On Compositionality and Emergence in Physical Systems Generativie Modeling

Justin Diamond
University of Basel
Basel, Switzerland
justin.diamond@unibas.ch

Abstract

The principle of compositionality plays a pivotal role in both machine learning and physical sciences but remains under-explored, particularly in the context of synthetic data derived from physical energy potentials. This study aims to bridge this gap by examining the compositional nature of synthetic datasets generated using composite energy potentials. By combining established Lennard-Jones and Morse potentials into a composite potential, we generate synthetic datasets using Markov Chain Monte Carlo (MCMC) techniques. These datasets serve as training grounds for machine learning models, specifically Neural Ordinary Differential Equations (ODEs). Our primary focus is to investigate whether the properties of the composite datasets retain the characteristics of their individual components, effectively testing the principle of compositionality. The findings not only shed light on the compositional integrity of synthetic physical datasets but also lay the groundwork for more robust and interpretable machine learning models applied to complex physical systems by using the formalism of Category Theory.

1 Introduction

The intersection of machine learning (ML) and physical sciences has evolved into a fertile ground for pioneering research [Noé+20; KW16]. Notably, one area that is being critically examined is the role of compositionality in both ML models and physical systems [GV22], which seems under-explored. Compositionality is essential for enabling simple interpretations and the reverse process, decomposition, of complex systems into simpler sub-systems. When composition holds, it allows for more efficient and interpretable models in both physics and machine learning. In line with this, the current study is devoted to a exploration of compositionality in the context of synthetic datasets and the energy potentials from which they were derived. Whereas in [DL23] where structured processes were imposed on generative models, we hope to uncover implicit structures in physical systems generative modeling.

The key objectives of this study are twofold:

- **Composing Energy Potentials:** We create a composite energy potential, drawing upon established energy potentials such as Lennard-Jones and Morse. These are relatively simple potentials with different physical interpretations. This enables us to investigate whether certain properties, like performance invariance and physical system distributions, hold (remain unchanged) when moving from individual to compositional structures [Che+18].
- **Composing Synthetic Datasets:** We also produce synthetic datasets through Markov Chain Monte Carlo (MCMC) methods, based on the individual and composite energy potentials. These datasets serve as a platform for training machine learning models, specifically Neural

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Ordinary Differential Equations (ODEs), to assess whether performance measures stay invariant under different forms of composition [HJA20; Hoo+23].

The cornerstone of this work is to establish under what conditions compositionality can be considered a valid principle in the realm of machine learning applied to physical systems. Through this, we aim to contribute insights that could potentially simplify the modeling of more complex systems in physics and machine learning. We also formalise this inquiry in the setting of Category Theory [SGW21] as it is increasingly used to study compositional problems in machine learning.

1.1 Significance to Machine Learning in Physical Sciences

Understanding the invariance properties of machine learning models has profound implications for the physical sciences, where the quest for universal laws often requires the composition of simpler systems to form more complex ones. In our study, we use relatively simple energy potentials to minimize the issue of sampling inefficiencies, making the results more interpretable and generalizable. Our approach paves the way to developing robust and versatile machine learning models that can adapt to composed systems without significant loss in performance, enabling more accurate simulations and predictions for physical systems by utilizing machine-learned approximations of energy potentials, and establishing a framework for evaluating the reliability of machine learning models when applied to complex, composed physical systems, thereby bridging the gap between true physical laws and derived synthetic datasets.

1.1.1 Interpretations in Category Theory

In the language of category theory, the entities we are dealing with can be conceptualized as follows:

- **Categories**: Our individual energy potentials (LJ, Morse) and their datasets (DLJ, DM) can be thought of as objects in separate categories. These categories house the structures we're interested in: either energy potentials or datasets generated from these potentials.
- **Functors**: The Neural ODE training process serves as a functor, mapping the objects in our categories (energy potentials or datasets) to another category of performance measures. For instance, the functor might take the LJ energy potential and map it to a performance measure PLJ.
- **Natural Transformations**: The composition procedures we're investigating—whether combining energy potentials to form a Joint potential, or concatenating datasets derived from independent runs of MCMC—can be viewed as natural transformations between these functors. They essentially encapsulate how varying the underlying structure (be it in terms of energy potentials or datasets) results in a transformation of performance measures.

The real power of this category-theoretical viewpoint lies in its ability to formalize our scientific problem neatly. Our study revolves around the question of whether different paths through these categories and functors lead to the same outcome—that is, whether the diagrams commute. If they do, it implies a certain robustness and invariance in how Neural ODEs respond to compositional procedures, whether it's in terms of energy potentials or datasets derived from them.

More explicitly, commutativity of these diagrams would imply that the order of operations—whether you compose first and then train, or train first on individual components and then compose—does not matter. This would be a powerful result, indicating that Neural ODEs trained on simpler potentials or their corresponding datasets can be naturally extended to more complex, composite systems without loss of performance.

2 Category-Theoretical Framework in Neural ODEs

This section presents a category-theoretical approach to understanding Neural ODE parameterizations with respect to energy potentials and datasets.

2.1 Categories and Objects

Energy Potential Category ($\mathcal{E}$):

- Objects (E): Different energy potentials.
- Morphisms: Monoidal addition of energy potentials.

Dataset Category ($\mathcal{D}$):

- Objects (D): Datasets derived from MCMC based on energy potentials.
- Morphisms: Concatenation of datasets.

2.2 Functor for Neural ODE Training

Let $T : \mathcal{E} \cup \mathcal{D} \to \Phi$ be a functor that maps objects and morphisms from both $\mathcal{E}$ and $\mathcal{D}$ to a space of neural network parameterizations Φ, updating parameters ϕ through Neural ODE training.

2.3 Commutative Diagrams

$$\begin{array}{ccc} E_{LJ} & \xrightarrow{T} & T(E_{LJ}) \\ {\scriptstyle +}\downarrow & & \downarrow{\scriptstyle +} \\ E_{LJ+M} & \xrightarrow[T]{} & T(E_{LJ+M}) \end{array} \tag{1}$$

$$\begin{array}{ccc} D_{LJ} & \xrightarrow{T} & T(D_{LJ}) \\ {\scriptstyle \oplus}\downarrow & & \downarrow{\scriptstyle \oplus} \\ D_{LJ+M} & \xrightarrow[T]{} & T(D_{LJ+M}) \end{array} \tag{2}$$

2.4 Addition of Energy Functions and Concatenation of Datasets

Addition in $\mathcal{E}$:

$$E_{LJ+M} = E_{LJ} + E_M \tag{3}$$

Concatenation in $\mathcal{D}$:

$$D_{LJ+M} = D_{LJ} \oplus D_M \tag{4}$$

These diagrams represent the fundamental inquiry: whether the Neural ODE training functor T preserves the structure of the compositions in both $\mathcal{E}$ and $\mathcal{D}$. Specifically, it questions if $T(E_{LJ+M})$ is analogous to $T(D_{LJ+M})$, thus exploring the relationship between the parameterization changes due to different energy potential compositions and dataset concatenations.

2.5 Transformation between Categories

The transformation from $\mathcal{E}$ to $\mathcal{D}$ is facilitated through MCMC simulations, where energy potentials are used to generate corresponding datasets. The reverse transformation is not defined in this framework by may be recoverable via if we can define an Adjoint Functor encompassing MCMC.

3 Methodology

We employ Neural Ordinary Differential Equations (Neural ODEs) trained by energy for the composed energy potential and by maximum likelihood for the composed synthetic datasets. The key question this study aims to answer is 'under what circumstances does performance and energy distributions of the generated systems remain invariant when subject procedures of composition of the energy potential or the synthetic datasets'?

3.1 Datasets and Objectives

The following datasets are derived: Dataset_LJ: Generated using MCMC with Lennard-Jones (LJ) potential, Dataset_Morse: Generated using MCMC with Morse potential, Dataset_Joint: Generated using MCMC with a composite of LJ and Morse potentials (referred to as U_{Joint}), and Dataset_Concat: Concatenation of Dataset_LJ and Dataset_Morse.

The primary objective is to investigate the invariance of the Neural ODE's performance under three different types of training conditions. Composition of Synthetic Datasets: To evaluate if the performance of a model trained on concatenated datasets (Dataset_Concat) is relatively similar to that of a model trained on the composed dataset (Dataset_Joint). Composition of Energy Potentials: To evaluate if the Neural ODE's performance when trained directly on U_{Joint} is comparable to when trained on the individual energy potentials. Cross-Comparison: To compare the Neural ODE's performance when trained directly on U_{Joint} against both the joint and concatenated datasets. This offers a more nuanced understanding of how well the model generalizes across different compositional procedures.

The study aims to provide a comprehensive analysis on the interplay between compositional synthetic datasets and energy potentials. Specifically, we investigate whether performance invariance under different types of composition holds similarly, thereby deepening our understanding of the role of true physical laws in machine learning applications for the physical sciences.

4 Results

We trained three different Neural ODE models, **Model_Energy**: Trained using the energy-based loss function, with Adam optimizer and learning rate 1×10^{-5}. **Model_MLE**: Trained using Maximum Likelihood Estimation (MLE) with datasets, with Adam optimizer and learning rate 1×10^{-4}. **Model_MLE_Cat**: Trained using MLE with concatenated datasets, with Adam optimizer and learning rate 1×10^{-4}.

Datasets were selected such that for composed energy functions the last 1000 elements from a simulation run of 10,000 timesteps, while for concatenated datasets only the last 500 elements from each individual run were concatenated.

4.1 Discussion on Statistical Measures

Table 1: Statistical Summary of Training and Generated Data (Filtered by specific criteria in Appendix)

Data Source	Mean	Max	Min	Var	Mode
Training Data	32.17	112.23	13.02	507.67	21.27
Generated (Energy)	20.51	23.58	17.91	0.82	19.59
Generated (MLE)	21.76	93.81	4.37	131.53	16.28
Generated (MLE Concatenated)	25.58	170.24	4.52	552.09	17.59

Figure Description: The table presents a statistical summary of the energy distributions for various data sources. All values have been filtered to include only those instances where the energy is less than 1000 units. The columns represent the mean, maximum, minimum, variance, and mode of these energy values. The data sources include the original training data, a model trained to minimize energy (Generated Energy), a Maximum Likelihood Estimate (MLE) model, and an MLE model trained on concatenated datasets from the Lennard-Jones and Morse potentials (MLE Concatenated). Analyzing the table, several insights can be gleaned regarding the behavior of the different models: **High Variance in MLE Concatenated:** Among all the models, the MLE Concatenated version has a high variance, which is closer to that of the training data. This suggests that it is able to explore a broader range of configurations, hinting at better compositionality at the data level. **Low Variance in Energy-Based Learning:** The model trained by energy exhibits a much lower variance compared to the other models and the training data. This could indicate that the energy-based model is narrowly focused on a specific region of the energy landscape, possibly representing overfitting, or mode collapse, to lower-energy configurations. **Challenges in MLE:** The MLE model displays a mean and variance that are neither too close nor too far from the training data, possibly indicating challenges in modeling

joint energies effectively. **Limited Energy Range in Energy-Based Learning:** Both the minimum and maximum energy values for the model trained by energy are constrained within a narrow range, which could indicate limited exploration capabilities. This is in contrast to MLE models, which show a broader range of energy values. **Emergence of New Configurations:** Interestingly, Both MLE generated models have minimum energy values that are lower than the training data. This suggests that the models are capable of exploring new lower energy configurations beyond what is present in the training set.

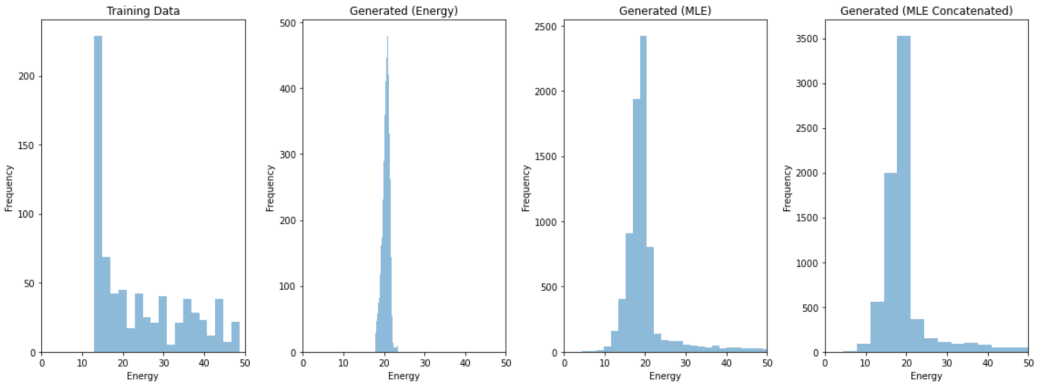

Figure 1: Histograms showing the energy distribution of different models and the training data. Each model's output is filtered to include energies less than 1000 kcal/mol.

5 Discussions

Our analyses indicate that Maximum Likelihood Estimation (MLE) based methods demonstrate strong evidence of compositional behavior, especially when compared to the training data derived from Monte Carlo Markov Chain (MCMC) methods. Notably, the MLE Concatenated model performs surprisingly well, displaying statistical measures close to the training set. This goes beyond mere compositionality to indicate emergence, which is particularly intriguing as the MLE Concatenated model outperforms what would be expected from merely summing the synthetic datasets from the Lennard-Jones and Morse potentials. This revelation calls for a deeper exploration of the compositional nature of concatenated synthetic datasets.

Our results suggest that while energy-based models excel in learning the underlying data distribution, they may lack the compositional richness observed in MLE models. This poses an avenue for future research, particularly in the domain of making energy-based models more compositional.

References

[RC04] Christian P. Robert and George Casella. "The Metropolis—Hastings Algorithm". In: *Monte Carlo Statistical Methods*. New York, NY: Springer New York, 2004, pp. 267–320. ISBN: 978-1-4757-4145-2. DOI: 10.1007/978-1-4757-4145-2_7. URL: https://doi.org/10.1007/978-1-4757-4145-2_7.

[KW16] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[Che+18] Ricky T. Q. Chen et al. "Neural Ordinary Differential Equations". In: *arXiv preprint arXiv:1806.07366* (2018). URL: https://arxiv.org/abs/1806.07366.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *arXiv preprint arXiv:2006.11239* (2020). URL: https://arxiv.org/abs/2006.11239.

[Noé+20] Frank Noé et al. "Machine learning for molecular simulation". In: *Annual review of physical chemistry* 71 (2020), pp. 361–390.

[Wan+20] Xipeng Wang et al. "The Lennard-Jones potential: when (not) to use it". In: *Physical Chemistry Chemical Physics* 22.19 (2020), pp. 10624–10633.

[Pin+21] Redi Kristian Pingak et al. "Accuracy of Morse and Morse-like oscillators for diatomic molecular interaction: A comparative study". In: *Results in Chemistry* 3 (2021), p. 100204. ISSN: 2211-7156. DOI: https://doi.org/10.1016/j.rechem.2021.100204. URL: https://www.sciencedirect.com/science/article/pii/S2211715621001090.

[SGW21] Dan Shiebler, Bruno Gavranović, and Paul Wilson. "Category theory in machine learning". In: *arXiv preprint arXiv:2106.07032* (2021).

[GV22] Bruno Gavranović and Mattia Villani. "Graph Convolutional Neural Networks as Parametric CoKleisli morphisms". In: *arXiv preprint arXiv:2212.00542* (2022).

[DL23] Justin Diamond and Markus Alexander Lill. "Geometric Constraints in Probabilistic Manifolds: A Bridge from Molecular Dynamics to Structured Diffusion Processes". In: *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*. 2023.

[Hoo+23] Emiel Hoogeboom et al. "Equivariant Diffusion for Molecule Generation in 3D". In: *International Conference on Machine Learning*. 2023, pp. 8867–8887. URL: https://arxiv.org/pdf/2203.17003.pdf.

A Methods

A.1 Energy Potentials

We employ two energy potentials: the Lennard-Jones (LJ) and the Morse potentials. The LJ potential [Wan+20] is given by:

$$U_{LJ}(r) = 4C\left(\left(\frac{A}{r}\right)^{12} - \left(\frac{A}{r}\right)^{6}\right) \tag{5}$$

and the Morse potential [Pin+21] is:

$$U_{Morse}(r) = C\left(1 - e^{-A(r-r_e)}\right)^2 \tag{6}$$

A.2 Neural Ordinary Differential Equations (Neural ODEs)

We use Neural ODEs [Che+18] to model the dynamics of the system, particularly for capturing the continuous-time transformations in the data. A Neural ODE is represented as:

$$\frac{dz}{dt} = f(z(t), t; \theta) \tag{7}$$

where $z(t)$ is the state of the system at time t and f is a neural network parameterized by θ.

A.2.1 Latent Space and Analytic Expression

The latent space of our Neural ODE model is assumed to follow a unit Gaussian distribution. The negative log-likelihood (NLL) of this distribution, with unit variance, has a simple analytic expression given by:

$$NLL = \frac{1}{2}\left(z^2 + \log(2\pi)\right) \tag{8}$$

This analytic form allows for efficient and precise calculations during the training and evaluation stages.

A.2.2 Training Objective for Neural ODE

The Neural ODE model is trained using different objectives depending on the type of data used for training. Two primary training methods are used: Training by Energy and Training by Maximum Likelihood Estimation (MLE).

Training by Energy: When using the compositional energy function, the loss function is the absolute difference between the predicted energy and the target energy:

$$\mathcal{L}_{energy}(\theta) = |U_{Joint}(z(t)) - U_{Target}| \tag{9}$$

Here U_{Joint} represents the composite energy potential, and U_{Target} is the ground truth energy.

Training by MLE: When working with compositional synthetic datasets, the training is performed using Maximum Likelihood Estimation. The model aims to minimize the negative log-likelihood (NLL) between the generated data and the ground truth, which is modeled as a Gaussian distribution with unit variance:

$$\mathcal{L}_{MLE}(\theta) = \frac{1}{2}\left((z(t) - \mu)^2 + \log(2\pi)\right) \tag{10}$$

In this expression, μ represents the mean of the Gaussian distribution, which is assumed to be zero in our case.

By training in both directions, we aim to explore the invariance of performance under different compositional procedures, either in the form of energy functions or synthetic datasets.

A.3 Metropolis-Hastings Algorithm for Dataset Generation

Datasets are generated using the Metropolis-Hastings algorithm [RC04], with specific energy potentials serving as the target distributions. These potentials include the Lennard-Jones (LJ) potential, the Morse potential, and a combined (Joint) energy potential. Given an initial state x_0 and a specified energy potential $U(x)$ (which could be U_{LJ}, U_{Morse}, or U_{Joint}), the algorithm iteratively performs the following steps:

1. Generate a candidate x' from a proposal distribution $q(x'|x_t)$.
2. Compute the acceptance probability α as:

$$\alpha = \min\left(1, \frac{e^{-U(x')/T}}{e^{-U(x_t)/T}}\right)$$

   where T is the temperature parameter of the Metropolis-Hastings algorithm.
3. Accept or reject the candidate based on α.

A.3.1 Algorithm Parameters

For our experiments, the following algorithmic parameters were set:

- Number of atoms (N_{atoms}) = 10
- Initial state randomly generated in 3D for N_{atoms}
- Number of iterations = 10,000
- Temperature (T) = 300.0
- Energy scaling factor (k) = 1.0

Different datasets are generated by setting $U(x)$ to U_{LJ}, U_{Morse}, or U_{Joint} for different runs of the algorithm.

B Justification for Using Lennard-Jones and Morse Potentials

B.1 Lennard-Jones Potential

The Lennard-Jones (LJ) potential is a mathematically simple model that captures the essential features of the interaction between neutral atoms and molecules. The potential captures both the Pauli repulsion at short distances and the van der Waals attraction at longer distances, making it an ideal choice for a broad range of molecular systems. It has been widely used in the literature for molecular dynamics simulations and other types of modeling.

Parameters:

- $C = 1.0$: This parameter specifies the depth of the potential well, serving as a scaling factor for the energy. In this study, it is set to 1.0 for simplicity and computational convenience.
- $A = 1.0$: The distance-related constant is also set to 1.0, enabling the potential to serve as a normalized benchmark.

B.2 Morse Potential

The Morse potential serves as another critical approach for modeling the interactions between atoms, especially in situations where bonds can be formed or broken. This potential captures the anharmonicity of real molecular bonds more effectively than a harmonic oscillator model would. The Morse potential is particularly useful for describing vibrational energy levels and is often used in more complex models like QM/MM simulations.

Parameters:

- $C = 1.0$: As with the LJ potential, C serves as a scaling factor for the energy and is set to 1.0 for normalized comparison.
- $A = 1.0$: This dimensionless parameter is related to the width of the potential well. It is set to 1.0 for computational simplicity.
- $r_e = 1.0$: The equilibrium bond distance, set to 1.0 in the same unit as the coordinates, serves as a normalized value for simplified calculations.

B.3 Composite Energy Function

The composite energy function U_{Joint} combines both Lennard-Jones and Morse potentials, scaled by parameters α and β. This allows the model to capture both the non-bonded interactions (via the LJ potential) and the bond formation/breaking behaviors (via the Morse potential).

By choosing normalized and dimensionless parameters for these energy potentials, we aim to study the underlying relationships and invariances in a generalized setting, abstracting away from the specifics of any particular molecular system.

B.4 Simplicity and Sampling Considerations for Compositional Studies

The choice of Lennard-Jones and Morse potentials serves a dual purpose: not only do these potentials represent physical systems with reasonable fidelity, but they also allow for efficient sampling. In many physical systems, sampling is often the most computationally expensive step and could introduce various irregularities that confound the study.

B.4.1 Enabling Efficient Compositional Studies

The relative mathematical simplicity of these potentials significantly mitigates the computational demands of Monte Carlo simulations. This enables us to generate synthetic datasets efficiently, and more importantly, allows us to focus on the study of compositional properties without the interference of sampling-related issues.

This is crucial for a study aiming to examine the compositional nature of different procedures, as it ensures that our findings are not skewed by the complexities or irregularities introduced by the

sampling process. One limitation of this study is the relative simplicity of the physical systems under investigation. Both the Lennard-Jones and Morse potentials are chosen for their computational convenience rather than their ability to model complex physical phenomena. Future work should extend these findings to more complicated systems, paying close attention to the sampling methods employed to ensure an accurate representation of the energy landscape.

C Data Filtering Methodology

To ensure a more reliable comparison between the generated data and the training set, the generated samples were subjected to an outlier removal procedure. This step is crucial to remove extreme energy values that may not be representative of the typical configuration spaces that we are interested in studying. The method used for outlier removal is based on the Interquartile Range (IQR). Specifically, we calculate the first quartile $Q1$ and the third quartile $Q3$ of the data, and then determine the IQR as $IQR = Q3 - Q1$.

```
def remove_outliers(data):
    Q1 = np.percentile(data, 25)
    Q3 = np.percentile(data, 75)
    IQR = Q3 - Q1
    filtered_data = data[(data >= Q1 - 1.5 * IQR) & (data <= Q3 + 1.5 * IQR)]
    filtered_data = filtered_data[filtered_data <= 1000]
    return filtered_data
```

Generated samples that lie outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are considered as outliers and are removed from the dataset. Additionally, to focus on configurations that are physically meaningful and computationally tractable, we also filter out any samples with energy values greater than 1000. It should be noted that this filtering is applied only to the generated data and not to the training data, to rigorously assess the capacity of the models to generate physically relevant and comparable configurations.