# AdS/VAEBM - theory from data: AdS/DL correspondence with flexible functional forms for bulk action

Minseong Kim*

(Dated: March 6, 2025)

## Abstract

The AdS/DL (AdS/deep learning) correspondence was implemented as the AdS/DBM correspondence in Hashimoto (2019). We extend AdS/DBM as AdS/VAEBM in general contexts of holographic renormalization such that the bulk theory action is no longer constrained by DBM, using the adversarial VAEBM framework offered in Han et al. (2020). The best of both supervised AdS/DL and self-supervised generative AdS/DL (as in Hashimoto (2019)) are offered by AdS/VAEBM - reconstruction of the classical background spacetime is feasible, while the bulk theory itself remains quantum. AdS/VAEBM then provides a framework to analyze how the holographic bulk theory is modified by perturbations to the IR limit and imposition of a braneworld scenario (where a d-dimensional gravitational theory lives within d+1-dimensional AdS). It is also noted that a single UV theory state is consistent with multiple IR theory states in generative AdS/DL (and therefore AdS/CFT), clearing a common misunderstanding in this regard.

* mkimacad@gmail.com; ORCiD:0000-0003-2115-081X

**CONTENTS**



## I. INTRODUCTION

In recent years, AdS/CFT has been analyzed within the AdS/DL correspondence [1], with supervised and self-supervised generative versions [2]. Conventional supervised learning cases of AdS/DL tend to be fairly limited - one does not attempt to completely recover the bulk theory and the bulk metric from the boundary theories by deep learning and instead

attempts to recover parts of the bulk metric and a few bulk observables that remain fairly constrained.

Although the AdS/DBM correspondence in [2] removes these constraints by understanding the DBM weights as corresponding to the bulk metric, the DBM (deep Boltzmann machine) nevertheless limits the possible functional forms of the bulk action. This paper generalizes AdS/DBM to AdS/VAEBM such that both the bulk theory and the metric are flexible. The VAEBM framework adopted in this paper follows [3] with the adversarial chasing game between encoder, decoder, and EBM (energy-based model).

In AdS/DBM, the bulk is understood as a variational encoder that compresses information on the UV boundary. We extend such encoder analysis in terms of the holographic renormalization framework in [4, 5] to formulate AdS/VAEBM. The adversarial game in AdS/VAEBM induces the encoder to compress information on the UV boundary (at $z = \epsilon \approx 0$) due to the presence of the EBM as a critic model to the decoder, while the decoder tries to recover much of the original information on the UV boundary [3]. We note that the EBM energy is understood as referring to the bulk action in generative AdS/DL [2]. Encoding is also guided by AdS regularization, which aims to minimize bulk deviations from AdS/CFT even when the IR limit is perturbed away from the AdS/CFT prediction. In order to recover the semiclassical picture of the bulk, the EBM structure must be carefully designed, which is carried out in Section III B.

With AdS/VAEBM, a braneworld scenario is considered in which a $d$-dimensional (classical) gravitational theory lives within $d+1$-dimensional AdS [6–8]. This is achieved by constraining coarse EBM self-energy $\mathcal{E}_{coarse}(\mathbf{h}, \mathbf{h})$ to approximate the $d$-dimensional gravitational theory action. The resulting lesson in AdS/DL is that IR physics may have minimal effects on UV physics and vice versa.

A generic feature within generative AdS/DL - that multiple IR theory field configurations ('states') are consistent with the same UV state - is also described in this paper. This clears a common misunderstanding that a single UV field configuration must have a single IR field configuration - no such classical-to-classical coarse-graining is typically available, and epistemic uncertainty is mostly inevitable if we are condemned to IR observables.


FIG. 1: The goal is to find out the physical law behind the black box, with inputs given by $(x_i, v_i)$ and outputs given by $(x_f, v_f)$, with $x$ understood as position and $v$ understood as velocity.

## II. CONVENTIONAL ADS/DL REVISITED

### A. Supervised AdS/DL

In [9], a simple classical example is used to outline the principles behind AdS/DL, which we partially replicate. For the setup described in Figure 1, it is not much of work to simply invoke conventional supervised learning with inputs $(x_i, v_i)$ and outputs $(x_f, v_f)$. The real issue is to translate this as a physical law by imposing constraints - this requires more works, and the following discretized flexible equation of motion is assumed:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + v^{(k)}\,\Delta t\,, \\ v^{(k+1)} &= v^{(k)} + f(x^{(k)}, v^{(k)})\,\Delta t\,, \end{aligned} \tag{1}$$

Input $(x_i, v_i)$ corresponds to $\mathbf{x}^{(0)} = (x^{(0)}, v^{(0)})^T$, while output $(x_f, v_f)$ corresponds to $\bar{\mathbf{x}}^{(\text{out})} = (x^{(\text{end})}, v^{(\text{end})})^T$ ideally. This system of equations is cast into a form resembling a neural network:

$$\mathbf{x}^{(k+1)} = \varphi^{(k)}\left(W^{(k)}\mathbf{x}^{(k)}\right)\,, \tag{2}$$

where

$$W^{(k)} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}\,, \qquad \varphi^{(k)}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ b + f(x^{(k)}, v^{(k)})\,\Delta t \end{pmatrix}\,. \tag{3}$$

and training is now modified to learning activation function $\varphi$, which is also about learning $f(x^{(k)}, v^{(k)})$ at each layer $k$. We can safely assume that $f$ is modeled as a neural network, with the loss function is changed to:

$$L = \frac{1}{n_{\text{batch}}}\sum_{\text{batch}} |\bar{\mathbf{x}}^{(\text{out})} - \mathbf{x}^{(T)}| + L_{\text{reg}}\,, \tag{4}$$

Typically, it is not an activation function that is learned - the setup of linear trainable weights with fixed activation functions is more usual, as is the case for the AdS/DL correspondence in [1]. There, it is metric that is a black box, with metric determining field equations of motion completely, assuming the large-N limit. A flexible functional form for (asymptotically) AdS metric goes as:

$$ds^2 = -f(\eta)dt^2 + d\eta^2 + g(\eta)(dx_1^2 + \cdots + dx_{d-1}^2) \tag{5}$$

The assumed scalar field theory goes as:

$$S = \int d^{d+1}x\sqrt{-\det g}\left[-\frac{1}{2}(\partial_\mu\phi)^2 - \frac{1}{2}m^2\phi^2 - V(\phi)\right]. \tag{6}$$

Discretizing the resulting equations of motion, we obtain:

$$\begin{aligned}\phi(\eta+\Delta\eta) &= \phi(\eta) + \Delta\eta\,\pi(\eta)\\ \pi(\eta+\Delta\eta) &= \pi(\eta) - \Delta\eta\left(h(\eta)\pi(\eta) - m^2\phi(\eta) - \frac{\delta V(\phi)}{\delta\phi(\eta)}\right)\end{aligned} \tag{7}$$

where:

$$\pi \equiv \partial_\eta\phi, \quad h(\eta) \equiv \partial_\eta \log\sqrt{f(\eta)g(\eta)^{d-1}} \tag{8}$$

Therefore, the goal is to train weight $h(\eta)$ at each discrete $\eta$, given input $(\phi(\eta_{ini}), \pi(\eta_{ini}))$ at $\eta_{ini} \approx \infty$ matched with binary output $y \in \{0, 1\}$ that labels whether the boundary condition at $\eta_{fin} \approx 0$ is satisfied, output zero labeling satisfaction. The boundary condition satisfaction refers to $F = 0$, with $F$ being used as the final output activation function:

$$F(\phi, \pi) \equiv \left[\frac{2}{\eta}\pi - m^2\phi - \frac{\delta V(\phi)}{\delta\phi}\right]_{\eta=\eta_{\rm fin}} \tag{9}$$

As a neural network form with $\mathbf{x^{(k)}} = (\phi^{(k)}, \pi^{(k)})^T$ and $\eta^{(k)} = \eta_{ini} + k\Delta\eta$,

$$\begin{aligned}W^{(n)} &= \begin{pmatrix} 1 & \Delta\eta \\ \Delta\eta\, m^2 & 1 - \Delta\eta\, h(\eta^{(n)}) \end{pmatrix},\\ \varphi_1(x_1) &= x_1\\ \varphi_2(x_2) &= x_2 + \Delta\eta\,\tfrac{\delta V(x_1)}{\delta x_1},\\ y(\mathbf{x^{(0)}}) &= F(\varphi(W^{(N-1)}\varphi(W^{(N-2)}\cdots\varphi(W^{(0)}\mathbf{x^{(0)}})))).\end{aligned} \tag{10}$$

A loss function example goes as:

$$L = \frac{1}{n_{batch}}\sum_{batch}|\bar{y} - y| + L_{reg} \tag{11}$$

where $\bar{y}$ is actual output data. $h(\eta)$ is then updated via backpropagation and some variant of gradient descent. We note several papers that are in a similar vein [10–13].

| AdS/CFT | AdS/DL (AdS/DBM, AdS/VAEBM) |
|---|---|
| Bulk radial coordinate $z$ | Hidden layer label $k$ |
| QFT source $J(x)$ | Input value $v_i$ |
| Bulk field $\phi(x,z)$ | Hidden variables $h_i^{(k)}$ |
| QFT generating functional $Z[J]$ | Probability distribution $P(\mathbf{v})$ |
| Bulk action $S[\phi]$ | Energy function $\mathcal{E}(\mathbf{v},\mathbf{h}^{(\mathbf{0<k\leq f})})$ |

TABLE I: From [2].

### B. Hashimoto (2019): (generative) AdS/DL as deep Boltzmann machines (DBM)

In [1], while the metric is slightly flexible, a bulk theory is already given. The bulk theory part and the metric can be made more flexible, as described in [2]. Indeed, in spirit of bidirectional holography, we should let the bulk theory completely emerge from the boundary theory as well. The cost involved is that we now move from conventional supervised learning involving input-output pairs to training generative models - in the case of [2], DBMs.

Instead of fields propagating with classical equations of motion as in [1], we now let fields be mostly unconstrained, and action $S$ - which defines a Wick-rotated quantum theory with the help of path integrals - is learned instead. Some functional form on action $S$ still needs to be implicitly given in [2], which is provided by the very structure of DBMs - this limits the class of theories that can be probed. **This restriction is what we intend to eliminate in this paper by moving away from DBMs to VAEBMs.**

The DBM energy $\mathcal{E}(\mathbf{v})$ is given by, with $v_i$ being part of visible layer $\mathbf{v}$:

$$\mathcal{E}(\mathbf{v}) = \sum_{i,j} w_{ij}^{(0)} v_i h_j^{(1)} + \sum_{k=1}^{N-1} \left[ \sum_{i,j} w_{ij}^{(k)} h_i^{(k)} h_j^{(k+1)} \right] \tag{12}$$

with unnormalized probability of $\mathbf{v}$ then given as:

$$\tilde{P}(\mathbf{v}) = \sum_{h_i^{(k)}} \exp\left(-\mathcal{E}(\mathbf{v})\right) \tag{13}$$

For later purposes, the following unnormalized probability is stated as well:

$$\tilde{P}(\mathbf{v},\mathbf{h}^{(\ell)}) = \sum_{h_i^{(k)},0<k<\ell} \exp\left(-\mathcal{E}(\mathbf{v})|_{\mathbf{h}^{(\ell)}}\right) \tag{14}$$

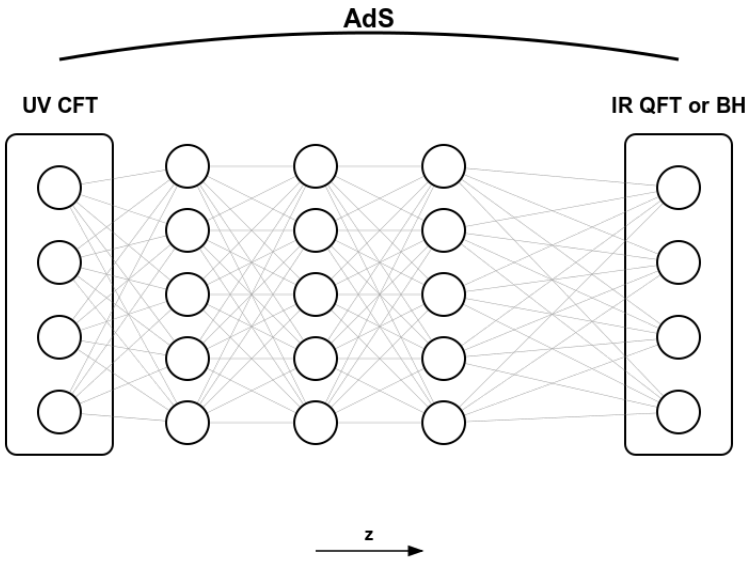


FIG. 2: Generative AdS/DL as deep Boltzmann machine (DBM). AdS space here is asymptotically AdS space. It resembles a fully connected feedforward network, but each node (neuron) output $h_i^{(k)}$ (node $0 \leq i < w$ at layer $0 \leq k \leq \ell$) is not deterministically dependent on nodes of preceding layers. Rather, DBM weights determine the probability by which some neuron value arises. In this paper, we utilize holographic renormalization further to consider the last layer to be an IR QFT, or an effective field theory (EFT) - we do not require it to be a fixed point (CFT) or to satisfy conventional boundary conditions such as black hole (BH) boundary conditions, replaced by an IR boundary theory. IR CFT can be used in place of IR QFT, nevertheless.

Note that within a single configuration, $e^{-S} = e^{-S|_{z<z'}} e^{-S|_{z>z'}} = e^{-(S|_{z<z'}+S|_{z>z'})}$, but across multiple configurations (denoted as paths), $\sum_{path} e^{-S} \neq e^{-\sum_{path} S}$. A coarse-grained action is then given as:

$$\mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\ell)}) = -\ln \tilde{P}(\mathbf{v}, \mathbf{h}^{(\ell)}) \tag{15}$$

This is the main action (or the EBM energy) that VAEBMs would seek to recover in Section III. In [2], the DBM is then trained against the probability distribution $P_{ev}(\mathbf{v})$ provided by the partition function $Z[J]$ usually with the KL divergence $D_{KL}$ as the loss function. An Einstein-Hilbert action-inspired additional regularization is used to avoid multiple local minima in [2].

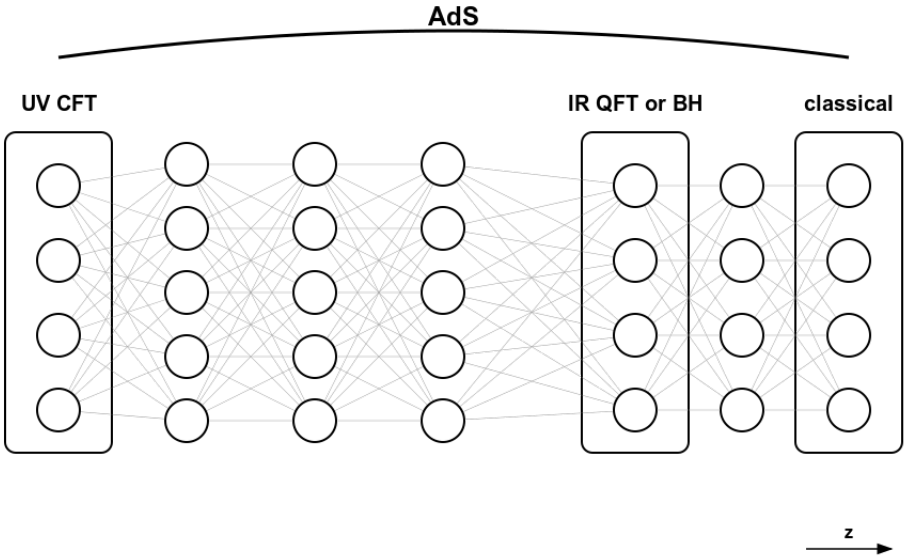


FIG. 3: An extended generative AdS/DL model, which builds on Figure 2. EBM refers to either DBM or VAEBM. The goal is to recover classical $d$-dimensional gravity as self-action (or EBM self-energy) in $d+1$-dimensional AdS space when only the 'IR QFT to classical' EBM is considered - braneworld scenario. Classical $d$-dimensional gravity is also the first-order correction without involving quantization to a classical theory. In this paper, DBMs are eventually replaced by VAEBMs. To recover the semiclassical picture in AdS/VAEBM, the EBM structure has to be carefully modified, which is carried out in Section III B.

In contrast to [2], we do not intend to rediscover AdS/CFT through neural networks; rather, our aim is to see how the bulk theory needs to be modified from the AdS bulk theory in AdS/CFT if a different IR limit is considered without modifying the UV boundary CFT. Since we seek to minimize deviations from AdS/CFT as much as possible, general but robust regularization mechanisms that penalize deviations from AdS/CFT predictions must be considered. This requires specifying AdS/CFT predictions, which are visited in Section III C.

### C. AdS/DL as holographic renormalization

It is natural to interpret AdS/DL in the context of holographic renormalization, where the UV CFT fixed point is reached with the inverse renormalization flow of $z \to 0$, and the IR CFT fixed point is reached with the renormalization flow of $z \to \infty$. Alternatively, we may introduce $z = z_\ell$ in place of $z \to \infty$, with which we understand Euclidean partition function as follows [4, 5] (note that [5] is mostly followed without content-wise modifications):

$$\begin{aligned} Z[J] &= \int \mathcal{D}\tilde{\phi} \int \mathcal{D}\phi|_{z>z_\ell} e^{-S|_{z>z_\ell}} \int \mathcal{D}\phi|_{z<z_\ell}\, e^{-S|_{z<z_\ell}} \\ &\equiv \int \mathcal{D}\tilde{\phi}\, \Psi_{IR}[\tilde{\phi}; z_\ell] \Psi_{UV}[J, \tilde{\phi}; \epsilon, z_\ell]. \end{aligned} \tag{16}$$

where $\phi(x, \epsilon) = \epsilon^{d-\Delta} J(x)$, $d+1$ being AdS spacetime dimension, $\Delta = \frac{d}{2} + \sqrt{\frac{d^2}{4} + m^2}$ and boundary conditions $\phi(x, z_\ell) = \tilde{\phi}(x)$. We can further decompose $\Psi_{IR}$ and $\Psi_{UV}$ as:

$$\Psi_{UV}[J, \tilde{\phi}; \epsilon, z_\ell] = \int \mathcal{D}\phi_\epsilon\, \Psi_{UV}[J, \phi_\epsilon, \tilde{\phi}; \epsilon, z_\ell] \tag{17}$$

$$\Psi_{UV}[J, \phi_\epsilon, \tilde{\phi}; \epsilon, z_\ell] = \int_{\phi(x,\epsilon)=\phi_\epsilon(x)}^{\phi(x,z_\ell)=\tilde{\phi}(x)} \mathcal{D}\phi\, e^{-S|_{z<z_\ell}} \tag{18}$$

$$\Psi_{IR}[\tilde{\phi}; z_\ell] = \int \mathcal{D}\phi_f\, \Psi_{IR}[\tilde{\phi}, \phi_f; z_\ell] \tag{19}$$

$$\Psi_{IR}[\tilde{\phi}, \phi_f; z_\ell] = \int_{\phi(x,z_\ell)=\tilde{\phi}(x)}^{\phi(x,z_f)=\phi_f} \mathcal{D}\phi\, e^{-S|_{z>z_\ell}} \tag{20}$$

$$\tilde{P}[J, \phi_\epsilon, \tilde{\phi}, \phi_f; \epsilon, z_\ell] = \Psi_{UV}[J, \phi_\epsilon, \tilde{\phi}; \epsilon, z_\ell] \Psi_{IR}[\tilde{\phi}, \phi_f; z_\ell] \tag{21}$$

$$Z[J] = \int \mathcal{D}\phi_\epsilon \mathcal{D}\tilde{\phi} \mathcal{D}\phi_f\, \tilde{P}[J, \phi_\epsilon, \tilde{\phi}, \phi_f; \epsilon, z_\ell] \tag{22}$$

In Section III, the UV part is interpreted in terms of the 'UV to IR' VAEBM, while the IR part is interpreted in terms of the 'IR to classical' VAEBM. Training for two VAEBMs ('UV to IR', 'IR to classical') can be done separately as long as they share the same intermediate layer (IR physics) at $z = z_\ell$ (layer $\ell$), and the final network (or theory) that sews together two VAEBMs produces the right IR and classical limits, while also being consistent with the assumed UV physics behavior. This implies that UV physics cast as a boundary theory is largely indifferent to IR physics and vice versa in AdS/DL - only the bulk metric (that is, gravity) adapts to accommodate different IR or UV physics.

In [5], this renormalization framework is pushed further in the $z_\ell \to 0$ limit to prove the equivalence of two AdS/CFT dictionaries, GKPW [14, 15] and BDHM dictionaries [16]. For the purpose of this paper, this is unnecessary.

### D. On probability in AdS/DL

In the DBM formalism of AdS/DL as in [2], it is clear that the nodes $h_i^{(\ell)}$ (nodes together, $\mathbf{h}^{(\ell)}$) in the intermediate IR layer at $z = \ell$ (with $z_\ell = z_f$ in [2]) almost always satisfy the strict inequality:

$$0 < P\left(\mathbf{h}^{(\ell)}|\mathbf{v}\right) < 1 \tag{23}$$

where $v_i$ is a node (as part of $\mathbf{v}$) in the UV 'visible' layer. This means that different IR field configurations are compatible with a single UV field configuration, which may be somewhat unexpected. (Different UV field configurations are compatible with a single IR field configuration as well and as expected.)

In other words, it is not each 'classically deterministic' field configuration that renormalizes; it is actual physics (dynamics or theory) that renormalizes. In reality, measurements are expected to only have access to bulk effective field theory (EFT) observables or 'IR theory' observables, which implies that even if a UV theory is free of probabilistic experiences, epistemic uncertainty in the IR theory can be inevitable - but this is due to observable limitations, not fundamental limitations. Indeed, some of this is already implied by [17] - ignoring gravitational connections, every measurement and observation arises probabilistically in an irreversible fashion, whereas the full Einstein picture suggests that some measurements can be reversed by following a protocol with appropriate observables.

### E. Holographic renormalization: braneworld scenario

With the holographic renormalization formalism of Section II C, we now consider the classical case for the 'IR to classical' EBM, where an observer does not notice effects of radial coordinate $z$ - this means that the IR layer field configuration is equivalent to the classical layer field configuration: $\mathbf{h}^{(\ell)} = \mathbf{h}^{(\mathbf{f})}$. In such a circumstance, coarse-grained EBM self-energy $\mathcal{E}_{coarse}(\mathbf{h}^{(\ell)} = \mathbf{h}^{(\mathbf{f})}, \mathbf{h}^{(\mathbf{f})})$ must be consistent with action $S_{cl-gr}(\mathbf{h}^{(\mathbf{f})})$ of the underlying classical $d$-dimensional Euclidean gravitational theory:

$$\mathcal{E}_{coarse}(\mathbf{h}^{(\mathbf{f})}, \mathbf{h}^{(\mathbf{f})}) \approx S_{cl-gr}(\mathbf{h}^{(\mathbf{f})}) \tag{24}$$

We may obtain $S_{cl-gr}$ by treating the $d$-dimensional non-gravitational IR quantum theory classically, coupled with $d$-dimensional general relativity or obtain it separately. Whenever

Equation (24) is satisfied, we have classical $d$-dimensional gravity embedded into a $d+1$-dimensional AdS bulk theory - for example, $d$-dimensional dS space can be embedded within $d+1$-dimensional AdS space, resembling braneworld scenarios [6, 7]. Equation (24) is therefore used as a regularization term for EBM training, resulting in the loss function of Equation (33) that balances the underlying $d+1$-dimensional AdS bulk, consistency with $d$-dimensional classical gravity and the fundamental theory being a non-gravitational IR quantum theory.

## III. ADS/VAEBM

### A. General idea

The main motivation to move toward AdS/VAEBM from AdS/DBM is to allow a flexible form for the bulk theory action. Another idea behind AdS/VAEBM is to strengthen the concept of bulk being an encoder of a UV boundary theory. Ideally, encoded IR outputs would be decoded back to the UV theory states with some decoder. Although regularization by the dual holographic AdS bulk theory assures some accuracy, it is theoretically more solid to consider the encoder-decoder (or inference-generator [3]) setup that naturally recovers holography without heavily relying on AdS/CFT.

The name 'VAEBM' is coined in [18], but the setup we utilize, for analytical simplicity, comes from [3]. The adversarial KL divergence triangle goes as the following objective function:

$$\begin{aligned}\min_{w_{enc},w_{dec}} \max_{w_{ebm}} & D_{KL}(P_{enc}(\mathbf{v},\mathbf{h}^{(\ell)})||P_{dec}(\mathbf{v},\mathbf{h}^{(\ell)})) + D_{KL}(P_{dec}(\mathbf{v},\mathbf{h}^{(\ell)})||P_{ebm}(\mathbf{v},\mathbf{h}^{(\ell)})) \\ & -D_{KL}(P_{enc}(\mathbf{v},\mathbf{h}^{(\ell)})||P_{ebm}(\mathbf{v},\mathbf{h}^{(\ell)}))\end{aligned} \tag{25}$$

with *enc* referring to encoder, *dec* referring to decoder and *ebm* refers to EBM (energy-based model). $w$ refers to the weights of the corresponding neural networks. $D_{KL}$ refers to the KL divergence. The encoder neural network models $P_{enc}(\mathbf{h}^{(\ell)}|\mathbf{v})$, the decoder neural network models $P_{dec}(\mathbf{v}|\mathbf{h}^{(\ell)})$ and the EBM neural network models $\mathcal{E}_{coarse}(\mathbf{v},\mathbf{h}^{(\ell)})$ relating to $P_{ebm}(\mathbf{v},\mathbf{h}^{(\ell)}) \propto e^{-\mathcal{E}_{coarse}(\mathbf{v},\mathbf{h}^{(\ell)})}$.

This is a chasing game: EBM seeks to be close to the encoder and move away from the decoder, but the decoder seeks to be close to the EBM and the encoder. This allows us to maintain critical balances so that the decoder is able to recover the original data minus some

noise. In contrast to [3], we do not assume that conditional probability follows Gaussian distributions, and let the encoder and the decoder determine unnormalized probability distributions from their own feedforward neural network. To sample from these distributions, MCMC is typically utilized - we do not go into this matter further, as it is conceptually unimportant.

In AdS/VAEBM, we already know the expected probability distribution $P_{ev}(\mathbf{v})$ and $P_{ev}(\mathbf{h}^{(\ell)})$. Then:

$$\begin{aligned} P_{enc}(\mathbf{v}, \mathbf{h}^{(\ell)}) &= P_{enc}(\mathbf{h}^{(\ell)}|\mathbf{v}) P_{ev}(\mathbf{v}) \\ P_{dec}(\mathbf{v}, \mathbf{h}^{(\ell)}) &= P_{dec}(\mathbf{v}|\mathbf{h}^{(\ell)}) P_{ev}(\mathbf{h}^{(\ell)}) \\ P_{enc}(\mathbf{h}^{(\ell)}) &= \sum_{\mathbf{v}} P_{enc}(\mathbf{h}^{(\ell)}|\mathbf{v}) P(\mathbf{v}) \end{aligned} \tag{26}$$

Reformulate the objective to the four-way KL divergence chasing game such that the encoder is to faithfully capture $P_{ev}(\mathbf{h}^{(\ell)})$:

$$\begin{aligned} \min_{w_{enc}, w_{dec}} \max_{w_{ebm}} \lambda_{enc,d} D_{KL}(P_{enc}(\mathbf{v}, \mathbf{h}^{(\ell)}) || P_{dec}(\mathbf{v}, \mathbf{h}^{(\ell)})) + \lambda_{dec,e} D_{KL}(P_{dec}(\mathbf{v}, \mathbf{h}^{(\ell)}) || P_{ebm}(\mathbf{v}, \mathbf{h}^{(\ell)})) \\ -\lambda_{enc,e} D_{KL}(P_{enc}(\mathbf{v}, \mathbf{h}^{(\ell)}) || P_{ebm}(\mathbf{v}, \mathbf{h}^{(\ell)})) + \lambda_{enc,h} D_{KL}(P_{enc}(\mathbf{h}^{(\ell)}) || P_{ev}(\mathbf{h}^{(\ell)})) \end{aligned} \tag{27}$$

### B. EBM neural network structure

We now constrain the structure of the feedforward EBM neural network as consisting of $\ell$ feedforward sub-networks $M_k$, with input-output structure as follows:

- $M_0$ input: $(\mathbf{v}, \mathbf{h}^{(\ell)})$
- $M_0$ EBM energy output: $(\mathbf{h}^{(\mathbf{1})}_{\mathbf{semi}}, \mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\mathbf{1})}_{\mathbf{semi}}))$
- $M_k$ input $(1 \leq k \leq \ell - 2)$: $(\mathbf{v}, \mathbf{h}^{(\mathbf{k})}_{\mathbf{semi}}, \mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\mathbf{k})}_{\mathbf{semi}}))$
- $M_k$ output $(1 \leq k \leq \ell - 2)$: $(\mathbf{h}^{(\mathbf{k+1})}_{\mathbf{semi}}, \mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\mathbf{k+1})}_{\mathbf{semi}}))$
- $M_{\ell-1}$ input: $(\mathbf{v}, \mathbf{h}^{(\ell)}, \mathbf{h}^{(\ell-\mathbf{1})}_{\mathbf{semi}}, \mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\ell-\mathbf{1})}_{\mathbf{semi}}))$
- $M_{\ell-1}$ output: $(\mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\ell)}), \mathbf{h}^{(\ell)}_{\mathbf{semi}})$
- Semiclassical mean-field approximation: $\mathbf{h}^{(\mathbf{k})}_{\mathbf{semi}} = \langle \mathbf{h}^{(\mathbf{k})} \rangle_{\mathbf{v}}$

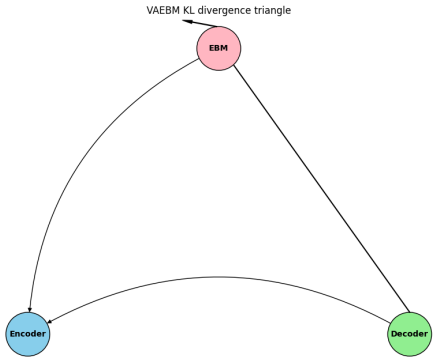


FIG. 4: The VAEBM KL divergence triangle (between neural networks in a VAEBM). EBM seeks to be closer to the encoder and attempts to move away from the decoder, while the decoder seeks to be closer to the encoder. Arrows reflect the chasing game. This triangle chasing game is appended with AdS regularizations and/or braneworld modifications. The encoder (feedforward) neural network (with weights $w_{enc}$) models conditional probability $P_{enc}(\mathbf{h}^{(\ell)}|\mathbf{v})$ and the decoder (feedforward) neural network (with weights $w_{dec}$) models conditional probability $P_{dec}(\mathbf{v}|\mathbf{h}^{(\ell)})$. Feedforward networks may be connected as residual networks [19].

The *coarse* part of $\mathcal{E}_{coarse}$ refers to the following fact:

$$\mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\mathbf{j})}) = -\ln \left( \sum_{\mathbf{h}^{(\mathbf{1} \leq \mathbf{k} \leq \mathbf{j-1})}} e^{-\mathcal{E}(\mathbf{v}, \mathbf{h}^{(\mathbf{1} \leq \mathbf{k} \leq \mathbf{j-1})}, \mathbf{h}^{(\mathbf{j})})} \right) \tag{28}$$

with EBM energy $\mathcal{E}$ for a fully-specified field configuration interpreted as bulk theory action $S$. Additional regularizations can be done with the classical AdS bulk theory dual to the UV CFT (which provides $P_{ev}(\mathbf{v})$), which we describe in Section III C.

### C. Regularization and classical bulk predictions in AdS/DL

We now discuss regularization and classical bulk predictions in AdS/DL. Some limitations are noted - we initially assume the large-N limit in order to turn the initial AdS bulk to be classical such that actions can be approximated:

$$S_{AdS,coarse}(\phi_\epsilon, \tilde{\phi}) \approx S_{AdS}(\Phi) \tag{29}$$

with $\Phi$ being a classical field configuration according to the classical bulk theory, with field configuration $\phi_\epsilon$ and $\tilde{\phi}$ within $\Phi$ at $z = \epsilon$ and $z = z_\ell$ respectively. $S_{AdS}$ refers to the AdS

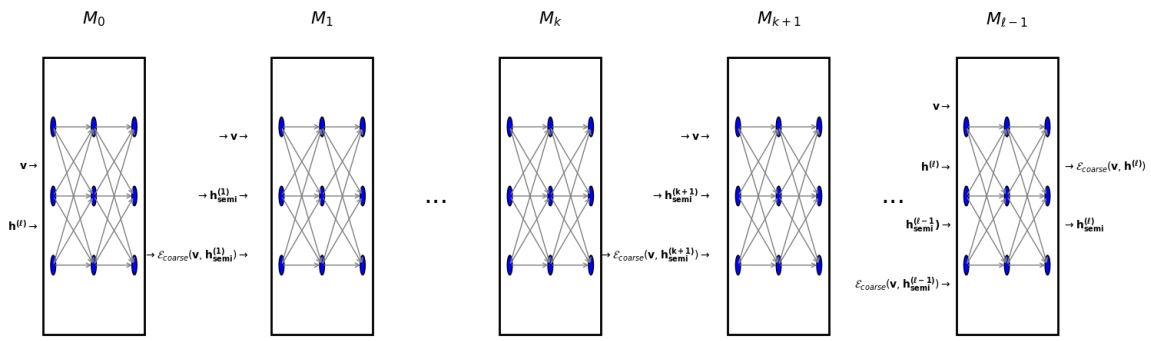


FIG. 5: The EBM neural network structure (out of VAEBM encoder/decoder/EBM neural networks), which consists of $\ell$ subnetworks. For the subnetwork structure, see Figure 6.

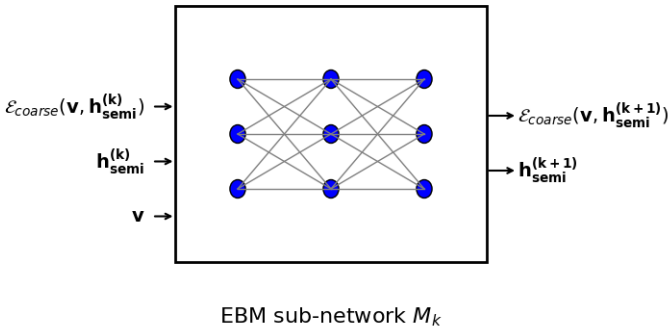


FIG. 6: The structure of subnetwork $M_k$ in the EBM neural network in Figure 5. $M_k$ is a feedforward neural network, fully connected or connected in fashion of a residual network [19]. (Resnets may be preferable for training performances - see [20].) The depth and the width of $M_k$ is unrestricted, not limited to 3.

bulk theory action (of AdS/CFT) with $\Phi$ required to be classical bulk field data. We now denote $S_{AdS,coarse}$ simply as $S_{AdS}$. Missing initial conditions relating to some observables $O$ when computing the classical bulk prediction for initial condition $\phi_\epsilon$ at $z = \epsilon$ are completed by assuming the semiclassical mean field prediction $O_\epsilon$:

$$O_\epsilon = \langle \phi_\epsilon | O | \phi_\epsilon \rangle \tag{30}$$

This allows us to compute field data at all $z$ from UV $z = \epsilon$ to IR $z = z_\ell$, as well as classical bulk action $S_{AdS}(\epsilon, z)|_{\phi_\epsilon, O_\epsilon}$, where we replaced field variables with radial coordinates. The regularization $L_{reg}$ part to the objective, which only involves the EBM neural network (with weights $w_{ebm}$), goes as:

$$L_{reg} = \frac{\lambda_{AdS}}{\ell \, n_{batch}} \sum_{batch} \sum_{k=1}^{\ell} \Big( \lambda_{\mathcal{E}} | S_{AdS}(\mathbf{v}, \mathbf{h}_{AdS}^{(k)}) - \mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}_{\mathbf{semi}}^{(\mathbf{k})}) | + \lambda_h | \mathbf{h}_{AdS}^{(k)} - \mathbf{h}_{\mathbf{semi}}^{(\mathbf{k})} | \Big) \tag{31}$$

where as aforementioned for classical action $S_{AdS}$, we use the semiclassical mean field approximation $O_\epsilon = O_\epsilon(\mathbf{v})$, noting that $S_{AdS}(\mathbf{v}, \mathbf{h}_{AdS}^{(k)}) = S_{AdS}(\mathbf{v})$ since the terminal point $\mathbf{h}_{AdS}^{(k)}$ is determined deterministically and classically within the initial AdS bulk theory. The full VAEBM training objective then is:

$$\begin{aligned} \min_{w_{enc}, w_{dec}} \max_{w_{ebm}} & \lambda_{enc,d} D_{KL}(P_{enc}(\mathbf{v}, \mathbf{h}^{(\ell)}) || P_{dec}(\mathbf{v}, \mathbf{h}^{(\ell)})) + \lambda_{dec,e} D_{KL}(P_{dec}(\mathbf{v}, \mathbf{h}^{(\ell)}) || P_{ebm}(\mathbf{v}, \mathbf{h}^{(\ell)})) \\ & + \lambda_{enc,h} D_{KL}(P_{enc}(\mathbf{h}^{(\ell)}) || P_{ev}(\mathbf{h}^{(\ell)})) - \lambda_{enc,e} D_{KL}(P_{enc}(\mathbf{v}, \mathbf{h}^{(\ell)}) || P_{ebm}(\mathbf{v}, \mathbf{h}^{(\ell)})) - L_{reg} \end{aligned} \tag{32}$$

where $L_{reg} = L_{reg}(w_{ebm})$.

### D. An additional braneworld constraint in case of the 'IR to classical' VAEBM

In the 'IR to classical' VAEBM that was discussed in Section II E, the $d$-dimensional gravitational braneworld constraint (within $d+1$-dimensional gravitational bulk theory) needs to be imposed, if following the braneworld scenario. For the 'IR to classical' VAEBM, the following notational change is made to prepare for joining two VAEBMs:

- $\mathbf{v}$ of the original VAEBM notation becomes $\mathbf{h}^{(\ell)}$ for the IR-classical VAEBM,
- $\mathbf{h}^{(\ell)}$ of the original VAEBM notation becomes $\mathbf{h}^{(\mathbf{f})}$ for the IR-classical VAEBM,
- $\mathbf{h}^{(\mathbf{k})}$ of the original notation becomes $\mathbf{h}^{(\ell+\mathbf{k})}$ for the IR-classical VAEBM.

Furthermore, for the 'IR to classical' VAEBM, we do not directly impose a particular classical theory and let the classical theory emerge. The following addition to the objective function can then be made:

$$L_{EH} = \frac{\lambda_{EH}}{n_{batch}} \sum_{batch} |S_{cl-gr}(\mathbf{h^{(f)}_{semi}}) - \mathcal{E}_{coarse}(\mathbf{h}^{(\ell)} = \mathbf{h^{(f)}_{semi}}, \mathbf{h^{(f)}_{semi}})| \tag{33}$$

where $S_{cl-gr}$ refers to the $d$-dimensional classical gravitational theory. ($EH$ refers to the Einstein-Hilbert action.) The full training objective for the 'IR to classical' VAEBM then goes as:

$$\begin{aligned} \min_{w_{enc}, w_{dec}} \max_{w_{ebm}} & \lambda_{enc,d} D_{KL}(P_{enc}(\mathbf{h}^{(\ell)}, \mathbf{h^{(f)}}) || P_{dec}(\mathbf{h}^{(\ell)}, \mathbf{h^{(f)}})) \\ & + \lambda_{dec,e} D_{KL}(P_{dec}(\mathbf{h}^{(\ell)}, \mathbf{h^{(f)}}) || P_{ebm}(\mathbf{h}^{(\ell)}, \mathbf{h^{(f)}})) \\ & - \lambda_{enc,e} D_{KL}(P_{enc}(\mathbf{h}^{(\ell)}, \mathbf{h^{(f)}}) || P_{ebm}(\mathbf{h}^{(\ell)}, \mathbf{h^{(f)}})) \\ - \frac{\lambda_{AdS}}{\ell \, n_{batch}} \sum_{batch} \sum_{k=\ell+1}^{f} & \Big( \lambda_{\mathcal{E}} |S_{AdS}(\mathbf{h}^{(\ell)}, \mathbf{h}^{(k)}_{AdS}) - \mathcal{E}_{coarse}(\mathbf{h}^{(\ell)}, \mathbf{h^{(k)}_{semi}})| + \lambda_h |\mathbf{h}^{(k)}_{AdS} - \mathbf{h^{(k)}_{semi}}| \Big) \\ & - \frac{\lambda_{EH}}{n_{batch}} \sum_{batch} |S_{cl-gr}(\mathbf{h^{(f)}_{semi}}) - \mathcal{E}_{coarse}(\mathbf{h}^{(\ell)} = \mathbf{h^{(f)}_{semi}}, \mathbf{h^{(f)}_{semi}})| \end{aligned} \tag{34}$$

### E. AdS/VAEBM as a synthesis of supervised AdS/DL and self-supervised AdS/DL

Conventional AdS/DL tends to be supervised AdS/DL with training data being input-output pairs and the bulk field theory assumed to be classical in the large-N limit [1, 13]. In contrast, AdS/DBM in [2] understands the bulk theory in a quantum way (though the bulk metric is classical), at the cost of self-supervised (or 'unsupervised') learning. AdS/DBM imposes a significant structure on the functional form of the bulk action, and this restriction is removed by transitioning to AdS/VAEBM.

Fundamentally, AdS/VAEBM remains self-supervised, as with AdS/DBM. However, its EBM 'semiclassical' regularization by the initial AdS bulk theory features supervised learning, where the semiclassical bulk theory (as arising from the full bulk theory) is made to be as close to the initial AdS bulk theory (that arises from AdS/CFT) - see Equation (31). In this case, training data input is $\mathbf{v}$, whereas training data outputs are $\mathbf{h^{(k)}_{AdS}}$ (with $1 \leq k \leq \ell$) and $S_{AdS}(\mathbf{v}, \mathbf{h^{(k)}_{AdS}})$. These training data outputs are compared to actual network outputs $\mathbf{h^{(k)}_{semi}}$ and $\mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h^{(k)}_{semi}})$, which makes AdS/VAEBM a synthesis of supervised and self-supervised AdS/DL.

The main advantage of this synthesis is that the bulk remains quantum, but the semi-classical picture is simultaneously available such that we may reconstruct the background spacetime.

### F. The full bulk theory as the UV-IR VAEBM and the IR-classical VAEBM connected by the shared IR layer $\ell$

The 'UV to IR' VAEBM and the 'IR to classical VAEBM' that were trained separately are now combined into the full bulk theory. Their intermediate outputs remain valid, and the only things that need to be appended relate to $\mathcal{E}_{coarse}$:

$$\mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\mathbf{k})}) = -\ln\left(\sum_{\mathbf{h}^{(\ell)}} \exp(-\mathcal{E}_{coarse}(\mathbf{v}, \mathbf{h}^{(\ell)}) - \mathcal{E}_{coarse}(\mathbf{h}^{(\ell)}, \mathbf{h}^{(\mathbf{k})}))\right) \quad (\text{for } k > \ell) \tag{35}$$

This completes defining the full bulk theory that deviates from the initial AdS bulk theory due to the IR limit perturbations. Converting a discretized (in radial coordinate $z$) neural network model into a neural ODE [21], we obtain a continuous AdS/VAEBM model.

### G. Brief digressions

#### *1. On expected neural network generalization error*

In the past decades, a common theoretical criticism of over-parameterized neural networks was that they are bound to suffer from generalization error eventually. The recent decade saw the reversal of this viewpoint with the rise of neural tangent kernel (NTK) analysis [22, 23]. The idea has been that infinite-width neural networks with gradient descent can be analyzed as linear models, with the networks exponentially converging to interpolants of minimum RKHS norm, which roughly translate to the ones with least complexity consistent with training data. This allows the expected generalization error to become smaller as the training data accumulate. In essence, this implies that practical applicability of AdS/DL is tied to complexity of physics involved, as over-parameterized neural networks are biased toward less complexity.

*2. On Pareto weights $\lambda$*

In Equation (32), different $\lambda$'s are used to place varying emphasis on different sub-objective functions. It is also possible to think of the optimized value of each sub-objective function as being a function of $\lambda$'s. The optimized sub-objective values can then be interpolated against the input $\lambda$'s, such as by neural networks, which then allows us to estimate the optimal choice of $\lambda$.

## IV. CONCLUSION

AdS/VAEBM generalizes AdS/DBM (AdS/DL in terms of DBM) of [2] by allowing for unrestricted functional forms of the bulk theory action. This is done without sacrificing a quantum treatment of the bulk theory (though not metric) in AdS/DBM, which has bulk fields excited probabilistically.

AdS/VAEBM follows a chasing game framework of [3], where the decoder tries to recover much of the original data from the encoded data, but EBM acts as a critic and adversary to the decoder, while EBM itself acts to be close to the encoder. This reflects the idea that there must be some encoding (and compression) ongoing along the radial coordinate $z$, but such encoding must not be detached from the original UV CFT data.

Meanwhile, AdS regularization of the bulk is done so that deviations of the bulk theory from AdS/CFT are kept minimal. This gives us benefits of both supervised AdS/DL and self-supervised AdS/DL, as described in Section III E.

The holographic renormalization framework of [4, 5] naturally provides two EBMs based on UV and IR limits. This framework is shared by both AdS/DBM and AdS/VAEBM. A braneworld scenario can then be imposed upon the 'IR to classical' EBM, where a $d$-dimensional gravitational theory lives within a $d+1$-dimensional gravitational AdS bulk theory via constraining self-EBM energy (self-action) $\mathcal{E}_{coarse}(\mathbf{h}, \mathbf{h})$ from $z = z_\ell$ (where the IR theory is defined) to $z = z_f$ (considered to be the classical regime) to approximate the given $d$-dimensional gravitational action. This self-action is then understood as the classical gravitational interpretation of the IR theory that matches (or equalizes) an IR theory state with a state in the $z = z_f$ limit.

Whether AdS/VAEBM or AdS/DBM, they share the same lesson about probability in

AdS/CFT: a UV CFT field configuration (that is, state) can correspond to multiple coarse-grained IR field configurations, probabilistically weighted. If physical observables are constrained to IR observables, then epistemic uncertainty is inevitable, even if the UV CFT does not have epistemic uncertainty.

---


[1] Koji Hashimoto, Sotaro Sugishita, Akinori Tanaka, and Akio Tomiya, "Deep learning and the AdS/CFT correspondence," Physical Review D **98** (2018), 10.1103/physrevd.98.046019.

[2] Koji Hashimoto, "AdS/CFT correspondence as a deep Boltzmann machine," Physical Review D **99** (2019), 10.1103/physrevd.99.106017.

[3] Tian Han, Erik Nijkamp, Linqi Zhou, Bo Pang, Song-Chun Zhu, and Ying Nian Wu, "Joint training of variational auto-encoder and latent energy-based model," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020) pp. 7975–7984.

[4] Idse Heemskerk and Joseph Polchinski, "Holographic and Wilsonian renormalization groups," Journal of High Energy Physics **2011** (2011), 10.1007/jhep06(2011)031.

[5] Daniel Harlow and Douglas Stanford, "Operator Dictionaries and Wave Functions in AdS/CFT and dS/CFT," (2011), arXiv:1104.2621 [hep-th].

[6] Lisa Randall and Raman Sundrum, "Large mass hierarchy from a small extra dimension," Physical Review Letters **83**, 3370–3373 (1999).

[7] Lisa Randall and Raman Sundrum, "An alternative to compactification," Physical Review Letters **83**, 4690–4693 (1999).

[8] Andreas Karch, Hao-Yu Sun, and Christoph F. Uhlemann, "Double holography in string theory," Journal of High Energy Physics **2022** (2022), 10.1007/jhep10(2022)012.

[9] Mugeon Song, Maverick S. H. Oh, Yongjun Ahn, and Keun-Young Kima, "AdS/Deep-Learning made easy: simple examples," Chinese Physics C **45**, 073111 (2021).

[10] Tetsuya Akutagawa, Koji Hashimoto, and Takayuki Sumimoto, "Deep learning and AdS/QCD," Physical Review D **102** (2020), 10.1103/physrevd.102.026020.

[11] Yu-Kun Yan, Shao-Feng Wu, Xian-Hui Ge, and Yu Tian, "Deep learning black hole metrics from shear viscosity," Phys. Rev. D **102**, 101902 (2020).

[12] Kai Li, Yi Ling, Peng Liu, and Meng-He Wu, "Learning the black hole metric from holographic conductivity," Phys. Rev. D **107**, 066021 (2023).

[13] Byoungjoon Ahn, Hyun-Sik Jeong, Keun-Young Kim, and Kwan Yun, "Holographic reconstruction of black hole spacetime: machine learning and entanglement entropy," Journal of High Energy Physics **2025** (2025), 10.1007/jhep01(2025)025.

[14] Edward Witten, "Anti-de Sitter space and holography," Adv. Theor. Math. Phys. **2**, 253–291 (1998).

[15] S.S. Gubser, I.R. Klebanov, and A.M. Polyakov, "Gauge theory correlators from non-critical string theory," Physics Letters B **428**, 105–114 (1998).

[16] Tom Banks, Michael R. Douglas, Gary T. Horowitz, and Emil Martinec, "AdS Dynamics from Conformal Field Theory," (1998), arXiv:hep-th/9808016 [hep-th].

[17] Leonard Susskind, "Copenhagen vs Everett, Teleportation, and ER=EPR," Fortschritte der Physik **64**, 551–564 (2016).

[18] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat, "VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models," in *International Conference on Learning Representations* (2021).

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) pp. 770–778.

[20] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, "Gradient descent finds global minima of deep neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 97, edited by Kamalika Chaudhuri and Ruslan Salakhutdinov (PMLR, 2019) pp. 1675–1685.

[21] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, Vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018).

[22] Arthur Jacot, Franck Gabriel, and Clément Hongler, "Neural tangent kernel: convergence and generalization in neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (Curran Associates Inc., Red Hook, NY, USA, 2018) p. 8580–8589.

[23] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington, "Wide neural networks of any depth evolve as linear

models under gradient descent," Journal of Statistical Mechanics: Theory and Experiment **2020**, 124002 (2020).