Neural Embeddings Evolve as Interacting Particles

Rohan Mehta
California Institute of Technology
rmehta2@caltech.edu

Ziming Liu
Massachusetts Institute of Technology
zmliu@mit.edu

Max Tegmark
Massachusetts Institute of Technology
tegmark@mit.edu

Abstract

We consider drawing a parallel between the training dynamics of neural embeddings and the dynamics of physical systems, in that embeddings can be regarded as interacting particles called "repons". We investigate this intuitive picture on neural networks performing addition, where each number is associated with a learnable d-dimensional embedding vector, interpreted as the spatial coordinates of a repon. We find that the evolution of repons can be successfully modeled by a conservative force (i.e., a force defined from a potential energy) with both attractive and repulsive components, and learn this potential field with a second neural network. The attractive potential is locally quadratic, while the repulsive potential resembles the famous Higgs potential, revealing an intriguing symmetry breaking mechanism. Beyond these preliminary results, this work also proposes a novel paradigm whereby physics ansatzes make it tractable to use auxiliary neural networks to interpret the training dynamics of another neural network.

1 Introduction

Neural embeddings demonstrate rich structures, e.g., the parallelogram formed by man-woman-king-queen [1, 2], the emergence of semantically meaningful directions [3, 4], and the arrangement of numbers in a circle or lattice in various algorithmic tasks [5, 6, 7]. These structures are reminiscent of crystals: atoms or molecules are arranged in certain structured ways due to their interactions. We are interested in whether the formation of structured embeddings in neural networks is also similar to the process of crystallization. More concretely, the central question of this paper is:

> Q: Can we view embedding vectors as the spatial coordinates of particles, such that their dynamics are only governed by their interactions?

We consider a multi-layer perceptron (MLP) trained on the task of addition. The learning process of a neural network involves following the negative gradient of a potential function (or loss function) $U(\theta)$, where θ includes both the embedding parameters $\theta_{\rm emb}$ and the decoder parameters $\theta_{\rm dec}$, i.e., $\theta = \theta_{\rm emb} \cup \theta_{\rm dec}$. If the training dynamics of the embeddings are only dependent on the embeddings themselves, it follows that $U(\theta) = \hat{U}(\theta_{\rm emb}) + \tilde{U}(\theta_{\rm dec})$ – or that the interaction among embeddings is a conservative one. This may seem too strong of an assumption, however we surprisingly find that this ansatz can still capture many features of embedding dynamics.

We propose learning the potential $\hat{U}(\theta_{\rm emb})$ with a second neural network. Such a strategy may appear circular – how can we use one black-box system to shed light on another? The key is: inductive

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

biases inspired by physics are built into the second neural network, hence greatly simplifying it. If we are able to learn a low-dimensional approximation of the potential, we can "throw away" the neural network that generated it, and simply study the resulting low-dimensional function using more traditional techniques. The fact that the approximation comes from a process we do not understand does not matter, so long as we can quantify the accuracy of the approximation and make progress on understanding *it*. This perspective is a novel contribution of our work - although understanding the evolution of representations through physical dynamics has been previously studied [6, 8, 9, 10, 11], to the best of our knowledge, we are the first to *learn* the physical theory in a data-driven way, rather than *manually design* the theory, which may limit its expressive power. Remarkably, we find that the learned potentials resemble physically realistic potentials to first-order, such as the well-known Higgs potential [12].

This paper is structured as follows: Section 2 introduces the problem setup and the method, Section 3 shows empirical results in $d = 1, 2$ embedding dimensions, followed by conclusions in Section 4.

2 Method

We study the dynamics of MLPs on a toy algorithmic dataset, where the model must learn addition from a dataset of (a, b, c) triplets, where $c = a + b$. The model architecture encodes a and b separately into $\mathbf{E}_a, \mathbf{E}_b \in \mathbb{R}^d$, sums their embeddings together, and passes the sum to the decoder – i.e., the model architecture is $(a, b) \mapsto \text{Dec}(\mathbf{E}_a + \mathbf{E}_b)$. In keeping with the physics bent of our analysis, we will henceforth refer to embeddings as "repons", to evoke the idea that they can be thought of as particles with positions in space. We denote the embedding $\mathbf{E}_i$ by the repon $\mathbf{r}_i$.

Let us note that since the problem we are considering is discrete, there is a finite set of possible inputs, which form the dataset $D = \{(\mathbf{r}_i, \mathbf{r}_j) \mid 0 \leq i \leq j < n\}$. Given two repon pairs $(\mathbf{r}_i, \mathbf{r}_j)$ and $(\mathbf{r}_k, \mathbf{r}_l)$, it is our intuition that they repel if their labels are different (i.e., $i + j \neq k + l$) but attract if their labels are the same (i.e., $i + j = k + l$). Thus, we suppose that

$$\hat{U}(\mathbf{r}_1, \ldots, \mathbf{r}_n) = \sum_{(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) \in \binom{D}{2}} f_{ijkl}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) \tag{1}$$

where

$$f_{ijkl}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) = \begin{cases} f_a(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) & \text{if } i + j = k + l, \\ f_r(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) & \text{if } i + j \neq k + l \end{cases} \tag{2}$$

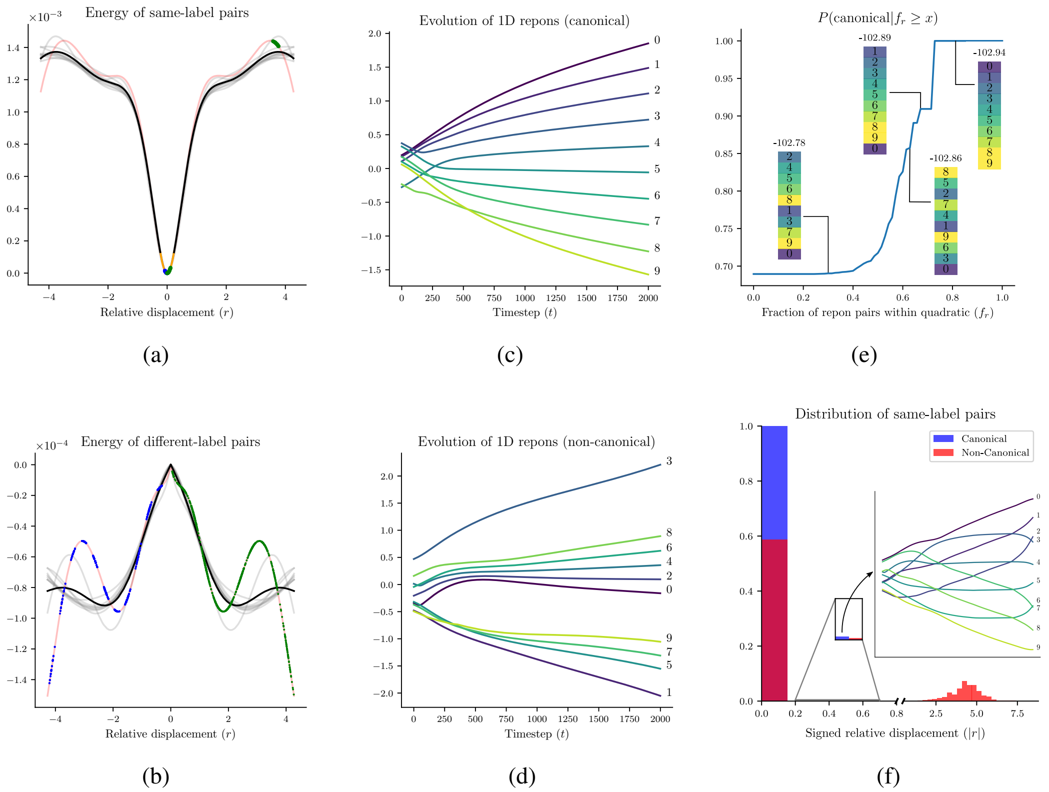
We have then parameterized the approximate potential with two $4d$-dimensional functions. We may also imagine introducing a few variants on this model, which impose further restrictions to drive down the dimensionality further:

1. *Translational invariance.* $f_{a/r}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) = f_{a/r}(\mathbf{r}_i + \mathbf{r}_j - \mathbf{r}_k - \mathbf{r}_l)$,
2. *Parity symmetry.* $f_{a/r}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) = f_{a/r}(|\mathbf{r}_i + \mathbf{r}_j - \mathbf{r}_k - \mathbf{r}_l|)$,
3. *Rotational symmetry.* $f_{a/r}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) = f_{a/r}(\|\mathbf{r}_i + \mathbf{r}_j - \mathbf{r}_k - \mathbf{r}_l\|)$,

where $|\cdot|$ is elementwise: $\mathbb{R}^d \to \mathbb{R}^d$, whereas $\|\cdot\|$: $\mathbb{R}^d \to \mathbb{R}$. Under these variants, both f_a and f_r are functions of either d variables (parity symmetry) or a single variable (rotational symmetry), making them extremely interpretable. To compare these different forms of the potential, we consider two metrics: the trajectory loss and the representation loss. The trajectory loss uses the learned potential to define a gradient flow, updates the repons according to this flow, and compares the simulated trajectory T' with the actual trajectory T in the dataset. It is a score between 0 and 1, with 0 representing a perfect prediction and 1 representing the trivial prediction (the zero vector)

$$\texttt{trajectory\_loss}(T, T') = \frac{\|T - T'\|}{\|T\| + \|T'\|}. \tag{3}$$

In contrast to the trajectory loss, the representation loss evaluates the local behavior of a potential, by computing the Kendall-Tau distance [13] between the final predicted configuration of the repons and their actual final configuration along each dimension, and then averaging.



(a) (c) (e)

(b) (d) (f)

Figure 1: Ten learned **(a)** attractive and **(b)** repulsive potentials plotted in their 99% data range. The potential with lowest representation loss is highlighted red; the average potential is plotted in black; the range of the quadratic fit is plotted in orange. The relative displacement of same-label repon pairs is plotted in blue, while different-label repon pairs are plotted in green. The best performing potentials are used to define a gradient flow, to simulate the evolution of the repons. There is a **(c)** "canonical" ground state as well as several **(d)** "non-canonical" metastable states. **(e)** and **(f)** establish that canonical configurations occur when a majority of same-label pairs are in the quadratic region.

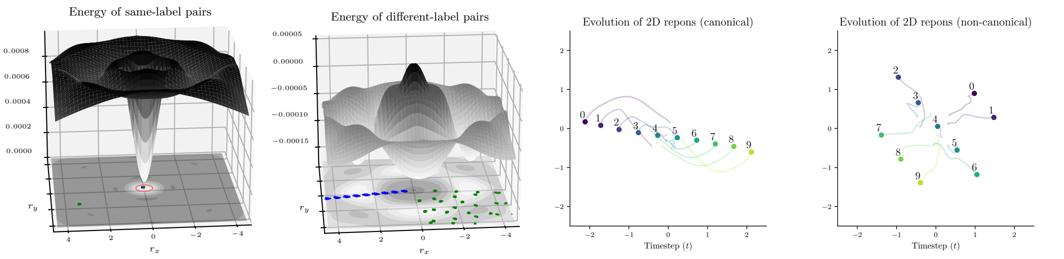
3 Experiments

We consider the task of scalar addition over the integers 0–9. We vary the embedding dimension – the number of spatial coordinates per repon vector – over $d = 1$ and $d = 2$. Each task is run $5,000$ times with a fixed neural net architecture and hyperparameter profile, but varying initialization seed. The matrix of repon positions $\mathbf{R} = \{\mathbf{r}_0, \cdots, \mathbf{r}_9\}$ and gradients $\mathbf{F} = \{\frac{d\mathbf{r}_0}{dt}, \cdots, \frac{d\mathbf{r}_9}{dt}\}$ are saved at each time step. Two neural networks are then trained to parameterize the functions f_a and f_r such that the quantity $(\nabla_{\mathbf{R}}\hat{U}(\mathbf{R}) - \mathbf{F})^2$ is minimized. Training all models required approximately 2 GPU days.

3.1 Scalar addition with $d = 1$

We first consider the case of scalar addition with $d = 1$. Ten attractive and repulsive parity-symmetric potentials are learned. Previous work [6] suggests that the attractive potential should be approximately quadratic by an informal Taylor expansion argument. However, the expansion is only feasible around $\mathbf{r} = 0$, and was not applied to the repulsive case. The learned attractive potentials validate the quadratic hypothesis, but clearly demonstrate that it is a local behavior. In Figure 1 (a), the potentials have been aligned so that their quadratic regions overlap[1], and the average potential has been plotted. The average potential takes the form of a central well that extends from $[-2, 2]$, however the average extent of the quadratic (as measured by a polynomial fit with $R^2 \geq 0.999$) is only $[-0.5, 0.5]$.

[1] Despite parity-symmetry, we plot the potentials over $\pm r$, for ease of visualization.



(a) (b) (c) (d)

Figure 2: Average **(a)** attractive and **(b)** repulsive potentials plotted in their 99% data range for the $d = 2$ case. The relative displacement of same-label repon pairs is plotted in blue, while different-label repon pairs are plotted in green. The repulsive potential resembles the Higgs potential, and is responsible for symmetry breaking of linear configurations. Canonical configurations become **(c)** arbitrarily oriented lines, while non-canonical configurations become **(d)** lattices of varying regularity.

The potential becomes roughly linear after the quadratic region, before bowing in slightly and greatly reducing its slope after exiting the central well. While the feature is not consistent across all potentials, the best-performing potential (measured by representation loss over the test set) has a much more significant bow inwards, leading to an apparent local minimum.

Interestingly, the repulsive potential exhibits a similar – albeit inverted – structure. Around zero, it has a sharp peak that can locally be fit by the polynomial $ax^2 + bx^2$. Around ± 2, this peak reverses course, leading to the creation of a local minima. Curiously, both the attractive and repulsive potentials have minima-like features emerge at the same value of $|\mathbf{r}|$, suggesting some characteristic length scale. We find that it is possible to fully understand these potentials with a theory of statistics, deferred to Appendix A.

3.2 Scalar addition with $d = 2$

We also consider the case of two-dimensional repons. We learn ten parity- and rotationally- symmetric potentials, and find that they are very similar in structure to their one-dimensional counterparts. The parity- and rotationally- symmetric potentials learn the same form for the attractive potential – a central well around the origin, which is surrounded by a shallower crater, as shown in Figure 2 (a). As before, the lowest reaches of this well are approximately quadratic, and canonical configurations occur when a majority of same-label repon pairs are situated in this quadratic. The canonical configuration is similarly a line with the repons for 0 and 9 at either end, but with arbitrary angular orientation. This orientation is controlled by the repulsive potential, shown in Figure 2 (b). Its Higgs-like shape leads to symmetry breaking, such that the direction in which different-label repons align along its surface determines the direction in which the canonical line is oriented. Interestingly, the parity-symmetric repulsive potential is not rotationally symmetric, and has local minima along the x and y axes, indicating that it would favor the formation of perfectly vertical and horizontal lines[3].

However, many of the predictions made by these potentials are not as accurate as those in the $d = 1$ case, as reflected by considerably worse loss metrics. We elaborate further on this limitation as well as the corresponding theory of statistics for the $d = 2$ case in Appendix B.

4 Conclusions

We show that interacting particle dynamics are able to capture several salient features of the trajectory of a neural network's embeddings. This evidence implies that a first-order approximation of the embeddings of the network as obeying a conservative force is quite reasonable. However, there are still observations left unexplained by our model. Future work includes approximating dynamics with

[2] The informal Taylor series argument provides some loose evidence that this should be the case.

[3] There is no significant difference between the trajectory and representation losses of the parity- and rotationally-symmetric models though, so we cannot be entirely sure this is a meaningful feature.

multiple potentials that hold locally in different regions of configuration space or learning a more general potential with both scalar and vector components, as well as better verifying the extent to which the assumption of conservativeness holds. Ultimately, the connection between potentials and training dynamics appears to be a deep and meaningful one, highlighting the emergence of simple effective dynamics in the seemingly complex training dynamics of neural networks.

References

[1] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014.

[2] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[3] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[4] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

[5] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[6] Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022.

[7] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Hangfeng He and Weijie J Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(36):e2221704120, 2023.

[9] Loek van Rossem and Andrew M Saxe. When representations align: Universality in representation learning dynamics. *arXiv preprint arXiv:2402.09142*, 2024.

[10] Tiberiu Musat. Clustering and alignment: Understanding the training dynamics in modular addition. *arXiv preprint arXiv:2408.09414*, 2024.

[11] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Peter W Higgs. Broken symmetries and the masses of gauge bosons. *Physical review letters*, 13(16):508, 1964.

[13] M. G. Kendall. A new measure of rank correlation. *Biometrika, vol. 30, no. 1/2, 1938, pp. 81–93.*, 1938.

Appendices

A More analysis of the $d = 1$ case

We find that it is possible to fully understand the potentials with a theory of statistics, by mapping the spatial distribution of repon pairs to the topography of the potential, and then correlating this topography with the overall configuration of the repons. We identify two broad classes of structure, which we term canonical configurations (two degenerate ground states) and non-canonical configurations (several metastable states).

The two canonical configurations occur when the repons line up in order from 0 to 9 or 9 to 0, and represent the global minimum of the system. As shown in Figure 1 (e), the cumulative frequency of canonical configurations rises monotonically with the fraction of same-label repons situated within the quadratic, f_r. This figure also plots a few specific configurations with different values of f_r. Configurations with values of f_r closer to 1 have lower energy, and have longer stretches of canonical ordering (even if they are not fully canonical themselves). Furthermore, if we intervene on the potential, and replace it with the fit of its quadratic region, all simulations yield a canonical configuration. All this evidence points to the conclusion that the quadratic well leads to canonical configurations.

As such, we would expect non-canonical configurations to have a sufficient fraction of same-label repons outside the quadratic well. The histogram in Figure 1 (f) shows that this is indeed the case. It also shows that the best-performing configurations are not just those with the correct ordering, but also the correct spacing. An outlier canonical configuration is zoomed in on with some of its same-label pairs further from the quadratic than is typical. This configuration turns out to have the right ordering, but just barely, with the repons for each number not having a uniform spacing.

We similarly notice distinctive patterns in the repulsive potential that correlate with canonical configurations. Different-label repon pairs form 18 different clusters, spaced out along the length of the potential, such that the first cluster contains all repon pairs whose labels differ by 1, the second all pairs whose labels differ by 2, etc. This can be seen in the blue dots plotted on the repulsive potential in Figure 1 (b). The different-label repon pairs for non-canonical configurations (plotted in green) are more chaotic, without any apparent pattern.

Under these facts, a clear picture emerges – repon pairs are initialized somewhere within the central well, but in certain initial configurations the repulsive force is too strong and pushes them upwards. Because the well is a local feature, it is possible for these repon pairs to escape and reach a much shallower region of the potential which acts as an effective minima. This seems to imply the network is operating by a principle of least work – rather than always arranging the repons in the canonical ground state, it will sometimes opt for a higher-energy state, if it is easier to reach from the initial configuration. It also suggests that initializing the repons closer to 0 might encourage them to stay within the quadratic region, and indeed we find that by decreasing the initialization scale of the repons by an order of magnitude, the network learns only canonical configurations. This is a very simple application of using the potential to optimize dynamics, but highlights why such a tool could be powerful, as it is much easier to think about and analyze functions than neural networks.

We also consider a set of potentials without translational invariance enforced, as shown in Figure 3. We analyze the potentials by plotting all possible two-dimensional slices (six, in this case) and fixing the other variables to 0. We find that the local behavior of these slices is incredibly simple, learning functions of the form $(x_1 \pm x_2)^2$, such that all the proper terms for a potential that is quadratic in the relative displacement of same-label repon pairs are learned. While these models do obtain slightly lower trajectory and representation losses, the difference does not appear to be meaningful, but rather due to much less interpretable behavior further from the origin. Thus, it seems that the quadratic behavior observed is not simply an artifact of making the potential a function of relative displacement, but is the optimal first-order approximation for a potential that assumes superposition of two-body interactions.

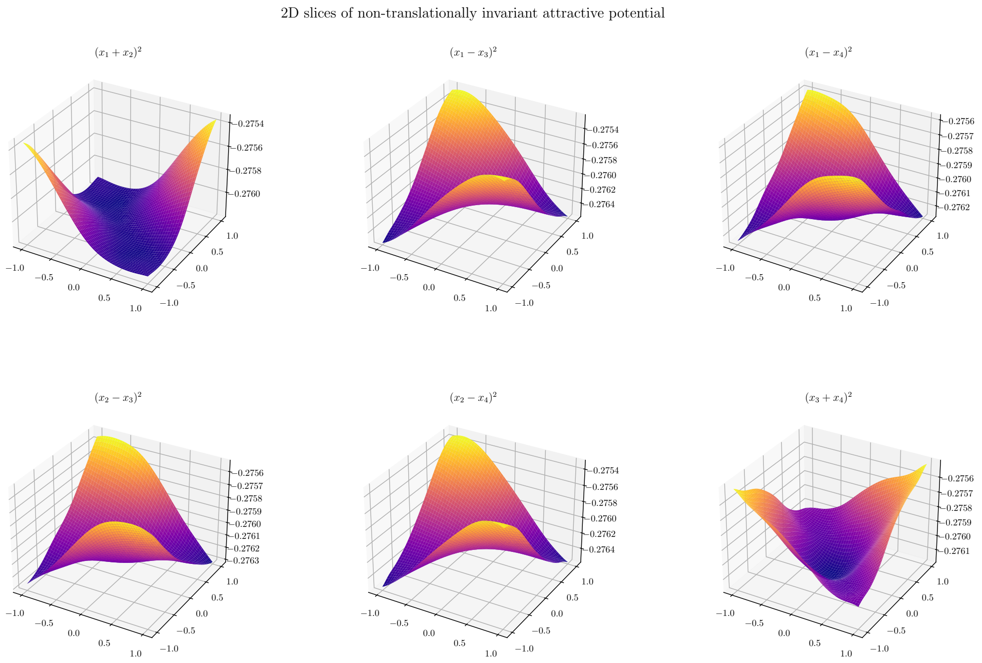


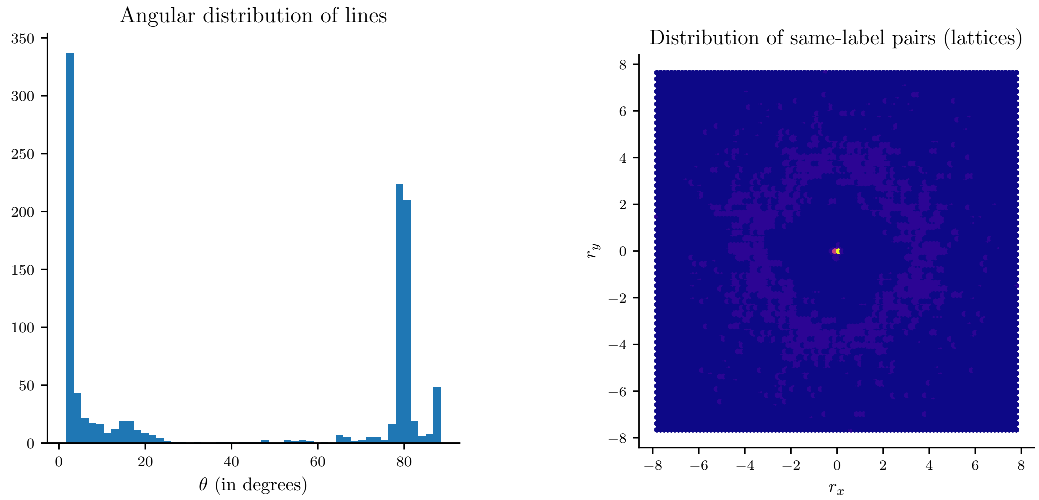
Figure 3: All possible 2D slices of the non-translationally invariant attractive potential. The axes x_1 and x_2 represent the first repon pair, while x_3 and x_4 represent the second repon pair. Note that all the proper terms are predicted for a potential of the form $(x_1 + x_2 - x_3 - x_4)^2$, which is what is locally observed for the cases where translational invariance is assumed.

B More analysis of the $d = 2$ case

The $d = 2$ models perform considerably worse than the $d = 1$ models, with trajectory and representation losses of ~ 0.4, compared to the ~ 0.2 of the $d = 1$ case. This is largely because one of the commonly predicted configurations by the potential, the canonical configuration of a line, is non-existent in the underlying data. This discrepancy prompted further exploration of the dataset, and it was found that none of the configurations in the dataset are generalizing ones (i.e., ones with perfect validation accuracy). It appears that the only generalizing solution is the canonical configuration of a line. By training the underlying model with weight decay and a constrained hidden layer, it was possible to recover a linear configuration. Thus, in this case, the potential predicted dynamics that were more optimal than the ones actually occurring in the training data. A possible hypothesis is that the generalizing dynamics are the ones for which the conservative force assumption is most valid, and thus easiest for the potential to learn. However this phenomenon requires further investigation.

It is also possible to analyze the statistics of the repons in the $d = 2$ case. The quadratic analysis remains the same. As the symmetry breaking argument predicted, Figure 4 (a) shows that linear configurations tend to be either horizontal or vertical. The case of non-canonical configurations is more interesting. All non-linear configurations tend to be lattices of varying degrees of order and regularity. The most ordered lattices tend to have one cluster of same-label repon pairs within the quadratic and one cluster within the shallower crater outside (see the green dots plotted in Figure 2 (a)). This hypothesis is further validated by 4 (b) which shows that the distribution of same-label pairs for lattice-like structures is highly concentrated in the quadratic and more disperse throughout the crater region, but zero everywhere else. Interestingly, the location of the cluster in the crater appears to align with one of the principal components of the lattice, indicating that the attractive potential may also contain information about direction in this case.

Perhaps most curiously, the crater also occurs at a radius of $r = 2$, mirroring the appearance of local minima in the $d = 1$ case. Once again, this seems to point to a coherent characteristic length scale.



(a)

(b)

Figure 4: The distribution **(a)** of the angle at which canonical configurations are oriented. There is a large bias towards perfectly vertical and horizontal lines. The distribution **(b)** of same-label repon pairs for lattice-like structures shows that lattices occur when some fraction of same-label pairs are in the quadratic, while the rest are in the shallower crater surrounding it.