A machine learning approach to duality in statistical physics

Prateek Gupta
Max Planck Institute
gupta@mpib-berlin.mpg.de

Andrea E. V. Ferrari
Deutsches Elektronen-Synchrotron DESY, Germany
School of Mathematics, The University of Edinburgh
andrea.e.v.ferrari@gmail.com

Nabil Iqbal
Department of Mathematical Sciences, Durham University.
Amsterdam Machine Learning Lab, University of Amsterdam.
nabil.iqbal@durham.ac.uk

Abstract

The notion of *duality* – that a given physical system can have two different mathematical descriptions – is a key idea in modern theoretical physics. Establishing a duality in lattice statistical mechanics models requires the construction of a dual Hamiltonian and a map from the original to the dual observables. By using neural networks to parameterize these maps and introducing a loss function that penalises the difference between correlation functions in original and dual models, we formulate the process of duality discovery as an optimization problem. We numerically solve this problem and show that our framework can rediscover the celebrated Kramers-Wannier duality for the 2d Ising model, numerically reconstructing the known mapping of temperatures.* We discuss future directions and prospects for discovering new dualities within this framework.

1 Background

A key concept in physics is *duality*, i.e. the idea that the same physical system can have two different mathematical descriptions. Duality sits at the heart of modern theoretical physics. In this work we seek to formalize the notion of duality in statistical physics in a manner that allows modern machine learning techniques to be used to systematically search for dualities.

Background on duality: Consider a statistical physics model with microstates σ and Hamiltonian functional $H[\beta, \sigma]$, where β are macroparameters such as the temperature. The model is determined by its partition function $Z = \sum_\sigma e^{-H[\beta,\sigma]}$. However, in nature we often have access to sets of expectation values of observables $O_\alpha(\sigma)$ (some real-valued functions of the microstates, e.g. correlation functions, with α being an arbitrary label)

$$\langle O_\alpha(\sigma) \rangle_H = \frac{1}{Z} \sum_\sigma O_\alpha(\sigma) \exp(-H[\beta, \sigma]). \tag{1}$$

It is a profound physical fact that occasionally there are alternative representations of these sets of correlation functions (see e.g. [1; 2; 3; 4; 5; 6; 7] for influential examples). That is, there exists another set of microstates $\tilde{\sigma}$, another Hamiltonian $\tilde{H}[\tilde{\beta}, \tilde{\sigma}]$ and for each observable $O_\alpha(\sigma)$ a dual

*Our implementation is publicly available at https://github.com/pg2455/duality_exploration

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

observable $\tilde{O}_\alpha(\tilde{\sigma}_i)$ such that

$$\langle O_\alpha(\sigma)\rangle_H = \langle \tilde{O}_\alpha(\tilde{\sigma})\rangle_{\tilde{H}}. \tag{2}$$

When this happens, we have a *duality* –the same physical system has at least two distinct mathematical descriptions, which may be useful for different reasons.

A prototypical example of such a duality is Kramers-Wannier duality for the 2d Ising model [2]. The 2d Ising model†consists of spins $\sigma_i = \pm 1$ living on the sites of a square lattice at temperature β^{-1}, with Hamiltonian that sums over neighbouring spins $\langle ij\rangle$

$$H[\beta,\sigma] = -\beta \sum_{\langle ij\rangle} \sigma_i \sigma_j. \tag{3}$$

It is a remarkable fact that the model described by (3) is precisely equivalent to a different 2d Ising model model with spins $\tilde{\sigma}_i = \pm 1$ living on the *dual* lattice, with a dual Hamiltonian of the same functional form $\tilde{H}[\tilde{\beta},\tilde{\sigma}] = H[\beta,\tilde{\sigma}] = -\tilde{\beta}\sum_{\langle ij\rangle} \tilde{\sigma}_i\tilde{\sigma}_j$, but with $\tilde{\beta}$ satisfying $\sinh(2\beta)\sinh(2\tilde{\beta}) = 1$. The fact that the functional form of the Hamiltonian is the same is exceptional, and in this case one can call the duality a *self*-duality.

Importantly, all observables constructed from the σ_i can be mapped to observables of the $\tilde{\sigma}_i$. Consider for instance two neighbouring spins σ_i and σ_j. We can build an observable $O_{ij} = \sigma_i\sigma_j$, which we call a *link product*. Then the KW duality implies that

$$\langle O_{ij}\cdots\rangle_H = \left\langle \tilde{O}_{ij}(\tilde{\sigma})\cdots\right\rangle_{\tilde{H}}, \qquad \tilde{O}_{ij}(\tilde{\sigma}) = e^{-2\tilde{\beta}\tilde{\sigma}_{i*}\tilde{\sigma}_{j*}} \tag{4}$$

where the notation $\tilde{\sigma}_{i*}$ refers to sites on the dual lattice such that the link connecting sites i^* and j^* intersects the the link connecting i and j. The $\cdots$ indicate that this is an operator equation which holds for arbitrary insertions of operators and thus can be used to construct any expectation value of an even number of the σ_i. Appropriate products of the link products determine all correlation functions.‡

In this work we tackle the problem of finding dualities using machine learning. In particular, starting from the original model (H, O_{ij}) we formulate an optimization problem whose solution recovers *both* (H, O_{ij}) as well as $(\tilde{H}, \tilde{O}_{ij})$. This constitutes an automated discovery of a duality.

Previous work: The problem of learning the parameters in a Hamiltonian from data is precisely that of training a Boltzmann machine, and has a very long history. Our case differs from the classical situation in that are we simultaneously learning a mapping of observables. Other work on dualities involving machine learning includes [9], [10], but none is aimed at recovering the full dual descriptions as we do here.

2 Methodology

We now explain how, starting from the Hamiltonian $H[\beta,\sigma]$ of some statistical model on a lattice, we can learn candidates $\tilde{H}[\tilde{\beta},\tilde{\sigma}]$ for dual models as well as a dictionary between original and dual observables. This includes learning the fact that the dual model is defined on the dual lattice.

Framework and loss function: We assume that $\tilde{H}$ can be written in terms of local couplings of spins:

$$\tilde{H}[\tilde{\beta},\tilde{\sigma}_i] = -\tilde{\beta}\sum_{\langle ij\rangle}\tilde{\sigma}_i\tilde{\sigma}_j - \tilde{\beta}_4 \sum_{\langle ijkl\rangle}\tilde{\sigma}_i\tilde{\sigma}_j\tilde{\sigma}_k\tilde{\sigma}_l - \cdots \tag{5}$$

where the couplings $\tilde{\beta}_a$ are parameters to be learned. We would like to find dual representations of the link products O_{ij} we described for the Ising model. We assume that the link product in the

†See e.g. [8] for a textbook treatment.

‡Due to a $\mathbb{Z}_2$ symmetry the expectation value of a moment of an odd number of spins *formally* vanishes in a finite model, though as usual in the symmetry spontaneously broken phase this might not be observed in a simulation that uses local updates. We also note that the relation is modified if we consider precisely the same two-spin operator *twice*, i.e. $(\sigma_i\sigma_j)^2 = 1$, when a careful derivation of the duality shows that the right-hand side must be modified and is also identically 1.

original model is mapped to *some* functions of *nearby* link products in the dual model, more precisely

$$\tilde{O}_{ij}(\tilde{\sigma}) = G(\{\tilde{\sigma}_k \tilde{\sigma}_l\}) \tag{6}$$

where $\{\tilde{\sigma}_k \tilde{\sigma}_l\}$ is a set of link products such as the one shown in Figure 1.

G is designed to be sufficiently flexible to recover models on lattices related in various ways to the original one. Note that a choice must be made about how to relate the assignment of link products neighbouring a horizontal link to the assignment of link products neighbouring a vertical link, as multiple choices are consistent with rotational invariance. In Figure 1 we display the choice used, which relates them by a rotation composed with a reflection. As we will see later, this choice is important for recovering the geometry of the dual lattice.

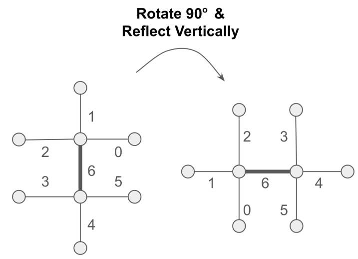

Figure 1: We parametrize G as a neural network that takes neighboring links of a given link (in this case # 6) as its input. The assignment on horizontal links is related to that on vertical ones by a rotation and reflection.

We now construct a loss function $\mathcal{L}$ that is minimized when all correlation functions of O_{ij} and $\tilde{O}_{ij}$ agree on the two sides of the duality. This is similar to the matching of moments of two distributions, which is a standard problem, and for which one can construct general kernels that are minimized only when all of the moments of two distributions agree (see e.g. [11]). Unfortunately, in the present case we cannot use kernels because of one conceptual and one technical problem: 1) certain moments need not be matched, as per Footnote 2, and 2) no notion of locality is embedded in standard moment matching (in the present case, correlation functions of faraway spins carry little information).

Instead we explicitly match *features* – i.e. moments of a small number of nearby link products, as shown in Figure 2 – which we then spatially average over the lattice. Denoting these features as ϕ^a with a running over features, we then construct the loss

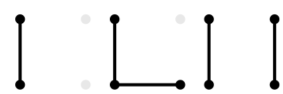

Figure 2: Examples of three features showing link products considered.

$$\mathcal{L}(G, \tilde{H}) = \sum_a \ell^a \ell^a \qquad \ell^a = \langle \phi^a[\sigma_i] \rangle_H - \langle \phi^a[G(\tilde{\sigma}_i)] \rangle_{\tilde{H}} \tag{7}$$

ℓ^a can be thought of as a vector in feature space indicating how far apart the two theories are.

It is clear that this loss can be minimized in two scenarios: (a) $\tilde{H} = H$ and $G(\tilde{\sigma}_i \tilde{\sigma}_j) = \tilde{\sigma}_i \tilde{\sigma}_j$, i.e., the original model is rediscovered, or (b) $\tilde{H} \neq H$ and $G(\tilde{\sigma}_i \tilde{\sigma}_j) \neq \tilde{\sigma}_i \tilde{\sigma}_j$, representing a nontrivial dual model where (selected) moments nevertheless perfectly match those of the original model.

Optimization: We now need to solve the following optimization problem:

$$G^*, \tilde{H}^* = \arg\min_{G, \tilde{H}} \mathcal{L}(G, \tilde{H}) \tag{8}$$

G is represented by a neural network with parameters θ, $G = G_\theta$.

Algorithm 1 outlines the procedure for optimization. Given a trial set of parameters θ and couplings for the dual Hamiltonian $\tilde{\beta}_a$, we simultaneously perform Markov Chain Monte Carlo (MCMC) sampling from the original and dual Hamiltonians using a standard Metropolis algorithm to obtain spin configurations σ_i and $\tilde{\sigma}_i$ drawn from the appropriate distributions respectively. We can then evaluate the expectation values in (7), and compute the loss $\mathcal{L}$.

To minimize it we also need to compute gradients $\partial_\theta \mathcal{L}$ and $\partial_{\tilde{\beta}_a} \mathcal{L}$. For θ this can be done straightforwardly using conventional automatic differentiation techniques. For the $\tilde{\beta}_a$ we cannot backpropagate through a stochastic sampler, but explicit differentiation shows that we can relate the gradients to expectation values that can be evaluated through MCMC sampling from the dual Hamiltonian (see

Appendix C for derivation), e.g. §

$$\partial_{\tilde{\beta}}\mathcal{L} = 2\left\langle \sum_a \ell^a \left(\sum_{\langle ij \rangle} \langle \sigma_i \sigma_j \rangle_{\tilde{H}} - \sum_{\langle ij \rangle} \tilde{\sigma}_i \tilde{\sigma}_j \right) \phi^a[G_\theta(\tilde{\sigma})] \right\rangle_{\tilde{H}} . \tag{9}$$

Algorithm 1 Machine learning for finding statistical mechanics duality

Inputs: $beta$, η (learning rate), N (number of samples)
Initialize: $\tilde{\beta}_0 \in \mathbb{R}$, $\theta \in \mathbb{R}^d$
for each epoch $t = 1, 2, \ldots, T$ **do**
    Draw N samples $\{\sigma_i\}_{i=1}^N \sim p(\sigma|\beta)$
    Draw N samples $\{\tilde{\sigma}_i\}_{i=1}^N \sim p(\tilde{\sigma}|\tilde{\beta})$ where $\tilde{\beta} \neq \beta$
    Compute the loss $\mathcal{L} = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(\sigma_i, G_\theta(\tilde{\sigma}_i))$
    Compute the gradients $\partial_{\tilde{\beta}}\mathcal{L}$ and $\partial_\theta \mathcal{L}$
    Update the parameters:

$$\tilde{\beta}_{t+1} \leftarrow \tilde{\beta}_t - \eta \partial_{\tilde{\beta}}\mathcal{L}$$
$$\theta_{t+1} \leftarrow \theta_t - \eta \partial_\theta \mathcal{L}$$

    if $\mathcal{L}$ has not improved for the last X epochs **then**
        Stop the optimization
    end if
end for

3 Experiments

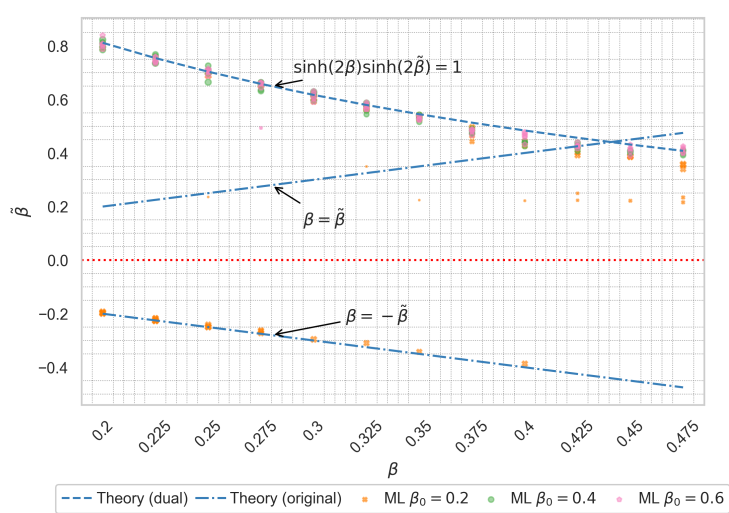

Figure 3: Final $\tilde{\beta}$ as found by the deep learning framework closely matches that of the theoretical results. Points are scaled by the negative logarithm of the best loss such that the **size of the points is inversely proportional to the loss**. We cap the minimum size so that smaller points are visible. The loss is a minimum along two fronts, i.e, original frame $\beta = \pm\tilde{\beta}$ and the dual frame along the lines $\sinh(2\beta)\sinh(2\tilde{\beta}) = 1$.

In this section, we describe some simple experiments using the above machinery. We take our original Hamiltonian H to be that of the 2d Ising model (3), and we take the dual Hamiltonian $\tilde{H}$ in (5) to have only one non-zero parameter $\tilde{\beta}$ (and so $\tilde{\beta}_4 = 0$, etc.).

Neural Network architecture for G_θ: For a given link product in the dual frame we assemble the 7 nearby links shown in Fig. 1 into a 7-dimensional vector $\mathbf{f}_{\langle ij \rangle} \in (\mathbb{Z}_2)^7$, where each element of the vector is the product of the two spins living on the two ends of the link. We consider a simplistic neural network acting on this input, with parameters formed by $\theta_1 \in \mathbb{R}^7$, and

§We note that this evaluation is computationally expensive, as each gradient step requires us to equilibrate an MCMC chain. For training conventional Boltzmann machines one can use more efficient approaches such as contrastive divergence [12]. Due to the presence of the mapping G, we are not aware of a similarly efficient algorithm in our case, and indeed all likelihood-based approaches seem conceptually difficult.

scalars θ_2 and θ_3. We opt for hard attention using Gumbel-Softmax [13] so that only a few of the seven nearby links are utilized in the prediction task. Thus, the mapping is defined by,

$$G_\theta(\mathbf{f}_{\langle ij \rangle}) = \theta_2 \cdot \text{Gumbel-Softmax}(\theta_1)^T \mathbf{f}_{\langle ij \rangle} + \theta_3 \tag{10}$$

As the elements of $\mathbf{f}_{\langle ij \rangle}$ are ± 1, a very simple network provides a very expressive function.

Rediscovery of the 2d Ising duality. In Figure 3, we show the result of deploying the above machinery on different model values of β on an 8×8 lattice with periodic boundary conditions. For each value of the input β, we ran a total of 15 optimizations, five from each of three initializations of $\tilde{\beta}$, i.e., $\tilde{\beta}_0 = 0.2$, $\tilde{\beta}_0 = 0.4$ and $\tilde{\beta}_0 = 0.6$. Due to the randomness involved in MCMC sampling, each seed is expected to be an independent run.

We record the value of $\tilde{\beta}$ obtained. There are three branches of solutions: the original model $\tilde{\beta} = \beta$, the dual model $\sinh(2\beta)\sinh(2\tilde{\beta}) = 1$, and an antiferromagnetic analogue of the original model $\tilde{\beta} = -\beta$. The latter is equivalent to the original frame, and is obtained by making the change of variables $\sigma_i \to -\sigma_i$ on every other site, thus flipping the sign of $\beta \to -\beta$. Note that the existence of the dual branch of solutions can be viewed as a numerical "rediscovery" of the KW duality line

$$\sinh(2\beta)\sinh(2\tilde{\beta}) = 1 \tag{11}$$

Further details on the experiments (including an exploration on how they depend on the system size) are shown in the Supplementary Material.

It is interesting to ask how the model recovers the structure of the dual *lattice*, as well as the dual observables. The attention mechanism used encourages the model to use only a single link of the input, and for the runs that find the dual temperature this ends up using the links numbered either 2 or 5 instead of the original 6 in Figure 1. As we show in an example in Figure 4, this is equivalent to finding the dual lattice from the original. Note that here it is important that we relate horizontal to vertical links by the composition of a rotation *and* reflection as shown in Figure 1; other choices will not result in the possibility of finding the dual lattice, and indeed in our experiments they do not find a duality. The optimized values of G_θ closely match theoretical results $\tilde{O}_{ij}(\tilde{\sigma}) = e^{-2\tilde{\beta}\tilde{\sigma}_{i*}\tilde{\sigma}_{j*}}$, as shown in more detail in the supplementary material.

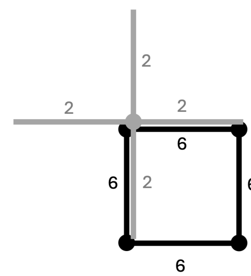

Figure 4: Emergence of dual lattice: e.g. if four original links (marked by 6) form a square, the corresponding four links that are referenced by the neighbour mapping (marked by 2) in Figure 1 form a cross, as expected for the dual lattice.

In this approach the 1-1 mapping of β to $\tilde{\beta}$ is only found numerically; one could possibly supplement this numerical determination with symbolic regression [14] to obtain an analytic formula such as (11), but in more complicated examples of the duality we do not expect there to necessarily exist a simple analytic formula and thus have not explored this.

4 Conclusions

Above we have explained how the process of finding dualities can be automated, demonstrating the mechanism by "rediscovering" the well-known Kramers-Wannier duality of the 2d Ising model. This is only a proof of principle, and much work remains to be done.

For example, as discussed in Section 3, at present we match a number of features which are constructed by hand. It would be ideal to find a kernel that allows matching of all the required moments while simultaneously giving lower weight to those involving faraway spins. On the operational side, it would be helpful to have a more efficient way of training; contrastive divergence fails here as there appears to be no simple way to map the likelihood of a single spin configuration across the duality.

On the physics side, we hope to use such techniques to find entirely new dualities. One concrete direction is to search for Kramers-Wannier duals of deformed Ising models, where extra spin-spin couplings have been added to the action: while some results exist for specific models [15; 16; 17], we are not aware of a completely general approach. In the setup above our preliminary experiments show that adding new couplings generically significantly hurts performance, and a more robust network architecture and training dynamics is desirable. We are currently investigating approaches which explicitly use more of the known symmetry structure of Kramers-Wannier duality. A less concrete but far more exciting direction would be if one could use the approach to find entirely new dualities, unconnected to any existing ones. We hope to return to this in the future.

Acknowledgments: We are very grateful to Roberto Bondesan, Arkya Chaterjee, Tarun Grover, Theo Jacobson, John McGreevy, Takuo Matsubara, Salvatore Pace and Tin Sulejmanpasic for helpful discussions. NI is supported by a Simons Pivot Fellowship and in part by the STFC under grant number ST/T000708/1. AEVF was in part supported by the EPSRC Grant EP/W020939/1 “3d N=4 TQFTs".

References

[1] R. Savit, “Duality in Field Theory and Statistical Systems,” *Rev. Mod. Phys.* **52** (1980) 453.

[2] H. A. Kramers and G. H. Wannier, “Statistics of the two-dimensional ferromagnet. part i,” *Phys. Rev.* **60** (Aug, 1941) 252–262. `https://link.aps.org/doi/10.1103/PhysRev.60.252`.

[3] H. A. Kramers and G. H. Wannier, “Statistics of the Two-Dimensional Ferromagnet. Part II,” *Phys. Rev.* **60** (1941) 263–276.

[4] M. E. Peskin, “Mandelstam ’t Hooft Duality in Abelian Lattice Models,” *Annals Phys.* **113** (1978) 122.

[5] C. Dasgupta and B. Halperin, “Phase transition in a lattice model of superconductivity,” *Physical Review Letters* **47** no. 21, (1981) 1556.

[6] S. R. Coleman, “The Quantum Sine-Gordon Equation as the Massive Thirring Model,” *Phys. Rev. D* **11** (1975) 2088.

[7] F. J. Wegner, “Duality in generalized ising models and phase transitions without local order parameters,” *Journal of Mathematical Physics* **12** no. 10, (1971) 2259–2272. `http://scitation.aip.org/content/aip/journal/jmp/12/10/10.1063/1.1665530`.

[8] M. Kardar, *Statistical physics of fields*. Cambridge University Press, 2007.

[9] P. Betzler and S. Krippendorf, “Connecting Dualities and Machine Learning,” *Fortsch. Phys.* **68** no. 5, (2020) 2000022, `arXiv:2002.05169 [physics.comp-ph]`.

[10] J. Bao, S. Franco, Y.-H. He, E. Hirst, G. Musiker, and Y. Xiao, “Quiver Mutations, Seiberg Duality and Machine Learning,” *Phys. Rev. D* **102** no. 8, (2020) 086013, `arXiv:2006.10783 [hep-th]`.

[11] Y. Li, K. Swersky, and R. Zemel, “Generative moment matching networks,” in *International conference on machine learning*, pp. 1718–1727, PMLR. 2015.

[12] M. A. Carreira-Perpinan and G. Hinton, “On contrastive divergence learning,” in *International workshop on artificial intelligence and statistics*, pp. 33–40, PMLR. 2005.

[13] E. Jang, S. Gu, and B. Poole, “Categorical reparameterization with gumbel-softmax,” *arXiv preprint arXiv:1611.01144* (2016) .

[14] M. Schmidt and H. Lipson, “Distilling free-form natural laws from experimental data,” *science* **324** no. 5923, (2009) 81–85.

[15] A. Strycharski and Z. Koza, “The dual model for an ising model with nearest and next-nearest neighbors,” *Journal of Physics A: Mathematical and Theoretical* **46** no. 29, (2013) 295003.

[16] E. Cobanera, G. Ortiz, and Z. Nussinov, “The Bond-Algebraic Approach to Dualities,” *Adv. Phys.* **60** (2011) 679–798, `arXiv:1103.2776 [cond-mat.stat-mech]`.

[17] D. Aasen, R. S. K. Mong, and P. Fendley, “Topological Defects on the Lattice I: The Ising model,” *J. Phys. A* **49** no. 35, (2016) 354001, `arXiv:1601.07185 [cond-mat.stat-mech]`.

[18] J. Haah, R. Kothari, and E. Tang, "Learning quantum Hamiltonians from high-temperature Gibbs states and real-time evolutions," *Nature Phys.* **20** no. 6, (2024) 1027–1031, arXiv:2108.04842 [quant-ph].

[19] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch,".

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014) .

A Supplementary material

We provide some further details on our experimental results. A rough measure of the uncertainty of our results can be obtained from the spread of the learned $\tilde{\beta}$ points in Figure 3, which grows as we approach the phase transition at $\beta_c \approx 0.44$. Interestingly, the method does not perform reliably for $\beta > \beta_c$, when the original frame is in the symmetry-broken phase. This is somewhat reminiscent of known difficulties in learning parameters of Hamiltonians at high β (see e.g. Appendix B of [18]) and deserves further study.

Further, from Figure 3 we observe that the initialization of β has an impact on the frame that is chosen. We find that $\beta_0 = 0.4$ almost always drifted away from the original frame as indicated by the absence of green points on the $\beta = -\tilde{\beta}$ line. Note for some β (e.g., $\beta = 0.35$), we see some suboptimal runs resulting in the final β far away from the dual or original frame. These have high loss and can be easily identified as failed runs.

Figure 5 shows runs from $\beta = 0.2$ grouped by β_0 and $\tilde{\beta}^*$, illustrating how the training progresses under different scenarios. For the seeds where either the dual or original frame is recovered, the loss goes to 0. Further, we track the entropy of Gumbel-Softmax(θ_1) to assess how the algorithm is weighing each feature. A value of 0 corresponds to a strong preference for one out of the seven input links.

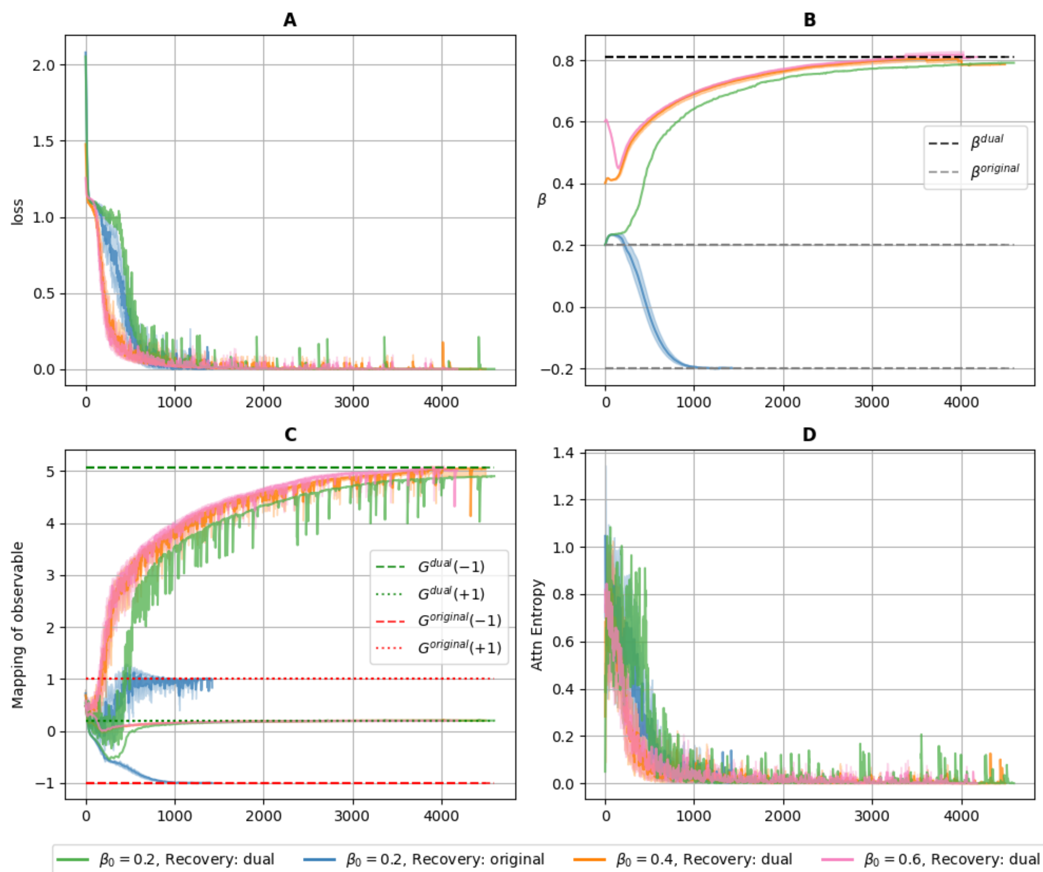

Figure 5: Training progress for runs from $\beta = 0.2$, grouped by β_0 and $\tilde{\beta}^*$ to showcase the trajectory of various metrics. We show exponentially smoothed moving average of the following metrics: (A) Loss, (B) β, (C) Mapping of observables, (D) Entropy of Gumbel-Softmax(θ_1) For (B) and (C) we denote theoretically expected values in original and dual frames by the dashed lines.

Finally, in Figure 6 we see which link is selected as a map to the observable, using the numbering in Figure 1; links 2 and 5 correspond to a mapping to the dual lattice (and are found when we recover a map to the dual frame) and link 6 corresponds to recovering the original link (and are found when we recover the original frame). The small amounts of other dimensions correspond to failed runs which generically have a higher loss.

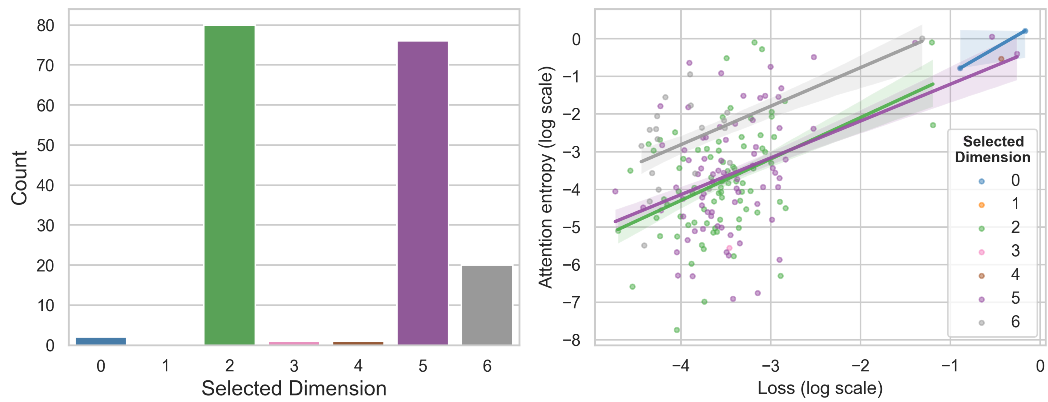

Figure 6: (left) Corresponding to Figure 3, we plot the frequency with which the dimension of the feature vector $\mathbf{f}_{\langle ij \rangle}$ was selected. Note that 2 and 5 pertain to the dual frame, while 6 relates to the discovery of the original frame. (right) Loss values (log scale) and attention entropy (log scale) are positively correlated such that lower loss increasingly prefers a single dimension of feature vectors. Note that both the loss and attention entropy are very low on the three features corresponding to the dual and original frames, as expected for a successful run.

B Neural Network Training

Our models are all implemented in PyTorch [19]. We used the Adam [20] optimizer with the learning rate of 0.005 for β and 0.01 for θ. [NI: Note I changed this so that it is correct – the effective learning rate from the algorithm was off by a factor of 2 for β due to the error previously]. Moreover, we used the early stopping criterion to stop the training if the loss didn't improve over 200 epochs. We ran the sampler in each experiment to generate 1000 samples for the lattice. We ran the training for a maximum of 25000 epochs, and our runs took about 1-3 hours each. Our experiments are run on the lattice size of 8×8.

C Derivation

Here we derive (9). We seek to compute the gradients with respect to $\tilde{\beta}$ of $\mathcal{L}$ defined in (7). Recall that for any function of spins $\mathcal{O}[\tilde{\sigma}]$ we have

$$\langle \mathcal{O} \rangle_{\tilde{H}} \equiv \frac{1}{Z(\tilde{\beta})} \sum_{\{\tilde{\sigma}_i\}} \mathcal{O}[\tilde{\sigma}] e^{\left(\sum_{\langle ij \rangle} \tilde{\beta} \tilde{\sigma}_i \tilde{\sigma}_j\right)} \qquad Z(\tilde{\beta}) \equiv \sum_{\{\tilde{\sigma}_i\}} e^{\left(\sum_{\langle ij \rangle} \tilde{\beta} \tilde{\sigma}_i \tilde{\sigma}_j\right)} \tag{12}$$

where the sum over $\{\sigma_i\}$ runs over all spin configurations. Now we have

$$\partial_{\tilde{\beta}} \mathcal{L} = -2 \sum_a \ell^a \partial_{\tilde{\beta}} \langle \phi^a [G(\tilde{\sigma}_i)] \rangle_{\tilde{H}}, \tag{13}$$

where we have used the definition of ℓ^a in (7). From (12) the gradient of any observable with respect to $\tilde{\beta}$ is

$$\partial_{\tilde{\beta}} \langle \mathcal{O} \rangle_{\tilde{H}} = -\langle \mathcal{O} \rangle_{\tilde{H}} \sum_{\langle ij \rangle} \langle \tilde{\sigma}_i \tilde{\sigma}_j \rangle_{\tilde{H}} + \sum_{\langle ij \rangle} \langle \tilde{\sigma}_i \tilde{\sigma}_j \mathcal{O} \rangle_{\tilde{H}} \tag{14}$$

where the first term comes from differentiating $Z(\tilde{\beta})$ and the second from differentiating inside the Boltzmann measure weighting each configuration in (12). Using this expression to evaluate (14) for $\mathcal{O} = \phi^a[G(\tilde{\sigma}_i)]$ we find (9).

D Scaling results

Figure 7 shows the fraction of instances in which either $\tilde{\beta}$, β, or $-\beta$ were successfully recovered. We observe that as the lattice size increases to 10×10 and 12×12, the recovery rate improves.

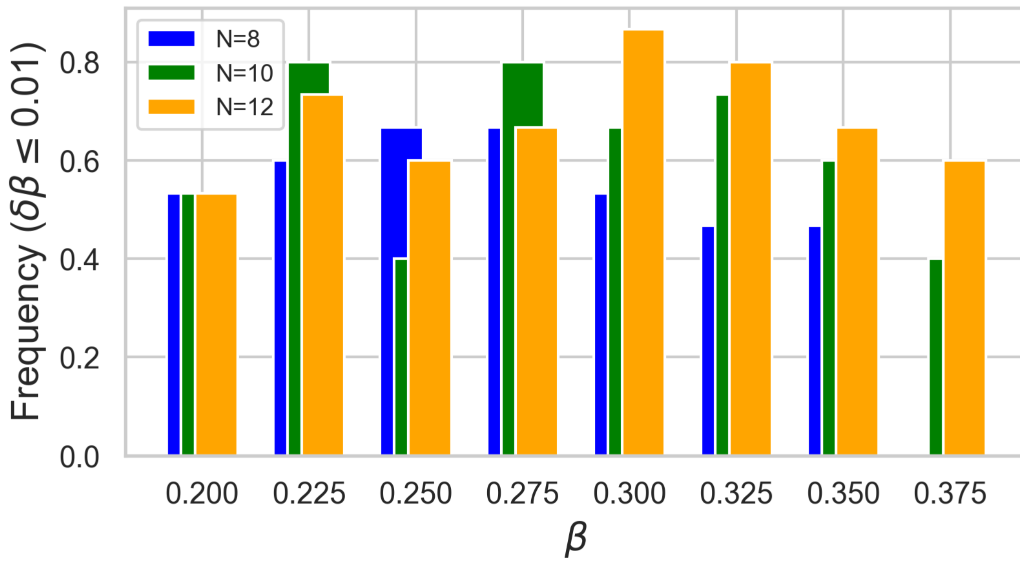

Figure 7: Fraction of times relevant beta was recovered within the tolerance of 0.01 as the lattice size is increased.

E Broader Impact

Our work, on further scaling and addressing the current limitations, hold significant potential to advance knowledge in statistical physics. At this stage, we do not anticipate any negative societal impacts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We empirically demonstrate the framework's ability to discover dualities.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss it in Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not have any new theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We outlined the algorithm as well as the training parameters. We have also provided a repository with the training code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code in the supplemental materials. The script: figure_3.sh runs the experiments needed to reproduce our main result, i.e., Figure 3.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide this information in Appendix B along with the Experiments section in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our main purpose is to show that the framework successfully discovers the original and dual frame. We do that by showing the results of 180 optimizations, the results of which are displayed in Figure 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide that in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We don't foreseen any negative societal impact of this work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include the statement in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the code with README.md that explains how to run it. We will also release the Github repo in the future.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.