Overcoming Data Scarcity in Deep Learning of Scientific Problems

by

Charlotte Chang Le Loh

BSc, Imperial College London (2014)
MASt, University of Cambridge (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author . . . . . .
Department of Electrical Engineering and Computer Science
August 25, 2021

Certified by . . . . . .
Marin Soljačić

Professor of Physics
Thesis Supervisor

Accepted by . . . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Overcoming Data Scarcity in Deep Learning of Scientific Problems

by

Charlotte Chang Le Loh

Submitted to the Department of Electrical Engineering and Computer Science
on August 25, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Data-driven approaches such as machine learning have been increasingly applied to the natural sciences, e.g. for property prediction and optimization or material discovery. An essential criteria to ensure the success of such methods is the need for extensive amounts of labeled data, making it unfeasible for data-scarce problems where labeled data generation is computationally expensive, or labour and time intensive. Here, I introduce surrogate- and invariance- boosted contrastive learning (SIB-CL), a deep learning framework which overcomes data-scarcity by incorporating three “inexpensive" and easily obtainable auxiliary information. Specifically, these are: 1) abundant unlabeled data, 2) prior knowledge of known symmetries or invariances of the problem and 3) a surrogate dataset obtained at near-zero cost either from simplification or approximation. I demonstrate the effectiveness and generality of SIB-CL on various scientific problems, for example, the prediction of the density-of-states of 2D photonic crystals and solving the time-independent Schrödinger equation of 3D random potentials. SIB-CL is shown to provide orders of magnitude savings on the amount of labeled data needed when compared to conventional deep learning techniques, offering opportunities to apply data-driven methods even to data-scarce problems.

Thesis Supervisor: Marin Soljačić

Title: Professor of Physics

Acknowledgments

I would first like to acknowledge my research advisor, Professor Marin Soljačić, for his continuous support of my research. He has allowed me to explore ideas freely, while always being there whenever I need feedback or advice, both technically and personally.

I would also like to thank everyone who has contributed to the success of this project; in particular, to Thomas Christensen for his technical guidance, dedicated involvement and support throughout this project. I also thank Rumen Dangovski and Samuel Kim for their helpful feedback, suggestions and fruitful discussions.

I would also like to acknowledge DSO National Laboratories for providing me the opportunity, and funding, to further my education at MIT. I further acknowledge all other funding sources (i.e. Air Force Research Laboratory and Artificial Intelligence Accelerator[1], Army Research Office[2] and National Science Foundation[3]) for sponsoring part of the research presented in this thesis.

Finally, I am grateful to my family and friends for their unwavering and unconditional support; especially to my husband, Jing Jie, for his love and encouragement for me to pursue my passion.

[1]Research was sponsored in part by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained here should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

[2]This material is also based upon work supported in part by the U. S. Army Research Office through the Institute for Soldier Nanotechnologies at MIT, under Collaborative Agreement Number W911NF-18-2-0048.

[3]This work is also supported in part by the the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, there has been enormous advances in the field of machine learning, in particular deep learning via neural networks, revolutionizing fields across computer science, image recognition [1, 2], natural language processing [3, 4] and decision making [5, 6]. Spurred by these advances, there has also been surging interest to apply deep learning to various problems in the natural sciences that have been traditionally guided by theory-driven analytical methods [7, 8, 9, 10]. Examples of such problems include predictive modelling [10, 11], property optimization [12, 13] and knowledge discovery [14, 15, 16]. Contrary to traditional analytical methods, deep learning relies on massive amounts of data to quantitatively discover hidden patterns and correlations among them in order to perform the desired tasks; its success is thus largely contingent on the amount of data available and a lack of sufficient data will render deep learning approaches futile.

For dominant fields in deep learning such as natural language processing and image recognition, datasets required to train the neural network are comparatively easier to curate since they are often unlabelled (e.g. in training language models like BERT [3] or GPT3 [4]) or can be annotated by means of crowd-sourcing (e.g. ImageNet [17] and CIFAR [18]). In contrast, the collection of a large labeled dataset for problems in the natural sciences is frequently more challenging and often requires resource-, time- or labour-intensive computational or experimental efforts, making the data scarcity issue particularly acute. An excellent example is density functional theory, an extremely important tool used to derive complex electronic structures for applications across chemistry, materials science and

biomedical research [19, 20], but also a highly computationally intensive calculation.

Nevertheless, problems in the natural sciences often benefit from a rich and deep array of codified insights which can be used to offset this scarcity of training data. Examples include exact or approximate analytical insights into the underlying problems, or data caches from related problems. The challenge, however, lies in combining multiple sources of information into a single deep learning approach, especially where models are inherently considered as "black-boxes". Here, I introduce surrogate- and invariance- boosted contrastive learning (SIB-CL), a novel approach using self-supervised and transfer learning techniques to incorporate various sources of auxiliary information into a single deep learning framework (Fig. 1-1), enabling effective and high-quality network training despite data scarcity. In order for this approach to be practical, the auxiliary information sources need to be already known or accessible a priori, or are easily obtainable via inexpensive methods. Specifically, the information sources used in SIB-CL that satisfy the aforementioned criteria are: (i) abundant unlabeled data; (ii) prior knowledge in the form of invariances of the physical problem, which can be governed by geometric symmetries of the inputs or general non-symmetry related invariances of the problem; and (iii) a surrogate dataset on a similar problem that is cheaper to generate, e.g. by invoking simplifications or approximations to the labelling

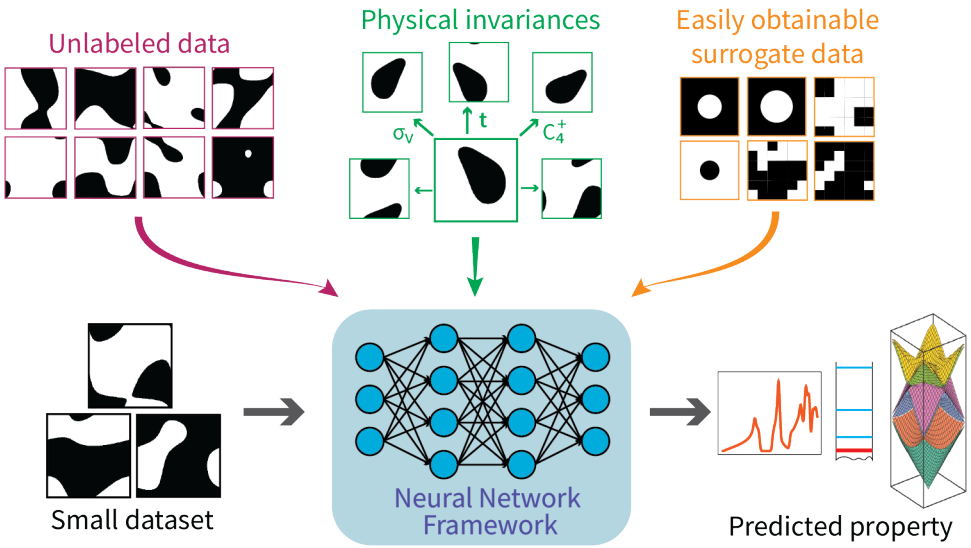

Figure 1-1: Approach used to overcome data scarcity. Examples here show periodic unit cells of photonic crystals as inputs; refer to Fig. 3-1 for a complete illustration.

process.

Here, I demonstrate the effectiveness of SIB-CL on various problems in the natural sciences, in particular, in the fields of photonics and quantum calculations. I focus on problems involving predictive modelling, where the goal is to learn a model that can accurately output some property or solution from some given inputs (i.e. solving the forward problem) at a fraction of the cost compared to the original approach. Having a quick forward model will further enable other useful applications of solving the inverse problem, such as property optimization, since they often require a huge number of invocations of the forward model.

Chapter 2

Related Work

Perhaps most prominently among techniques aimed at overcoming data scarcity in deep learning is transfer learning [21, 22, 23, 24], whose key idea involves re-purposing a model trained on an particular task (the source task) to a different but related task (the target task). This approach can be used to overcome a lack of labeled data in the target task by first pre-training a neural network on a larger and more easily obtainable labeled source dataset, and then retaining either all or part of the network for fine-tuning on the target task. Transfer learning is widely adopted in computer vision (CV) and natural language processing (NLP) applications; prominently, models pre-trained on ImageNet [17] such as ResNet [2] and VGGNet [25] are used for fine-tuning on a variety of CV tasks, while powerful language models like BERT [3] and GPT3 [4] pre-trained on large-scale corpus have been used for fine-tuning on a variety of downstream NLP tasks. Within the natural sciences, transfer learning has also been explored and proven effective in various works [26, 27, 28, 29]; most of these works, however, make use of source data from a different problem [26, 27] or even from a different field [28, 29], e.g. using models pre-trained on ImageNet [17] to accelerate training of material prediction tasks. Since the efficacy of transfer learning is largely dependent on the similarity between the source and target problems [30, 31], using source data from a drastically different problem or field may provide limited improvement to model accuracies for the data-scarce target task. Here, SIB-CL improves the efficacy of transfer learning by utilizing a surrogate/source dataset on the *same* problem, but trained instead with approximate labels or particularly simple classes of inputs. Perhaps most

closely aligned to this idea is the work of [32] where simple, auxiliary estimations were exploited to accelerate training of material properties. The approach is not deep learning-based, however, and thus requires delicate feature engineering and the auxiliary estimations were incorporated simply as additional descriptors in the feature space. Here, SIB-CL uses a deep learning approach and incorporate information of the surrogate dataset via transfer learning, enabling access to other merits of deep neural networks such as scalability and flexibility to integrate with other techniques (e.g. generative methods [33, 34, 35]).

Another approach to deal with scarce training data is to leverage a large quantity of unlabeled data; one such technique gaining considerable attention recently is self-supervised learning (SSL). Building upon transfer learning, SSL is a technique where the pre-training stage uses only unlabeled data. Specifically, "pretext tasks" are invented for the data to provide its own supervision, which include tasks like context prediction [36], image rotation prediction [37] and jigsaw puzzle solving [38]. In particular, there has been a recent surge of interest in contrastive self-supervised learning [39], or contrastive learning, where the pretext task is constructed as contrasting between two "views" of a sample and/or with other samples, where "views" are generated via a pre-optimized transformation strategy. The core objective is for the pre-trained model to output embeddings, where similar instances are grouped closer together and differing instances are pushed further apart (when measured in the embedding metric space). Examples of recent contrastive learning techniques include MoCo [40], SimCLR [41], BYOL [42], SwAV [43] and SimSiam [44]; these techniques differ mostly in their strategy used to generate the instances. These techniques have achieved unprecendented successes in CV; yet, there has been few applications to the natural sciences [45, 46], which can be partly attributed to the difficulty in designing ideal transformation strategies for scientific problems. Here, SIB-CL demonstrates that enforcing invariances can be an effective transformation strategy for contrastive learning on scientific problems.

The invocation of invariance or symmetry knowledge is an appealing technique in deep learning and is often adopted for scientific problems [47, 48]. Fundamentally, this is because the exploitation of symmetry knowledge provides a strong inductive bias, or constraint in the space of possible models, allowing the network to achieve better predictive accuracies. A well-known example is the superiority of convolutional neural networks over fully-connected

ones for image data, arising from the translation equivariance of convolutional layers. There have been numerous prior works aimed at developing strategies to incorporate knowledge of invariances/symmetries [49, 50, 51, 52]; however, they often require highly specialized, task-specific kernels and architectures. SIB-CL adopts contrastive learning, an appealing and broadly-applicable alternative to such strategies, to effectively learn invariances during pre-training rather than building them in by architecture choice.

Chapter 3

Dataset Generation

To evaluate the effectiveness of SIB-CL in data-scarce settings, various scientific problems are designed. Specifically, problems involving two-dimensional (2D) photonic crystals (e.g. predicting photonic density-of-states and band structures) and problems involving the three-dimensional (3D) non-interacting Schrödinger equation are used. Their construction and data generation are described in this chapter.

3.1 Photonic Crystals

Photonic crystals (PhCs) are wavelength-scale structured materials, whose dielectric profiles are engineered to create exotic optical properties not found in bulk materials, such as photonic band gaps and negative refractive index, making them useful for applications in photonic integrated circuits and flat lenses [53, 54]. The design of exotic properties in PhCs is often guided by the density of states (DOS) of photonic modes, which captures the number of modes accessible in a spectral range. Computing the DOS is expensive, however, as it requires dense integration across the full Brillouin zone (BZ) of the PhC and summation over bands. Thus, the task of predicting the DOS in 2D PhCs is an appropriate data-scarce scientific problem and will be used to evaluate SIB-CL's effectiveness at reducing the data requirements for neural network training.

3.1.1 Unit cells

PhCs are characterised by a periodically varying permittivity, $\varepsilon(\mathbf{r})$, whose tiling makes up the PhC's structure (see Fig. 3-1a). For simplicity, 2D square lattices of periodicity a with a "two-tone" permittivity profile, i.e. $\varepsilon \in \{\varepsilon_1, \varepsilon_2\}$, with $\varepsilon_i \in [1, 20]$ are considered, and lossless isotropic materials are assumed so that $\varepsilon(\mathbf{r})$ and the resultant eigenfrequencies are real. The permittivity profile $\varepsilon(\mathbf{r})$ of each unit cell is parameterized by choosing a level set of a Fourier sum function ϕ, defined as a linear sum of planed waves with frequencies evenly spaced in the reciprocal space (up to some cut-off). i.e.

$$\phi(\mathbf{r}) = \mathrm{Re}\left[\sum_{k=1}^{9} c_k \exp(2\pi \mathrm{i} \mathbf{n}_k \cdot \mathbf{r})\right], \tag{3.1}$$

where each $\mathbf{n}_k$ is a 2D vector (n_x, n_y) and 3 Fourier components are used per dimension, i.e. $n_x, n_y \in [-1, 0, 1]$ (and thus the summation index k runs over 9 terms). c_k is a complex coefficient, $c_k = re^{i\theta}$ with r, θ separately sampled uniformly in $[0, 1)$. Finally, a filling fraction, defined as the fraction of area in the unit cell occupied by ε_1, is uniformly sampled in $[0, 1)$ to determine the level set Δ so as to obtain the permittivity profile:

$$\varepsilon(\mathbf{r}) = \begin{cases} \varepsilon_1 & \phi(\mathbf{r}) \leq \Delta \\ \varepsilon_2 & \phi(\mathbf{r}) > \Delta \end{cases}. \tag{3.2}$$

This procedure produces periodic unit cells, shown in Fig. 3-1a, with features of uniformly varying sizes due to the uniform sampling of the filling ratio and without strongly divergent feature scales and thus corresponds to fabricable designs. 25 000 of such unit cells are generated and each discretized to result in a 32×32 pixel image; they form the inputs to the neural network, i.e. $\mathbf{x} \in \mathbb{R}^{32\times32}$.

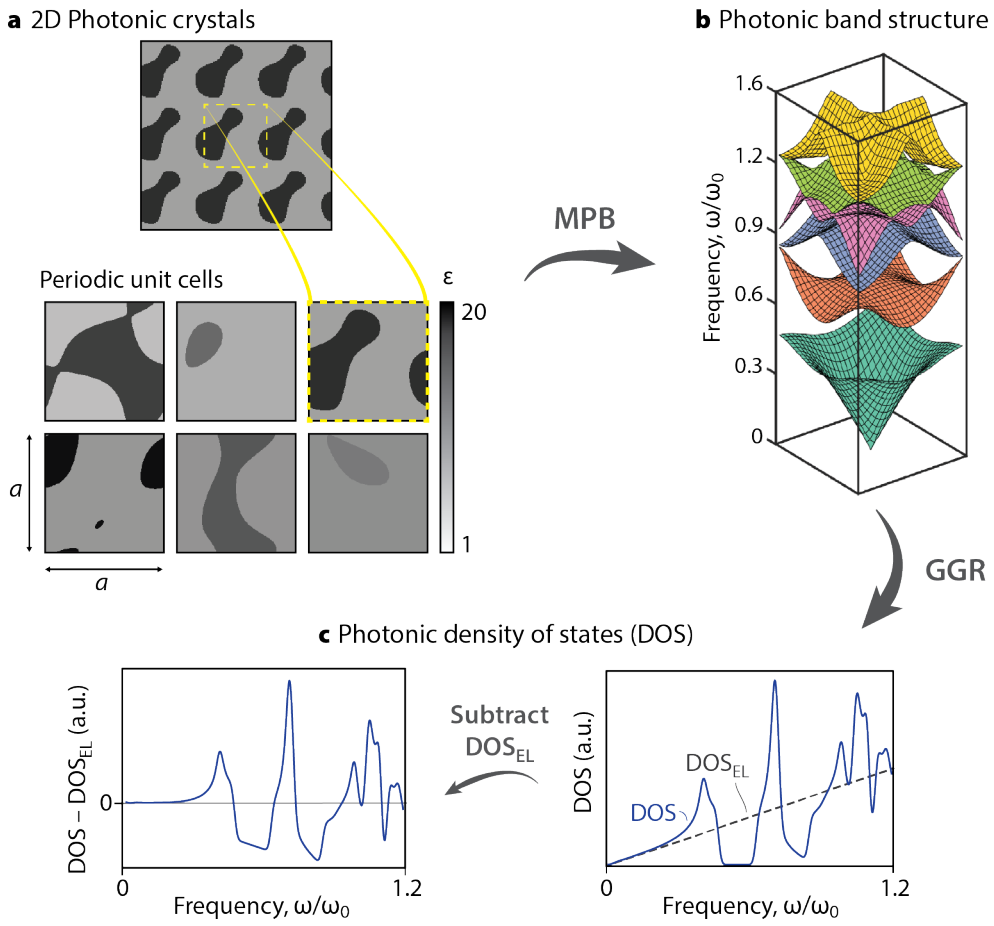

Figure 3-1: 2D photonic crystals density of states (DOS) calculation procedure, where (a) square unit cells are generated using a level set of Fourier sums; (b) their transverse magnetic band structures are computed using the MIT photonic bands (MPB) package; and (c) used to compute their DOS via the Generalized Gilat-Raubenheimer (GGR) method; network labels are derived by substracting the "empty-lattice" DOS. Quantities are made dimensionless by normalizing with $\omega_0 = 2\pi c/an_{\text{avg}}$ (with n_{avg} being the average index).

3.1.2 Density-of-states calculation

The DOS of a 2D PhC can be defined as [55]

$$\text{DOS}(\omega) = \frac{A}{(2\pi)^2} \sum_n \int_{\text{BZ}} \delta(\omega - \omega_{n\mathbf{k}})\, \mathrm{d}^2\mathbf{k}, \tag{3.3}$$

with ω denoting the considered frequency, $\omega_{n\mathbf{k}}$ the PhC band structure, n the band index, $\mathbf{k}$ the momentum in the BZ and $A = a^2$ the unit cell area. Fig. 3-1 summarizes the procedure of evaluating the DOS and is described in more detail here. Using the MIT Photonic Bands

(MPB) package [56], the transverse magnetic (TM) polarized band structure $\omega_n(\mathbf{k})$ of each unit cell is computed, up to the lowest 10 bands. A resolution of 25×25 plane waves (or equivalently, 25×25 $\mathbf{k}$-points) is used over the BZ, $-\pi/a < k_{x,y} \leq \pi/a$. The group velocities at each $\mathbf{k}$-point are also extracted and used to compute the photonic density-of-states (DOS) using an extrapolative technique based on the Generalized Gilat–Raubenheimer (GGR) method [57], in an implementation adapted from [58]. The DOS spectrums, computed to a resolution of 20 000 points, are then subjected to three simple post-processing steps: (i) spectral smoothing using a narrow Gaussian kernel S_ω of size 100, (ii) shifting by the DOS of the "empty-lattice", i.e. uniform lattice of index n_{avg} (where $n_{\text{avg}} = \frac{1}{A}\int_A \sqrt{\varepsilon(\mathbf{r})}\,\mathrm{d}^2\mathbf{r}$), $\text{DOS}_{\text{EL}}(\omega) = \omega a^2 n_{\text{avg}}^2/2\pi c^2$, and (iii) rescaling *both* the DOS- and the frequency-values by $\omega_0 = 2\pi c/an_{\text{avg}}$. More explicitly, the labels for network training are defined as

$$\mathbf{y} \triangleq \omega_0[(S_\omega * \text{DOS}) - \text{DOS}_{\text{EL}}](\omega/\omega_0), \tag{3.4}$$

where $\mathbf{y} \in \mathbb{R}^{400}$ is uniformly interpolated over the normalized spectral range $0 \leq \omega/\omega_0 \leq 0.96.$, Step (i) accounts for the finite spectral width of physical measurements and regularizes singularities associated with e.g. van Hove points; step (ii) counteracts the linear increase in average DOS that otherwise leads to a bias at higher frequencies, emphasizing instead the local spectral features of the DOS; and step (iii) ensures comparable input- and output-ranges across distinct unit cells, regardless of the cell's average index.

In this work, I consider two target problems associated with 2D PhCs, namely, predicting the DOS ($\mathbf{y} \in \mathbb{R}^{400}$) and predicting the band structure (lowest 6 TM bands, $\mathbf{y} \in \mathbb{R}^{6\times25\times25}$) of 2D periodic unit cells ($\mathbf{x} \in \mathbb{R}^{32\times32}$).

3.1.3 Surrogate dataset

One of the auxiliary information sources exploited by SIB-CL is a surrogate dataset which can be easily obtained via inexpensive means. For this purpose, a simple dataset of 10 000 PhC cylindrical structures, i.e. the unit cells of circles with periodicity a and varying radii, is generated. For each unit cell, the radius is uniformly sampled in $(0, a/2]$ and the two permittivities inside and outside the circle are each uniformly sampled in $[1, 20]$. This

simple structure is chosen since there exist semi-analytical solutions (e.g. using Korringa–Kohn–Rostoker equations or multiple scattering theory [59, 60, 61, 62]) and thus accurate calculations of their band structures can be obtained rather quickly. Consequently, such a dataset can be curated at a very low computational cost and qualifies as a surrogate dataset of *inexpensive* data. Here, since the MPB computations and DOS calculations were relatively fast for the resolution considered and to avoid having to separately implement one of the above methods, the same procedure as in Section 3.1.2 is used to compute the surrogate labels $\tilde{\mathbf{y}}$, for convenience, and post-processed similarly according to Eq. (3.4).

3.2 Time-independent Schrödinger equation

In order to test the generality of SIB-CL, I also considered second scientific problem of solving the time-independent Schrödinger Equation (TISE). Solving for the electronic structure of molecules and materials is of fundamental importance in many disciplines across physics, chemistry and material science. For simplicity, I consider the toy problem of predicting the ground state energy of single electrons in random 3D potentials, also explored in [63, 64] (in 1D and 2D).

3.2.1 Unit cells and ground state energy calculation

The energy states, E_n, of a (non-relativistic) quantum system can be found by solving the TISE:

$$\hat{H}\psi_n = (\hat{T} + \hat{U})\psi_n = E_n\psi_n, \tag{3.5}$$

where $\hat{H}$ is the Hamiltonian and is equal to the sum of the kinetic energy $\hat{T} = -\frac{\hbar^2}{2m_e}\nabla^2$ and potential energy $\hat{U} = U(\mathbf{r}) \triangleq U(x, y, z)$ operators. This is essentially an eigenvalue problem, where E_n and ψ_n are the nth eigenvalues and eigenfunctions of $\hat{H}$ respectively. For simplicity, I use Hartree atomic units (h.a.u.), i.e. $\hbar = m_e = 1$.

To generate samples of $U(\mathbf{r})$, the same procedure in Eqs. (3.1) and (3.2) is first used to create two-tone potential profiles in 3D, i.e. $\mathbf{r} = (x, y, z)$ and $\mathbf{n}_k = (n_x, n_y, n_z)$ are now 3D vectors. Finer features are created by increasing the number of Fourier components

to $n_x, n_y, n_z \in [-2, -1, 0, 1, 2]$ (and hence the summation in Eq. (3.1) now runs over 125 terms). ε_1 in Eq. (3.2) is set to 0 here, while ε_2 is uniformly sampled in $(0, 1]$. To remove the periodicity, 20% of the unit cell is truncated from *each* edge. Next, a gaussian filter (with a kernel size 8% of the current unit cell dimension) is applied to smooth the potential profile and, finally, the unit cells are discretized to a $32 \times 32 \times 32$ resolution. This process is illustrated in Fig. 3-2, in 2D for simplicity. Examples of unit cells in the 3D dataset are depicted in Fig. 3-4.

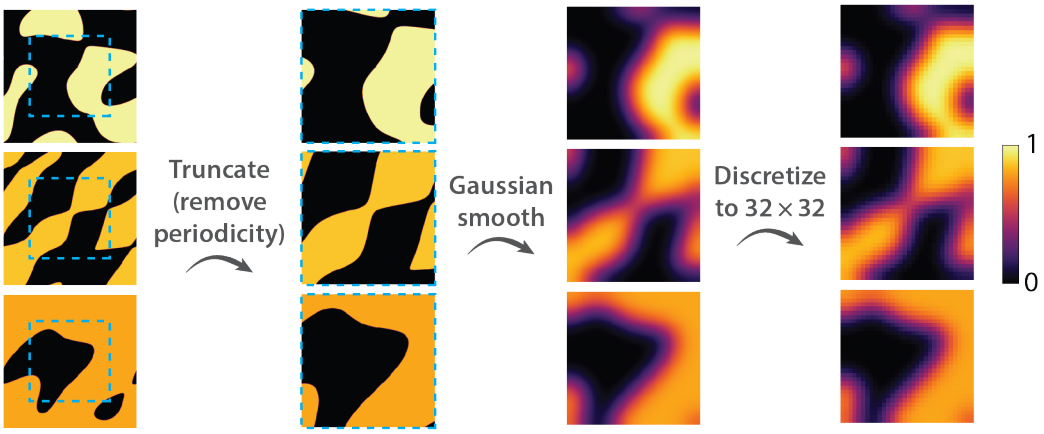

Figure 3-2: Processing steps used to generate random (non-periodic) potentials for the inputs to the Schrödinger equation, illustrated here in 2D.

The physical length of the 3D box per dimension is set at 10 and infinite potentials are assumed outside this box. The dynamic range of $U(x, y, z)$ within the box is set to be in $[0, 1]$ (i.e. for $0 < x, y, z < 10$, and $U(x, y, z) \approx \infty$ otherwise); this range is selected to produce non-trivial wavefunctions. A dataset with trivial solutions is one where the samples have their wavefunctions either all (i) resemble the ground-state solution of a "particle in a box", i.e. the variations in the potential are too weak relative to the confinement energy of the box, or all (ii) highly localized in local minimas of the potential, i.e. the variations in the potentials are too strong relative to the confinement energy. Both extremes would not create a meaningful deep learning problem and they are avoided by carefully controlling the ratio between the potential range and length scale. Some examples of potentials in the dataset and their corresponding ground-state solution are displayed in Fig. 3-3, in 2D for simplicity, which show that they indeed avoid the two aforementioned extremes.

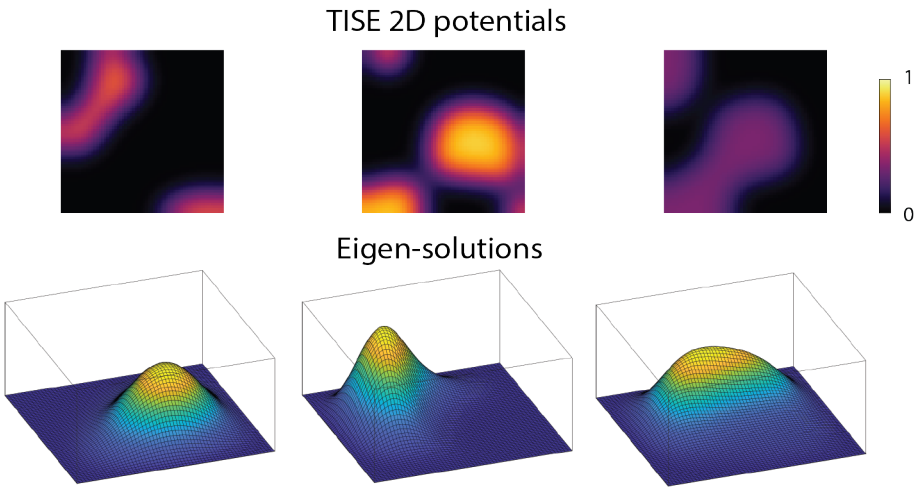

Figure 3-3: Random potentials (shown here in 2D for convenience) and their corresponding eigen-solution show that non-trivial wavefunctions are obtained.

To obtain the ground state energies, E_0, a (central) finite difference method with implicit Dirichlet boundary conditions is implemented to approximate $\hat{H}$ and Eq. (3.5) is solved numerically using standard eigensolvers (from the SciPy packages in Python). The data pair of $U(x, y, z)$ and E_0 are the inputs $\mathbf{x} \in \mathbb{R}^{32\times32\times32}$ and labels $y \in \mathbb{R}$ respectively; examples of input-label pairs are illustrated in Fig. 3-4.

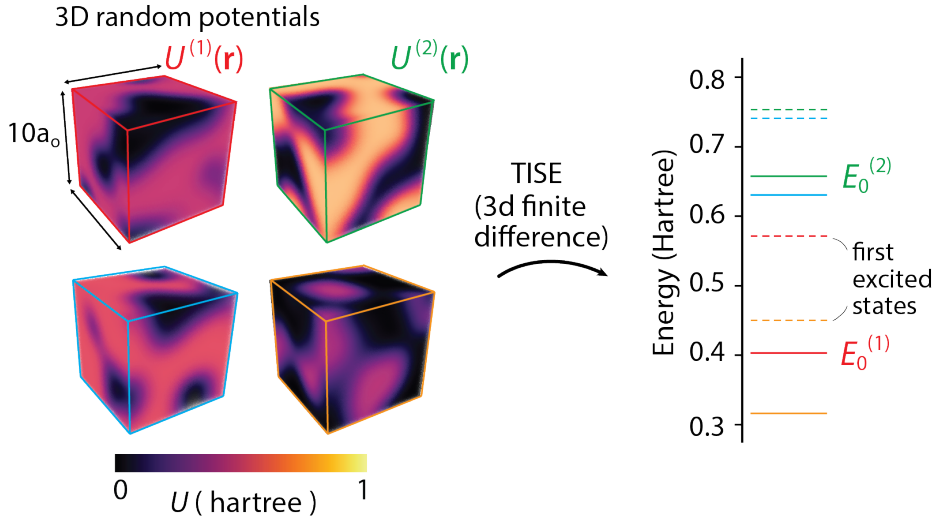

Figure 3-4: Examples of 3D random potentials $U(\mathbf{r})$ and their corresponding ground-state energies E_0 obtained from solving the time-independent Schrödinger equation.

3.2.2 Surrogate datasets

Two types of surrogate datasets are considered for the TISE problem; (i) where the ground-state energies are computed using a lowered resolution of the original inputs (Fig. 3-5a), and (ii) quantum harmonic oscillator (QHO) potentials with closed-form ground-state solutions (Fig. 3-5b).

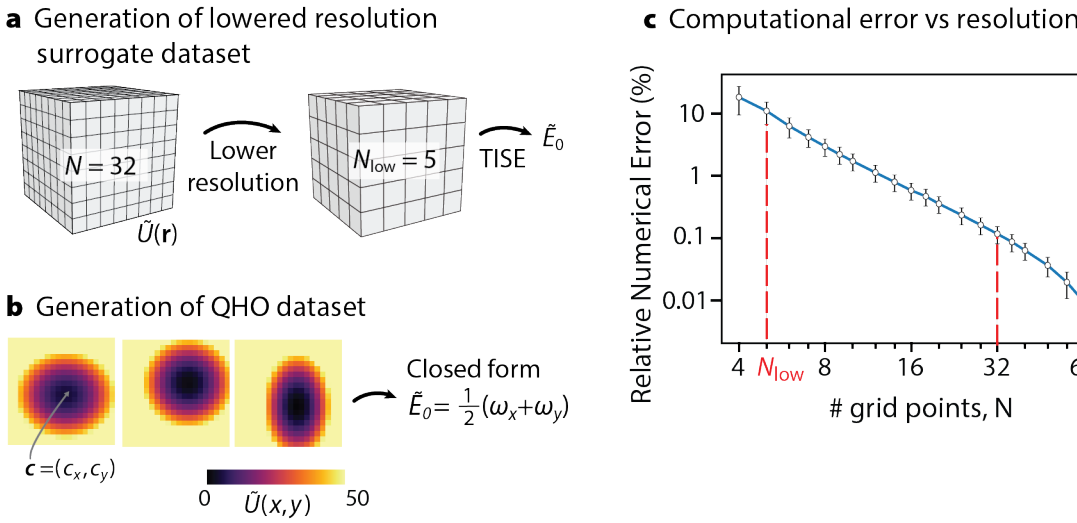

Figure 3-5: Generation of two surrogate datasets, (a) where ground-state energies are calculated from reduced resolution input and (b) of quantum harmonic oscillator (QHO) potentials with closed-form solutions. (c) Numerical error against number of grid points N (per dimension) used during computation, where error is relative to the solution at $N = 128$.

(i) is relevant for many computational problems in science and engineering, where fine meshing of the computational grid in finite element or finite difference methods are used to produce accurate, but computationally intensive results, and approximate data can often be more quickly obtained from using a coarser computational grid. Specifically, to generate surrogate dataset (i), the input resolution is reduced from $32\times32\times32$ to $5\times5\times5$ via bi-linear interpolation and the same numerical solver is used to calculate the ground-state energy, $\tilde{E}_0$. A $5 \times 5 \times 5$ resolution is selected to create labels with a relatively high numerical error of around 10% compared to the target dataset which has a 0.1% error (see Fig. 3-5c), to enable the investigation of whether approximate calculations can be useful. Assuming a primitive eigensolver (for dense matrices) with time complexity $O(n^3)$, and $n = N^3$ in 3D where N is the linear dimension, the reduced resolution computation will in theory be 10^7 times faster

for the resolutions considered in this problem. (In practice, this factor is smaller since one can make use of more efficient sparse matrix algorithms). For surrogate dataset (i), the inputs $\tilde{\mathbf{x}} = U(x, y, z) \in \mathbb{R}^{32\times32\times32}$ are still the original unit cells with non-reduced resolution (since the input to the encoder $\mathbf{H}$ needs to be of the same dimension across different steps) and labels $\tilde{y} = \tilde{E}_0 \in \mathbb{R}$ are the approximated ground-state solutions.

Surrogate dataset (ii) consists only of quantum harmonic oscillator (QHO) potentials given by:

$$\tilde{U}(\mathbf{r}) = \tfrac{1}{2}\boldsymbol{\omega}^{\circ 2} \cdot (\mathbf{r} - \mathbf{c})^{\circ 2}, \tag{3.6}$$

where $(\mathbf{A}^{\circ n})_i = A_i^n$ is the Hadamard (element-wise) power operator, and $\boldsymbol{\omega}$ and $\mathbf{c}$ are varied to obtain different unit cells. For quicker training of the neural networks, I work with the TISE problem in 2D for this surrogate dataset. Due to its simple form, there exists a simple analytical solution for the ground-state energy,

$$\tilde{E}_0 = \tfrac{1}{2}\left(\omega_x + \omega_y\right); \tag{3.7}$$

hence surrogate dataset (ii) can be generated without even using the numerical solver and so incurs near-zero computation cost. In fact, since infinite potentials are imposed at the boundaries (while the QHO potential never actually reaches infinity), $\tilde{E}_0$ given by Eq. (3.7) is merely an approximation of the QHO ground-state energies. This error is numerically computed to be around 2.6%. The sampling intervals of $\boldsymbol{\omega} \triangleq (\omega_x, \omega_y)$ and $\mathbf{c} \triangleq (c_x, c_y)$ are also chosen such that this error is kept low; the chosen intervals are: $\omega_x, \omega_y \in [0.3, 3.2]$ and $c_x, c_y \in [0, 4.5]$.

In this work, I consider two problem scenarios associated with the TISE; specifically, predicting the ground-state energy ($y \in \mathbb{R}$) from random 3D potentials ($\mathbf{x} \in \mathbb{R}^{32\times32\times32}$) and 2D potentials ($\mathbf{x} \in \mathbb{R}^{32\times32}$) when using surrogate datasets (i) and (ii) respectively.

Chapter 4

Proposed Model

4.1 Invariant Transformations

SIB-CL exploits various sources of auxiliary information to enable high-quality network training for data-scarce problems. One of these include prior knowledge in the form of invariances of the physical problem; the associated transformation operations used in this work are listed in this section. Specifically, here I will describe the group of invariant transformations, $\mathcal{G}$, involved for DOS prediction in 2D PhCs; the transformations involved in all other problems studied in this work can be obtained from a subset of this group. $\mathcal{G}$ can be derived as the product group $\mathcal{G} = \mathcal{G}_0 \times \mathcal{G}_t \times \mathcal{G}_s$, where $\mathcal{G}_0$, $\mathcal{G}_t$ and $\mathcal{G}_s$ represent the 4mm symmetry point group, the group of pixel-discrete translations and the group of scaling transformations respectively. In more detail,

1. **Point group symmetry,** 4mm ($\mathcal{G}_0$): this includes the identity operation (1), 2- and 4-fold rotations (C_2 and $C_4^{\pm}$), and horizontal, vertical, and diagonal mirrors (σ_h, σ_v, and $\sigma_d^{(\prime)}$), i.e. $\mathcal{G}_0 = \{1, C_2, C_4^-, C_4^+, \sigma_h, \sigma_v, \sigma_d, \sigma_d'\}$. For convenience, rotations are collectively defined as $C = \{C_2, C_4^-, C_4^+\}$ and mirrors as $\sigma = \{\sigma_h, \sigma_v, \sigma_d, \sigma_d'\}$ in later sections.

2. **Translation symmetry ($\mathcal{G}_\mathbf{t}$)**: While the DOS is invariant under all continuous translations $\mathbf{t}$, the pixelized unit cells are compatible only with pixel-discrete translations; i.e., here I consider the (factor) group $\mathcal{G}_\mathbf{t} = \{iN^{-1}a\hat{\mathbf{x}} + jN^{-1}a\hat{\mathbf{y}}\}_{i,j=0}^{N-1}$ with $N = 32$.

3. **Refractive scaling** ($\mathcal{G}_\text{s}$): Due to the scale-invariance [53] endowed by Maxwell equations, the set of (positive) amplitude-scaling transformations of the refractive index $g(s)n(\mathbf{r}) = sn(\mathbf{r})$ define a group $\mathcal{G}_\text{s} = \{g(s) \mid s \in \mathbb{R}_+\}$ and maps the PhC eigenspectrum from $\omega_{n\mathbf{k}}$ to $s^{-1}\omega_{n\mathbf{k}}$. Equivalently, $g(s)$ maps DOS(ω) to sDOS($s\omega$) and thus leaves the $\mathbf{y}$-labels of Eq. (3.4) invariant under the ω_0-normalization.

For band structure prediction in PhCs, since they are no longer integrated over the BZ, $\mathcal{G}$ is reduced to just the choice of freedom of unit cell origin (i.e., to translation invariance, $\mathcal{G} = \mathcal{G}_\mathbf{t}$). For the prediction of ground-state energies for the TISE problem, there is no periodicity (i.e. no translation symmetry); relevant transformations are rotations and mirrors, i.e. $\mathcal{G} = \mathcal{G}_\mathbf{0}$ for 2D. In 3D, we instead have the $\mathrm{m}\bar{3}\mathrm{m}$ point group, which has 48 unique symmetry operations (instead of just 8, listed above for the 4mm point group in 2D).

4.2 Surrogate- and invariance- boosted contrastive learning (SIB-CL)

The goal is to train a neural network to predict desired properties (or labels) $\mathbf{y}$ from inputs $\mathbf{x}$ using minimal training data. More precisely, for a target problem $D_\text{t} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_\text{t}}$ consisting of N_t input–label pairs, the focus is on problem spaces where N_t is too small to successfully train the associated network. To overcome this, two auxiliary data sets are introduced: (1) a set of zero-cost unlabeled inputs $D_\text{u} = \{\mathbf{x}_i\}_{i=1}^{N_\text{u}}$ and (2) a surrogate data set $D_\text{s} = \{\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i\}_{i=1}^{N_\text{s}}$ consisting of inexpensively computed labels $\tilde{\mathbf{y}}_i$ (e.g. from approximation or semi-analytical models) with associated input $\tilde{\mathbf{x}}_i$ (possibly, but not necessarily, different from $\mathbf{x}$). The quantity of each of these auxiliary data sets are assumed to far exceed the target problem, i.e. $\{N_\text{u}, N_\text{s}\} \gg N_\text{t}$ (and, typically, $N_\text{u} > N_\text{s}$).

On the basis of these auxiliary datasets, here, I introduce a novel framework—Surrogate and Invariance-Boosted Constrastive Learning (SIB-CL)—which significantly reduces the data requirements on D_t (Fig. 4-1). SIB-CL achieves this by splitting the training process into two stages: an interleaved two-step pre-training stage using the auxiliary data sets D_u and D_s (Fig. 4-1a,b), followed by a fine-tuning stage using the target data set D_t (Fig. 4-

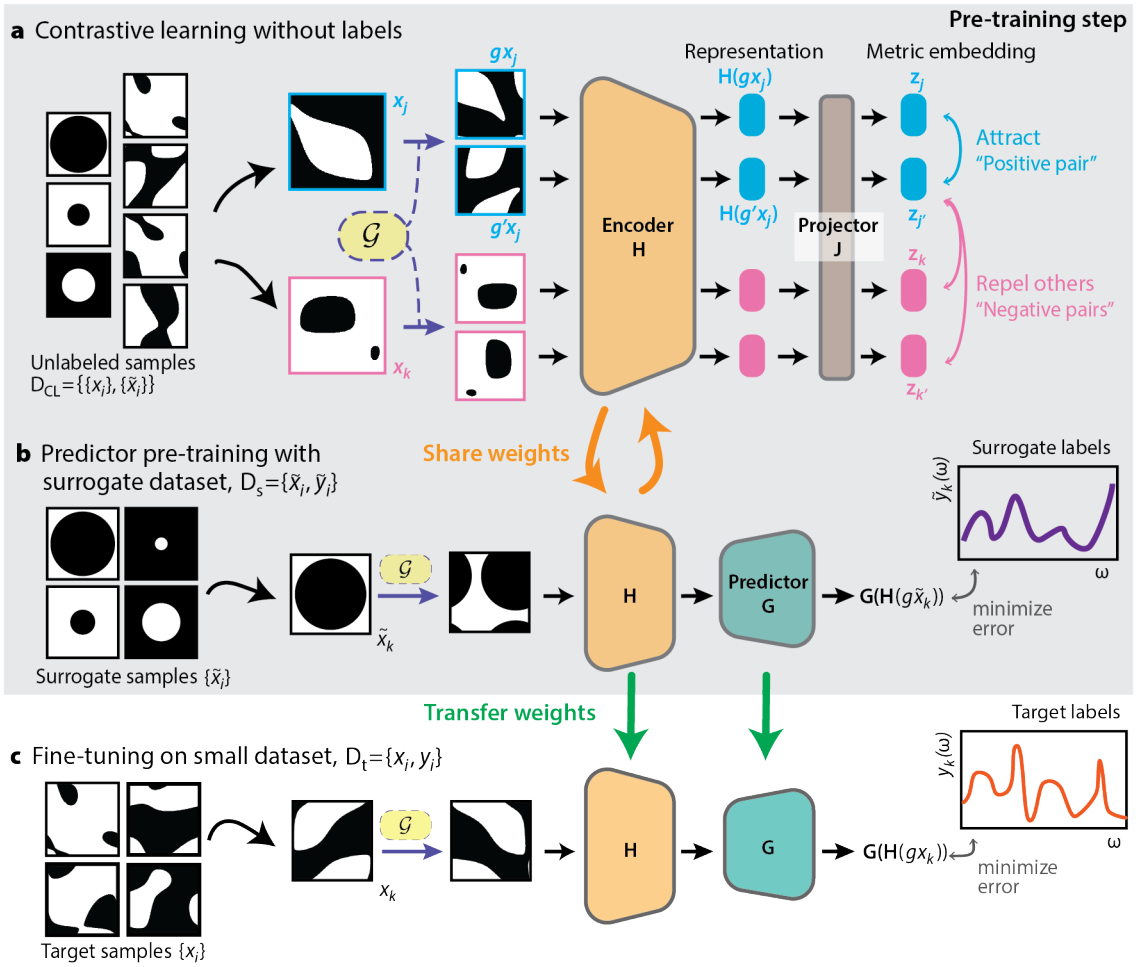

Figure 4-1: Surrogate- and Invariance-Boosted Contrastive Learning (SIB-CL) framework. Network training proceeds via a pre-training stage (a,b) followed by a fine-tuning stage (c). The pre-training stage alternates a contrastive learning step (a) using unlabeled data D_{CL} with a supervised learning step (b) using surrogate data D_{s}. The weights are then transferred for fine-tuning step (c) on the small target dataset D_{t}.

1c). In the first step of the pre-training stage (Fig. 4-1a), contrastive learning is used to learn invariances in the problem space using unlabelled inputs aggregated from the target and surrogate data sets $D_{\text{CL}} = \{\mathbf{x}_i\}_{i=1}^{N_{\text{u}}} \cup \{\tilde{\mathbf{x}}_i\}_{i=1}^{N_{\text{s}}}$. D_{CL} is complemented by a set of physics-informed invariance relations $\{g\}$ (i.e. elements of the group $\mathcal{G}$) which maps input $\mathbf{x}_i$ to $g\mathbf{x}_i$ while retaining its label $\mathbf{y}_i$. In SIB-CL, this step is based on the SimCLR technique [41], though alternative techniques like BYOL [42] are also explored here (see Section 4.2.1). Specifically, for each input $\mathbf{x}_i$ in D_{CL} (sampled in batches of size B), two derived variations $g\mathbf{x}_i$ and $g'\mathbf{x}_i$ are created by randomly sampling two concrete mappings g and g' from $\mathcal{G}$. The

resultant $2B$ inputs are then fed into encoder and projector networks, $\mathbf{H}$ and $\mathbf{J}$ respectively, producing metric embeddings $\mathbf{z}_{i^{(\prime)}} = \mathbf{J}(\mathbf{H}(g^{(\prime)}\mathbf{x}_i))$. A positive pair $\{\mathbf{z}_i, \mathbf{z}_{i'}\}$ is the pair of metric embeddings derived from the two variations of $\mathbf{x}_i$, i.e. $g\mathbf{x}_i$ and $g'\mathbf{x}_i$; all other pairings in the batch are considered negative. At each training step, the weights of $\mathbf{H}$ and $\mathbf{J}$ are simultaneously updated according to a contrastive loss function defined by the normalized temperature-scaled cross entropy (NT-Xent) loss [41]:

$$\mathcal{L}_{ii'} = -\log \frac{\exp(s_{ii'}/\tau)}{\sum_{j=1}^{2B}[i \neq j]\exp(s_{ij}/\tau)}, \tag{4.1}$$

where $s_{ii'} = \hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_{i'}$ (and $\hat{\mathbf{z}}_i = \mathbf{z}_i/\|\mathbf{z}_i\|$) denotes the cosine similarity between two normalized metric embeddings $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}_{i'}$, $[i \neq j]$ uses the Iverson bracket notation, i.e. evaluating to 1 if $i \neq j$ and 0 otherwise, and τ is a temperature hyperparameter. The total loss is taken as the sum across all positive pairs in the batch. In the batch sampling of D_{CL}, one-third of each batch is sampled from D_{s} and two-thirds from D_{u}. Conceptually, the NT-Xent loss acts to minimize the distance between embeddings of positive pairs (numerator of Eq. (4.1)) while maximizing the distances between embeddings of negative pairs in the batch (denominator of Eq. (4.1)), essentially "pulling together" representations related by invariance relations and "pushing apart" representations that are not. Consequently, representations $\mathbf{H}(\mathbf{x}_i)$ that respect the underlying invariances of the problem are obtained.

Each epoch of contrastive learning (i.e. each full sampling of D_{CL}) is followed by a supervised learning step—the second step of the pre-training stage (Fig. 4-1b)—on the surrogate dataset D_{s}, with each input from D_{s} also subjected to a random invariance mapping. This supervised step shares the encoder network $\mathbf{H}$ with the contrastive step but additionally features a predictor network $\mathbf{G}$, both updated via a task-dependent supervised training loss function (which will be separately detailed later). This step pre-conditions the weights of $\mathbf{G}$ and further tunes the weights of $\mathbf{H}$ to suit the target task.

The pre-training stage is performed for 100 to 400 epochs and is followed by the fine-tuning stage (Fig. 4-1c). This final stage uses D_{t} to fine-tune the networks $\mathbf{H}$ and $\mathbf{G}$ to the actual problem task—crucially, with significantly reduced data requirements on D_{t}. The associated supervised training loss function is again problem-dependent and may even differ

from that used in the pre-training stage.

4.2.1 Other contrastive learning techniques

Methods like SimCLR are based on instance discrimination since the network is trained to discriminate between other "negative" samples in the batch. Intuitively, such methods may not seem useful for regression problems, where the latent space is often continuous rather than discrete/clustered like in classification problems. Thus, as an alternative to SimCLR [41] for the contrastive learning step, I also explore a second technique, BYOL [42], which notably does not use explicit negative pairs in its contrastive loss function and is illustrated in Fig. 4-2.

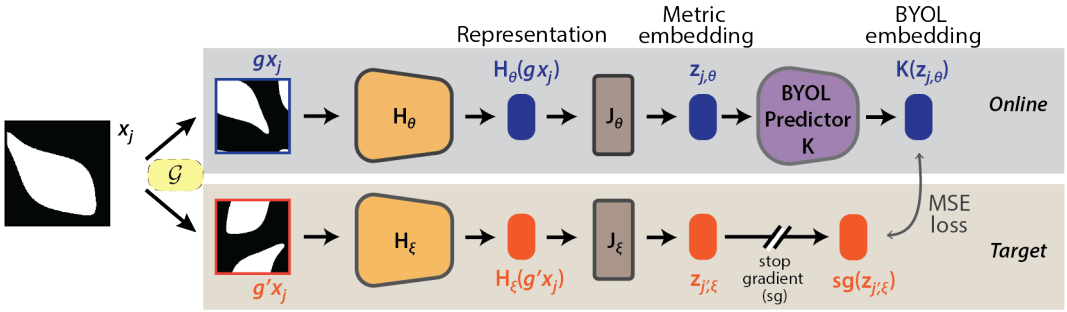

Figure 4-2: Illustration of the BYOL technique, which can be used as an alternative contrastive learning method to replace the SimCLR technique in Fig. 4-1a.

In BYOL [42] as illustrated in Fig. 4-2, instead of a single network like in SimCLR [41], there are two networks of the same architecture: a trainable online network (with weights θ) and a fixed target network (with weights ξ), and thus the following modifications are made to Fig. 4-1a. The two variations of a unit cell obtained from invariance relations are separately fed through the online and target network to obtain their embeddings; for the online embedding, this is the embedding after an additional BYOL predictor network that is not present in SimCLR. The distance between the (normalized) online and target embeddings are minimized via mean squared error (MSE) loss,

$$\mathcal{L}^{BYOL}_{jj',\theta\xi} = \left\| \bar{K}(z_{j,\theta}) - \bar{z}_{j',\xi} \right\|^2 = 2 - 2 \cdot \frac{\langle K(z_{j,\theta}), z_{j',\xi} \rangle}{\left\| K(z_{j,\theta}) \right\| \cdot \left\| z_{j',\xi} \right\|}, \tag{4.2}$$

where $\|\cdot\|$ indicates the L2 norm; this loss is used to update only the trainable online network. As for the target network, at each training step, it is updated using an exponential moving average of the online network weights, i.e.

$$\xi \leftarrow \tau\xi + (1-\tau)\theta, \tag{4.3}$$

where $\tau \in [0, 1]$ is a target decay rate. Unlike SimCLR's loss function (Eq. (4.1)), where negative pairs are explicitly being "pushed apart" (due to the denominator of the cross entropy loss function), the BYOL loss (Eq. (4.2)) does not explicitly involve negative pairs. Both contrastive learning methods are explored in SIB-CL and will be compared in the results section.

4.3 Baselines

To evaluate its effectiveness, SIB-CL is benchmarked against four baselines:

1. **Direct supervised learning (SL)**: randomly initialized networks **H** and **G** are trained using supervised learning on the target dataset D_{t}. This reflects the performance of direct supervised learning, i.e. without exploiting auxiliary data sources.

2. **Conventional transfer learning (TL)**: networks **H** and **G** are first pre-trained using supervised learning on the surrogate dataset D_{s} and then subsequently fine-tuned on D_{t}. This reflects the performance of surrogate-boosted supervised learning, wherein information from the surrogate task is incorporated using the conventional transfer learning technique.

3. **Supervised learning with invariances (SL-I)**: similar to SL, except that each input in D_t undergoes a random invariance transformation sampled from $\mathcal{G}$ before entering **H**. This reflects the performance of invariance-boosted supervised learning, wherein the invariances are incorporated via a conventional data augmentation approach [65].

4. **Transfer learning with invariances (TL-I)**: similar to TL, except that each input in D_s and D_t undergoes a random invariance transformation sampled from $\mathcal{G}$ before

entering **H** during *both* the pre-training and fine-tuning stages. This reflects the performance of surrogate- and invariance- boosted supervised learning via the use of conventional TL and data augmentation techniques. Crucially, this baseline critically benchmarks SIB-CL's effectiveness at incorporating these auxiliary information sources, since they both involve similar auxiliary sources.

4.4 Training Details

The network architectures used for all the problems in this work and further training details (e.g. hyperparameters, regularization approaches, etc) are detailed in Appendix A. In this section, I describe the loss metrics used during training and evaluation for the various problems.

DOS prediction. In the second step of the pre-training stage for SIB-CL, TL and TL-I (where supervised learning is performed on D_s) (Fig. 4-1b), the pre-training loss function,

$$\mathcal{L}^{PT} = \mathrm{mean}_{\omega/\omega_0}\left(\log(1 + |\mathbf{y}^{\mathrm{pred}} - \mathbf{y}|)\right) \tag{4.4}$$

is used, where $\mathbf{y}^{\mathrm{pred}}$ and $\mathbf{y}$ are the network prediction and the true label of each sample respectively and $|\cdot|$ gives the element-wise absolute value. The mean over the (ω_0-normalized) frequency axis is taken to get a scalar for $\mathcal{L}^{PT}$. The purpose of using this (over the standard L1 or MSE) loss function is to encourage the network to learn from the surrogate dataset the general features in the DOS spectrum and underemphasize the loss at places where the DOS diverges, i.e. at the Van Hove singularities. After the pre-training step, the standard L1 loss function is used during fine-tuning on D_t (Fig. 4-1c) for SIB-CL and all the baselines.

During evaluation, the "empty-lattice" DOS is first added back and a relative error measure also used in [58] is used for easier interpretation:

$$\mathcal{L}^{\mathrm{eval}} = \frac{\sum_{\omega/\omega_0}\left|\mathrm{DOS}_*^{\mathrm{pred}} - \mathrm{DOS}_*\right|}{\sum_{\omega/\omega_0}\mathrm{DOS}_*}, \tag{4.5}$$

where $\mathrm{DOS}_*^{\mathrm{pred}} = \omega_0^{-1}\mathbf{y}^{\mathrm{pred}} + \mathrm{DOS}_{\mathrm{EL}}$ and $\mathrm{DOS}_* = \omega_0^{-1}\mathbf{y} + \mathrm{DOS}_{\mathrm{EL}}$ are the predicted and actual

DOS respectively. The subscript $(*)$ is used to indicate the gaussian-smoothened DOS, i.e. $\mathrm{DOS}_* = S_\omega * \mathrm{DOS}$ (see Section 3.1.2).

Band structure prediction. During supervised training, for *both* pre-training and fine-tuning (Fig. 4-1bc), the mean squared error (MSE) loss function is used. For evaluation, a relative error measure (for easier interpretation) is used and is given by,

$$\mathcal{L}^{\mathrm{eval}} = \mathrm{mean}_{\mathbf{k}} \left(\frac{1}{6} \sum_{n=1}^{6} \frac{|\omega_n^{\mathrm{pred}}(\mathbf{k}) - \omega_n(\mathbf{k})|}{\omega_n(\mathbf{k})} \right), \tag{4.6}$$

where $\omega_n(\mathbf{k})$ are the eigen frequencies indexed over band numbers $n = 1, 2, ..., 6$ and $\mathbf{k}$ are the wave vectors restricted to the Brillouin zone, i.e. $-\pi/a < k_{x,y} \leq \pi/a$. The evaluation loss is taken as the mean over all 6 bands and over all $\mathbf{k}$-points.

Ground-state energy prediction. Analogous to the band structure prediction problem, the mean squared error (MSE) loss function is used, for *both* the pre-training and fine-tuning stage (Fig. 4-1bc). During evaluation, a simple relative error measure,

$$\mathcal{L}^{\mathrm{eval}} = |y^{\mathrm{pred}} - y|/y \tag{4.7}$$

is used, where y^{pred} is the network prediction and $y = E_0$ is the actual ground-state energy for each sample in the test set. The metric given by Eq. (4.7) is also used to express the computation error of the numerical method used to solve the Schrödinger equation.

Chapter 5

Results and Discussion

5.1 Photonic Crystals

5.1.1 DOS prediction

The performance of SIB-CL relative to the four baselines, using the evaluation error function Eq. (4.5), is depicted in Fig. 5-1a. Fig. 5-1a shows the average error on a fixed test set over multiple fine-tuning runs, with each run featuring a randomly selected dataset of size N_t. Averaging over different target datasets of the same size provides a better assessment of the effectiveness of the pre-training stage in the different methods by removing fluctuations in the choice of the training dataset. The error bars in Fig. 5-1a represent the one-standard-deviation level of uncertainty between these runs. A significant reduction of prediction error (e.g. from 7.6% in SIB-CL to 4.6% in SL when $N_t = 100$) is achieved when using SIB-CL compared to the baselines, especially at low sample numbers. More notably, we see a large reduction in the number of fine-tuning samples N_t needed to achieve the same level of prediction error, which directly illustrates the data savings in the data-scarce problem. Specifically, a 12× (7×) savings in N_t is achieved when compared to the SL (TL) baseline at the same level of prediction error. For better visualization of the network predictions, the predicted DOS spectrums are plotted against their true spectrums in Fig. 5-1b for samples in the test set (illustrated in the inset) at various error levels.

Also notable from the results is SIB-CL's performance edge over the TL-I baseline,

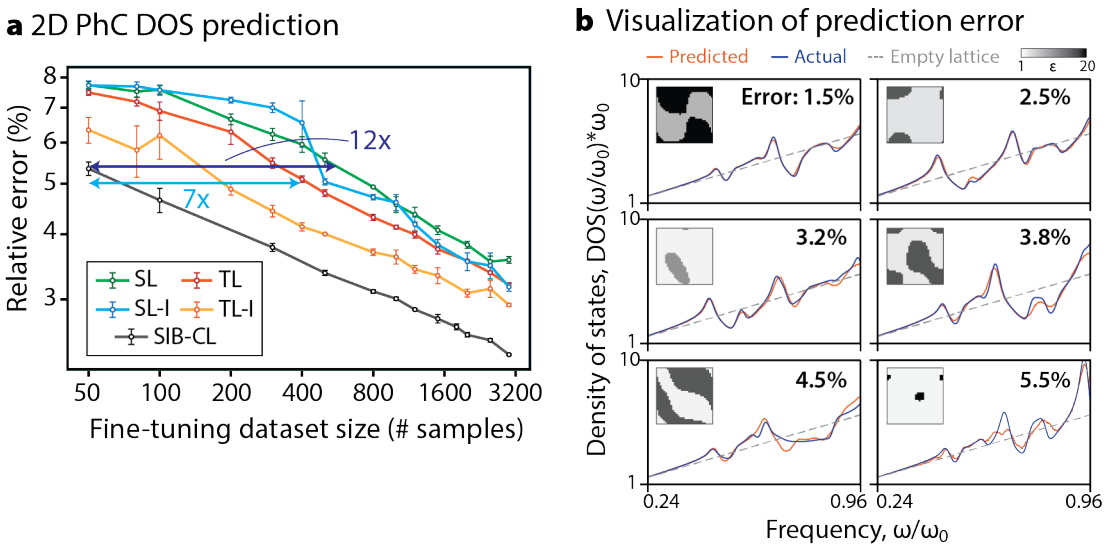

Figure 5-1: (a) Performance of SIB-CL contrasted against the four baselines described in Section 4.3 for the density of states (DOS) prediction of 2D photonic crystals (PhC). (b) Visualization of the network prediction of the (gaussian-smoothened) DOS, compared against their actual values, at various prediction error levels.

which is significant since both methods incorporate the same set of auxiliary information, i.e. of both the surrogate dataset and prior knowledge of invariances. This performance edge could come from enforcing the invariances using contrastive learning (as opposed to the more primitive data augmentation approach). When invariant transformations are removed from the contrastive learning step, SIB-CL gives very similar prediction accuracies as TL (see Section 5.3), consistently across different contrastive learning techniques (SimCLR or BYOL). This observation underscores the interplay between invariances and the contrastive loss mechanism, and that the performance edge comes from using both in conjunction. It is also possible that SIB-CL benefits from learning through massive amounts of unlabeled data in a self-supervised manner, i.e. mediated by the pretext task of invoking invariances. This large amount of unlabeled data used during the contrastive learning stage might help to improve generalization in the encoder $\mathbf{H}$, preventing the network from over-fitting (possibly to both the surrogate and target datasets). Over-fitting is an especially relevant problem in this work since the fine-tuning target dataset is small.

The SimCLR technique used in SIB-CL is an example of contrastive learning methods based on instance discrimination which, intuitively, may not seem useful for regression

problems (see Section 4.2.1). Indeed, several observations are made empirically that supports this claim—notably, the widely corroborated finding [41, 66, 67], in contrastive learning for classification problems, that a larger batchsize is always more beneficial is not echoed here. This could be evidence that instance discrimination is not highly appropriate in regression problems since a larger batchsize is desirable to create more negative examples. On this ground, an alternative contrastive learning method not based on instance discrimination, BYOL [42], is explored. However, BYOL is found to give comparable, and sometimes poorer, model accuracies as shown in Fig. 5-2. This warrants a deeper investigation; despite their incredible success, contrastive learning remains poorly understood and lacks a solid theoretical explanation [68, 69, 70, 71] on why and when these algorithms work well. The development of a solid theoretical framework will be crucial to provide better insights on the workings of contrastive learning methods for regression tasks.

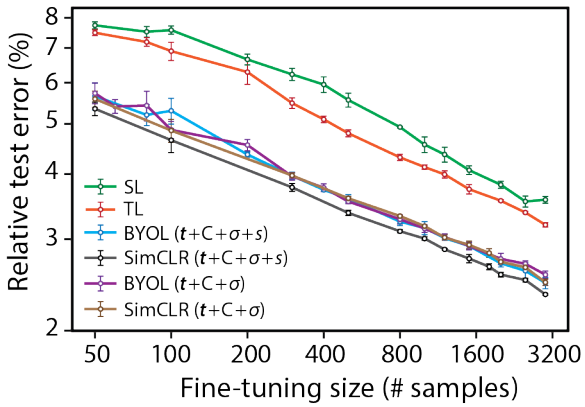

Figure 5-2: Performance of SIB-CL when using the BYOL technique, compared to the SimCLR technique, as well as to the SL and TL baselines. Various combinations of transformations from the group $\mathcal{G} = \{\mathbf{t}, C, \sigma, s\}$ are considered, including translations ($\mathbf{t}$), rotations (C), mirrors (σ) and scaling (s).

5.1.2 Band structure prediction

To verify that SIB-CL's effectiveness is not limited to the DOS prediction problem, SIB-CL is also evaluated on the problem of predicting (the lowest 6 TM) photonic band structures of 2D PhCs. Fig. 5-3a compares SIB-CL's performance against the baselines, with the prediction error defined according to the evaluation metric in Eq. (4.6). SIB-CL is seen to

enable more than 60× data savings when compared to *all* the baselines. Visualizations of the band structure predicted by SIB-CL at various error levels are also depicted in Fig. 5-3b, where the markers indicated the network prediction and the surface plots indicate the true band structure.

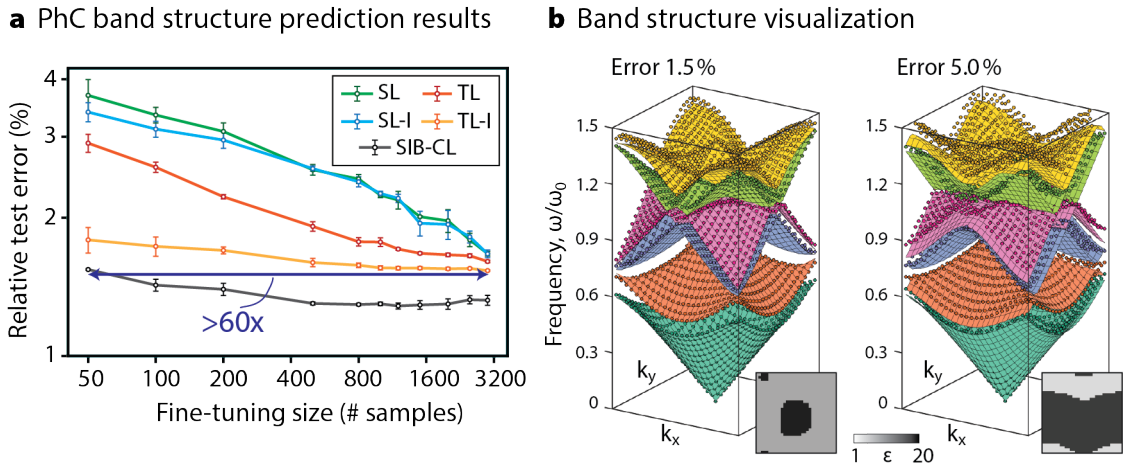

Figure 5-3: (a) Performance of SIB-CL, compared against the baselines described in Section 4.3, for the band structure prediction of 2D photonic crystals. (b) Visualizations of the network predictions of the band structure (indicated by markers) relative to the actual band structure (indicated by the surface plots) as various network prediction error levels.

While the data savings achieved by SIB-CL is highly impressive, I caution here that the choice (and associated implications) of the evaluation metric for the band structure prediction task is more subtle (than the DOS task). This is because most of the unique features in the band structure of each unit cell are contained only at some **k**-points and the band structure at the remaining **k**-points can be very well approximated using the "empty-lattice" (a featureless unit cell filled entirely by the average permitivitty); by taking the mean over **k**, the prediction error is undesirably under-emphasized. This subtlety is further illustrated in Table 5.1, which shows the relative numerical error (according to metric Eq. (4.6)) of the band structures when computed using a reduced resolution (i.e. approaching the "empty-lattice") of the unit cell. At a resolution of 4×4 we have an error amounting to only 3.5%, suggesting that the "empty-lattice" band structure is a rather good approximation. In other words, to evaluate the band structure prediction problem more critically, an asymmetric loss function that emphasizes the differences at the "special"

k-points (and de-emphasizes the differences elsewhere) should ideally be used. For the sake of brevity, this is not explored here; thus, the relative differences in the prediction error produced by SIB-CL and the baselines should be interpreted with caution. Nevertheless, Fig. 5-3a serves as clear evidence that SIB-CL outperforms all baselines.

Resolution	**Band structure error**	**DOS error**
4×4	3.5%	18.7%
8×8	1.7%	3.1%
16×16	0.5%	1.0%

Table 5.1: Computational accuracy of band structure compared to DOS.

5.2 Time-independent Schrodinger equation

5.2.1 Reduced resolution surrogate

Fig. 5-4a compares the performance of SIB-CL against the baselines, when using the surrogate dataset of reduced resolution data (Section 3.2.2) for the target task of predicting the ground-state energies in 3D random potentials. Up to 35× data savings is obtained when using SIB-CL as compared to SL, thus validating SIB-CL's effectiveness on problems spanning across different scientific domains. This result also demonstrates that prior knowledge of invariances used in SIB-CL can be arbitrarily simple, just like the simple rotation and mirror transformations used here, yet can still lead to significant data savings.

As a validation step, from Fig. 5-4a, the network prediction accuracies are noted to be in the orders of $\approx$ 1%; this makes the surrogate dataset with $\approx$ 10% error (Fig. 3-5c) an appropriate design choice as *approximate* data, and the original dataset with $\approx$ 0.1% numerical error (Fig. 3-5c) an appropriate design choice for target data. All errors here are defined using the same network evaluation metric of Eq. (4.7).

5.2.2 QHO surrogate

When the QHO surrogate data (Section 3.2.2) is used in place of the reduced resolution surrogate data, up to 4× data savings is obtained when using SIB-CL compared to the

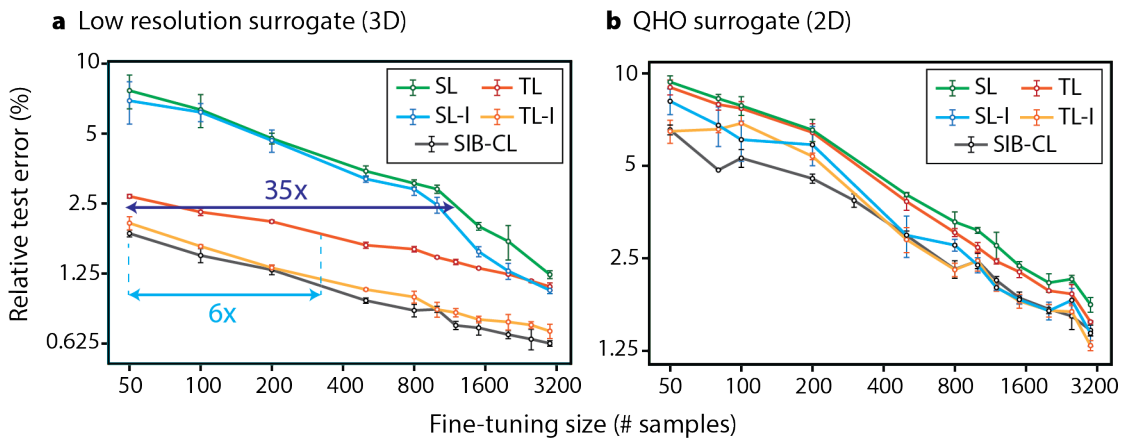

Figure 5-4: Performance of SIB-CL, compared against the baselines described in Section 4.3, for the ground-state energy prediction of (a) 3D random potentials when using the lowered resolution surrogate dataset and (b) 2D random potentials when using the quantum harmonic oscillator (QHO) surrogate dataset.

baselines, as shown in Fig. 5-4b. The benefits from transfer learning (i.e. in SIB-CL and the surrogate-boosted baselines) are diminished when compared with the experiments using the reduced resolution surrogate dataset. This is expected, since the QHO surrogate dataset is way simpler and contains less information that can be used to aid the network training on the target task. Notably, SIB-CL still outperforms the baselines, especially at low sample numbers; this further highlights SIB-CL's flexibility, specifically, in the choice and simplicity of the surrogate dataset.

5.3 Ablation studies

The exploitation of invariance information via contrastive learning is a critical part of SIB-CL. To investigate this concept further, a series of ablation studies is performed and presented here. Specifically, the DOS prediction problem in 2D PhCs is used for these studies.

First, invariant transformations used for the contrastive learning step are selectively removed. The complete set of invariant transformations used for the PhC DOS prediction problem is $\mathcal{G} = \{\mathbf{t}, C, \sigma, s\}$, using notation defined in Section 4.1. Fig. 5-5a shows the result of selectively, and increasingly, removing transformations from $\mathcal{G}$. The prediction ac-

curacy is observed to *monotonically worsen (improve)* as more transformations are removed (added) during contrastive learning. This is expected, since all of these transformations are *true* invariances of the (DOS prediction) problem, and thus including more types of transformation is equivalent to incorporating more prior knowledge of the problem which should improve the model accuracy. (This observation was not echoed in standard contrastive learning on computer vision applications, however, since the data augmentation strategies used to generate the transformations are not *true* invariances and thus some may prove ineffective [70].) This observation also validates SIB-CL's effectiveness in using contrastive learning as a strategy to incorporate prior knowledge of physical invariances.

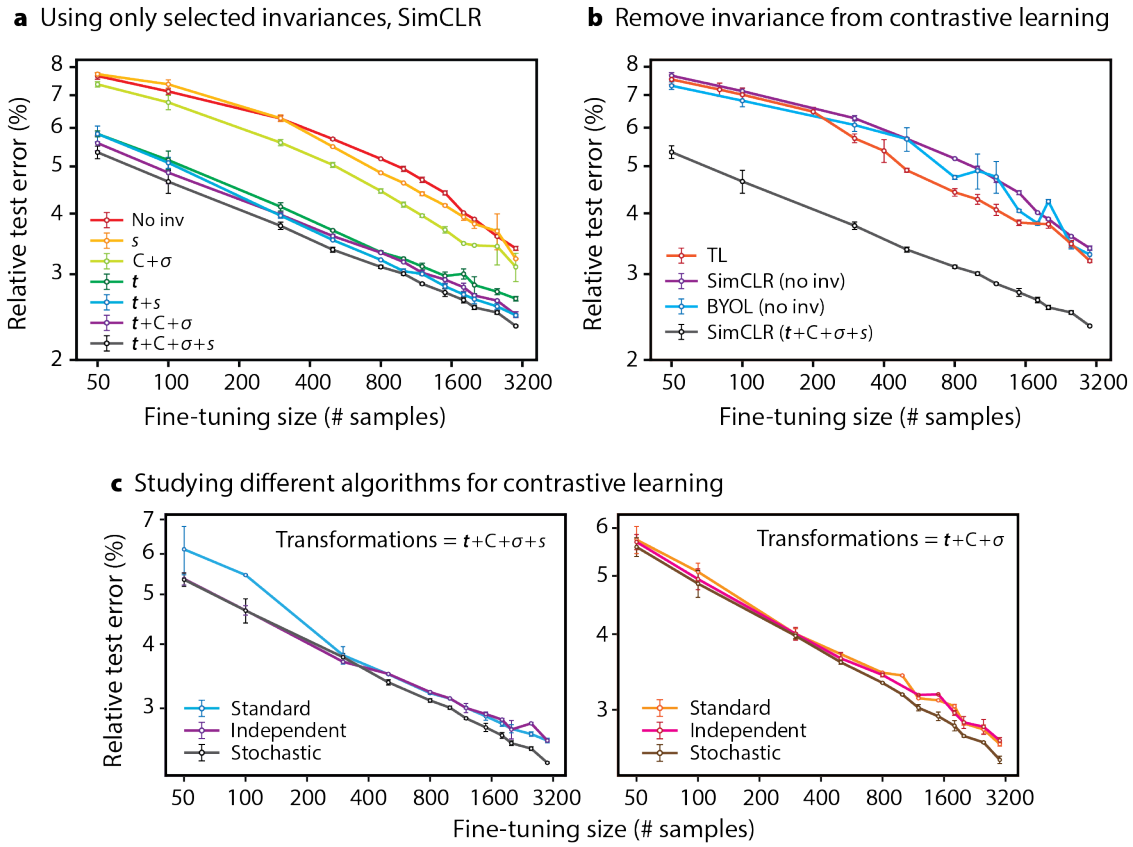

Figure 5-5: Results from ablation experiments where (a) selected transformations are removed from the complete group of invariances $\mathcal{G}$, and (b) all transformations are removed altogether from contrastive learning, for both SimCLR and BYOL. (c) Different algorithms are also experimented for contrastive learning, when all transformations are included (left) and all but scaling transformation are included (right).

Next, ablation experiments where the transformations are removed from the contrastive

learning step altogether, for both SimCLR and BYOL, are also performed. To do so, I remove the first step of sampling from $\mathcal{G}$ in both methods and set the two variations to be the original input (i.e. making $g\mathbf{x}_i = g'\mathbf{x}_i = \mathbf{x}_i$ in Fig. 4-1); I then (pre-)train the networks via contrastive loss as usual according to Fig. 4-1a for SimCLR and Fig. 4-2 for BYOL. The results are depicted in Fig. 5-5b, where both SimCLR and BYOL are observed to give very similar prediction accuracies to the TL baseline (which includes only information from the surrogate dataset and not invariances). This further validates SIB-CL's approach of using contrastive learning to invoke knowledge of invariances—once this knowledge is removed, doing contrastive learning as a pre-training step provides *no benefits*.

Finally, I also study various algorithms for performing contrastive learning on the invariances; specifically these are;

1. **Standard**: Each transformation in $\mathcal{G}$ is uniformly sampled and applied sequentially to the input, in the following order: $[\mathbf{t}, C + \sigma, s]$. Rotations ($C$) and mirrors ($\sigma$) are together considered a single transformation since they constitute the point group together. Due to sequential application of multiple transformations, this sampling procedure gives rise to instances which have "highly compounded" invariances; in other words, the two variations (of a positive pair) can look drastically different.

2. **Independent**: Contrastive learning is performed independently for each transformation and their losses are summed; i.e. we create a positive pair for *each* transformation, compute the NT-Xent loss for each pair using Eq. (4.1) and sum them to get the total contrastive loss. This algorithm is slower (though this slow-down can be avoided via parallel computing) than the rest since the number of forward passes per iteration increases according to the number of transformations present. However, this sampling procedure creates positive instances that differ only by a single transformation, hence their differences are less drastic.

3. **Stochastic**: Similar to standard, except there is some probability, p, as to whether each transformation is applied. In the experiments, I set $p = 0.5$, i.e. there is a 50% chance of not applying each transformation.

These three algorithms are compared in Fig. 5-5c, where the stochastic approach is seen

to produce the best model accuracies. While the differences are small, the stochastic method is still strictly and consistently better than the other two for various combinations of transformations (Fig. 5-5c left and right). I conjecture that this is because the standard method uses instances with highly compounded invariances which are difficult to learn, whereas the independent method uses simple invariances governed by single transformations which are easily learnt but omits useful knowledge of compounded invariances. The stochastic method presents an ideal middle-ground between the two methods.

Chapter 6

Conclusion and Outlook

In summary, I have proposed a novel framework, surrogate- and invariance- boosted contrastive learning (SIB-CL), which enables effective and high-quality network training despite data scarcity for scientific applications. SIB-CL achieves this by effectively leveraging sources of auxiliary information, specifically, prior knowledge of invariances in the problem and a more easily obtainable surrogate dataset. Using problems designed across various scientific domains, I have demonstrated SIB-CL's flexibility and broad-applicability—invariances can be highly simple (e.g. using rotations and mirrors only) and diverse options can be used for the surrogate dataset. Specifically, examples of these options are; simplified structures where (semi-)analytical solutions exist (e.g. circular structures of PhCs), approximate calculations of the target problem (e.g. reduced resolution computations of TISE), or even a combination of the two (e.g. approximated energies of QHO potentials in the TISE problem).

There are several opportunities to further improve SIB-CL's performance. Due to the vast number of possible training hyperparameters involved in SIB-CL, I did not study all. For example, contrastive learning and supervised learning on the surrogate dataset are simply (pre-)trained in alternate steps; one could in principle vary the interval between the two steps, perform the two steps entirely sequentially, or consider joint-training (i.e. updating the weights of $\mathbf{H}$ using the averaged losses of both steps). As such, performances better than the ones presented here may be attainable.

This work is highly complementary to the growing body of work in developing equivari-

ant networks for various groups of symmetries [50, 51, 52, 72], particularly for applications in the natural sciences [47, 48, 47]. Similar to equivariant networks, SIB-CL also aims to create a network that respects the underlying symmetries of the problem, albeit rather than "hard-coding" this into the model architecture, the process is facilitated via contrastive learning. The main drawback of SIB-CL, however, is that invariance to the desired group of symmetries is not guaranteed (e.g. on unseen test data); nevertheless, SIB-CL has the advantage of being simple and highly general to *any* known invariance, i.e. without designing kernels specialized for each symmetry. Besides, SIB-CL also allows one to exploit invariances that are not governed by symmetries (i.e. invariances that cannot be defined by specialized kernels such as the scaling invariance used in the DOS prediction problem). The performance edge of SIB-CL over TL-I also serves as evidence that using contrastive learning to enforce invariances is likely to be more effective than naive data augmentation. Given these merits, potential future directions for this work include assessing and analyzing SIB-CL's competitiveness against these equivariant networks.

While self-supervised learning, specifically contrastive learning, has gained considerable attention in recent years, it has been highly focused on computer vision applications (e.g. image classification, segmentation and detection); their application in regression problems is seldom explored. Here, SIB-CL's effectiveness is demonstrated for regression problems, suggesting optimistic potential for contrastive learning methods to be used for regression tasks which encompasses a wide range of real world problems. Nevertheless, several questions about, and gaps in the understanding of, the workings of contrastive learning on regression problems have been highlighted here. This can be partly attributed to the lack of a solid theoretical understanding of contrastive learning methods. With increasing research efforts and progress in this area, improved understanding of contrastive learning can aid future investigations of the questions posed here and, crucially, provide insights to tailor and improve contrastive learning methods for regression tasks.

Appendix A

Additional training details

A.1 Model architecture

The encoder network, i.e. **H** in Fig. 4-1, consists firstly of $3-4$ layers of convolutional neural network (CNN) followed by 2 fully connected (FC) layers, where the embedding after the CNNs is flattened before being fed into the FCs layers. The channel dimensions in the CNN layers and number of nodes in the FC layers vary for the different problems, and are listed in Table A.1. For the 3D Schrödinger equation problem, the CNN layers have 3D kernels to cater for the 3D inputs, while the CNNs for the remaining problems use regular 2D kernels. Between the CNN layers, BatchNorm [73], ReLU [74] activations and MaxPooling are included, while only ReLU activations are included between the FC layers.

Problem	**Channel dim per CNN layer**	**# nodes per FC layer**
DOS	[64, 256, 256] (2D)	[1024, **1024**]
Band structure	[64, 256, 256] (2D)	[256, **1024**]
TISE	[64, 256, 256, 256] (3D)	[256, **256**]

Table A.1: Network architecture for **H**. Bold values indicate the dimension of the representation (see Fig. 4-1 for definition) for the different problems.

For the predictor network, **G** in Fig. 4-1, four FC layers (with ReLU [74] activations between them) are used for all the problems, with the number of nodes listed in Table A.2. The predictor network for the band structure problem consists of six sub-networks of the same architecture, each sub-network leading to each of the six bands and separately updated

using the loss from each band during training. A similar architecture was used in previous work [11].

Problem	# nodes per FC layer
DOS	$[1024, 1024, 512, \mathbf{400}]$
Band structure	$[256, 512, 512, \mathbf{625}] * 6$
TISE	$[256, 256, 32, \mathbf{1}]$

Table A.2: Network architecture for **G**. Bold values indicate the dimension of the network output which matches the label dimension for that problem

For the projector network, **J** in Fig. 4-1, two FC layers are used (which is the standard architecture used in most contrastive learning methods [41, 42, 43]) with hidden dimension 1024, ReLU [74] activation, and the final metric embeddings have dimension 256. **J** is fixed across all problems.

A.2 Training hyperparameters

For training the networks in all the problems, Adam optimizers [75] are used, with learning rates for the different steps specified in Table A.3. An adaptive learning rate scheduler is also used for the finetuning stage. Even though standard contrastive learning methods implement a cosine annealing scheduler [76], this is found to be not beneficial for SIB-CL and so is omitted.

Problem	CL	PT of G	FT
DOS	$B \in \{192, 768\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$	$B \in \{16, 32, 64\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$	$B \in \{16, 32, 64\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$
Band structure	$B \in \{192, 768\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$	$B \in \{16, 32, 64\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$	$B \in \{16, 32, 64\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$
TISE	$B \in \{128\}$ $\alpha \in \{10^{-6}, 10^{-5}\}$	$B \in \{32, 64, 128\}$ $\alpha \in \{10^{-5}, 10^{-4}\}$	$B \in \{32, 64, 128\}$ $\alpha \in \{10^{-4}, 10^{-3}\}$

Table A.3: Hyperparameters used in the contrastive learning (CL), pre-training (PT) and fine-tuning (FT) steps, i.e. steps (a), (b) and (c) of Fig. 4-1 respectively. The main hyperparameters I varied are the batchsize (B) and the learning rates (α).

To prevent networks **H** and **G** from overfitting to the surrogate dataset, various conventional regularization techniques are explored for the pre-training stage, such as weight decay

and dropout. These techniques are found to be not beneficial in practice; thus instead, I use early-stopping where the pre-trained model is saved at various epochs and thereafter, fine-tuning is performed on all of them and the best result among them is used as the final performance. For SIB-CL, the pre-trained model is saved at $\{100, 200, 400\}$ epochs; for the TL baselines (both with and without invariances), the pre-trained model is saved at $\{40, 100, 200\}$ epochs.

Finally, another important hyperparameter is the kernel size (n_k) of the CNN layers. Apart from aiding with training optimization, n_k can be used to adjust the network size. This is important in this work since training and fine-tuning are performed on varying sizes N_t of the target dataset; a smaller (bigger) dataset is likely to need a smaller (bigger) network for optimal results. For DOS prediction, I vary $n_k \in \{3, 5, 7\}$; for band structures prediction, $n_k \in \{5, 7, 9\}$; and for ground-state energy prediction, $n_k \in \{5, 7\}$. In every problem, the same set of n_k is applied for SIB-CL and all baselines.

Bibliography

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. page 9.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*, May 2019.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A.A Rusu, J. Veness, M.G Bellemare, A. Graves, M. Riedmiller, A.K Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and S. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[6] D. Silver, A. Huang, Maddison C.J, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[7] Ankit Agrawal and Alok Choudhary. Deep materials informatics: Applications of deep learning in materials science. *MRS Communications*, 9(3):779–792, September 2019.

[8] Adam C. Mater and Michelle L. Coote. Deep Learning in Chemistry. *Journal of Chemical Information and Modeling*, 59(6):2545–2559, June 2019.

[9] Akinori Tanaka, Akio Tomiya, and Koji Hashimoto. *Deep Learning and Physics*. Springer Nature, March 2021.

[10] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020.

[11] Thomas Christensen, Charlotte Loh, Stjepan Picek, Domagoj Jakobović, Li Jing, Sophie Fisher, Vladimir Ceperic, John D. Joannopoulos, and Marin Soljačić. Predictive and generative machine learning models for photonic crystals. *Nanophotonics*, 9(13):4183–4192, October 2020.

[12] Samuel Kim, Peter Y. Lu, Charlotte Loh, Jamie Smith, Jasper Snoek, and Marin Soljačić. Scalable and Flexible Deep Bayesian Optimization with Auxiliary Information for Scientific Problems. *arXiv:2104.11667*, April 2021.

[13] Sungsoo Ahn, Junsu Kim, Hankook Lee, and Jinwoo Shin. Guiding Deep Molecular Optimization with Genetic Exploration. *arXiv:2007.04897*, October 2020.

[14] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. ElemNet : Deep Learning the Chemistry of Materials From Only Elemental Composition. *Scientific Reports*, 8(1):17593, December 2018.

[15] Peter Y. Lu, Samuel Kim, and Marin Soljačić. Extracting Interpretable Physical Parameters from Spatiotemporal Systems Using Unsupervised Learning. *Physical Review X*, 10(3):031056, September 2020.

[16] Alexandre Tkatchenko. Machine learning for chemical discovery. *Nature Communications*, 11(1):4125, August 2020.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[18] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009.

[19] Tanja van Mourik, Michael Bühl, and Marie-Pierre Gaigeot. Density functional theory across chemistry, physics and biology. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 372(2011):20120488, March 2014.

[20] Á Nagy. Density functional. Theory and application to atoms and molecules. *Physics Reports*, 298(1):1–79, 1998.

[21] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. *arXiv:1808.01974*, August 2018.

[22] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[23] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 513–520, 2011.

[24] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[26] Yurui Qu, Li Jing, Yichen Shen, Min Qiu, and Marin Soljačić. Migrating Knowledge between Physical Scenarios Based on Artificial Neural Networks. *ACS Photonics*, 6(5):1168–1174, May 2019.

[27] Hironao Yamada, Chang Liu, Stephen Wu, Yukinori Koyama, Shenghong Ju, Junichiro Shiomi, Junko Morikawa, and Ryo Yoshida. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Central Science*, 5(10):1717–1730, October 2019.

[28] Nicholas Lubbers, Turab Lookman, and Kipton Barros. Inferring low-dimensional microstructure representations using convolutional neural networks. *Physical Review E*, 96(5):052111, November 2017.

[29] Xiaolin Li, Yichi Zhang, He Zhao, Craig Burkhart, L. Catherine Brinson, and Wei Chen. A Transfer Learning Approach for Microstructure Reconstruction and Structure-property Predictions. *Scientific Reports*, 8(1):13461, September 2018.

[30] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To Transfer or Not To Transfer. In *NIPS'05 Workshop, Inductive Transfer: 10 Years Later*, 2005.

[31] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv:1411.1792*, November 2014.

[32] Ying Zhang and Chen Ling. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, 4(1):1–8, May 2018.

[33] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring GANs: Generating Images from Limited Data. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11210, pages 220–236. Springer International Publishing, Cham, 2018.

[34] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On Leveraging Pretrained GANs for Generation with Limited Data. *arXiv:2002.11810*, August 2020.

[35] Marouan Belhaj, Pavlos Protopapas, and Weiwei Pan. Deep Variational Transfer: Transfer Learning through Semi-supervised Deep Generative Models. *arXiv:1812.03123*, December 2018.

[36] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. *arXiv:1505.05192*, January 2016.

[37] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv:1803.07728*, March 2018.

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. *arXiv:1603.09246*, August 2017.

[39] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A Survey on Contrastive Self-supervised Learning. *arXiv:2011.00362*, February 2021.

[40] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv:1911.05722*, March 2020.

[41] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709*, June 2020.

[42] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv:2006.07733*, September 2020.

[43] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv:2006.09882*, January 2021.

[44] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. *arXiv:2011.10566*, November 2020.

[45] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. MolCLR: Molecular Contrastive Learning of Representations via Graph Neural Networks. *arXiv:2102.10056*, February 2021.

[46] Sebastian J. Wetzel, Roger G. Melko, Joseph Scott, Maysum Panju, and Vijay Ganesh. Discovering symmetry invariants and conserved quantities by interpreting siamese neural networks. *Physical Review Research*, 2(3):033499, September 2020.

[47] Simon Batzner, Tess E. Smidt, Lixin Sun, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. SE(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *arXiv:2101.03164*, January 2021.

[48] Zhantao Chen, Nina Andrejevic, Tess Smidt, Zhiwei Ding, Yen-Ting Chi, Quynh T. Nguyen, Ahmet Alatas, Jing Kong, and Mingda Li. Direct prediction of phonon density of states with Euclidean neural networks. *Advanced Science*, page 2004214, March 2021.

[49] Julia Ling, Reese Jones, and Jeremy Templeton. Machine learning strategies for systems with invariance properties, 2016.

[50] Taco S. Cohen and Max Welling. Group Equivariant Convolutional Networks. *arXiv:1602.07576*, June 2016.

[51] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv:1802.08219*, May 2018.

[52] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *arXiv:1807.02547*, October 2018.

[53] J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade. *Photonic Crystals: Molding the Flow of Light*. Princeton University Press, 2 edition, 2008.

[54] Eli Yablonovitch. Inhibited Spontaneous Emission in Solid-State Physics and Electronics. *Physical Review Letters*, 58(20):2059–2062, May 1987.

[55] L. Novotny and B. Hecht. *Principles of Nano-Optics*. Cambridge University Press, 2012.

[56] Steven G. Johnson and J. D. Joannopoulos. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Optics Express*, 8(3):173–190, January 2001.

[57] G Gilat and L.J. Raubenheimer. Accurate numerical method for calculating frequency-distribution functions in solids. *Phys. Rev.*, 144(2):390, 1966.

[58] Boyuan Liu, Steven G. Johnson, John D. Joannopoulos, and Ling Lu. Generalized Gilat-Raubenheimer method for density-of-states calculation in photonic crystals. *Journal of Optics*, 20(4):044005, April 2018.

[59] K. Ohtaka. Energy band of photons and low-energy photon diffraction. *Physical Review B*, 19(10):5057–5067, May 1979.

[60] Xindong Wang, X.-G. Zhang, Qingliang Yu, and B. N. Harmon. Multiple-scattering theory for electromagnetic waves. *Physical Review B*, 47(8):4161–4167, February 1993.

[61] Alexander Moroz. Density-of-states calculations and multiple-scattering theory for photons. *Physical Review B*, 51(4):2068–2081, January 1995.

[62] Alexander Moroz. Metallo-dielectric diamond and zinc-blende photonic crystals. *Physical Review B*, 66(11):115109, September 2002.

[63] Cesar Lema and Anna Choromanska. Approximating Ground State Energies and Wave Functions of Physical Systems with Neural Networks. *arXiv:2011.10694*, November 2020.

[64] Kyle Mills, Michael Spanner, and Isaac Tamblyn. Deep learning and the Schrödinger equation. *Physical Review A*, 96(4):042113, October 2017.

[65] Connor Shorten and T. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019.

[66] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*, January 2019.

[67] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. *arXiv:1906.05849*, December 2020.

[68] Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *arXiv:2005.10242*, November 2020.

[69] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. *arXiv:1902.09229*, February 2019.

[70] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? *arXiv:2005.10243*, December 2020.

[71] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On Mutual Information in Contrastive Learning for Visual Representations. *arXiv:2005.13149*, June 2020.

[72] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *arXiv:2006.10503*, November 2020.

[73] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167*, March 2015.

[74] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010.

[75] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, January 2017.

[76] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv:1608.03983*, May 2017.