, Jia P, Dai Y, Tao C, Zhao Z and Zhi D 2019 Gene2vec: distributed representation of genes based on co-expression *BMC Genomics* **20** 82
[34] Wu F, Yang R, Zhang C and Zhang L 2021 A deep learning framework combined with word embedding to identify DNA replication origins *Sci. Rep.* **11** 844
[35] Hie B, Zhong E D, Berger B and Bryson B 2021 Learning the language of viral evolution and escape *Science* **371** 284–8
[36] Ji Y, Zhou Z, Liu H and Davuluri R V 2021 DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome *Bioinformatics* **37** 2112–20
[37] Finlay C, Jacobsen J-H, Nurbekyan L and Oberman A M 2020 How to train your neural ODE: the world of Jacobian and kinetic regularization (arXiv:2002.02798)
[38] Grathwohl W, Chen R T Q, Bettencourt J, Sutskever I and Duvenaud D 2018 Ffjord: free-form continuous dynamics for scalable reversible generative models (arXiv:1810.01367)
[39] van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[40] Doud M B, Lee J M and Bloom J D 2018 How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin *Nat. Commun.* **9** 1–12
[41] Costello K 2011 *Renormalization and Effective Field Theory, 170 of Mathematical Surveys and Monographs* (Providence, RI: American Mathematical Society)