# Entangling Solid Solutions: Machine Learning of Tensor Networks for Materials Property Prediction

David E. Sommer and Scott T. Dunham

Department of Electrical Engineering, University of Washington, Seattle, WA, 98195, USA

Progress in the application of machine learning techniques to the prediction of solid-state and molecular materials properties has been greatly facilitated by the development state-of-the-art feature representations and novel deep learning architectures. A large class of atomic structure representations based on expansions of smoothed atomic densities have been shown to correspond to specific choices of basis sets in an abstract many-body Hilbert space. Concurrently, tensor network structures, conventionally the purview of quantum many-body physics and quantum information, have been successfully applied in supervised and unsupervised learning tasks in computer vision and natural language processing. In this work, we argue that architectures based on tensor networks are well-suited to machine learning on Hilbert-space representations of atomic structures. This is demonstrated on supervised learning tasks involving widely available datasets of density functional theory calculations of metal and semiconductor alloys. In particular, we show that certain standard tensor network topologies exhibit strong generalizability even on small training datasets while being parametrically efficient. We further relate this generalizability to the presence of complex entanglement in the trained tensor networks. We also discuss connections to learning with generalized structural kernels and related strategies for compressing large input feature spaces.

#### I. INTRODUCTION

Rapid advances in electronic structure methods and computational resources have enabled high-throughput ab initio calculations of solids and molecules over broad classes of chemistries and structures. Much of this work has been motivated by a pressing need to expand the tools for engineering new materials with desirable properties, with applications ranging from drug design to optoelectronics, energy storage, and quantum computing. First-principles calculations, in particular, play a crucial role in mapping the atomistic structure and composition of a material to fundamental properties such as ground state energies, potential energy surfaces, band structures and optical excitation spectra. However, a full exploration of materials design space is plagued by various well-known curses of dimensionality. These include the exponential scaling of many-body Hilbert spaces and the associated computational complexity of solving the many-body Schrödinger equation, as well as the combinatorics of decorating finite and periodic lattices by various chemical species. While density functional theory (DFT) and many-body perturbation theory (MBPT) have seen numerous successes in addressing the former, they can still present large computational barriers to address the latter.

In recognition of this challenge, the development and application of machine learning techniques to produce accurate and computationally efficient surrogate models of *ab initio* calculations has become a very active area of research [1–12]. Broadly speaking, the success of these techniques depends on the identification of a sufficiently descriptive feature space which captures the variance of the data one wishes to model. Much progress has been made recently in identifying and engineering feature spaces which accurately represent local atomic environments and can be used as inputs to standard machine

learning algorithms. Efficient representations are often achieved by constructions that respect physical symmetries, such as invariance to global rotations, translations and permutations of identical chemical species. For the class of atom-centered representations based on expansions of atomic densities, symmetry is incorporated either by directly constructing invariant polynomials, as is the case for atom-centered symmetry functions (ACSFs) [1] and moment tensor potentials (MTPs) [4], or by explicit integration over relevant symmetry groups, as exemplified by the smooth overlap of atomic positions (SOAP) [2] and spectral neighbor analysis potentials (SNAP) [3]. To a large extent, instances in this class of atom-centered representations correspond to different choices of basis sets in an abstract N-body Hilbert space [8], and the hierarchy of N-body features has recently been shown to be organized in the so-called atomic cluster expansion (ACE) framework [7, 13–15].

It is worth mentioning that an alternative to extensive feature engineering is to employ a so-called end-to-end approach, in which inputs to the model are minimally processed, and the relevant feature space is learned by the model architecture during the course of training. Examples of this approach in materials informatics include graph neural network methods [16–19] and explicitly symmetry-equivariant architectures [20, 21]. However, in the absence of prior feature engineering, training accurate deep learning models becomes difficult when the amount of data is limited.

Thus, from a practical standpoint, balancing the bias and complexity of a model between the two extremes of extensive feature engineering and highly flexible model architectures is often influenced by the availability of training data. Strictly end-to-end deep learning can require substantially more data when correlations between features are complex, while naive application of parametrically efficient features can lead to a model which

generalizes poorly on new inputs. Of course, the strict distinction between these two extremes can be somewhat arbitrary: the latent space encoded by the hidden layers of a deep learning architecture can be thought of as a kind of renormalized feature space [22, 23], subject to its own forms of bias through choices of hyperparameters, regularization schemes and architectural topology. To that end, certain deep model architectures may possess better so-called *inductive bias* [24] over a given dataset, more efficiently prioritizing the search space of solutions when their structure reflects the pattern of correlations in the feature space.

Interesting connections between inductive bias and feature correlations have grown out of recent efforts to introduce techniques from the study of quantum entanglement and strongly correlated quantum many-body systems to the field of machine learning. Specifically, tensor network methods have been successfully applied to supervised and unsupervised learning tasks in computer vision and natural language processing [25–34] and have provided a basis for analyzing the expressiveness of common deep learning architectures based on their entanglement properties [35, 36]. Indeed, a general argument for the effectiveness of tensor networks in machine learning contexts is that the pattern of entanglement encoded in the network can efficiently represent the pattern of correlation in the feature space of the data.

In this work, we argue for the applicability of tensor networks in machine learning structure-property models of materials. This problem is addressed from two directions: we show how both the construction of an input feature space and of an associated machine learning architecture can be formulated in the language of tensor networks. In Section II, a large class of atomic structure representations, corresponding to the SO(3)invariant tensor basis set of the (smoothed) atomic cluster expansion (ACE), are shown to admit a natural tensor network description. An immediate consequence of this rewriting is that the equivalence classes and the recursive construction of the hierarchy of N-body ACE basis tensors become transparent in the graphical notation of the tensor network. In Section III, we show how common tensor network factorizations for the weights of machine learning models can be naturally built on top of individual ACE basis tensors. We discuss the relationship between this approach and kernel learning, and in particular how certain tensor networks can realize a form of alchemical learning by coupling information between the local atomic structure and the chemical elements within it. We also comment on how the introduction of copying and merging operations in tensor network structures can be used to introduce higher-order correlations from fixed N-body features, akin to the construction of nonlinear kernels.

In Section IV, we benchmark some of the proposals from Sections II and III using widely available datasets of density functional theory calculations of metal and semi-conductor alloys. For concreteness, we use input fea-

tures corresponding to the commonly used SOAP power spectrum and study the learning performance of different tensor network factorizations of the model weights on training sets of various sizes. Compared to standard kernel learning methods and fully-connected neural networks, we find that models based on matrix product states (MPS) and matrix product operators (MPO) show strong generalizability and parametric efficiency, with notable performance on small training sets. We subsequently study the entanglement properties of the trained networks and find signatures of high entanglement complexity consistent with the models' strong generalizability. We also provide some evidence that the latent spaces learned by the hidden layers of the networks are able to capture physically relevant structural and chemical information, and we utilize this insight to effectively compress the input basis tensors.

# II. SO(3)-INVARIANT ATOMIC REPRESENTATIONS

In this work, we are concerned with the prediction of some atomic property V given a configuration  $\{\mathbf{r}_i\}$  of N atomic species. For simplicity, we will restrict to the case where V is a scalar, such as the total energy or band gap of the system, but extensions to vectorial and tensorial properties are discussed in Appendix C. Following the discussions by Drautz [7] and Lysogorskiy et al [15], a natural starting point is to formulate a coarse-grained model given by a sum of local terms,  $V = V^{(0)} + \sum_i V_i$ , where the local property  $V_i$  associated with atom i can be approximated by expanding in terms of local n-body interactions  $V^{(n)}$ :

$$V_i = V^{(1)}(\mathbf{r}_i) + \frac{1}{2} \sum_j V^{(2)}(\mathbf{r}_i, \mathbf{r}_j)$$
$$+ \frac{1}{3!} \sum_{jk} V^{(3)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \cdots$$
(1)

Requiring consistency with fundamental symmetries introduces constraints on the form of the interactions. For instance, translation invariance leads to atom-centered functions of the form  $V^{(\nu+1)}(\mathbf{r}_{j_1i},\ldots,\mathbf{r}_{j_\nu i})$ , where  $\mathbf{r}_{ji}=\mathbf{r}_j-\mathbf{r}_i$  can be thought of as an effective bond from the central site i to a site j in its relative environment. Interaction terms can be further decomposed by projecting onto an appropriate  $\nu$ -order basis set  $\Phi_{s_1,s_2,\ldots,s_\nu}$ ,

$$V^{(\nu+1)}(\mathbf{r}_{j_1i},\dots,\mathbf{r}_{j_\nu i}) = \sum_{\{s\}} J_{s_1,s_2,\dots,s_\nu} \Phi_{s_1,s_2,\dots,s_\nu}(\mathbf{r}_{j_1i},\dots,\mathbf{r}_{j_\nu i}) ,$$
(2)

with general interaction coefficients  $J_{s_1,s_2,...,s_{\nu}}$ . A crucial insight from the recent development of atom-centered descriptors [2, 7, 8, 14] is that an efficient representation of

atomic environments can be obtained by a low-rank approximation of the  $\Phi_{s_1,s_2,...,s_{\nu}}$  cluster basis in terms of a single-bond basis  $\phi_{s_k}(\mathbf{r}_{j_k i})$ ,

$$\Phi_{s_1, s_2, \dots, s_{\nu}}(\mathbf{r}_{j_1 i}, \dots, \mathbf{r}_{j_{\nu} i}) = \prod_{k=1}^{\nu} \phi_{s_k}(\mathbf{r}_{j_k i}) , \qquad (3)$$

combined with a reorganization of the summations in (1) and (2),

$$\sum_{j_1\cdots j_{\nu}} \Phi_{s_1,s_2,\dots,s_{\nu}}(\mathbf{r}_{j_1i},\dots,\mathbf{r}_{j_{\nu}i}) = \prod_{k=1}^{\nu} A_{i,s_k} . \tag{4}$$

We note that the unrestricted summation in (4) contains products over identical bonds. In the ACE framework [7], these self-interaction terms can be canceled by lower order terms in the expansion (1).

Formally speaking, the atom-centered descriptor,  $A_{i,s} := \sum_j \phi_s(\mathbf{r}_{ji})$ , can be understood as the expansion coefficients of the local atomic density  $\sigma_i$  in an abstract atom-centered Hilbert space  $|s\rangle \in \mathcal{V}$ ,

$$|\sigma_i\rangle = \sum_s A_{i,s}|s\rangle$$
 (5)

In practice, the real-space atomic density for a given atomic environment is approximated by a superposition of localized functions h,

$$\langle \alpha \mathbf{r} | \sigma_i \rangle = \sum_{j \neq i} \delta_{\alpha \alpha_j} f_c(r_{ji}) h(\mathbf{r} - \mathbf{r}_{ji}) ,$$
 (6)

where species  $\alpha_j$  occupies site j and  $f_c(r_{ji})$  is a smooth cutoff function that restricts the sum to some local environment of atom i. In the original ACE formalism [7], h is chosen to be a delta function, whereas in the SOAP and SNAP formalisms [2], h is chosen to be a Gaussian.

The atom-centered basis (5) is not generally invariant under action of the rotation group  $\mathcal{G} = \mathrm{SO}(3)$ . Rather, according to Maschke's theorem, the atom-centered Hilbert space  $\mathcal{V}$  decomposes into a direct sum of irreducible representations ("irreps") of  $\mathrm{SO}(3)$ ,  $\mathcal{V} \cong \bigoplus_l \mathcal{D}_l \otimes \mathcal{V}_l$ , where  $\mathcal{V}_l$  is the subspace of the irrep  $l = 0, 1, 2, \ldots$ , and the degeneracy space  $\mathcal{D}_l$  contains additional degrees of freedom (e.g., atomic species  $\alpha$  and purely radial components n) which are untouched by the group action. The generic indices, s, of the single-bond basis functions can thus be replaced by the set  $(\alpha nlm)$ , where  $-l \leq m \leq l$  labels components of the irrep subspace  $\mathcal{V}_l$ , and the angular momentum channels of the real-space single-bond basis correspond to spherical harmonics  $Y_l^m(\hat{\mathbf{r}})$ ,

$$\langle \mathbf{r} | \alpha n l m \rangle = \phi_{\alpha n l m}(\mathbf{r}) | \alpha \rangle = R_{n l}(r) Y_l^m(\hat{\mathbf{r}}) | \alpha \rangle .$$
 (7)

Common choices for radial basis functions  $R_{nl}(r)$  are spherical Gaussian-type orbitals and orthogonal polynomials [2, 7].

The expansion of the local atomic density centered on site i can thus be written as

$$|\sigma_i\rangle = \sum_{\alpha nlm} A_{i,\alpha nlm} |\alpha nlm\rangle ,$$
 (8)

where the atom-centered descriptors are given by

$$A_{i,\alpha nlm} = \langle \alpha nlm | \sigma_i \rangle$$

$$= \int_{\mathbb{T}^2} d\Omega \ R_{nl}(r) Y_l^m(\hat{\mathbf{r}}) \langle \alpha \mathbf{r} | \sigma_i \rangle \ .$$
(9)

Furthermore, the reorganized  $(\nu+1)$ -body expansion (4) is equivalent to a product state formed by repeated copies of the density  $|\sigma_i^{\otimes \nu}\rangle := \bigotimes^{\nu} |\sigma_i\rangle$ ,

$$\prod_{k=1}^{\nu} A_{i,\alpha_k n_k l_k m_k} = \langle \alpha_1 n_1 l_1 m_1 \cdots \alpha_{\nu} n_{\nu} l_{\nu} m_{\nu} | \sigma_i^{\otimes \nu} \rangle . \quad (11)$$

The expressions (8) and (11) suggest that each term in the  $\nu$ -order series expansion admits a natural description in terms of a tensor network (TN) for the states  $|\sigma_i^{\otimes \nu}\rangle$ . In the following, we will consider an associated graphical calculus which incorporates the SO(3) representations carried by the tensors,  $A_{i,\alpha_k n_k l_k m_k}$ , which will be useful in explicitly constructing SO(3)-invariant descriptors. In particular, the formalism presented below closely follows recent treatments of symmetric tensor networks [37–40] and the earlier, related development of spin networks [41, 42] from angular momentum recoupling theory [43].

#### A. Graphical calculus for atomic descriptors

Working with explicit expressions for tensor product spaces can quickly become cumbersome as the order  $\nu$ increases, and this is particularly true when the tensor space possesses a complex internal structure, as is the case for working with SO(3) symmetry. However, much of the algebraic structure can be encapsulated in a consistent graphical notation, which we utilize in the following. In the standard diagrammatic notation of tensor networks (Appendix A), the basic atom-centered descriptor  $A_{\alpha nlm}$ , as a 4-index tensor, can be represented by a shape with 4 open edges (Figure 1). Note that an arrow is added to the edge carrying the irrep space,  $\mathcal{V}_l$ , of the symmetry group to distinguish it from its dual vector space,  $\mathcal{V}_{l}^{*}$ , and we will subsequently drop the atom site index i for simplicity. Regular representations of SO(3)are given by the unitary Wigner D-matrices,  $D_{m'm}^{(l)}(g)$ , and the action of  $D_{m'm}^{(l)}(g)$  on the atom-centered basis is

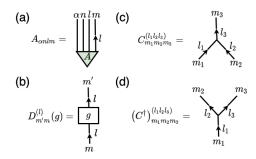


FIG. 1: Basic graphical components for the construction of invariant descriptors, including (a) atom-centered descriptors (10), (b) Wigner *D*-matrices, and (c,d) Clebsch-Gordan coefficients as fusion and splitting nodes.

given by contraction with A along the irrep edge, i.e., by  $\sum_{m} D_{m'm}^{(l)}(g) A_{\alpha n l m} \text{ (cf. Appendix A)}.$  A  $\nu\text{-order product state of multiple $A$-tensors (11)}$ 

A  $\nu$ -order product state of multiple A-tensors (11) is depicted simply by a row of disconnected A's. This  $\nu$ -order product state is not invariant under three-dimensional rotations,  $U_g^{\otimes \nu}|\sigma_i^{\otimes \nu}\rangle \neq |\sigma_i^{\otimes \nu}\rangle$ . However, it can be made so by explicitly taking the Haar integral over  $g \in SO(3)$ ,  $|\sigma_i^{\otimes \nu}\rangle_g := \int_g U_g^{\otimes \nu}|\sigma_i^{\otimes \nu}\rangle$ , since the homomorphism  $U_{g_1}U_{g_2} = U_{g_1g_2}$  extends to the tensor product space. In tensor components, this symmetrized descriptor is given by

$$\langle \alpha_1 n_1 l_1 m_1 \cdots \alpha_{\nu} n_{\nu} l_{\nu} m_{\nu} | \sigma_i^{\otimes \nu} \rangle_g$$

$$= \int_{g \in SO(3)} dg \prod_{k=1}^{\nu} \langle \alpha_k n_k l_k m_k | U_g | \sigma_i \rangle , \quad (12)$$

where, upon insertion of resolutions of identity, one obtains a tensor product of Wigner D-matrices,  $D_{mm'}^{(l)}(g) =$  $\langle lm|U_q|lm'\rangle$ , acting on the A-tensors. An explicit formula for (12) can be computed by introducing the intertwiners of the symmetry group, the Clebsch-Gordan (CG) coefficients, into the basic building blocks of the graphical calculus (Figure 1). We outline this derivation in Figure 2 for the case  $\nu = 3$ , but it can be extended by induction to all  $\nu$ . In particular, repeated application of the CG identities and their equivariance (Figure 20) reduces (12) to a single Wigner D-matrix contracted with recursively constructed fusion trees of CG coefficients. A single, invariant irrep edge must transform as the trivial representation, hence the Haar integral projects the final D-matrix-decorated edge onto the trivial irrep space. We distinguish an edge carrying the trivial representation (l = 0) by a dashed line. Because the trivial representation is one-dimensional (m = 0) [44], symmetrization vields two disconnected diagrams, which we label Q and B (see below).

A few comments are in order. First, the pattern of contractions encapsulated by the structure of the fusion

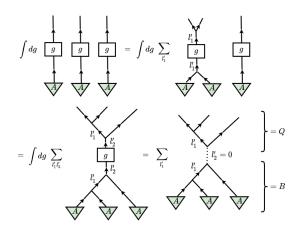


FIG. 2: The explicit symmetrization (12) of a  $(\nu + 1)$ -body product state (11) derived from the algebraic rules of SO(3)-recoupling theory. For simplicity, the edges  $(\alpha nl)$  are not shown.

trees in Figure 2 represents an implicit choice of recoupling scheme. Different choices of recoupling scheme are possible and are related to each other by unitary transformations, i.e., the so-called F-symbols (cf. Appendix B). Second, symmetrization of  $\nu$ -order product states for  $\nu > 2$  introduces summations over intermediate angular momenta l', which we represent explicitly, while contractions of irrep edges imply summation over the corresponding magnetic numbers m'. Finally, this direct symmetrization yields the so-called Jucys-Levinson-Vanagas (JLV) theorems [45], also referred to as the generalized Wigner-Eckhart theorem [37–39, 46]. Accordingly, a symmetric, n-index tensor T decomposes into a tensor product,  $T = B \otimes Q$ , of a structural tensor Q determined completely by the symmetry group and of a degeneracy tensor B. The structural tensor Q, given by a fusion tree of CG coefficients, is known in other contexts as a spin network [41, 42]. In the spirit of the original Wigner-Eckhart theorem, the degeneracy tensor B is equivalent to the "reduced matrix element" of the decomposition and encapsulates the degrees of freedom not fixed by the symmetry group [46]. This is shown explicitly in Figure 4, where the  $\{\alpha_k n_k l_k\}$  edges remain

The B-tensors constitute the set of SO(3)-invariant descriptors which form the basis of the ACE framework. In the recoupling scheme chosen in Figure 2, the  $\nu$ -order series of invariant tensors is given by the contraction of  $\nu$  atom-centered tensors A with the appropriate fusion tree (Figure 3), and invariance is thus manifested by mapping the tensor product of  $\nu$  irreps to the trivial representation. It is readily apparent that invariant descriptors with trivial intermediate irreps factorize, e.g.,  $B_{l'_1=0}^{(4)}=B^{(2)}\otimes B^{(2)}$ . Moreover, this construction can be generalized [13] to SO(3)-equivariant descriptors  $B_L^{(\nu)}$  by allowing the tensor product of  $\nu$  irreps to fuse to a non-trivial representation L, where the open L edge of

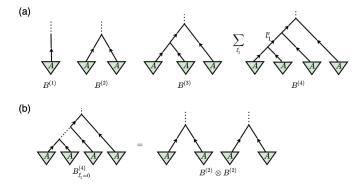


FIG. 3: (a) The ν-order hierarchy of SO(3)-invariant descriptors in terms of recursively constructed fusion trees. (b) Higher order tensors with trivial intermediate irreps factorize into products of lower order tensors.

the fusion tree carries the dimension of the irrep, 2L+1. Higher-order SO(3)-equivariant descriptors,  $B_L^{(\nu+1)}$ , can be built recursively from lower orders  $B_{L'}^{(\nu)}\otimes B_{L''}^{(1)}$  by contracting with the so-called cup and cap tensors (normalized 2jm symbols), which diagrammatically enable the orientation of irrep edges to be reversed. This recursive construction is discussed in more detail in Appendix C. In the following sections, we will represent the invariant ACE basis tensors  $B^{(\nu)}$ , as in Figure 4, by suppressing the fusion trees.

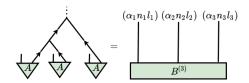


FIG. 4: Simplified representation of an invariant descriptor with fixed SO(3)-recoupling scheme.

We give explicit formulae [7, 13] for the first few invariant descriptors, while higher order  $B^{(\nu)}$  can be constructed via the recursive procedure outlined in Appendix C or simply read off of the corresponding fusion tree:

$$B_{\alpha n}^{(1)} = A_{\alpha n00} \tag{13}$$

$$B_{\alpha_1 \alpha_2}^{(2)} = \sum_{m=-l}^{l} \frac{(-1)^{l-m}}{\sqrt{d_l}} A_{\alpha_1 n_1 l m} A_{\alpha_2 n_2 l - m} \quad (14)$$

$$(B^{(3)})_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}^{\alpha_1 \alpha_2 \alpha_3} = \sum_{m_1 m_2 m_3} \frac{(-1)^{l_3 - m_3}}{\sqrt{d_{l_3}}} C_{m_1 m_2 - m_3}^{(l_1 l_2 l_3)} \times A_{\alpha_1 n_1 l_1 m_1} A_{\alpha_2 n_2 l_2 m_2} A_{\alpha_3 n_3 l_3 m_3}$$
(15)

Here,  $d_l = 2l + 1$  is the dimension of the irrep, and the angular momentum coupling in (15) can be rewritten in terms of the Wigner 3jm symbol, as in [7]. As expected,

the  $\nu = 1$  term loses all angular information after integrating over the SO(3) rotation group. Hence, accurate expansions of atom-centered properties typically require higher-order terms. In the  $\nu = 2, 3$  terms, one finds the SOAP power spectrum and bispectrum, respectively [2, 7]. In a real-space basis, these  $\nu$ -order invariants can be understood as spherical moments of the bond distribution centered on atom i. For  $\nu = 1$ , this corresponds to spherically averaging over a single bond  $\mathbf{r}_{j_1i}$ , which yields a completely isotropic function depending only on the bond length. The power spectrum,  $\nu = 2$ , measures the correlation between pairs of bonds  $(\mathbf{r}_{j_1i}, \mathbf{r}_{j_2i})$ , where spherical averaging over the local environment results in a function depending on the bond lengths  $(r_{j_1i}, r_{j_2i})$  and the relative angle,  $\hat{\mathbf{r}}_{j_1i} \cdot \hat{\mathbf{r}}_{j_2i}$ , between bond vectors. Similarly, a real-space projection of the bispectrum,  $\nu = 3$ , depends only on three bond lengths  $(r_{j_1i}, r_{j_2i}, r_{j_3i})$  and the relative angles between each pair of bond vectors.

# III. TENSOR NETWORK LEARNING

The SO(3)-invariant, atom-centered descriptors  $B_i^{(\nu)}$  can be used as input features for a variety of machine-learning methods. In the ACE formalism, this amounts to recasting the expansion (1) in the symmetrized basis,

$$V_{i} = \sum_{\nu} w_{i}^{(\nu)} \cdot B_{i}^{(\nu)}$$

$$= \sum_{\nu} \sum_{\{\alpha_{k} n_{k} l_{k}\}} w_{i\{\alpha_{k} n_{k} l_{k}\}}^{(\nu)} B_{i\{\alpha_{k} n_{k} l_{k}\}}^{(\nu)} , \qquad (16)$$

where the model weights  $w_i^{(\nu)}$  are fully contracted with the free edges,  $\{\alpha_k n_k l_k\}$ , of the invariant descriptors. As dense tensors, the size of the weights  $w_i^{(\nu)}$  scale as  $\mathcal{O}(|\alpha|^{\nu}|n|^{\nu}|l|^{\nu})$  with the number of chemical species  $|\alpha|$  and the number of radial |n| and angular momentum |l| channels included in the input descriptors.

FIG. 5:  $(\nu = 2)$ -order term in the ACE expansion as an inner product between the descriptor state,  $|B_i^{(\nu)}\rangle$ , and a learnable state,  $|\psi_i^{(\nu)}(w)\rangle$ .

It is clear from (16) that each  $\nu$ -order term can be understood as an inner product,  $\langle \psi_i^{(\nu)}(w)|B_i^{(\nu)}\rangle$ , between the invariant basis states  $|B_i^{(\nu)}\rangle$  (the reduced part of the symmetrized state  $|\sigma_i^{\otimes \nu}\rangle_g$ ) and a state  $|\psi_i^{(\nu)}(w)\rangle$  which depends on a set of learnable parameters w (Figure 5). We can generalize this construction by considering a tensor network ansatz for  $|\psi_i^{(\nu)}(w)\rangle$ , where the topology of the network encodes an entanglement structure in the

state. Taking inspiration from the application of tensor networks in quantum physics as low-rank approximations of many-body ground states, tensor network factorizations of the model weights will be used to constrain the correlation structure between the tensor elements  $\{\alpha_k n_k l_k\}$  of the input descriptor. We will find that this acts as an implicit method to regularize the model fitting [25].

Since there is no inherent geometric relationship between the positions of the descriptor indices, we consider factorizations which preserve their order. To maintain generality, we will primarily examine factorizations built from matrix product states (MPS),

$$|\psi^{(N)}\rangle = \sum_{s_1,\dots,s_N} a_1^{s_1} \cdots a_N^{s_N} |s_1 \cdots s_N\rangle , \qquad (17)$$

and matrix product operators (MPO),

$$\hat{T}^{(N)} = \sum_{\substack{s_1, \dots, s_N \\ s'_1, \dots, s'_N}} M_1^{s'_1 s_1} \cdots M_N^{s'_N s_N} |s'_1 \cdots s'_N\rangle \langle s_1 \cdots s_N |$$
(18)

where  $a_k^{s_k}$  and  $M_k^{s_k's_k}$  are  $\chi_k \times \chi_{k+1}$  matrices, which constitute the learnable model parameters. Note that because the input descriptors and target properties are real-valued, the tensors  $a_k$  and  $M_k$  will also be real-valued. The internal bond dimensions  $\chi_k$  are often referred to as virtual dimensions, and the physical indices  $s_k$  denote either a descriptor index  $\{\alpha_k, n_k, l_k\}$  or a generic, internal vertical bond. MPS and MPO tensor networks are shown in Figure 6 for the case of finite virtual boundary conditions,  $\chi_1 = \chi_N = 1$ . For simplicity, we will set all virtual dimensions to be equal,  $\chi_k = \chi$ , except for the boundary edges. As discussed further in Section IV, the virtual dimensions impose upper bounds on the bipartite entanglement entropies of the state  $|\psi^{(\nu)}\rangle$ .

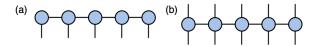


FIG. 6: Tensor network representations of (a) a matrix product state (MPS) and (b) a matrix product operator (MPO).

It is useful to note that the sequential application of several MPOs to an arbitrary product state  $|0\rangle^{\otimes N}$  is equivalent to evolving the state by a finite-depth, variational quantum circuit. If each MPO possesses bond dimension  $\chi$ , then contraction of  $n_d$  MPOs with an arbitrary product state yields an MPS with bond dimension  $\chi^{n_d}$  (Figure 7a). This can be an efficient way to achieve an MPS with large effective bond dimensions while mitigating the growth in the number of parameters. For the cases studied in this work, the bond dimensions required to achieve accurate models remain relatively small, so we

take as our starting ansatz a single MPS whose bond dimensions are treated as hyperparameters. Furthermore, since the MPS tensor elements are real-valued, we optimize them directly, rather than relying on unitary embeddings, using classical (e.g., stochastic gradient descent (SGD)) rather than quantum algorithms. We will return to this tensor network / quantum circuit correspondence in our discussion in Section IV. To make use of high-performance automatic differentiation and back propagation algorithms commonly employed in deep learning, the bond dimensions remain fixed during training. However, performing local tensor updates based on the density matrix renormalization group (DMRG) [26, 47], which adaptively evolve the bond dimensions, may provide an interesting alternative training method.

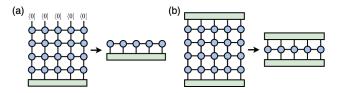


FIG. 7: The (a) MPS and (b) MPO models used in this work viewed as the contraction of several MPOs. In particular, the upper layer on the LHS of the MPS model (a) defines an embedding of an MPS as an MPO. Learnable model weights are shown in blue, while input descriptors are shown in green.

While we have focused so far on the construction of learnable states  $|\psi^{(\nu)}\rangle$ , we can alternatively formulate the task of learning scalar contributions to the target property  $V_i$  in terms of an expectation value of an MPO,  $\langle B_i^{(\nu)}|\hat{T}^{(3\nu)}|B_i^{(\nu)}\rangle$ , with respect to the invariant basis states  $|B_i^{(\nu)}\rangle$ . This can be thought of as a relaxed version of the Born rule,  $p(B) = |\langle B|\psi\rangle|^2$ , where the pure state density matrix,  $|\psi\rangle\langle\psi|$ , is replaced by a general mixed state, described by a matrix product density operator (MPDO),  $\hat{\rho}$ . Indeed, it was shown in [48] within the context of probabilistic graphical models that locally-purified approximations of MPDOs can represent a larger class of non-negative tensors than an MPS model of equivalent rank. Because we are concerned with the prediction of some target scalar which is not necessarily a probability p(B), we do not enforce non-negativity on the elements of the MPO,  $\hat{T}$ . Nonetheless, we will provide empirical evidence in Section IV that a similar rank-efficiency relationship may hold for regression tasks when measured against their performance on unseen data. Again, while we could in general train a sequence of MPOs, akin to mapping an invariant basis state back to itself [49], we will instead consider a single MPO with adjustable bond dimensions (Figure 7b).

As the order  $\nu$  and the dimensions of the individual indices of the dense descriptor tensor increase, its contraction with the variational states  $|\psi^{(\nu)}\rangle$  and operators  $\hat{T}^{(3\nu)}$  can become computationally demanding. However,

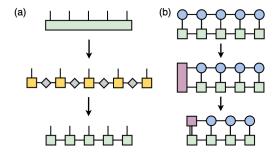


FIG. 8: (a) A MPS factorization of a dense tensor can be constructed by the sequential application of singular value decompositions between groups of edges. The grey diamonds are matrices of singular values which are subsequently contracted with their neighboring tensors.

(b) An efficient contraction order for an MPS model with factorized inputs, where the current contracted tensor is shown in purple.

this issue can be alleviated by analogously constructing low-rank, MPS approximations of the states  $|B_i^{(\nu)}\rangle$ . A standard procedure [47, 50-53], shown in Figure 8a, is to recursively take singular value decompositions (SVDs) between neighboring pairs of edges, discard a subset of the lowest singular values (SV), and contract the corresponding SV matrices with their neighboring tensors. The virtual bond dimensions of the subsequent MPS is equal to the number of singular values retained at each edge (i.e., the Schmidt rank), which controls the accuracy of the approximation. For small enough bond dimensions, this constitutes a kind of sparsification procedure on the inputs, and we find in practice (cf., Section IVD) that model accuracy can be maintained with surprisingly small bonds. With this MPS decomposition, a more efficient contraction scheme (e.g., Figure 8b), can be implemented.

# A. Relationship to other methods

In practice, a balance must be sought between the computational efficiency of the model and the number of input features required to accurately represent the atomic environment. A common simplification is to restrict the expansion (16) to many-body descriptors of relatively low order. This is, for instance, the strategy employed in Behler-Parrinello (BP) neural network potentials [1], which use 1-, 2- and 3-body symmetry functions. It is often the case that features derived from order  $\nu \leq 3$  expansions are able to differentiate the relevant structural characteristics in a given sample, although certain counterexamples exist [54]. Furthermore, as noted in Section II, higher-order descriptors contain products of lower order, which enables accurate models to be built on a fixedorder descriptor [14, 55]. This is the case for kernel-based models utilizing either the SOAP power spectrum ( $\nu = 2$ ) or bispectrum ( $\nu = 3$ ).

While the expansion (16) motivates the systematic introduction of SO(3)-invariant descriptors, one is not limited to a linear model for the prediction of some atomic property  $V_i$ . Indeed, kernel methods based on SOAP features and neural networks using ACSFs incorporate general forms of nonlinearity, whether through the choice of covariance kernel  $K(\mathbf{x}, \mathbf{x}')$  in the former, or the structure of the learning architecture in the later. It is worth dwelling on some essential aspects of kernel-based and neural network-based methods, as they will serve as further motivation for the tensor network methods introduced above.

In supervised learning, where the goal is to find a function  $f(\mathbf{x})$  which approximates a target quantity  $y \approx f(\mathbf{x})$ , a nonlinear function on the inputs  $x_i$  can be constructed via a mapping  $\varphi$  into a higher-dimensional feature space. Such a mapping allows the function f to be formulated as a linear model on the feature space, but the computational complexity of working directly with very large feature vectors  $\varphi(\mathbf{x})$  often prohibits their explicit implementation in machine learning tasks. However, since linear models constructed in a feature space can be rewritten in terms of an inner product  $\langle \varphi(\mathbf{x}) | \varphi(\mathbf{x}') \rangle$  on that space, Mercer's theorem allows one to replace the explicit feature map with a positive, semidefinite kernel function  $K(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}) | \varphi(\mathbf{x}') \rangle$ . This so-called kernel trick and the representer theorem leads to a model of the form  $f(\mathbf{x}) = \sum_a w_a K(\mathbf{x}, \mathbf{x}_a)$ , where the sum runs over a reference set of training instances. Note that the cost to evaluate the model scales linearly with the size of the training set, and thus smaller training sets may generalize poorly to out-of-sample data. Furthermore, the cost for training the model scales like  $\mathcal{O}(N^3)$  for N training instances.

An advantage of the expansion (8) of the local density in the basis of SO(3) irreps is that it comes equipped with a natural inner product, which induces an inner product on the SO(3)-invariant tensor product states  $|\sigma_i^{\otimes \bar{\nu}}\rangle_g$ . Under this induced inner product, the feature space can be identified with the descriptor space. As a tensor network, the associated kernel function is given by contracting the open edges of the SO(3)-invariant descriptors  $B_i^{(\nu)}$  and  $B_j^{\prime(\nu)}$  (e.g., Figure 10). Note that in kernel-based methods, pairs of local environments in the kernel function come from different structures, which we distinguish using a prime on the descriptors. The (smoothed) overlap  $k\left(B_i^{(\nu)}, {B'}_j^{(\nu)}\right) \coloneqq \langle B_i^{(\nu)}|{B'}_j^{(\nu)}\rangle$  quantifies the similarity between atomic environments, and raising this kernel to an integer power  $\zeta$  enhances its sensitivity to the differences between those environments. Up to a normalization of the descriptors, these overlaps yield the class of SOAP kernels upon summing over all sites. Introducing a power  $\zeta$  to the overlaps is equivalent to taking a  $\zeta$ -order tensor product of the invariant basis state,  $|B_i^{(\nu)}\rangle^{\otimes\zeta}$ , which increases the effective many-body character of the state [8]. We will return to this idea when discussing generalizations of our method below. Note

that these  $(\nu \cdot \zeta)$ -order states do not span the entirety of their SO(3)-invariant tensor product space but instead correspond to states with trivial intermediate irreps (cf. Figure 3).

A dense (fully-connected, feed-forward) neural network (NN) model approximates the predictor function  $f(\mathbf{x}) = \mathcal{FL} \cdots \mathcal{FL}(\mathbf{x})$  by an alternating composition of affine maps  $\mathcal{L}(\mathbf{x}) = w\mathbf{x} + \mathbf{b}$  and nonlinear activation functions  $\mathcal{F}(\cdot)$  applied element-wise to their input vectors. A dense neural network possesses a relatively simple tensor network representation (Figure 9), in which the linear weight matrices w are shown as 2-index tensors while the biases **b** and activation functions remain implicit. In the case where the SO(3)-invariant descriptor space is taken to be the input feature space and the activation functions and biases are chosen to be trivial, this neural network architecture provides a tensor network factorization of the model weights  $w_i^{(\nu)}$  in the ACE expansion (16), where the internal bond dimensions in the tensor network correspond to the number of nodes in the hidden layers of the neural network, and the output (open) bond has dimension 1 when the target property is a scalar. Restoring the biases and nonlinear activation functions thus generalizes the linear maps  $w_i^{(\nu)}$ . Since the descriptors are treated as input feature vectors with a multi-index  $(\alpha_1 n_1 l_1 \cdots)$ , the input bond dimension suffers from the same exponential scaling with  $\nu$  as the original ACE expansion. For deeper neural networks with many hidden nodes, this can lead to an exceptionally large number of model weights and a potential risk of overfitting when the number of training samples is small.



FIG. 9: A dense neural network as a simple tensor network, where the internal bond dimensions are determined by the number of nodes in each hidden layer, and the layer-wise activation functions and biases are implicit.

At this point we can draw some connections between the MPO/MPS model in Figure 7a and the neural network- and kernel-based methods discussed above. A sequence of  $n_d$  MPOs applied to the input descriptors can be viewed as a particular factorization of the weight matrices w in the  $n_d$  layers of a dense neural network. For sufficiently small virtual bond dimensions, this can lead to a substantial reduction in the number of model parameters [25, 32]. Hence, the tensor network structure can be understood as a form of regularization on the otherwise dense model weights, where the chosen factorization scheme can enforce a degree of sparsity in the parameters. Moreover, a single MPO applied to an input state  $|B_i^{(\nu)}\rangle$  constitutes a mapping from the symmetry-adapted, atom-centered Hilbert space to a potentially lower dimensional space,  $|B_i^{(\nu)}\rangle \to |J_i^{(\nu)}\rangle = \hat{T}^{(3\nu)}|B_i^{(\nu)}\rangle$ . Note that since this operator acts on the reduced tensor elements  $B^{(\nu)}$  of the SO(3)-invariant subspace, it commutes with the action of the SO(3) rotation group.

The overlap  $\langle J_i^{(\nu)}|J_j^{\prime(\nu)}\rangle$  is equivalent to a generalized kernel function which couples the structure of the atomic environments to the chemical species within them. This constitutes a very general form of low-rank approximation for the "nonfactorizable operators" discussed in [8]. Indeed, this overlap contains the class of so-called "alchemical" kernels built from the SOAP power spectrum (Figure 10), which have been shown to improve the accuracy of structural kernel-based methods [56, 57]. Adopting the notation in [8, 56, 57], we see that the alchemical couplings,  $\kappa_{\alpha\alpha'}$ , can be decomposed in a lowerdimensional elemental basis  $|s\rangle$ ,  $\kappa_{\alpha\alpha'} = \sum_s u_{\alpha s} u_{s\alpha'}$ . In previous applications of these generalized kernel methods, the values of these couplings were either chosen explicitly, for instance by incorporating physical intuition [56, 58], or learned via additional feature selection steps prior to training the model [57, 59]. Alternatively, by considering general tensor network factorizations of the the weights  $w_i^{(\nu)}$ , one can work directly with the atomcentered Hilbert space rather than with distinct pairs of local environments. The above discussion highlights the fact that the choice of network topology imposes a kind of inductive bias on the model which may be used as a way to prioritize a certain kind of solution (e.g., an alchemical one) for the predictor  $f(\mathbf{x})$ .

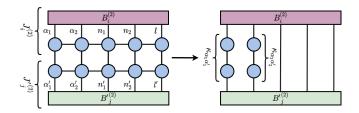


FIG. 10: A low-rank approximation of a generalized structural kernel can be represented in terms of MPOs. For the appropriate arrangement of physical bonds, this contains the class of alchemical kernels shown at right.

So far, our construction has focused on tensor decompositions of the weights  $w_i^{(\nu)}$  as multilinear maps. However, as mentioned above, there may be some advantage to introducing nonlinearity into the model architecture. This could be accomplished analogously to neural networks by introducing element-wise, nonlinear activation functions at every level. Alternatively, following the work of Stoudenmire and Schwab [26], we could construct an explicit higher-dimensional embedding via a feature map  $\varphi$  and work directly with tensor network factorizations of linear weights  $\omega$  on the embedding space,  $f(x) = \omega \cdot \varphi(\mathbf{x})$ ,

without, as with kernel methods, passing to the dual vector space. A convenient embedding is formed by a (unentangled) product state,

$$\varphi(x) = \varphi^{s_1}(x_1) \otimes \varphi^{s_2}(x_2) \otimes \cdots \otimes \varphi^{s_N}(x_N) , \qquad (19)$$

where each element of the input vector  $\mathbf{x}$  (i.e., the vectorization and standardization of the input descriptor  $B^{(\nu)}$ is encoded by a local feature map  $\varphi^{s_i}(x_i)$ . We found that in practice a simple linear embedding  $\varphi^{s_i}(x_i) =$  $[1-x_i,x_i]^{\mathsf{T}}$  performed well on vectorized atom-centered descriptors, outperforming the family of spin-coherent embeddings proposed in [26]. Because  $\varphi$  maps a ddimensional input vector to a feature space of dimension  $2^d$ , one is limited to tensor network factorizations of the weights that can be efficiently contracted. As in the original proposal, the simplest choice corresponds to a matrix product state (Figure 11). Still, to obtain an accurate model, one is often left with a large number of learnable parameters. Moreover, we have found that TN models built directly on the descriptor space outperform models with an additional feature map, while requiring fewer parameters.

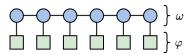


FIG. 11: Stoudenmire and Schwab MPS model [26], where each input component is explicitly mapped to an unentangled product state,  $\varphi$ .

Another interesting possibility is to mimic the construction of nonlinear structural kernels by forming  $\zeta$ order tensor products of the symmetrized descriptors  $B^{(\nu)}$ :

$$B^{(\nu)} \to B^{(\nu)} \otimes B^{(\nu)} \to \dots \to \left(B^{(\nu)}\right)^{\otimes \zeta}$$
 (20)

As inputs to a NN or TN model, this construction entails reuse of information in the subsequent network. Information reuse is a common characteristic of modern NN architectures, such as convolutional (CNN) and recurrent (RNN) neural networks, where several trainable filters act on the same input or subsets of input [30]. Recent work [35, 36] has shown that convolutional (CAC) and recurrent (RAC) arithmetic circuits, which share fundamental architectural components like overlapping filters and pooling with conventional CNNs and RNNs, can be mapped to so-called *generalized* tensor networks [30, 40] by introducing local copy operations into the tensor algebra [60]. It was shown in [35, 36] that a deep CAC with a local pooling scheme maps to a hierarchical, treestructured tensor network and, for maximally overlapping filters, can support volume-law scaling of the entanglement entropy when modeling the amplitudes of manybody quantum states. Moreover, the entanglement capacity [61] of the CAC tensor network was shown empirically to strongly influence the inductive bias of supervised image classification with a CNN [35].

In a similar vein, the MPO model introduced above can be viewed as a specific choice of learnable architecture acting on  $\zeta=2$  copies of the input descriptor,  $T^{(3\nu)}\left(|B^{(\nu)}\rangle^{\otimes 2}\right)$ . Since copy operations (20) applied to the input descriptors increase their effective many-body character [8], this method provides a potential route toward capturing higher-order correlations with lower order features.

#### IV. NUMERICAL BENCHMARKS

To validate the methods described in Section III, we will focus on a typical learning task encountered in materials property prediction using a widely available dataset (NMD18 [11]) of 3000 transparent conducting oxide (TCO) alloys  $(Al_xGa_yIn_z)_2O_3$ , a class of wide bandgap materials with high technological applicability in optoelectronic devices. The target property is the DFT mixing energy per cation (referred to as the formation energy in [11]) for a given configuration, referenced to the compositional endpoints of the material,

$$\Delta H_{\text{mix}}[(\mathrm{Al}_x \mathrm{Ga}_y \mathrm{In}_z)_2 \mathrm{O}_3] = E[(\mathrm{Al}_x \mathrm{Ga}_y \mathrm{In}_z)_2 \mathrm{O}_3] - xE[\mathrm{Al}_2 \mathrm{O}_3] - yE[\mathrm{Ga}_2 \mathrm{O}_3] - zE[\mathrm{In}_2 \mathrm{O}_3] . \quad (21)$$

Here x + y + z = 1, and E[c] is the ground state DFT energy per cation for the given compound c. The mixing energy characterizes the (zero-temperature) stability of the alloy configuration relative to the single cation phases. The underlying crystalline lattices of the alloy configurations span six distinct space groups (C/2m) $Pna2_1$ ,  $R\overline{3}c$ ,  $P6_3/mmc$ ,  $Ia\overline{3}$ , and  $Fd\overline{3}m$ ) and the number of sites in each structure can vary in integer multiples of the number of primitive cell sites. The reference compositional endpoints, however, are fixed to their ground-state lattices,  $R\overline{3}c$  for Al<sub>2</sub>O<sub>3</sub>, C/2m for Ga<sub>2</sub>O<sub>3</sub>, and  $Ia\overline{3}$  for  $In_2O_3$ . The inclusion of multiple lattice symmetries is difficult to handle with more conventional atomistic modeling methods, such as the standard cluster expansion (CE) [10, 11]. Moreover, allowing for chemically ordered/disordered sublattices presents challenges to deep end-to-end learning methods, where very large datasets are often required to obtain sufficient accuracy [62–64]. Depending on the application and the computational cost of the first-principles calculations, generating a large dataset can be prohibitively expensive. It is therefore useful to understand how the accuracy of these methods scales with the size of the training dataset.

To further emulate working in a data-constrained environment, we do not assume prior knowledge of the fully relaxed atomic positions and lattice constants. Instead, the input structures retain their ideal lattice positions,

while the lattice vectors are simply scaled according to Vegard's law. The target properties, however, are calculated from the relaxed geometries. This constitutes a particularly challenging scenario. Indeed, it was previously found that learning a mapping from unrelaxed structures to ground-state energies is generally more difficult than using relaxed input structures, at least at the level of kernel-based learning [12]. Moreover, the atomic descriptor was not found to be the limiting factor, but rather prediction error was dominated by implicit noise in the underlying set of atomic structures [12].

To construct a global descriptor suitable for the prediction of the global property  $\Delta H_{\rm mix}$ , we take an average over the local, atom-centered descriptors of the structure,

$$\overline{B}^{(\nu)} = \frac{1}{N} \sum_{i}^{N} B_{i}^{(\nu)} . \tag{22}$$

This is a special case of the general prescription of calculating  $\Delta H_{\rm mix}$  as a sum over (possibly distinct) contributions  $H_i^{(\nu)}$  from N local environments, where for (22) all local  $\nu$ -order contributions are contracted with the same weight tensor  $w^{(\nu)}$ . To simplify the comparison between different machine learning methods and architectures, we choose the SOAP power spectrum  $B^{(2)}$  with fixed radial and angular momentum cutoffs ( $R_c = 6 \text{ Å}, n_{\text{max}} = 4$ ,  $l_{\text{max}} = 3$ ) to be the input descriptor. This also allows us to compare our results to those reported in [11, 12], where the former tested the performance of both a deep NN with SOAP feature vectors and Gaussian process regression (GPR) with the SOAP kernel ( $\zeta = 2$ ), and the latter employed kernel ridge regression (KRR) using a Gaussian / radial basis function (RBF) kernel. Thus, in addition to systematically evaluating various tensor network architectures, we provide consistent benchmarks with respect to the performance of GPR with SOAP and RBF kernels, as well as the deep NN architecture described in [11].

#### A. Learning curves

We quantified the predictive performance of the MPS and MPO models, as well as GPR and NN models, using stratified k-fold cross-validation. In particular, we considered 10 splits of the NMD18 dataset into a testing set of 600 structures and a remaining pool for training and validation. Of this remaining pool of structures, 10 subsets of a fixed size were chosen with consistent distributions of volume, composition and energies, following [12], from which we constructed 80-20 splits into training and validation sets. Prediction errors were measured on the testing sets, which remain untouched during model training, while validation sets were used for hyperparameter tuning.

Figure 12 compares the average and standard deviations of the root mean square error (RMSE) and the

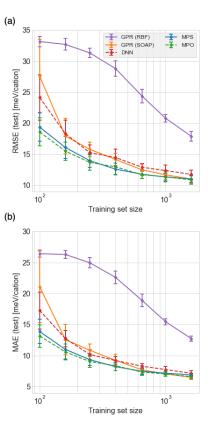


FIG. 12: Test RMSE (a) and MAE (b) learning curves on the NMD18 dataset. MPS and MPO models are compared to a fully-connected, deep neural network (DNN) and Gaussian process regression (GPR) with RBF and SOAP kernels.

mean absolute error (MAE) of the model predictions over the testing splits, as a function of the training set size. First, while it shows consistently small variance, GPR with an RBF kernel performs the worst among the models considered, likely due to over-localization of the atomic representations in the embedding feature space. By contrast, GPR with the SOAP kernel exhibits higher sensitivity to the differences between training structures, improving substantially as the number of reference structures increases. As observed in [11], we find that the DNN model and GPR with SOAP kernel perform similarly across training sets. However, the number of model weights in the DNN architecture remains fixed, while the number of weights increases with the training set size in the kernel methods. Importantly, the DNN is incrementally improvable with the acquisition of new data, while the kernel-based models must be retrained on the whole dataset. Moreover, for much larger datasets, the computational cost of kernel-based methods becomes prohibitively expensive.

The MPS and MPO models maintain strong generalizability even for very small ( $\mathcal{O}(10^2)$  structures) training sets. In this data-constrained regime, they outperform the other models considered, while converging with the

DNN and GPR (SOAP) models at the largest training set size. The advantage of the TN models is particularly notable in the RMSE, which is more sensitive to the presence of outliers than the MAE. As with the DNN model, the TN models are systematically improvable with new data without requiring retraining on the full dataset. Unlike the DNN model, they require substantially fewer parameters ( $\mathcal{O}(10^3)$ ) compared to  $\mathcal{O}(10^6)$ ) to achieve similar accuracy. This parametric efficiency may, however, be related to the observed onset of saturation in the model accuracy, particularly for the MAE of the MPS model, as the training set size increases. Yet, similar saturation is also apparent in the DNN model. Moreover, the MPO model marginally outperforms the MPS model, despite possessing a lower virtual bond dimension, and appears less prone to plateaued accuracies for increasing training set sizes. As we show below, the accuracies of the TN models do not suffer from increasing the bond dimensions for fixed datasets, and thus we expect that this saturation can be overcome by simply increasing the number of parameters. It is worth noting that we have additionally tested deeper TN factorizations with stronger entanglement scaling, specifically tree tensor networks (TTN) [65] and the Multi-Scale Entanglement Renormalization Ansatz (MERA) [66] (Figure 29), and found that they produce accuracies comparable to the MPS model. They are, however, more computationally expensive to contract. Finally, the computational cost of the TN models scales linearly with the number of samples, as opposed to the quadratic scaling of the kernel-based methods. Thus, we expect the application of these models to be practical across a broad range of dataset sizes.

# B. Entanglement measures

We have seen above that the MPS and MPO models can yield expressive supervised learning models even with a small number of training samples. To what extent does the choice of TN architecture influence this apparent inductive bias? Since the topology and bond dimensions of the TN constrain the local entanglement structure of the state  $|\psi^{(\nu)}\rangle$ , it is reasonable to expect that the entanglement in the learned TN parameters captures underlying correlations in the input data. Indeed, similar reasoning has been used to justify a priori the choice of a given TN machine learning architecture by characterizing the spatial scaling of the entanglement entropy [29, 35] or the mutual information [67, 68] for bipartitions of the input features. These studies were based primarily on the analysis of image data, which possess a precise notion of spatial arrangement. Similar analyses could be performed by measuring correlations between individual features  $x_i$ or their product state embedding  $\varphi(x)$ , in the original spirit of Stoudenmire and Schwab [26]. However, since our TN models operate directly on the tensor structure of the input descriptors, the number of available partitions of the tensor indices is small. Hence, an analysis

and interpretation of their scaling with subsystem size is limited, except perhaps at large order  $\nu$ . Instead, in this section, we scrutinize the entanglement learned by the models themselves, conditioned on the target property.

Let us recall that the entanglement entropy is a zerotemperature quantum analogue of the classical Shannon entropy. We will adhere to common practice by taking the entanglement entropy to mean specifically the bipartite, von Neumann entropy,

$$S_{\text{vN}}(\rho_{\mathcal{A}}) := -\operatorname{Tr}\left(\rho_{\mathcal{A}}\log\rho_{\mathcal{A}}\right) ,$$
 (23)

where  $\rho_{\mathcal{A}}(|\psi\rangle)$  is the reduced density matrix of a subsystem  $\mathcal{A}$  of dimension  $d_{\mathcal{A}}$ , defined by tracing out the degrees of freedom of a complementary subsystem  $\mathcal{B}$  in a pure state  $|\psi\rangle$ ,

$$\rho_{\mathcal{A}}(|\psi\rangle) = \operatorname{Tr}_{\mathcal{B}}(|\psi\rangle\langle\psi|) .$$
(24)

The entanglement entropy can be computed from the coefficients of a Schmidt decomposition of the pure state  $|\psi\rangle$ ,

$$|\psi\rangle = \sum_{k=0}^{\min(d_{\mathcal{A}}, d_{\mathcal{B}})} \lambda_k |\psi_{\mathcal{A}}^k\rangle |\psi_{\mathcal{B}}^k\rangle . \tag{25}$$

The Schmidt spectra  $\lambda_k$  are the singular values of the of matrix  $C_{\mathcal{AB}}$ , where  $|\psi\rangle = \sum_{\psi_{\mathcal{A}},\psi_{\mathcal{B}}} C_{\mathcal{AB}} |\psi_{\mathcal{A}}\rangle |\psi_{\mathcal{B}}\rangle$ . Since the eigenvalues of the reduced density matrix  $\rho_{\mathcal{A}}$  are  $\lambda_k^2$ , the entanglement entropy reduces to  $S_{\text{vN}}(\rho_{\mathcal{A}}) = \sum_k -\lambda_k^2 \log(\lambda_k^2)$ .

We can formulate an analogous entanglement entropy for the MPO models (18) by constructing a canonical purification [69, 70]. A vectorization  $|T\rangle$  of the MPO  $\hat{T}$  follows from the Choi isomorphism [70–72] (Figure 13), from which an effective pure state density matrix can be defined,  $Q = |T\rangle\langle T|$ . We note that under this isomorphism, an MPO with physical dimensions d is mapped to an MPS with physical dimension  $d^2$ . Equivalent definitions of the entanglement entropy and Schmidt spectra thus follow from (23) and (24), replacing  $\rho$  by Q and  $|\psi\rangle$  by  $|T\rangle$ .

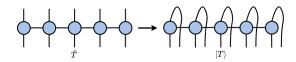


FIG. 13: The Choi isomorphism constitutes a vectorization  $|T\rangle$  for MPO  $\hat{T}$ .

Figure 14 compares average errors and entanglement entropies of the MPS and MPO models, as well as a TTN model (Figure 29), as a function of their virtual bond dimensions. The average is taken over 10 training-validation splits with a fixed testing set, and we test two

sizes (100 and 1000 structures) of the training set. For the entanglement entropies, we consider two bipartitions of the descriptor indices: (1) a contiguous subsystem  $\mathcal{A}$  consisting of chemical degrees of freedom  $\{\alpha_1\alpha_2\}$  and  $\mathcal{B}$  containing the structural components  $\{n_1n_2l\}$ , and (2) a noncontiguous subsystem  $\mathcal{A}$  consisting of a chemical and radial component  $\{\alpha_1n_1\}$  with  $\mathcal{B}$  containing the remaining degrees of freedom. These bipartitions are shown as insets in Figures 14c,d. For comparison, we also plot the errors and entanglement entropies for the corresponding models using a dense ACE tensor (Figure 5), and we indicate the bond dimension  $\chi \leq \chi_{\rm ct}$  for which the number of TN model parameters is less than the dense ACE tensor. We will refer to this bond dimension  $\chi_{\rm ct}$  as the "compression threshold."

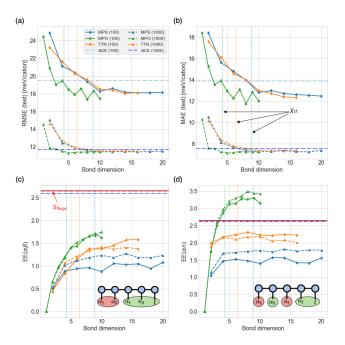


FIG. 14: Test RMSE (a) and MAE (b) as a function of the virtual bond dimension used in MPS, MPO and TTN models, for training sets containing either 100 or 1000 structures. This is compared to the performance of a model using a dense ACE tensor (dashed, horizontal lines), and the corresponding compression thresholds.  $\chi_{\rm ct}$ , for each TN model are marked with vertical, dotted lines. The associated entanglement entropies (c,d) are plotted for different bipartitions (insets; partition  $\mathcal{A}$  in red,  $\mathcal{B}$  in green), and the corresponding Page entropy,  $S_{\text{Page}}$ , is indicated by a solid, red line. Note that because the MPO model acts on two copies of the input tensor, the effective physical dimensions in each partition (as an MPS under the Choi isomorphism) are larger. The Page entropy associated with this larger Hilbert space is not shown.

For the three TN architectures, we generally find that improved test errors are strongly correlated with larger entanglement enabled by higher bond dimensions. This is particularly true for the smaller training set, which requires larger bond dimensions before reaching minimal errors. The TN models outperform the ACE models once the bond dimension reaches the compression threshold, that is, once the number of parameters exceeds that of the ACE tensor. Again, this improvement is more significant in the small training set than in the larger one. In the larger training set, the test errors of the TN models become competitive with the ACE model well before reaching the compression threshold. This behavior is somewhat counterintuitive in the context of classical statistical learning, where one expects the generalizability of an overparameterized model to degrade by overfitting on small amounts of training data, a consequence of the so-called bias-variance tradeoff [73]. However, this phenomena is not uncommon in deep neural networks, where models with trainable parameters greatly exceeding the training sample size nonetheless perform well on unseen data [74–76].

We observe that the average entanglement entropies converge to values well below the theoretical limits set by the min-cuts in the network [61] and that this convergence closely follows the formation of plateaus in test errors. Similar sub-maximal convergence in the entanglement entropy was also noted in [32] for MPO layers applied to the MNIST dataset, so we expect that this is a general phenomenon when the TN model is complex enough for the learning task. However, the converged entropies in our work appear to be nonuniversal, exhibiting strong dependence on the structure of the network. Indeed, we generally find that networks which can host stronger entanglement entropy scaling with subsystem size relative to others (for instance, logarithmic versus area law scaling between TTNs and MPSs [52, 65]) tend to converge to higher values. Furthermore, for fixed virtual bond dimension, we find that the MPO model outperforms the MPS model and achieves higher entanglement. This appears consistent with the findings [48] of greater representational efficiency of MPDOs versus MPSs in probabilistic graphical models, although, for this work, it is in the context of regression. Hence, under the same arrangement of virtual bonds, the entanglement entropy seems to provide a consistent measure of the TN model's generalizability.

Nonetheless, comparisons of the entanglement between trained models with distinct virtual bond topologies are difficult to make. For instance, the dense ACE tensors perform somewhat worse than the other TN models, but achieve much higher entanglement entropies. In this case, the entanglement entropies do approach a universal value, that of the Page entropy [77],

$$S_{\text{Page}} = \frac{1 - d_{\mathcal{A}}}{2d_{\mathcal{B}}} + \sum_{k=d_{\mathcal{B}}+1}^{d_{\mathcal{A}}d_{\mathcal{B}}} \frac{1}{k}$$
 (26)

defined as the average entanglement entropy of a pure state randomly drawn from the entire Hilbert space. This would not be entirely surprising for an untrained ACE tensor  $w^{(\nu)}$  treated as a length  $(|\alpha|^{\nu}|n|^{\nu}|l|^{\nu})$  vector since the individual elements are initialized according to a standard normal distribution  $\mathcal{N}(0,\sigma^2)$ . However, it is surprising that the high entanglement of this initial state is essentially preserved under SGD with a mean square error (MSE) loss function, at least up until early stopping. This may provide some explanation for the improved performance of the models with an explicit TN factorization: the entanglement constraints imposed by the network itself help drive the model toward minima in the loss landscape which are characterized by entanglement more finely tuned to the target property.

It has long been recognized that the entanglement entropy alone is insufficient to fully capture the entanglement of general quantum states and that much richer structure can be found in the full entanglement spectrum  $\{\lambda_k^2\}$  or its logarithms  $\{\xi_k := -\log(\lambda_k^2)\}$  [78]. The level spacing statistics between consecutive pairs of eigenvalues have proven useful in characterizing the irreversibility in quantum circuits [79, 80] and emergent entanglement complexity in many-body quantum dynamics [81– 84]. For entanglement spectra arranged in descending order  $\lambda_0^2 \geq \lambda_1^2 \geq \ldots$ , let  $s_k := \lambda_k^2 - \lambda_{k+1}^2$  denote the kth level spacing. To avoid unfolding the spectrum [85], a procedure sensitive to spurious finite size effects, it is common practice [81, 86–88] to study the ratio of consecutive level spacings  $r_k = s_k/s_{k+1}$  or the related quantity  $\widetilde{r}_k = \min(r_k, 1/r_k)$ , which are independent of the level density of states. It was shown in [79-81] that either the distribution of spacings P(s) or their ratios P(r), collectively referred to as the entanglement spectrum statistics (ESS), provides a measure of a quantum state's entanglement complexity, defined by the existence of an efficient algorithm that completely disentangles the state. Complexly entangled states, for which efficient disentangling algorithms fail, were found to feature ESS with Wigner-Dyson (WD) statistics,

$$P_{\text{WD}}(r) = \frac{1}{Z_{\beta}} \frac{(r+r^2)^{\beta}}{(1+r+r^2)^{1+(3/2)\beta}} , \qquad (27)$$

where  $Z_{\beta}$  is a normalization factor, and the Dyson index,  $\beta$ , specifies one of three Gaussian random matrix ensembles [88]. On the other hand, states which could be efficiently disentangled possessed Poisson-like ESS,  $P_{\text{Poisson}}(r) = 1/(1+r)^2$ . An important feature that distinguishes WD from Poisson level statistics is the presence of level repulsion,  $P_{\text{WD}}(r \to 0) \sim r^{\beta} \to 0$ , which reflects universal statistical correlations between adjacent levels [85]. We also note that there appears to be a deep connection between quantum circuits capable of universal computing and ESS [79–82], in that a universal gate set gives rise to WD statistics. However, a recent study [84] showed that the converse is not true in general, finding that a class of classically simulatable circuits can prepare states with WD ESS.

It is reasonable to expect the above picture to hold,

to some extent, for a randomly initialized MPS. Indeed, a random MPS, viewed as a unitary embedding [89–91], forms an approximate 2-design [92]: as a random quantum circuit, the first and second moments approximate those of a Haar distribution. Moreover, Haar-distributed random unitary circuits possess WD ESS corresponding to the Gaussian unitary ensemble (GUE),  $\beta = 2$  [82], although a precise relationship between ESS and k-designs remains an open question [84]. Nonetheless, we will explore the ESS of our MPO model, viewed as an MPS under the Choi isomorphism. For simplicity, we consider the entanglement spectrum for the  $\{\alpha n\}$  bipartition shown in Figure 14d and analyze its evolution under SGD for a fixed training set. To obtain sufficient statistics, we collect data over 100 independent runs of 1000 SGD epochs, and we study models with virtual bond dimensions for which the spectrum is either truncated  $(\chi = 12)$ , marginal  $(\chi = 16)$ , or full-rank  $(\chi = 20)$ .

Figure 15 shows both the logarithm of the average spectrum  $\langle \xi_k \rangle$  and the ESS at different time steps. First, we find that for all cases the entanglement spectrum deviates from the Marchenko-Pastur (MP) law for Haardistributed random states sampled from the full Hilbert space [93, 94]. Interestingly, close examination of the truncated ( $\chi = 12$ ) model reveals residual MP structure in the tail of the spectrum, similar to the two-component structure discussed in [95], and this MP tail persists under SGD up to a constant shift from normalization. For all models, upon random initialization of the MPS tensors, the spectrum is relatively flat and thus highly entangled. The largest changes in the spectrum occur at the earliest stages of training, and the spectrum quickly converges to a configuration in which the local level density has decreased. We also find that the largest spectra (low  $\xi_k$ ) approximately converge to the same values, regardless of the bond dimension, which is consistent with the prior observation that the model retains high accuracy for smaller bonds.

At initialization, all three cases display WD ESS corresponding to the Gaussian orthogonal ensemble (GOE),  $\beta = 1$ . This may not be entirely surprising given the above discussion, although in this case the generating random matrix ensemble is real-valued. However, as the state evolves under SGD, the ESS retains a substantial degree of its GOE character. To measure the difference between the GOE ESS,  $P_{\text{GOE}}(r)$ , and the observed ESS,  $P_{\text{obs}}(r)$ , we plot the Kullback-Leibler (KL) divergence,  $D_{\text{KL}}(P_{\text{GOE}}||P_{\text{obs}}) = \text{Tr}_r[P_{\text{GOE}}\log(P_{\text{GOE}}/P_{\text{obs}})], \text{ in the}$ insets to Figure 14. First, we note that the higher starting values of  $D_{\mathrm{KL}}$  in the truncated and marginal spectra relative to the full-rank case are likely due to finite-size effects imposed by the bond dimension. We see that at early times there is rapid growth in the KL divergence followed by a regime in which  $D_{KL}$  increases very slowly, accompanied by oscillations due to the stochasticity of the optimization. The onset of this slow growth regime prevents the ESS from significantly deviating from the GOE over the course of training. As we noted above, the

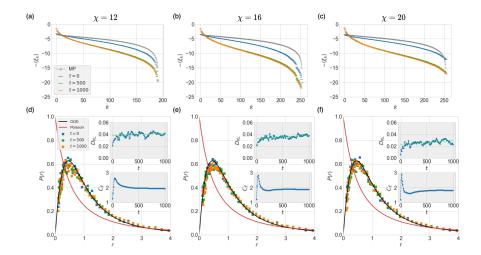


FIG. 15: (a-c) The average entanglement spectrum of the MPO model as a function of training step, t, compared to the Marchenko-Pastur (MP) law. (d-f) The distribution P(r) of the ratio r of level spacings in the entanglement spectrum (referred to as the entanglement spectrum statistics (ESS) in the main text) closely follows the Gaussian orthogonal ensemble (GOE). Under stochastic gradient descent (SGD), the KL divergence (upper insets) exhibits a short, initial regime of fast growth followed by a slow growth regime. A linear fit to the KL divergence in the slow growth regime (dashed, light blue line) is included to help guide the eye. The crossover between the fast and slow growth in the observed ESS relative to the GOE, coincides with a divergence in the capacity of entanglement,  $C_E$ , (lower insets) signaling a phase transition in the entanglement spectrum.

regime of rapid growth is accompanied by a decrease in the level density, which essentially stabilizes in the slow growth regime.

Recalling the level repulsion in the GOE, this expansion of the level density resembles the expansion and reequilibration of a one-dimensional Coulomb gas following a pertubation of its confining potential. Specifically, the joint probability distribution of N eigenvalues in the GOE is precisely a Boltzmann-Gibbs factor for a gas of N particles with pairwise, repulsive  $-\log(|\lambda_i - \lambda_j|)$  interactions in a quadratic potential [94]. This picture suggests that SGD evolves the entanglement spectrum and the ESS by evolving the confining potential of the joint level distribution. Furthermore, if this potential is driven far enough, the entanglement spectrum should undergo a phase transition. We can examine whether this is the case by calculating the so-called capacity of entanglement [96],

$$C_E := \langle H_{\text{ent}}^2 \rangle - \langle H_{\text{ent}} \rangle^2 ,$$
 (28)

where  $H_{\text{ent}} := -\log \rho_{\mathcal{A}}$  is the entanglement Hamiltonian [78, 97]. The capacity of entanglement measures fluctuations in the entanglement spectrum, analogous to the heat capacity in classical statistical mechanics. The insets to Figure 14 show the evolution of  $C_E$  for the three cases. We observe that the crossover from the fast to slow growth regimes in the KL divergence coincides with a divergence in the capacity of entanglement, signaling a phase transition in the entanglement spectrum. Again, this transition is smoothed out in the truncated and

marginal cases due to finite-size effects.

The above discussion provides a mechanism through which the model entanglement adapts to the learning task. Crucially, while the average entanglement spectrum converges to a specific configuration, particularly near its low- $\xi_k$  edge, the ESS maintains universal characteristics of WD statistics, namely level repulsion. Thus, the learned state retains the entanglement complexity from its random initialization. This is consistent with its strong generalizability in that the model is expressive enough to capture complex correlations in the highdimensional feature space conditioned on the target property. Moreover, this behavior appears to have close connections to overparameterized neural networks [98, 99] and quantum circuits [83, 100], where SGD applied to sufficiently overparameterized models finds solutions with small generalization error that are nonetheless close to their random initialization. The transition observed in the entanglement spectrum during training is reminiscent of the dynamical phase transition observed in SGDbased training of deep neural networks [101]. At early times, model evolution is dominated by the average gradient in the loss function, which rapidly minimizes the model error. At latter times, model training becomes dominated by stochastic fluctuations in the gradient, introducing diffusive behavior in the model evolution. This appears to be reflected in the dynamics of the capacity of entanglement, where in the slow growth regime the entanglement spectrum exhibits stronger fluctuations. In [101], this diffusion phase was associated with compression of the model's latent representation of the inputs.

We observe similar compressive behavior in the evolution of the entanglement spectrum, where contributions to the entanglement entropy become increasingly dominated by the low- $\xi_k$  edge of the entanglement spectrum.

#### C. Latent space encoding

As discussed in Section III A, TN factorizations of the learnable states  $|\psi^{(\nu)}\rangle$  can be understood as multilinear maps which internally couple structural and chemical degrees of freedom. In this section, we explore this idea further by visualizing the latent tensor product spaces learned by a deeper TN architecture (Figure 29) corresponding to a MERA with fixed virtual bond dimension. To do so, we use t-distributed stochastic neighbor embedding (t-SNE), a nonlinear dimensional reduction method that preserves the relative locality between points in the dataset [102–104]. Again, the MERA network is trained to predict the mixing energy (21) on a 1000-structure subset of the full NMD18 dataset, and we employ t-SNE to embed the tensor product spaces mapped by subsequent hidden layers.

Figure 16 shows the evolution of the latent space image of each datapoint as function of depth in the MERA network. Note that we apply t-SNE to the entire dataset, not just to the training set. The input average SOAP descriptors,  $\overline{B}^{(2)}$ , clearly exhibit a high degree of structure, to the extent that t-SNE surprisingly reproduces the ternary phase diagram of the alloy. The first layer of the network reorganizes the data in a hierarchical fashion, grouping structures with the same space group into local clusters. Within each cluster, the separation of struc-

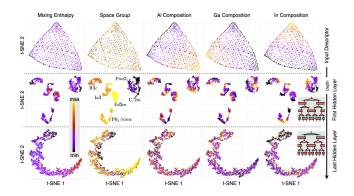


FIG. 16: Evolution of t-distributed stochastic neighbor embeddings (t-SNE) of the NMD18 dataset with increasing depth in a MERA tensor network model (insets). The regularity of the encoding of chemical composition in the input SOAP descriptor is reorganized by the first hidden layer according to the lattice symmetries of the structures. The features are subsequently coarse-grained by fitting the model to the mixing energy, yielding a quasi-one dimensional encoding ordered by high and low energy.

tures according to alloy composition is also maintained. As a byproduct, we also see that a significant number of high energy structures in the data occur for space groups  $Ia\overline{3}$  and  $Fd\overline{3}m$  which contain a significant amount of Al. At the final hidden layer, this hierarchical structure has been replaced by a single quasi-1D cluster, with an orientation determined by the mixing energy. This conforms with the intuition provided by a renormalization group interpretation of the network [22, 23, 66, 105], in the sense that lower levels in the network resolve finer details in the structure representations, and these are subsequently coarse grained in mapping to the final target quantities.

# D. Compression of large descriptors

While the expansion (16) is mathematically well-defined [14], the exponential scaling of the descriptors  $B^{(\nu)}$  with the number of distinct chemical species and the cutoffs of the radial and angular momentum channels can yield exceptionally large feature spaces. This can impose computational limitations on the application of machine learning to atomistic modeling, particularly when the structures in the dataset span compositions with a large number of elements. Thus, recent efforts [6, 57–59, 106, 107] have been made toward optimizing the efficiency of atomic representations while maintaining their symmetry invariance. This was, for example, a central motivation for the introduction of generalized kernel functions [8], alchemical or otherwise (cf. Section III A).

As discussed above, one way to alleviate the computational complexity of learning with high dimensional descriptors is to employ TN factorizations of the model architectures, which impose a kind of structured sparsity on the trainable model weights. Similar constructions can be applied to the input descriptors. In this section, we explore this possibility with two methods. First, we validate the MPS factorization scheme (Figure 8) discussed in Section III. Second, we implement an autoencoder using MPOs (Figure 18). In principle, these two methods can be combined, both in training the autoencoder and in subsequent supervised learning with the encoded input tensors, although we do not test this combination here.

To test these methods, we use the BA10 dataset [10], which consists of 15950 binary alloys and their mixing energies. The binary alloys span composition with 10 different metals and 3 different lattice symmetries, face-centered cubic (fcc), body-centered cubic (bcc) and hexagonal close-packed (hcp), with volumes up to 8 atoms per unit cell. It is worth noting that, like NMD18, the structures in BA10 retain their idealized lattice positions, while the volumes are scaled according to Vegard's law. However, the DFT mixing energies are computed for these unrelaxed structures, as opposed to the relaxed ground states in NMD18. Hence, the prediction accuracy in prior studies [10, 12] was found to be more limited

by the descriptor parameters than implicit noise in the dataset. In addition to the large number of distinct chemical species, accurate models based on the SOAP power spectrum,  $B^{(2)}$ , require a fairly large number of radial  $(n_{\text{max}} = 8)$  and angular momentum  $(l_{\text{max}} = 8)$  channels [10, 12]. As in [12], we will consider training-testing splits of 1600 and 1000 structures, respectively, for both the unsupervised task of training the autoencoder and the subsequent supervised task of predicting the mixing energies. For the latter supervised task, we use an MPS model with the encoded tensors as input. Given the exponential scaling of the descriptor dimensions, the input state  $|\overline{B}^{(2)}\rangle$  retains only the independent entries of the original descriptor tensor. Furthermore, zero-padding is added to the state  $|\overline{B}^{(2)}\rangle$  to allow for uniform physical dimensions in the subsequent MPS and MPO tensor networks. Note that this does not affect the symmetry invariance of the descriptor, but the physical dimensions of the MPS and MPO architectures in this analysis do not admit the same correspondence with the structural kernels discussed above. Nonetheless, the high accuracies obtained with these models underscores the flexibility in choosing the TN factorization and their broader applicability in machine learning [25, 26, 29, 30, 32, 67, 68].

Figure 17 shows the change in the test errors as a function of the maximum bond dimension,  $\chi_{in}$ , for a MPS factorization of the descriptor state  $|\overline{B}^{(2)}\rangle$ . In this case, the factorized descriptor is used as input to a supervised MPO model ( $\chi = 4$ ). This is compared to the performance of the unfactorized, dense  $\overline{B}^{(2)}$  tensor, and we also plot the corresponding number of input parameters,  $N_{\rm in}$ . As expected, retaining large bond dimensions results in small deviations from the baseline error, while the efficiency of contracting the tensor network is substantially improved. The test errors grow as the bond dimension decreases, but surprisingly, this growth is nonmonotonic. Thus, despite reducing the number of input parameters by up to two orders of magnitude, the test errors remain relatively stable, and there are certain smaller bond dimensions that outperform larger ones. The origin of this nonmonotonic behavior is unclear. Naively, one would expect that discarding lower singular values at each virtual bond would introduce noise into the inputs, eventually degrading the model performance. However, it appears that some competing process is present. One possibility is that compressing the input tensors can improve the quality of the loss landscape (i.e., its smoothness or convexity), similar to overparameterization in deep neural networks [98, 99]. If that is the case, then deciding how to balance the input factorization against the TN model architecture would be an interesting open question.

We mentioned previously that an MPO applied to an input state can be viewed as a multilinear map from the atom-centered Hilbert space to a potentially lower dimensional space. Moreover, the topology of certain hierarchical tensor networks, such as TTNs and MERAs,

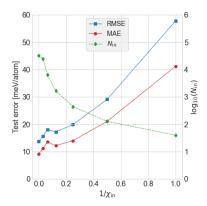


FIG. 17: The test performance of a supervised MPS model using MPS factorizations of the dense input descriptors on the BA10 dataset. The number of input parameters for a MPS factorization with maximum bond dimension  $\chi_{\rm in}$  is also plotted. The case where the exact input descriptors are used  $(1/\chi_{\rm in} \to 0)$  is included for comparison.

can be viewed as enforcing this kind of dimensional reduction (see Section IV C). For instance, it has recently been recognized that there is a fundamental correspondence between MERAs and wavelet transformations [108]. We can exploit this general idea to learn approximately faithful compressions of the input descriptors. To do so, we employ MPOs in an autoencoder setup, shown in Figure 18, where the output dimensions,  $d_{latent}$ , of the MPO tensors are less than the input dimensions,  $d_{in}$ . In this setting, the encoder MPO maps the input state to a lower dimensional latent space, while the decoder MPO attempts to invert this transformation. The autoencoder network is trained by SGD to minimize the MSE between the components of the input state and its reconstruction by the decoder. The trained encoder contracted with the input state (Figure 18b) yields a compressed tensor adapted to the data.

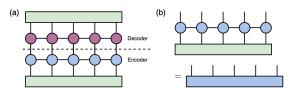


FIG. 18: (a) An autoencoder based on MPOs used to compress the physical dimensions of large input descriptor tensors. (b) A compressed tensor is given by the contraction of the encoder MPO with the original input.

In Table I, we record the test reconstruction errors (RMSE and MAE) for autoencoders with different  $d_{\text{latent}} < d_{\text{in}} = 8$ , as well as the test performance of the encoded inputs in predicting the mixing energies with an MPS model. The bond dimensions of the autoen-

TABLE I: The performance of a 5-site MPO-based autoencoder with virtual bond dimension  $\chi_{\rm enc/dec}=6$  is measured by the RMSE (MAE) for reconstructing input descriptors in a test set. This test reconstruction error (r-Error) is shown for varying the latent dimensions,  $d_{\rm latent}$ , of the encoder, and the subsequent compression of the input  $d_{\rm latent}^5/d_{\rm in}^5$  is also included. The test RMSE (MAE), denoted s-Error, characterizes the performance of a supervised MPS model ( $\chi_{\rm mps}=10$ ) applied to the compressed inputs.

$d_{ m latent}$	Compression	r-Error	s-Error [meV/atom]
-	1	-	14.8 (9.9)
6	0.237	0.060 (0.031)	16.3(11.2)
4	0.031	0.088(0.047)	19.3 (13.9)
2	$9.76 \times 10^{-4}$	0.137(0.070)	103.2 (77.8)

coder MPOs are fixed ( $\chi_{\rm enc} = \chi_{\rm dec} = 6$ ), as are the bond dimensions of the supervised MPS model ( $\chi_{\rm mps} = 10$ ). The compression mediated by the encoder is measured against the original size of the input state,  $d_{\rm latent}^n/d_{\rm in}^n$ . We find that the MPS model performance in predicting the mixing energies remains high even for encoded inputs retaining a few percent of the original number of parameters.

Thus, we have found that both an MPS decomposition and an MPO-based autoencoder are effective methods to reduce the computational complexity imposed by the input descriptors. It is again worth emphasizing the parametric efficiency of these methods compared to more standard deep learning architectures. For example, the deep convolution neural network in the original BA10 work [10] uses  $\mathcal{O}(10^8)$  parameters combined with manybody tensor representations (MBTR) [5], which lead to feature vectors with  $\mathcal{O}(10^5)$  entries. On the other hand, the high-performing MPS / MPO models possess  $\mathcal{O}(10^3)$  parameters with compressed input tensors of similar size.

#### V. DISCUSSION AND SUMMARY

In this work, we have demonstrated that tensor networks provide both a unifying formalism for a large class of atomic structure representations as well as robust machine learning architectures based on them. Specifically, we detailed how the ACE framework admits a natural description in terms of symmetric tensor networks. This formalism allows for the transparent application of SO(3) recoupling theory to the construction of a rotationally invariant basis and generalizes straightforwardly to SO(3)-equivariant cases. By taking seriously the Hilbert space structure of atom-centered descriptors [8], we have shown how tools commonly applied in quantum manybody physics can be repurposed in coarse-grained surrogate models of fundamental material properties.

Indeed, the map between a solid-state or molecular structure and a given property can be written as an inner

product between a learnable state and a sufficiently representative descriptor, both of which admit low-rank tensor network factorizations. We have found that, as learnable architectures, MPOs and MPSs exhibit strong generalizability in common learning tasks in material informatics, and their performance is particularly notable when the amount of available data is very limited. We provided two routes toward understanding this apparent inductive bias. First, we showed how learnable tensor networks, viewed as multilinear maps on the atom-centered Hilbert space, can be related to generalized kernel functions. An appropriate choice of network topology thus provides an explicit way of coupling structural and chemical components in the model. Unlike standard structural kernel methods, the evaluation of these TN models scales more efficiently with the number of structures. This is similar to the application of neural networks, and the MPS models benchmarked in this work can be viewed as parametrically efficient regularizations of a dense neural network. Indeed, the apparent compressibility of the descriptor space and the learning architecture can be used to obtain more efficient atomic representations, and we demonstrated this using both a standard MPS decomposition algorithm as well as a MPO-based autoencoder. The second way to understand the strong generalizability of these TN models is more fundamental and potentially a universal feature of tensor-network machine learning more broadly. In this case, the learned TN states possessed signatures of high entanglement complexity. In particular, the initial WD statistics of the level spacings in the entanglement spectrum were essentially preserved under stochastic gradient descent, while the spectrum itself adapted to the learning task.

The above discussion points to several directions for potential future work. First, general TN factorizations can be used to replace fully connected layers common in deep learning architectures, a program already undertaken in [25, 32]. In the context of materials prediction, adding TN layers in end-to-end architectures like graph neural networks is an effective way to reduce memory overhead and could potentially improve performance on smaller training sets. Similarly, TN layers could be implemented in the Behler-Parrinello neural network framework [1, 9] for modeling high-dimensional potential energy surfaces. In this case, separate tensor networks would replace the neural networks applied to each local environment. On a related note, it would be worth quantifying the performance of these networks on very large datasets and identifying, in particular, whether there is some advantage to using deep versus shallow network architectures [23].

Let us mention that training TNs with a large number of tensors, as would be the case (20) for high orders  $\nu$  and  $\zeta$ , using SGD suffers from the presence of barren plateaus in the loss function [91, 109, 110]. For a large MPS, this is intrinsically related to the fact that, under random initialization, the state constitutes an approximate 2-design for which the expected gradients vanish [91]. It may be

possible to overcome this problem by choosing a different initialization scheme [90], or by preconditioning the network with a different algorithm, such as DMRG. Formulating a local loss function should also alleviate the issue [91, 110]. Alternatively, one could utilize common strategies employed in deep learning for regulating gradients. For example, a local tensor network scanned across subpartitions of an input tensor provides a TN analogue of weight sharing in a convolutional neural network. Residual skip connections and batch normalization may also prove valuable in deeper networks. A related question underlying these approaches is the extent to which they lead to complexly entangled states. If WD statistics in the entanglement spectrum is a fundamental signature of a highly generalizable model, it would be worth characterizing, in general, how the entanglement spectrum in approximate k-designs evolves under stochastic gradient descent.

From a practical standpoint, the parametric efficiency of these methods as well as their high performance on limited training data makes a strong case for their application in large-scale materials simulation and high-throughput screening, particularly when the calculation of target parameters from first principles is computationally demanding. Furthermore, while we have characterized these methods on the important class of atom-density representations, we expect that they are broadly adaptable to generic material feature spaces [111].

# Appendix A: Tensor network graphical notation

An arbitrary state  $|\psi\rangle = \sum_{\{s\}} \psi_{s_1 s_2 \cdots s_n} |s_1 s_2 \cdots s_n\rangle$  in a tensor product space  $|\psi\rangle \in \mathcal{H}^{\otimes n}$ , where each  $|s_k\rangle$  forms a basis for a finite-dimensional vector space  $\mathcal{H}$ , can be described by the tensor  $\psi_{s_1 s_2 \cdots s_n}$  formed by its coefficients. In the standard diagrammatic notation of tensor networks, an arbitrary, dense tensor,  $\psi_{s_1 s_2 \cdots s_n}$ , of order n is represented by a shape or node with n open edges. Summation over a matching pair of tensor indices, otherwise referred to as contraction, is represented by a closed edge in the network. The familiar example of matrix multiplication is shown as a tensor network in Figure 19, as well as a less trivial network.

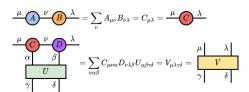


FIG. 19: Examples of graphical notation for tensor networks. Tensor indices associated with each tensor in the network are represented by open edges. Closed edges between pairs of tensors indicate summation / contraction over the matching tensor indices.

## Appendix B: Additional SO(3) recoupling relations

In this appendix, we collect some useful identities in the algebra of SO(3) representations. Recalling the diagrammatics from the main text, each line with an arrow is labelled by an irrep l and spans an associated subspace of dimension  $d_l=2l+1$ . Intertwiners between irreps are given by the usual Clebsch-Gordan (CG) coefficients (Figure 1c,d), which satisfy the orthogonality relations shown in Figure 20a,b. The CG tensors commute with the action of the group (Figure 20d for C and analogous for  $C^{\dagger}$ ); that is, they form a natural (i.e., equivariant) transformation. Note that the following three identities are sufficient to derive the generalized Wigner-Eckhart theorem discussed in the main text (Figure 2).

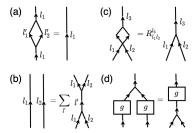


FIG. 20: Additional recoupling relationships between Clebsch-Gordan tensors and Wigner *D*-matrices, including (a) orthogonality, (b) completeness, and (d) naturality /equivariance. (c) Braiding between irrep edges is defined in terms of a set of *R*-symbols, which reduce to sign factors in SO(3).

The recoupling of three or more irreps can be described by the contraction of multiple CG tensors to form a fusion tree, and recoupling schemes represented by different fusion trees can be related to each other via unitary transformations known as F-symbols (Figure 21a,b). In SO(3), explicit values for the F-symbols can be derived from the Wigner 6j symbols (or alternatively from Racah W coefficients), which are given by the maximally irreducible diagram formed by the contraction of two 4-valent fusion trees [45, 46]. Additionally, pairs of irreps can be swapped (Figure 20c) using the socalled R-symbols. In SO(3), they take the explicit values  $R_{l_1l_2}^{l_3}=(-1)^{l_1+l_2-l_3}$  and can be used to constrain to the parity of the state. For instance, an SO(3)-invariant  $\nu$ order descriptor possesses inversion symmetry if  $\sum_{k=1}^{\nu} l_k$ is even, which corresponds to  $\left(\prod_{k=1}^{\nu-2}R_{l_k'l_{k+2}}^{l_{k+1}'}\right)R_{l_1l_2}^{l_1'}=1$  for  $\nu>2$  and  $l_{\nu-1}'=0$  in the recoupling scheme used in the main text.

Let us mention that additional coherence relations must be satisfied by the F- and R-symbols, namely the so-called pentagon and hexagon identities [112], where the pentagon identity is shown in Figure 21c. Indeed, the algebraic structure of finite-dimensional representations of SO(3) corresponds more generally to a symmetric tensor category. A useful aspect of this construc-

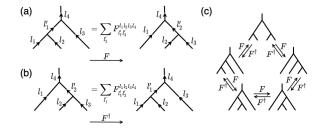


FIG. 21: Unitary transformations (a,b) known as F-symbols define mappings between recoupling schemes. They must satisfy the consistency condition (c) known as the pentagon identity.

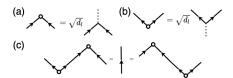


FIG. 22: (a) Cap and (b) cup tensors defined in terms of Wigner 2jm symbols. Diagrammatically, they allow for the reversal of an irrep edge orientation, and satisfy the inversion identity (c).

tion is the ability to encode nontrivial manipulations of tensors in the local deformations of diagrams. For example, the reversal an irrep arrow can be accomplished by contracting with normalized Wigner 2jm symbols,  $\epsilon_{mm'}^{(l)} \coloneqq \sqrt{d_l} C_{mm'0}^{(l0)} = (-1)^{l-m} \delta_{m,-m'},$  where the cup and cap tensors  $\epsilon^{(l)}, (\epsilon^{(l)})^{\dagger}$  are matrix inverses of each other,  $\epsilon^{(l)}(\epsilon^{(l)})^{\dagger} = (\epsilon^{(l)})^{\dagger} \epsilon^{(l)} = \mathbb{I}^{(l)}$  (Figure 22). The cup and cap tensors will be used to recursively construct equivariant descriptors in the following section.

# Appendix C: Generalizing to SO(3) equivariants and beyond

In this appendix, we show how the SO(3)-invariant tensor networks in Section II A can be generalized to SO(3)-equivariance and applied to the learning of vectorial and tensorial properties. In doing so, we will largely reproduce central results related to the general ACE formalism [13, 14], as well as the so-called  $\lambda$ -SOAP [113] descriptors and the recursive N-body iterative contraction of equivariants (NICE) framework [55].

First, we recall from the main text (Figure 2) that  $\nu$ -order SO(3)-invariant tensors can be explicitly constructed by taking the Haar integral over the action of the group. This can be understood as a projection onto the invariant subspace,  $\operatorname{Inv}(\mathcal{V}_{l_1} \otimes \cdots \otimes \mathcal{V}_{l_{\nu}}) \cong \operatorname{Hom}(\mathcal{V}_{l_1} \otimes \cdots \otimes \mathcal{V}_{l_{\nu}}, \mathbf{1})$ , of the tensor product of  $\nu$  irrep spaces  $\mathcal{V}_{l_k}$ . This projection operator,

$$P^{(l_1 \cdots l_{\nu})} = \int dg D_{m'_1 m_1}^{(l_1)}(g) \cdots D_{m'_{\nu} m_{\nu}}^{(l_{\nu})}(g)$$
 (C1)

$$= \sum_{\iota} \overline{\iota_{m_1' \cdots m_{\nu}'}} \iota_{m_1 \cdots m_{\nu}} \tag{C2}$$

is shown graphically in Figure 23, where for  $\mathcal{G}$  a semisimple group this projection decomposes, as before, into dual fusion trees (i.e., higher-order intertwiners) factored through the trivial irrep space [42, 114]. Note that this decomposition is a direct consequence of the Peter-Weyl theorem [42]. We denote these general intertwiners by  $\iota_{m_1\cdots m_n}$ , and the sum in (C2) runs over the intermediate irreps, which depend on the chosen recoupling scheme.

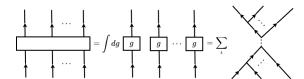


FIG. 23: General projection operator onto the invariant subspace  $\operatorname{Inv}(\mathcal{V}_{l_1} \otimes \cdots \otimes \mathcal{V}_{l_{\nu}})$  of an arbitrary tensor product space  $\mathcal{V}_{l_1} \otimes \cdots \otimes \mathcal{V}_{l_{\nu}}$ .

Under this projection, a  $\nu$ -order SO(3)-equivariant descriptor,  $B_L^{(\nu)}$ , transforming as a state with definite angular momentum L can be constructed by symmetrizing over the product of  $\nu$  atom-centered tensors,  $A_{\alpha nlm}$ , along with an identity operator carrying the target irrep space L, as shown in Figure 24. As before, the descriptor,  $B_L^{(\nu)}$ , corresponds to the reduced part of the decomposition. By bending upward the open L edge of the corresponding tensor network using cup and cap tensors, one can verify that the tensor product of  $\nu$  irreps is mapped to the target L irrep, and the fully invariant case from the main text can be recovered by taking L=0.

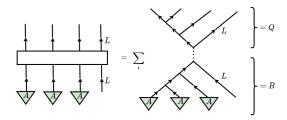


FIG. 24: Elements of the basic recursive construction of order- $(\nu + 1)$  equivariant descriptors, involving (a) the reorientation of the final momentum irrep edges using cup tensors and (b) the fusion of additional order-1 descriptors with an order- $\nu$  descriptor.

From this picture, it is clear that a  $(\nu + 1)$ -order descriptor can be obtained by fusing together lower order descriptors. This is shown in our chosen recoupling scheme in Figure 25, where an order-1 descriptor, equivalent to the unsymmetrized atom-centered tensor A, is

combined with an order-2 descriptor by contraction with the necessary CG, cup and cap tensors. Repeated application of this process yields a recursive formula for higher order descriptors.

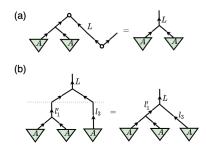


FIG. 25: Elements of the basic recursive construction of order- $(\nu+1)$  equivariant descriptors, involving (a) the reorientation of the final momentum irrep edges using cup tensors and (b) the fusion of additional order-1 descriptors with an order- $\nu$  descriptor.

The angular momentum L of an equivariant descriptor can be preserved by a tensor network model by acting only on the remaining  $\{\alpha_k, n_k, l_k\}$  degrees of freedom. For example, an MPS-like model which maps to a target quantity  $y_{sM}^{(L)}$  living in the space  $|s\rangle\otimes|LM\rangle$ , where s labels a possible additional degree of freedom, is shown in Figure 26a. Naively taking  $\zeta$ -order tensor products of L>0 equivariant descriptors destroys the angular momentum channel of the original state since products of irreps can be subsequently recoupled. Alternatively, as mentioned in [8], the effective many-body character of an equivariant model can be increased by taking tensor products of the equivariant state,  $B_L^{(\nu)}$ , with  $\zeta-1$  copies its invariant counterpart,  $B_{L=0}^{(\nu)}$ . An example  $\zeta=2$  MPO-type model is shown in Figure 26b.

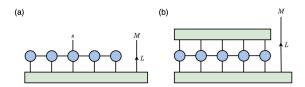


FIG. 26: Tensor network models which preserve the equivariance of the input descriptor. (a) An example MPS model which allows for additional output degrees of freedom s. (b) An example MPO model which implicitly increases the effective many-body character of the input descriptor by contracting with an equivariant descriptor and its invariant counterpart.

A useful aspect of this formalism is that the fusion trees that encode the equivariance of the model can be manipulated purely algebraically and computed independently from the atom-centered A-tensors [39]. As demonstrated in [14, 15] with the ACE formalism, this allows for potential speedups in applications, such as molecular dynamics or Monte Carlo, which constantly update the underlying

descriptors. In particular, the fusion trees can be recursively precomputed and contracted with the A-tensors along with the learned model weights at runtime.

Let us point out some implications for the constructions of equivariant structural kernels [113]. Figure 27a.b displays the diagrams for kernel functions corresponding to the  $\lambda$ -SOAP generalization of the power spectrum  $(\nu = 2)$  and bispectrum  $(\nu = 3)$ , respectively. The pairs of descriptors in these kernels are built from the same recoupling scheme, and for L=0 it can be seen that each kernel function is essentially given by a spin network equivalent to a Wigner 3nj symbol [46] weighted by the trace over the remaining  $\{\alpha_k, n_k, l_k\}$  degrees of freedom. For either  $(L=0, \nu > 3)$  or  $(L>0, \nu > 2)$ , there exist topologically distinct kernel functions determined by the application of F-symbols to the pairs of descriptor fusion trees. An example for  $\nu = 4$ , valid for all L, is shown in Figure 27c, and for L=0 the right-hand side is equivalent to the maximally irreducible diagram of a weighted Wigner 6*i* symbol. This raises the interesting question of whether certain nonequivalent inner products for higher  $\nu$  equivariants play a more privileged role than others, particularly as a basis for machine learning tasks.

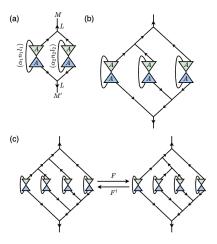


FIG. 27: Equivariant  $\lambda$ -SOAP generalization of the (a) power spectrum ( $\nu=2$ ) and (b) bispectrum ( $\nu=3$ ) structural kernels as tensor networks. (c) Topologically distinct kernel functions exists for either ( $L=0, \nu>3$ ) or ( $L>0, \nu>2$ ), determined by the repeated application of F-symbols to the pairs of descriptor fusion trees.

The above discussion has focused on the physically important case of global SO(3) covariance. However, this recoupling structure can be further generalized. For example, as described in [13], the inclusion of classical onsite magnetic moments,  $\mathbf{m}_i$ , either through spin or orbital degrees of freedom can be accommodated by extending the symmetry of the atom-centered descriptors to SO(3)  $\otimes$  SO(3). Note that while bond,  $\kappa = (nlm)$ , and magnetic,  $\tilde{\kappa} = (\tilde{n}l\tilde{m})$ , degrees of freedom transform independently in the atom-centered tensors,  $A_{\alpha\kappa\tilde{\kappa}}$ , they

must be recoupled to form a descriptor with definite, total angular momentum. The general construction of such a combined equivariant descriptor is shown in Figure 28a,b. Another example is provided by an extension of the SOAP method to the SO(4) group [2], otherwise known as the SNAP method [3], which circumvents the need to explicitly specify a radial basis function,  $R_{nl}(r)$ , by embedding the radial degree of freedom into the 3sphere,  $S^3$ . This amounts to treating the radial component as an additional polar angle that, along with the two angles inherited from  $S^2$ , describes rotations in  $\mathbb{R}^4$ . The hyperspherical harmonic functions,  $U^l_{m\widetilde{m}}$ , form a complete Fourier basis for functions on  $S^3$ . Since, at the level of Lie algebras, there exists an isomorphism SO(4)  $\sim SO(3) \otimes SO(3)$ , the SO(4) recoupling CG coefficients, H, decompose into products of SO(3) CG coefficients,  $H_{m_1\tilde{m}_1m_2\tilde{m}_2m_3\tilde{m}_3}^{(l_1l_2l_3)} = C_{m_1m_2m_3}^{(l_1l_2l_3)}C_{\tilde{m}_1\tilde{m}_2\tilde{m}_3}^{(l_1l_2l_3)}$  [2, 115, 116]. Recoupling in the SNAP bispectrum using this "parabolictype" decomposition [115] is shown in Figure 28c,d.

#### Appendix D: Additional computational details

SOAP power spectra were computed with the DSCRIBE software package [117]. GPR models were trained with the SCIKIT-LEARN library [118] using expectation maximization. The construction and training of deep neural networks and tensor network models were facilitated by the TensorNetwork [119] and TensorFlow [120]

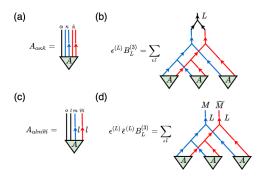


FIG. 28: Recoupling schemes for (a-b)  $SO(3) \otimes SO(3)$  and (c-d) SO(4) equivariant descriptors.

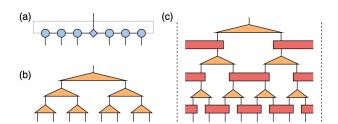


FIG. 29: Tensor network factorizations corresponding to (a) a matrix product state (MPS), (b) a tree tensor network (TTN), and (c) a Multi-Scale Entanglement Renormalization Ansatz (MERA).

libraries. These models were trained using the Adam optimization algorithm, a variant of stochastic gradient descent that adaptively updates the learning rates based on moments of the gradients [121].

#### ACKNOWLEDGMENTS

This research was supported through the UW Molecular Engineering Materials Center, a Materials Research Science and Engineering Center (Grant No. DMR-1719797) and was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

- J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98, 146401 (2007).
- [2] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, Physical Review B 87, 184115 (2013).
- [3] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, Journal of Computational Physics 285, 316 (2015).
- [4] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Modeling & Simulation 14, 1153 (2016).
- [5] H. Huo and M. Rupp, Unified representation for machine learning of molecules and crystals, arXiv preprint arXiv:1704.06439 13754 (2017).
- [6] A. Shapeev, Accurate representation of formation energies of crystalline alloys with many components, Computational Materials Science 139, 26 (2017).
- [7] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Physical Review B 99, 014104 (2019).
- [8] M. J. Willatt, F. Musil, and M. Ceriotti, Atomdensity representations for machine learning, The Journal of Chemical Physics 150, 10.1063/1.5090481 (2019), 1807.00408.
- [9] J. Behler, Four generations of high-dimensional neural network potentials, Chemical Reviews 121, 10037 (2021), pMID: 33779150, https://doi.org/10.1021/acs.chemrev.0c00868.
- [10] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csányi, D. W. Wingate, and G. L. W. Hart, Machine-learned multisystem surrogate models for materials prediction, npj Computational Materials 5, 10.1038/s41524-019-0189-9 (2019), 1809.09203.
- [11] C. Sutton, L. M. Ghiringhelli, T. Yamamoto, Y. Lyso-gorskiy, L. Blumenthal, T. Hammerschmidt, J. R. Golebiowski, X. Liu, A. Ziletti, and M. Scheffler, Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition, npj Computational Materials 5, 111 (2019).
- [12] M. F. Langer, A. Goeßmann, and M. Rupp, Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning, arXiv (2020), 2003.12081.
- [13] R. Drautz, Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer, Phys. Rev. B 102, 024104 (2020).
- [14] M. Bachmayr, G. Csanyi, R. Drautz, G. Dusson, S. Etter, C. v. d. Oord, and C. Ortner, Atomic Cluster Expansion: Completeness, Efficiency and Stability, arXiv (2019), 1911.03550.
- [15] Y. Lysogorskiy, C. v. d. Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, and R. Drautz, Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon, npj Computational Materials 7, 97 (2021).

- [16] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, Schnet—a deep learning architecture for molecules and materials, The Journal of Chemical Physics 148, 241722 (2018).
- [17] T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120, 145301 (2018).
- [18] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chemistry of Materials 31, 3564 (2019).
- [19] J. Klicpera, J. Groß, and S. Günnemann, Directional message passing for molecular graphs, arXiv preprint arXiv:2003.03123 (2020).
- [20] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds, arXiv preprint arXiv:1802.08219 (2018).
- [21] F. Fuchs, D. Worrall, V. Fischer, and M. Welling, Se(3)-transformers: 3d roto-translation equivariant attention networks, Advances in Neural Information Processing Systems 33, 1970 (2020).
- [22] P. Mehta and D. J. Schwab, An exact mapping between the Variational Renormalization Group and Deep Learning, arXiv (2014), 1410.3831.
- [23] H. W. Lin, M. Tegmark, and D. Rolnick, Why does deep and cheap learning work so well?, Journal of Statistical Physics, 1 (2016), 1608.08225.
- [24] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, Relational inductive biases, deep learning, and graph networks, arXiv (2018), 1806.01261.
- [25] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov, Tensorizing Neural Networks, arXiv (2015), 1509.06569.
- [26] M. E. Stoudenmire and D. J. Schwab, Supervised Learning with Quantum-Inspired Tensor Networks, arXiv (2016), 1605.05775.
- [27] Z.-Y. Han, J. Wang, H. Fan, L. Wang, and P. Zhang, Unsupervised generative modeling using matrix product states, Phys. Rev. X 8, 031012 (2018).
- [28] S. Efthymiou, J. Hidary, and S. Leichenauer, TensorNetwork for Machine Learning, arXiv (2019), 1906.06329.
- [29] J. Martyn, G. Vidal, C. Roberts, and S. Leichenauer, Entanglement and Tensor Networks for Supervised Image Classification, arXiv (2020), 2007.06082.
- [30] I. Glasser, N. Pancotti, and J. I. Cirac, From Probabilistic Graphical Models to Generalized Tensor Networks for Supervised Learning, IEEE Access 8, 68169 (2020).
- [31] J. Wang, C. Roberts, G. Vidal, and S. Leichenauer, Anomaly detection with tensor networks, arXiv preprint arXiv:2006.02516 (2020).
- [32] Z.-F. Gao, S. Cheng, R.-Q. He, Z. Y. Xie, H.-H. Zhao, Z.-Y. Lu, and T. Xiang, Compressing deep neural networks by matrix product operators, Phys. Rev. Research

- **2**, 023300 (2020).
- [33] J. Reyes and M. Stoudenmire, A multi-scale tensor network architecture for classification and regression, arXiv preprint arXiv:2001.08286 (2020).
- [34] S. Cheng, L. Wang, and P. Zhang, Supervised learning with projected entangled pair states, Phys. Rev. B 103, 125117 (2021).
- [35] Y. Levine, D. Yakira, N. Cohen, and A. Shashua, Deep Learning and Quantum Entanglement: Fundamental Connections with Implications to Network Design, arXiv (2017), 1704.01552.
- [36] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, Quantum Entanglement in Deep Learning Architectures, Physical Review Letters 122, 065301 (2019).
- [37] S. Singh, R. N. C. Pfeifer, and G. Vidal, Tensor network decompositions in the presence of a global symmetry, Phys. Rev. A 82, 050301 (2010).
- [38] S. Singh and G. Vidal, Tensor network states and algorithms in the presence of a global su(2) symmetry, Phys. Rev. B 86, 195114 (2012).
- [39] P. Schmoll, S. Singh, M. Rizzi, and R. Orús, A programming guide for tensor networks with global su(2) symmetry, Annals of Physics 419, 168232 (2020).
- [40] J. D. Biamonte, S. R. Clark, and D. Jaksch, Categorical Tensor Network States, AIP Advances 1, 042172 (2011), 1012.0531.
- [41] R. Penrose, Angular momentum: an approach to combinatorial space-time, Quantum Theory and Beyond, 151 (1971).
- [42] R. Oeckl, Discrete Gauge Theory (Imperial College Press, London, 2005).
- [43] Most of the graphical machinery discussed below carries over more generally to other symmetric tensor categories, of which finite-dimensional representations of SO(3) are an example.
- [44] More generally, the trivial representation is the unital object in category  $\text{Rep}(\mathcal{G})$  of representations of the group  $\mathcal{G}$ .
- [45] G. E. Stedman, Diagram Techniques in Group Theory (Cambridge University Press, Cambridge, 1990).
- [46] P. Cvitanović, Group Theory: Birdtracks, Lie's, and Exceptional Groups (Princeton University Press, New Jersey, 2008).
- [47] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, Annals of physics 326, 96 (2011).
- [48] I. Glasser, R. Sweke, N. Pancotti, J. Eisert, and I. J. Cirac, Expressive power of tensor-network factorizations for probabilistic modeling, with applications from hidden Markov models to quantum machine learning, arXiv (2019), 1907.03741.
- [49] This is similar to the treatment of thermal states as path integrals supported on a compact manifold [122].
- [50] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, Annals of physics 349, 117 (2014).
- [51] J. C. Bridgeman and C. T. Chubb, Hand-waving and interpretive dance: an introductory course on tensor networks, Journal of physics A: Mathematical and theoretical 50, 223001 (2017).
- [52] R. Orús, Tensor networks for complex quantum systems, Nature Reviews Physics 1, 538 (2019).
- [53] J. I. Cirac, D. Pérez-García, N. Schuch, and F. Verstraete, Matrix product states and projected entangled

- pair states: Concepts, symmetries, theorems, Rev. Mod. Phys. **93**, 045003 (2021).
- [54] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Incompleteness of atomic structure representations, Phys. Rev. Lett. 125, 166001 (2020).
- [55] J. Nigam, S. Pozdnyakov, and M. Ceriotti, Recursive evaluation and iterative contraction of N-body equivariant features, The Journal of Chemical Physics 153, 121101 (2020), 2007.03407.
- [56] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, Comparing molecules and solids across structural and alchemical space, Physical Chemistry Chemical Physics 18, 10.1039/c6cp00415f (2016), 1601.04077.
- [57] M. J. Willatt, F. Musil, and M. Ceriotti, Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements, Physical Chemistry Chemical Physics 20, 29661 (2018), 1807.00236.
- [58] N. Artrith, A. Urban, and G. Ceder, Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species, Phys. Rev. B 96, 014112 (2017).
- [59] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, Automatic selection of atomic fingerprints and reference configurations for machinelearning potentials, The Journal of Chemical Physics 148, 241730 (2018), 1804.02150.
- [60] The general copying of vectors and tensors by a standard tensor network on a finite-sized Hilbert space is forbidden by a no-cloning theorem.
- [61] S. X. Cui, M. H. Freedman, O. Sattath, R. Stong, and G. Minton, Quantum Max-flow/Min-cut, Journal of Mathematical Physics 57, 062206 (2016).
- [62] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, Benchmarking materials property prediction methods: the mathench test set and automatminer reference algorithm, npj Computational Materials 6, 1 (2020).
- [63] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, et al., Performance and cost assessment of machine learning interatomic potentials, The Journal of Physical Chemistry A 124, 731 (2020).
- [64] C. Chen, Y. Zuo, W. Ye, X. Li, and S. P. Ong, Learning properties of ordered and disordered materials from multi-fidelity data, Nature Computational Science 1, 46 (2021)
- [65] Y.-Y. Shi, L.-M. Duan, and G. Vidal, Classical simulation of quantum many-body systems with a tree tensor network, Phys. Rev. A 74, 022320 (2006).
- [66] G. Vidal, Entanglement renormalization, Phys. Rev. Lett. 99, 220405 (2007).
- [67] I. Convy, W. Huggins, H. Liao, and K. B. Whaley, Mutual Information Scaling for Tensor Network Machine Learning, arXiv (2021), 2103.00105.
- [68] S. Lu, M. Kanász-Nagy, I. Kukuljan, and J. I. Cirac, Tensor networks and efficient descriptions of classical data, arXiv (2021), 2103.06872.
- [69] E. van Nieuwenburg and O. Zilberberg, Entanglement spectrum of mixed states, Phys. Rev. A 98, 012327 (2018).
- [70] H. Weimer, A. Kshetrimayum, and R. Orús, Simulation methods for open quantum many-body systems, Rev. Mod. Phys. 93, 015008 (2021).

- [71] M.-D. Choi, Completely positive linear maps on complex matrices, Linear algebra and its applications 10, 285 (1975).
- [72] A. Jamiołkowski, Linear transformations which preserve trace and positive semidefiniteness of operators, Reports on Mathematical Physics 3, 275 (1972).
- [73] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning (Springer, New York, 2009).
- [74] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine learning practice and the bias-variance trade-off, arXiv (2018), 1812.11118.
- [75] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep Double Descent: Where Bigger Models and More Data Hurt, arXiv (2019), 1912.02292.
- [76] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma, Optimal Regularization Can Mitigate Double Descent, arXiv (2020), 2003.01897.
- [77] D. N. Page, Average entropy of a subsystem, Phys. Rev. Lett. 71, 1291 (1993).
- [78] H. Li and F. D. M. Haldane, Entanglement spectrum as a generalization of entanglement entropy: Identification of topological order in non-abelian fractional quantum hall effect states, Phys. Rev. Lett. 101, 010504 (2008).
- [79] C. Chamon, A. Hamma, and E. R. Mucciolo, Emergent irreversibility and entanglement spectrum statistics, Phys. Rev. Lett. 112, 240501 (2014).
- [80] D. Shaffer, C. Chamon, A. Hamma, and E. R. Mucciolo, Irreversibility and entanglement spectrum statistics in quantum circuits, Journal of Statistical Mechanics: Theory and Experiment 2014, P12007 (2014).
- [81] Z.-C. Yang, A. Hamma, S. M. Giampaolo, E. R. Mucciolo, and C. Chamon, Entanglement complexity in quantum many-body dynamics, thermalization, and localization, Phys. Rev. B 96, 020408 (2017).
- [82] L. Zhang, J. A. Reyes, S. Kourtis, C. Chamon, E. R. Mucciolo, and A. E. Ruckenstein, Nonuniversal entanglement level statistics in projection-driven quantum circuits, Phys. Rev. B 101, 235104 (2020).
- [83] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the hamiltonian variational ansatz, PRX Quantum 1, 020319 (2020).
- [84] J. Iaconis, Quantum state complexity in computationally tractable quantum circuits, PRX Quantum 2, 010329 (2021).
- [85] M. L. Mehta, Random Matrix Theory (Elsevier, New York, 2004).
- [86] V. Oganesyan and D. A. Huse, Localization of interacting fermions at high temperature, Phys. Rev. B 75, 155111 (2007).
- [87] A. Pal and D. A. Huse, Many-body localization phase transition, Phys. Rev. B 82, 174411 (2010).
- [88] Y. Y. Atas, E. Bogomolny, O. Giraud, and G. Roux, Distribution of the ratio of consecutive level spacings in random matrix ensembles, Phys. Rev. Lett. 110, 084101 (2013).
- [89] S. Garnerone, T. R. de Oliveira, S. Haas, and P. Zanardi, Statistical properties of random matrix product states, Phys. Rev. A 82, 052312 (2010).
- [90] J. Haferkamp, C. Bertoni, I. Roth, and J. Eisert, Emergent statistical mechanics from properties of disordered random matrix product states, PRX Quantum 2, 040308 (2021).

- [91] Z. Liu, L.-W. Yu, L. M. Duan, and D.-L. Deng, The Presence and Absence of Barren Plateaus in Tensor-network Based Machine Learning, arXiv (2021), 2108.08312.
- [92] A. W. Harrow and R. A. Low, Random Quantum Circuits are Approximate 2-designs, Communications in Mathematical Physics 291, 257 (2009), 0802.1919.
- [93] M. Žnidarič, Entanglement of random vectors, Journal of Physics A: Mathematical and Theoretical 40, F105 (2006).
- [94] P. J. Forrester, Log-Gases and Random Matrices (Princeton University Press, New Jersey, 2010).
- [95] Z.-C. Yang, C. Chamon, A. Hamma, and E. R. Mucciolo, Two-component structure in the entanglement spectrum of highly excited states, Phys. Rev. Lett. 115, 267206 (2015).
- [96] H. Yao and X.-L. Qi, Entanglement entropy and entanglement spectrum of the kitaev model, Phys. Rev. Lett. 105, 080501 (2010).
- [97] J. de Boer, J. Järvelä, and E. Keski-Vakkuri, Aspects of capacity of entanglement, Phys. Rev. D 99, 066012 (2019).
- [98] Z. Allen-Zhu, Y. Li, and Z. Song, A Convergence Theory for Deep Learning via Over-Parameterization, arXiv (2018), 1811.03962.
- [99] Y. Li and Y. Liang, Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data, arXiv (2018), 1808.01204.
- [100] B. T. Kiani, S. Lloyd, and R. Maity, Learning Unitaries by Gradient Descent, arXiv (2020), 2001.11897.
- [101] R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information, arXiv preprint arXiv:1703.00810 (2017).
- [102] G. Hinton and S. T. Roweis, Stochastic neighbor embedding, in *NIPS*, Vol. 15 (Citeseer, 2002) pp. 833–840.
- [103] L. van der Maaten and G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9, 2579 (2008).
- [104] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Physics Reports 810, 1 (2019), a highbias, low-variance introduction to Machine Learning for physicists.
- [105] G. Evenbly and G. Vidal, Class of Highly Entangled Many-Body States that can be Efficiently Simulated, Physical Review Letters 112, 240502 (2014), 1210.1895.
- [106] A. Glielmo, C. Zeni, and A. D. Vita, Efficient nonparametric n-body force fields from machine learning, Physical Review B 97, 184307 (2018), 1801.04823.
- [107] F. A. Faber, A. S. Christensen, B. Huang, and O. A. v. Lilienfeld, Alchemical and structural distribution based representation for universal quantum machine learning, The Journal of Chemical Physics 148, 241717 (2018).
- [108] G. Evenbly and S. R. White, Entanglement renormalization and wavelets, Phys. Rev. Lett. 116, 140403 (2016).
- [109] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nature communications 9, 1 (2018).
- [110] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nature Communications 12, 1791 (2021), 2001.00550.

- [111] B. Onat, C. Ortner, and J. R. Kermode, Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials, The Journal of Chemical Physics 153, 144106 (2020).
- [112] Z. Wang, Topological quantum computation, 112 (American Mathematical Soc., 2010).
- [113] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Symmetry-Adapted Machine-Learning for Tensorial Properties of Atomistic Systems, Physical Review Letters 120, 036002 (2017), 1709.06757.
- [114] A. Perez, The Spin-Foam Approach to Quantum Gravity, Living Reviews in Relativity 16, 3 (2013), 1205.2019.
- [115] A. V. Meremianin, Multipole expansions in fourdimensional hyperspherical harmonics, Journal of Physics A: Mathematical and General 39, 3099 (2006), math-ph/0510080.
- [116] M. A. Caprio, K. D. Sviratcheva, and A. E. McCoy, Racah's method for general subalgebra chains: Coupling coefficients of SO(5) in canonical and physical bases, Journal of Mathematical Physics 51, 093518 (2010), 1006.2875.
- [117] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, DScribe: Library of descriptors for machine learning in materials science, Computer Physics Communications 247, 106949 (2020).

- [118] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12, 2825 (2011).
- [119] C. Roberts, A. Milsted, M. Ganahl, A. Zalcman, B. Fontaine, Y. Zou, J. Hidary, G. Vidal, and S. Leichenauer, Tensornetwork: A library for physics and machine learning (2019), arXiv:1905.01330 [physics.compph].
- [120] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: Largescale machine learning on heterogeneous systems (2015), software available from tensorflow.org.
- [121] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [122] G. Evenbly and G. Vidal, Tensor network renormalization yields the multiscale entanglement renormalization ansatz, Phys. Rev. Lett. 115, 200401 (2015).