

Hamilton-Zero: A Neural Tensor-Network Foundation Model for Ground States of Arbitrary Quadratic Qubit Hamiltonians

Timothy Heightman^{1,2,*} Elena Orlova² Phillip Mantrov² Aleksei Ustimenko^{2,*}

¹ICFO – Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology,
08860 Castelldefels, Barcelona, Spain

²Simulacra Research Inc., London, UK & Chicago, USA
timothy.heightman@icfo.eu aleksei@simulacra-ai.com

Abstract

We formulate spin-1/2 quantum ground-state learning as manifold variational optimization over centrally odd scalar functions on $SU(2)^N$. This replaces explicit Hilbert-space vector amplitudes with manifold functions acted on by Lie-derivative Hamiltonians via automatic differentiation. We prove a novel variational principle in this function-space that preserves the spin-1/2 manifold variational principle, then train our foundation model on hundreds of thousands of different Hamiltonian systems, varying the connection topology, system size, and edge weights, bringing together 100 years of many-body literature to construct our own training datasets. We construct a state-of-the-art sampling scheme for $SU(2)$ replica-exchange Langevin sampling and sharded natural-gradient optimization, extending the KFAC optimizer to rank-four tensors in order to efficiently train the model on small systems. Our foundation model then generalises to unseen Hamiltonian system topologies *and* unseen system scales up to 100+ qubits. Preprint in preparation in which will open-source code and model weights.

Manifold Wavefunctions and Energies via Autograd

A central problem in quantum many-body physics is to approximate the ground state of a universal [1] spin-1/2 Hamiltonian

$$\hat{H} = \frac{1}{4} \sum_{i < j} \sum_{a,b} J_{ij}^{ab} \hat{\sigma}_i^a \hat{\sigma}_j^b + \frac{1}{4} \sum_i \sum_a h_i^a \hat{\sigma}_i^a, \quad (1)$$

where J_{ij}^{ab} encodes arbitrary quadratic Pauli interactions and h_i^a encodes local fields. Standard Neural Quantum States (NQS) [2–4] learn amplitudes on the discrete spin basis $\{\uparrow, \downarrow\}^N$. This makes each model tied to one system size and one Hamiltonian instance, so transfer across Hamiltonian structure is difficult [5–7]. Only recently researchers have tried to start generalising within a given Hamiltonian interaction topology [3, 5–8]. Hamilton-Zero instead represents the state as a smooth function on the product Lie group $\psi_\theta : SU(2)^N \rightarrow \mathbb{C}$. Each spin is parametrised by a unit quaternion $q_i \in SU(2) \cong S^3 \hookrightarrow \mathbb{R}^4$, so the network consumes $4N$ continuous coordinates rather than the discrete basis of NQS [2]. Thus spin operators can act on this manifold through left-invariant vector fields $L_i^a, \hat{\sigma}_i^a \longleftrightarrow -iL_i^a$ for first order and for second, $\hat{\sigma}_i^a \hat{\sigma}_j^b \longleftrightarrow -L_i^a L_j^b$. The Hamiltonian becomes a second-order differential operator on $SU(2)^N$, which we can compute with automatic differentiation of the model with respect to its input coordinates,

$$\hat{H} \longleftrightarrow -\frac{1}{4} \sum_{i < j, a, b} J_{ij}^{ab} L_i^a L_j^b - \frac{i}{4} \sum_{i, a} h_i^a L_i^a, \quad (2)$$

extending recent works on moment generating functions [9] to the manifold function directly via custom jax-primitives of these Lie derivatives. Hence, the Hamiltonian’s interaction data is a coupling tensor $(J, h) \in \mathbb{R}^{N \times N \times 9} \times \mathbb{R}^{3N}$, while its action on states is evaluated by autograd of the manifold wavefunction’s inputs.

*These two authors contributed equally.

Spin-1/2 Variational Principle in $L^2(\text{SU}(2)^N)$

By moving from amplitudes to functions on $\text{SU}(2)^N$, we have a space $L^2(\text{SU}(2)^N)$ which is larger than the physical spin-1/2 Hilbert space. However, the Peter–Weyl theorem [10, 11] gives us a spectrum for this function space,

$$L^2(\text{SU}(2)) = \bigoplus_{j=0, \frac{1}{2}, 1, \dots} V_j \otimes V_j^*, \quad (3)$$

decomposing it site-wise into irreducible spin sectors, V_j . The eigenfunctions in the physical subspace satisfy $\Delta_i \psi = \frac{3}{4} \psi$ for every site i , where Δ_i is the Casimir (Laplace–Beltrami) operator on each local set of $\text{SU}(2)$ coordinates [12, 13]. All physical spin-1/2 functions on this manifold satisfy this Casimir relation, however an unconstrained neural function on $\text{SU}(2)^N$ may leak into higher j sectors giving unphysical energies below the ground state. To resolve this we impose per-site oddness $\psi(\dots, -q_i, \dots) = -\psi(\dots, q_i, \dots)$.

This removes all integer- j sectors and leaves only half-integer sectors $j = 1/2, 3/2, 5/2, \dots$, with the physical spin-1/2 sector now the lowest allowed sector. Indeed, for arbitrary spin- j the eigenspectrum of Δ_i has eigenvalues $j(j+1)$, which increase quadratically in j [10, 13]. As such, by adding a sufficiently large energy penalty for support outside $j = 1/2$, we provably recover the variational principle of the spin-1/2 sector alone. Our overall energy then reads,

$$E_\lambda[\psi] = \frac{\langle \psi | \hat{H} | \psi \rangle}{\langle \psi | \psi \rangle} + \lambda C[\psi]. \quad (4)$$

where the Casimir penalty is $C[\psi] = \sum_i \langle \Delta_i \psi / \psi - 3/4 \rangle$, and we optimize the regularized objective. Each Hamiltonian in our datasets has a different $\lambda = \lambda(J, h)$, depending only on the interaction tensor and fields, and λ is efficiently computable for every (J, h) such that we can saturate the inequality, recovering true ground-state energies with no gap in principle.

Hamiltonian-Conditioned Foundation Ansatz

The architecture has two inputs: a sampled spin configuration $q \in \text{SU}(2)^N$ and a Hamiltonian (J, h) , passing through different routes of the model in order to preserve the parity symmetry described above, see Fig. 1.

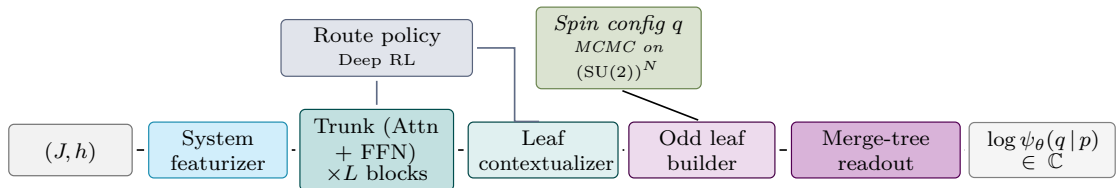


Figure 1: Hamilton-Zero information flow. The Hamiltonian is encoded first; spin coordinates enter only at the odd leaf builder, enforcing central oddness before the merge-tree readout returns $\log \psi_\theta(q | p)$.

The Hamiltonian data-stream takes the raw interaction tensors and creates learnable embedded features for L -blocks of residual attention and feed-forward blocks. Crucially, the spin coordinates q_i do not enter this trunk in order to preserve parity symmetry, meaning it acts only as an encoder for a given Hamiltonian itself. Readout of the model is then done through a deep-reinforcement learning algorithm which contracts the leaves of a tree of quaternionic coordinates that are merged into a complex scalar value. This is akin to tree-tensor networks [14–16], however the deep reinforcement learning (RL) component [17, 18] (see Fig. 1) enables us to entangle arbitrary spins since the leaves of the tree are positioned via a routing policy that is conditioned on the Hamiltonian data through a leaf contextualiser. We also use lateral self-attention during contraction, meaning the features arising from (J, h) propagate through the entire contraction.

Datasets, Training and Results

To train our foundation model, we collated hundreds of thousands of different Hamiltonian interaction topologies, spanning 100 years of quantum many-body and quantum computing literature [19–65]. Because the corpus also includes Hamiltonians native to superconducting, trapped-ion, Rydberg, spin-qubit, and molecular platforms, Hamilton-Zero can be evaluated as a classical foundation surrogate for hardware-relevant ground-state problems.

To train efficiently, we devised a state-of-the-art sampling scheme for MCMC on the configuration manifold $SU(2)^N$ which extends Metropolis Adjusted Langevin [66] and DEO replica exchange [67] with a novel online temperature adjustment. We also developed our own Jaxpr [68] evaluator implementing second order differential operator corresponding to energy that exploits TN structure, to achieve effectively near linear scaling of local energy evaluation. Furthermore, we developed our own multi-head self-attention Pallas kernels for JVP, in contrast to pre-existing VJP kernels [69–72]. Finally, we extended the rank of the KFAC [73] optimiser to take into account the merging strategy in the tree readout, adding custom blocks and a 3-way factorisation that accounts for stacked layers’ curvature.

Here we show some examples of how Hamilton-Zero amortizes training on unseen systems. Generalisation applies to (i) new weights in the interaction graphs of systems seen in training, (ii) larger system sizes for interaction graphs seen in training and (iii) completely unseen interaction topologies. Fig. 2 shows generalisation for (i) in the top-two rows, where we sweep the XXZ model over its critical point for $N = 32$ and 64 spins, and compare to the exact Bethe-ansatz solution. Fig. 2(c) shows generalisation of the XXZ model, Kagome and triangular lattices to large system sizes (> 64), not seen in the training data, whilst (d) shows a comparison between zero-shot evaluation, fine-tuning, and training from scratch on an in-distribution system (triangular lattice) and an unseen topology, the Mobius-Kantor graph at $N = 16$ spins.

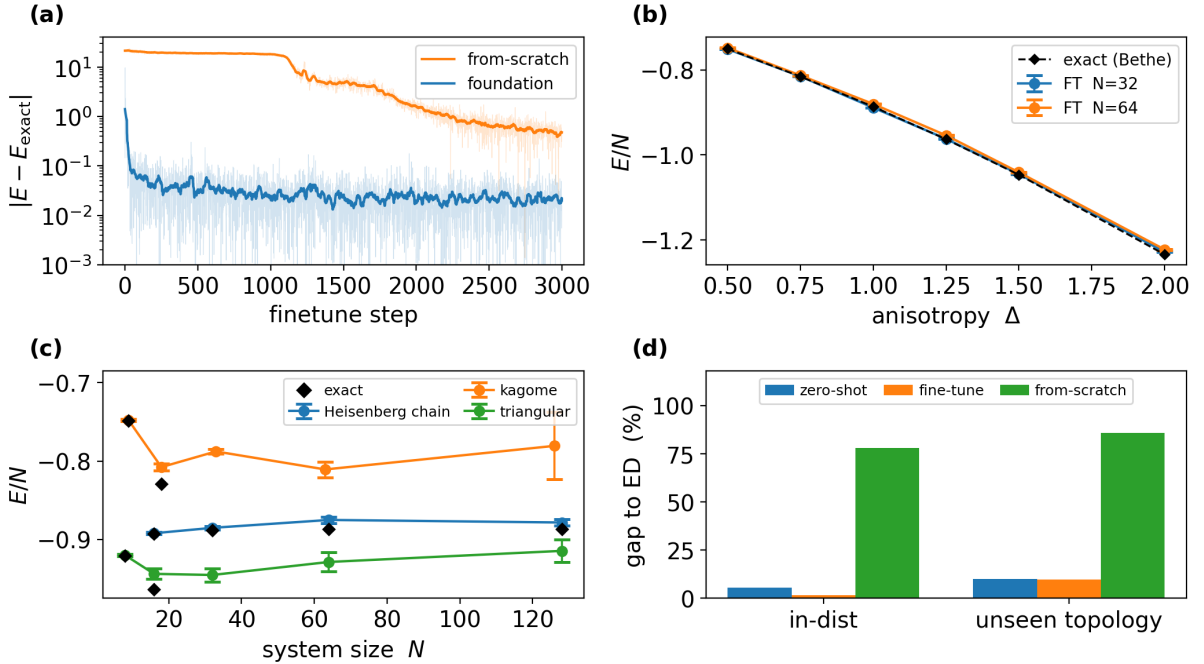


Figure 2: A recent checkpoint of the model’s weights with 3000 steps of fine-tuning and training from scratch for 3000 steps. (a) shows the gap to exact values for $N = 16$ in the Heisenberg XXZ model, (b) shows a sweep of the system for $N = 32$ and $N = 64$ through its phase transition at $\Delta = 1$. (c) shows generalisation to unseen system sizes, and (d) shows zero-shot evaluation for the triangular lattice (in-dist) and the Mobius-Kantor graph (unseen topology). Here zero-shot means the model was evaluated with no further training of the weights.

References

- [1] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, 10th anniversary edition, 2010. ISBN 9781107002173. doi: 10.1017/CBO9780511976667.
- [2] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. doi: 10.1126/science.aag2302.
- [3] Yuan-Hang Zhang and Massimiliano Di Ventra. Transformer quantum state: A multipurpose model for quantum many-body problems. *Physical Review B*, 107(7):075147, 2023. doi: 10.1103/PhysRevB.107.075147.
- [4] Sebastián Roca-Jerat, Manuel Gallego, Fernando Luis, Jesús Carrete, and David Zueco. Transformer wave function for quantum long-range models. *Physical Review B*, 110(20):205147, 2024. doi: 10.1103/PhysRevB.110.205147.
- [5] Riccardo Rende, Sebastian Goldt, Federico Becca, and Luciano Loris Viteritti. Fine-tuning neural network quantum states. *Physical Review Research*, 6(4):043280, 2024. doi: 10.1103/PhysRevResearch.6.043280.
- [6] Riccardo Rende, Luciano Loris Viteritti, Federico Becca, Antonello Scardicchio, Alessandro Laio, and Giuseppe Carleo. Foundation neural-networks quantum states as a unified ansatz for multiple hamiltonians. *Nature Communications*, 16(1):7213, 2025. doi: 10.1038/s41467-025-62098-x.
- [7] Luciano Loris Viteritti, Riccardo Rende, Giacomo Bracci Testasecca, Jacopo Niedda, Roderich Moessner, Giuseppe Carleo, and Antonello Scardicchio. Quantum spin glass in the two-dimensional disordered heisenberg model via foundation neural-network quantum states, 2025.
- [8] Timothy Zaklama, Daniele Guerri, and Liang Fu. Attention-based foundation model for quantum states, 2025.
- [9] Timothy Heightman, Edward Jiang, Ruth Mora-Soto, Maciej Lewenstein, and Marcin Płodzień. Quantum machine learning in multi-qubit phase-space part i: Foundations, 2025.
- [10] Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, Cham, 2 edition, 2015. doi: 10.1007/978-3-319-13467-3. Key retained as hall2013lie for compatibility with the manuscript.
- [11] F. Peter and H. Weyl. Die vollständigkeit der primitiven darstellungen einer geschlossenen kontinuierlichen gruppe. *Mathematische Annalen*, 97:737–755, 1927. URL <https://eudml.org/doc/182662>.
- [12] Gerald B. Folland. *A Course in Abstract Harmonic Analysis*. CRC Press, Boca Raton, 2 edition, 2016. ISBN 9781498727136. doi: 10.1201/b19172.
- [13] Anthony W. Knap. *Representation Theory of Semisimple Groups: An Overview Based on Examples*, volume 36 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 2001. ISBN 9780691090894. Paperback reprint with a new preface; originally published in 1986.
- [14] Yao-Yun Shi, Lu-Ming Duan, and Guifré Vidal. Classical simulation of quantum many-body systems with a tree tensor network. *Physical Review A*, 74(2):022320, 2006. doi: 10.1103/PhysRevA.74.022320.

- [15] Valentin Murg, Örs Legeza, Reinhard M. Noack, and Frank Verstraete. Simulating strongly correlated quantum systems with tree tensor networks. *Physical Review B*, 82(20):205105, 2010. doi: 10.1103/PhysRevB.82.205105.
- [16] Naoki Nakatani and Garnet Kin-Lic Chan. Efficient tree tensor network states (TTNS) for quantum chemistry: Generalizations of the density matrix renormalization group algorithm. *Journal of Chemical Physics*, 138(13):134113, 2013. doi: 10.1063/1.4798639.
- [17] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 1998. ISBN 9780262193986.
- [18] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *Foundations and Trends in Machine Learning*, 11(3–4):219–354, 2018. doi: 10.1561/22000000071.
- [19] P. W. Anderson. Absence of diffusion in certain random lattices. *Physical Review*, 109(5):1492–1505, 1958. doi: 10.1103/PhysRev.109.1492.
- [20] Elliott Lieb, Theodore Schultz, and Daniel Mattis. Two soluble models of an antiferromagnetic chain. *Annals of Physics*, 16(3):407–466, 1961. doi: 10.1016/0003-4916(61)90115-4.
- [21] J. Hubbard. Electron correlations in narrow energy bands. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 276(1365):238–257, 1963. doi: 10.1098/rspa.1963.0204.
- [22] Chanchal K. Majumdar and Dipan K. Ghosh. On next-nearest-neighbor interaction in linear chain. I. *Journal of Mathematical Physics*, 10(8):1388–1398, 1969. doi: 10.1063/1.1664978.
- [23] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925. doi: 10.1007/BF02980577.
- [24] W. P. Su, J. R. Schrieffer, and A. J. Heeger. Solitons in polyacetylene. *Physical Review Letters*, 42(25):1698–1701, 1979. doi: 10.1103/PhysRevLett.42.1698.
- [25] M. J. Rice and E. J. Mele. Elementary excitations of a linearly conjugated diatomic polymer. *Physical Review Letters*, 49(19):1455–1459, 1982. doi: 10.1103/PhysRevLett.49.1455.
- [26] Alexei Yu. Kitaev. Unpaired majorana fermions in quantum wires. *Physics-Uspekhi*, 44(10S):131–136, 2001. doi: 10.1070/1063-7869/44/10S/S29.
- [27] Xie Chen, Zheng-Cheng Gu, and Xiao-Gang Wen. Classification of gapped symmetric phases in one-dimensional spin systems. *Physical Review B*, 83(3):035107, 2011. doi: 10.1103/PhysRevB.83.035107.
- [28] Subir Sachdev. *Quantum Phase Transitions*. Cambridge University Press, Cambridge, 1999. ISBN 9780521582548. doi: 10.1017/CBO9780511622540.
- [29] Anders W. Sandvik. Evidence for deconfined quantum criticality in a two-dimensional heisenberg model with four-spin interactions. *Physical Review Letters*, 98(22):227202, 2007. doi: 10.1103/PhysRevLett.98.227202.
- [30] T. Senthil, Ashvin Vishwanath, Leon Balents, Subir Sachdev, and Matthew P. A. Fisher. Deconfined quantum critical points. *Science*, 303(5663):1490–1494, 2004. doi: 10.1126/science.1091806.

- [31] C. Monroe, W. C. Campbell, L.-M. Duan, Z.-X. Gong, A. V. Gorshkov, P. W. Hess, R. Islam, K. Kim, N. M. Linke, G. Pagano, P. Richerme, C. Senko, and N. Y. Yao. Programmable quantum simulations of spin systems with trapped ions. *Reviews of Modern Physics*, 93(2):025001, 2021. doi: 10.1103/RevModPhys.93.025001.
- [32] Antoine Browaeys and Thierry Lahaye. Many-body physics with individually controlled rydberg atoms. *Nature Physics*, 16:132–142, 2020. doi: 10.1038/s41567-019-0733-z.
- [33] T. Hensgens, T. Fujita, L. Janssen, Xiao Li, C. J. Van Diepen, C. Reichl, W. Wegscheider, S. Das Sarma, and L. M. K. Vandersypen. Quantum simulation of a fermi-hubbard model using a semiconductor quantum dot array. *Nature*, 548(7665):70–73, 2017. doi: 10.1038/nature23022.
- [34] U. Las Heras, A. Mezzacapo, L. Lamata, S. Filipp, A. Wallraff, and E. Solano. Digital quantum simulation of spin systems in superconducting circuits. *Physical Review Letters*, 112(20):200501, 2014. doi: 10.1103/PhysRevLett.112.200501.
- [35] P. W. Anderson. Resonating valence bonds: A new kind of insulator? *Materials Research Bulletin*, 8(2):153–160, 1973. doi: 10.1016/0025-5408(73)90167-0. Alias of anderson1973rvb kept because both keys occur in the manuscript.
- [36] Alexei Kitaev. Anyons in an exactly solved model and beyond. *Annals of Physics*, 321(1):2–111, 2006. doi: 10.1016/j.aop.2005.10.005.
- [37] P. W. Anderson. Resonating valence bonds: A new kind of insulator? *Materials Research Bulletin*, 8(2):153–160, 1973. doi: 10.1016/0025-5408(73)90167-0. Alias duplicate of anderson1973resonating kept because both keys occur in the manuscript.
- [38] Leon Balents. Spin liquids in frustrated magnets. *Nature*, 464(7286):199–208, 2010. doi: 10.1038/nature08917.
- [39] Lucile Savary and Leon Balents. Quantum spin liquids: a review. *Reports on Progress in Physics*, 80(1):016502, 2017. doi: 10.1088/0034-4885/80/1/016502. Published online 8 November 2016.
- [40] D. M. Basko, I. L. Aleiner, and B. L. Altshuler. Metal–insulator transition in a weakly interacting many-electron system with localized single-particle states. *Annals of Physics*, 321(5):1126–1205, 2006. doi: 10.1016/j.aop.2005.11.014.
- [41] Rahul Nandkishore and David A. Huse. Many-body localization and thermalization in quantum statistical mechanics. *Annual Review of Condensed Matter Physics*, 6(1):15–38, 2015. doi: 10.1146/annurev-conmatphys-031214-014726.
- [42] Dmitry A. Abanin, Ehud Altman, Immanuel Bloch, and Maksym Serbyn. Colloquium: Many-body localization, thermalization, and entanglement. *Reviews of Modern Physics*, 91(2):021001, 2019. doi: 10.1103/RevModPhys.91.021001.
- [43] Don N. Page. Average entropy of a subsystem. *Physical Review Letters*, 71(9):1291–1294, 1993. doi: 10.1103/PhysRevLett.71.1291.
- [44] H. J. Lipkin, N. Meshkov, and A. J. Glick. Validity of many-body approximation methods for a solvable model. I. exact solutions and perturbation theory. *Nuclear Physics*, 62(2):188–198, 1965. doi: 10.1016/0029-5582(65)90862-X.
- [45] Daniel M. Greenberger, Michael A. Horne, and Anton Zeilinger. Going beyond bell’s theorem. In Menas Kafatos, editor, *Bell’s Theorem, Quantum Theory and Conceptions of the Universe*, pages 69–72. Kluwer Academic Publishers, Dordrecht, 1989. doi: 10.1007/978-94-017-0849-4_10.

- [46] Kenneth G. Wilson. Confinement of quarks. *Physical Review D*, 10(8):2445–2459, 1974. doi: 10.1103/PhysRevD.10.2445.
- [47] John B. Kogut and Leonard Susskind. Hamiltonian formulation of wilson’s lattice gauge theories. *Physical Review D*, 11(2):395–408, 1975. doi: 10.1103/PhysRevD.11.395.
- [48] John B. Kogut. An introduction to lattice gauge theory and spin systems. *Reviews of Modern Physics*, 51(4):659–713, 1979. doi: 10.1103/RevModPhys.51.659.
- [49] Pascual Jordan and Eugene P. Wigner. Über das paulische äquivalenzverbot. *Zeitschrift für Physik*, 47(9–10):631–651, 1928. doi: 10.1007/BF01331938. English title: About the Pauli exclusion principle.
- [50] Sergey B. Bravyi and Alexei Yu. Kitaev. Fermionic quantum computation. *Annals of Physics*, 298(1):210–226, 2002. doi: 10.1006/aphy.2002.6254.
- [51] Andrew Tranter, Peter J. Love, Florian Mintert, and Peter V. Coveney. A comparison of the bravyi–kitaev and jordan–wigner transformations for the quantum simulation of quantum chemistry. *Journal of Chemical Theory and Computation*, 14(11):5617–5630, 2018. doi: 10.1021/acs.jctc.8b00450.
- [52] J. Goldstone. Field theories with superconductor solutions. *Il Nuovo Cimento*, 19:154–164, 1961. doi: 10.1007/BF02812722.
- [53] Yoichiro Nambu. Quasi-particles and gauge invariance in the theory of superconductivity. *Physical Review*, 117(3):648–663, 1960. doi: 10.1103/PhysRev.117.648.
- [54] Assa Auerbach. *Interacting Electrons and Quantum Magnetism*. Graduate Texts in Contemporary Physics. Springer, New York, 1994. ISBN 978-0-387-94286-5. doi: 10.1007/978-1-4612-0869-3.
- [55] Christopher J. Turner, Alexios A. Michailidis, Dmitry A. Abanin, Maksym Serbyn, and Zlatko Papić. Weak ergodicity breaking from quantum many-body scars. *Nature Physics*, 14:745–749, 2018. doi: 10.1038/s41567-018-0137-5.
- [56] Hannes Bernien, Sylvain Schwartz, Alexander Keesling, Harry Levine, Ahmed Omran, Hannes Pichler, Soonwon Choi, Alexander S. Zibrov, Manuel Endres, Markus Greiner, Vladan Vuletić, and Mikhail D. Lukin. Probing many-body dynamics on a 51-atom quantum simulator. *Nature*, 551(7682):579–584, 2017. doi: 10.1038/nature24622.
- [57] Ian Affleck, Tom Kennedy, Elliott H. Lieb, and Hal Tasaki. Valence bond ground states in isotropic quantum antiferromagnets. *Communications in Mathematical Physics*, 115: 477–528, 1988. doi: 10.1007/BF01218021. The related PRL titled "Rigorous results on valence-bond ground states in antiferromagnets" is 1987, DOI 10.1103/PhysRevLett.59.799.
- [58] C. N. Yang. Concept of off-diagonal long-range order and the quantum phases of liquid He and of superconductors. *Reviews of Modern Physics*, 34(4):694–704, 1962. doi: 10.1103/RevModPhys.34.694.
- [59] Subir Sachdev and Jinwu Ye. Gapless spin-fluid ground state in a random quantum heisenberg magnet. *Physical Review Letters*, 70(21):3339–3342, 1993. doi: 10.1103/PhysRevLett.70.3339.
- [60] Alexei Kitaev and S. Josephine Suh. The soft mode in the Sachdev–Ye–Kitaev model and its gravity dual. *Journal of High Energy Physics*, 2018(5):183, 2018. doi: 10.1007/JHEP05(2018)183.

- [61] Juan Maldacena and Douglas Stanford. Remarks on the sachdev-ye-kitaev model. *Physical Review D*, 94(10):106002, 2016. doi: 10.1103/PhysRevD.94.106002.
- [62] I. Dzyaloshinsky. A thermodynamic theory of “weak” ferromagnetism of antiferromagnetics. *Journal of Physics and Chemistry of Solids*, 4(4):241–255, 1958. doi: 10.1016/0022-3697(58)90076-3.
- [63] Tôru Moriya. Anisotropic superexchange interaction and weak ferromagnetism. *Physical Review*, 120(1):91–98, 1960. doi: 10.1103/PhysRev.120.91.
- [64] K. I. Kugel and D. I. Khomskii. The jahn-teller effect and magnetism: transition metal compounds. *Soviet Physics Uspekhi*, 25(4):231–256, 1982. doi: 10.1070/PU1982v025n04ABEH004537. The differently titled "Crystal structure and magnetic properties of substances with orbital degeneracy" is the related 1973 JETP paper.
- [65] Zohar Nussinov and Jeroen van den Brink. Compass models: Theory and physical motivations. *Reviews of Modern Physics*, 87(1):1–59, 2015. doi: 10.1103/RevModPhys.87.1.
- [66] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. doi: 10.2307/3318418.
- [67] Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-reversible parallel tempering: A scalable highly parallel MCMC scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350, 2022. doi: 10.1111/rssb.12464.
- [68] Roy Frostig, Matthew James Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. In *Conference on Systems and Machine Learning (SysML)*, 2018.
- [69] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. <http://github.com/google/jax>, 2018.
- [70] Philippe Tillet, H. T. Kung, and David Cox. Triton: An intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)*, 2019. doi: 10.1145/3315508.3329973.
- [71] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [72] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning, 2023.
- [73] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR, 2015. URL <https://proceedings.mlr.press/v37/martens15.html>.