Brigham Young University
BYU ScholarsArchive

Theses and Dissertations

2024-04-25

TrIP - Transformer Interatomic Potential Predicts Realistic Energy Surface Using Physical Bias

Bryce Eric Hedelius
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Physical Sciences and Mathematics Commons

BYU ScholarsArchive Citation

Hedelius, Bryce Eric, "TrIP - Transformer Interatomic Potential Predicts Realistic Energy Surface Using Physical Bias" (2024). *Theses and Dissertations*. 10362.
https://scholarsarchive.byu.edu/etd/10362

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

TrIP – Transformer Interatomic Potential Predicts Realistic Energy Surface Using Physical Bias

Bryce Eric Hedelius

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Dennis Della Corte, Chair
Gus Hart
Daniel H. Ess

Department of Physics and Astronomy

Brigham Young University

Copyright © 2024 Bryce Eric Hedelius

All Rights Reserved

ABSTRACT

TrIP – Transformer Interatomic Potential Predicts Realistic Energy Surface Using Physical Bias

Bryce Eric Hedelius
Department of Physics and Astronomy, BYU
Master of Science

Accurate interatomic energies and forces enable high-quality molecular dynamics simulations, torsion scans, potential energy surface mapping, and geometry optimization. Machine learning algorithms have enabled rapid estimates of energies and forces with high accuracy. Further development of machine learning algorithms holds promise for producing general potentials that support dozens of atomic species. I present my own Transformer Interatomic Potential (TrIP): a chemically sound potential based on the SE(3)-Transformer. TrIP's species-agnostic architecture—using continuous atomic representation and homogenous graph convolutions—encourages parameter sharing between atomic species for more general representations of chemical environments, keeps a reasonable number of parameters, serves as a form of regularization, and is a step towards accurate universal interatomic potentials. I introduce physical bias in the form of Ziegler-Biersack-Littmark-screened nuclear repulsion and constrained atomization energies to improve qualitative behavior for near and far interaction. TrIP achieves state-of-the-art accuracies on the COMP6 benchmark with an energy prediction error of just 1.02 kcal/mol MAE, outperforming all other models. An energy scan of a water molecule shows improved short- and long-range interactions compared to other neural network potentials, demonstrating its physical realism compared to other models. TrIP also shows stability in molecular dynamics simulations with a reasonable exploration of Ramachandran space.

Keywords: Equivariant, machine learning, interatomic potential, SE(3)-Transformer, Ramachandran plot

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 1

Introduction

Developing accurate chemical models is integral for understanding nature and developing innovative technologies. My research, originally represented in a paper currently in review, delves into my approach to modeling chemical environments using an interatomic potential I developed called Transformer Interatomic Potential (TrIP). This thesis is an expansion of the paper, aiming to delve deeper into the foundations and experimental design.

1.1 Motivation

Describing chemical reactions and transformations is vital to understanding biological and industrial processes. Computational modeling assists in predicting the outcomes of chemical reactions[1], elucidating reaction pathways[2], and identifying stable and transition states[3]. Modeling and simulating reactions offer a profound understanding of the underlying mechanisms, thereby facilitating the development of new processes and catalysts, including enzymes[4]. Advancements in computational tools further enhance these approaches by providing heightened accuracy and accessibility.

Molecular mechanics methods are instrumental in studying atomistic systems comprising thousands to millions of atoms due to their rapid energy and force approximations[5]. Employing

classical mechanics and featuring nearly linear complexity, these methods are particularly applicable to the study of complex biological macromolecules[6]. However, these methods have limitations, often relying on fixed atom types and predefined interaction rules. These classical force fields prove inadequate for exploring chemical reactions or understanding reactive systems[7], such as enzymatic reactions[8], where the chemistry of the system changes over time.

Another concern in classical force field approaches is generalizability. The parameters underpinning conventional force fields are typically derived from quantum mechanical calculations and fine-tuned to match experimental data, thus the quality and comprehensiveness of input data limits the accuracies of conventional force fields. While they perform well in simulating protein systems due to the availability of extensive data, they face limitations in scenarios like molten salt simulations, where input data is limited[9]. This variability in data availability results in force fields that work satisfactorily for one system but struggle to transfer the same level of accuracy to different systems[10]. This underscores the necessity for approaches that can accommodate a broader range of chemical environments and system complexities, enabling more universally applicable simulation tools.

The inclusion of polarization effects poses yet another formidable challenge for classical force fields[11]. Polarization plays a pivotal role in shaping molecular properties and intermolecular interactions, contributing to the accuracy of simulations. However, integrating polarization into force fields proves to be a complex endeavor. One approach is the utilization of Drude models[12], which introduce explicit 'Drude' particles to simulate the electronic cloud of atoms and capture polarization effects. While this technique offers a means to address polarization, Drude models present their own set of limitations and challenges, requiring careful consideration of trade-offs between accuracy and computational efficiency. As researchers seek

to improve the fidelity of simulations, finding effective methods to incorporate polarization effects remains a subject of ongoing investigation.

Although the constraints inherent in classical force fields might prompt consideration of more advanced strategies like *ab initio* molecular dynamics (AIMD), it is crucial to acknowledge that these alternative methods also present their own array of obstacles. AIMD, which incorporates first-principles quantum mechanical calculations into molecular dynamics simulations, offers the capacity to precisely model alterations in electronic structure, the dynamics of bond breaking and formation, and other intricate chemical processes. However, such precision comes at a significant computational expense, as AIMD mandates the iterative solution of the Schrödinger equation for the entire system at each computational time step, thereby increasing the computational demands and limiting the scale of feasible simulations.

The computational demands associated with AIMD render it impractical for simulating extensive systems or for conducting simulations over extended timescales[13]. Consequently, its application is primarily constrained to small systems, typically encompassing fewer than 100 atoms, and short simulation durations on the order of 100 picoseconds. These limitations consequently pose a formidable hurdle when attempting to explore more extensive biochemical systems[14] or phenomena unfolding over prolonged temporal scales, such as the intricate dynamics of protein folding or the progressive evolution of materials[15].

Consequently, an escalating interest has emerged in devising approaches that capitalize on the speed of classical force fields while also harnessing the precision of quantum mechanical calculations[16]. In this context, machine learning (ML) has garnered considerable attention for its potential to surmount the limitations inherent in these two techniques[17]. Machine learning provides a promising avenue for addressing the deficiencies associated with both methods,

thereby offering an amalgamation of their strengths. By utilizing machine learning techniques, it becomes feasible to train models using *ab initio* molecular dynamics (AIMD) data and subsequently incorporate these models into conventional molecular dynamics frameworks for rapid simulations[18]. This synthesis effectively harmonizes the respective merits of classical force fields and quantum mechanical calculations, delivering a powerful tool for comprehensively exploring complex molecular systems.

One of the initial forays into machine learning for molecular dynamics was undertaken by Behler and Parrinello[19]. Their approach employs symmetry functions to encapsulate the chemical environment surrounding each atom. Subsequently, atom-specific artificial neural networks predict atomic energies, which are then summed to calculate the total energy of the system. In a parallel vein, the DeePMD-Kit[20] adopts a similar strategy, streamlining the creation of customized machine learning potentials by providing a comprehensive framework. These endeavors have significantly influenced the trajectory of machine learning applications within molecular dynamics, paving the way for further developments in the field.

The ANI (Accurate Neuronal Network Potential) networks[21 22] constitute a significant advancement beyond the Behler-Parinello methodology. These networks adopt refined symmetry functions derived from the Behler-Parinello approach, resulting in enhanced feature representation. Additionally, ANI networks incorporate deeper neural network architectures and are trained on a substantial dataset of molecular conformers and energies. The ANI-1x dataset, for instance, encompasses approximately 5 million conformations across 50,000 molecular configurations. This training dataset is complemented by the Comprehensive Machine-learning Potential (COMP6) benchmarking set[23], which includes molecules larger than those found in the

ANI-1x dataset. The ANI-1x potential achieves an energy mean squared error of 3.37 kcal/mol on COMP6, highlighting its robust adaptability to diverse chemical environments.

A more modern approach is characterized by its departure from traditional symmetry functions to encode the chemical environment, instead opting for message-passing graph neural networks. As one of the first message-passing neural networks used in interatomic potentials, SchNet [24] [25] provided molecules in a new way as a graph representation and a more general approach for encoding the chemical environment through continuous-filter message-passing layers.

These methodologies previously discussed exhibit invariance, wherein both input and intermediate features remain unchanged during rotations or translations. However, our focus shifts to equivariant neural networks, which distinctly embrace symmetry—transforming features symmetrically under rotations[26]. In the realm of physics, numerous physical properties, encompassing forces, polarizations, dipoles, quadrupoles, and other geometric tensors, inherently demonstrate equivariance under rotations and translations[27]. Thus, an equivariant neural network can appropriately represent polarization effects internally and provide a more general approach for an interatomic potential.

Noteworthy instances of equivariant neural networks include NequIP[28], MACE[29], and NewtonNet[30], which effectively extend graph networks like SchNet by incorporating equivariant features, thereby amplifying the scope of understanding complex molecular interactions.

Despite the advancing sophistication of machine learning interatomic potentials, the current landscape of machine learning potentials is marked by a significant limitation in their generalizability[31]. The prevailing practice involves training and evaluating machine learning potentials on relatively straightforward datasets like QM9[32] and MD17[33], yielding accuracies that

notably outstrip those achievable by Density Functional Theory (DFT). While these models can be adeptly fine-tuned to emulate quantum mechanical potentials for specific systems with exceptional precision, their predictive power often wanes when extended to unfamiliar systems lying beyond the confines of their training data. For instance, ANI has been reported to inaccurately predict attraction instead of repulsion in near-range interactions, stemming from a paucity of pertinent training data in these regions. Consequently, while machine learning potentials might excel on these datasets, their performance might not translate effectively to scenarios involving intricate chemical reactions, diverse molecular environments, or materials with specific properties beyond the purview of the training data.

The application of machine learning models in computational chemistry encounters significant challenges, notably the capacity for their misuse. A lack of understanding of models' limitations and the nuances of their training process can lead to erroneous conclusions drawn from simulations. This issue was recently exemplified in a thorough review of machine learning potentials used for Molten Salt simulations[34]. Here, the authors revealed that the adoption of the Behler-Parinello method within DeepMD-Kit introduced a systematic error, rendering the accurate simulation of intricate salt systems problematic. This underscores the necessity for users to grasp the intricacies of the machine learning models they employ, ensuring a comprehensive comprehension of their boundaries and a careful interpretation of their results.

These limitations motivated the development of our Transformer Interatomic Potential (TrIP). TrIP is based around a sophisticated equivariant message-passing network called SE(3)-Transformers that enables it to develop a detailed description of its chemical environment. I introduce physical bias in the form of screen nuclear Coulombic interactions and exponentially decaying asymptotic constraints of isolated atoms to improve the behavior of the potential away

from chemical equilibrium. I train and test different molecular configurations, evaluating the network's ability to generalize rather than just interpolate. These approaches improve the physical behavior of TrIP's predictions and TrIP yields state-of-the-art accuracy.

Chapter 2

Background

In this chapter, I delve into the theoretical underpinnings that form the background of my research in developing the Transformer Interatomic Potential (TrIP). I establish the essential mathematical framework that enables TrIP to maintain symmetry, encode chemical neighborhoods, and ensure the continuity of derivatives. As a potential, TrIP belongs to a family of mathematical constructs that encode the interactions between atoms, dictating how forces and energy change with atomic positions. Furthermore, I explore the crucial role of molecular dynamics simulations, which provide the dynamic context for validating TrIP's predictions and understanding its performance in capturing the dynamic behavior of atomistic systems.

2.1 Mathematical Foundations

TrIP's methodology is deeply rooted in mathematics. I rely on graph theory, equivariance theory, tensor theory, and mathematical smoothness to build a robust framework for predicting atomistic interactions. These mathematical concepts enable TrIP to maintain symmetry, encode complex relationships, and ensure continuity in derivatives – crucial factors for accurate and reliable molecular simulations.

2.1.1 Graph Theory

Graphs serve as a fundamental framework for understanding and representing complex relationships and structures in diverse fields of study. In the realm of computational chemistry and molecular simulations, graph theory provides us with a versatile mathematical language to describe and analyze molecular systems. Graphs, consisting of nodes and edges, offer an elegant means to model and comprehend the intricate interactions between atoms and molecules. They serve as a bridge between the abstract world of mathematics and the concrete realm of molecular behavior.

A graph denoted G, is a pair $G = (V, E)$, where V is the set of vertices, also called nodes, and E is the set of paired vertices, also called edges. Each edge in E connects two distinct vertices in V, thereby defining a pairwise relationship between them. In the context of this work, I consider symmetric directed graphs, meaning that the edges connecting vertices have an inherent direction, but for each edge (v_1, v_2) in V there exists an edge (v_2, v_1) in V. Graphs may include attributes, weights, and labels associated with nodes and edges. These additional properties make graphs a useful type of data structure for representing connected information.

In this work, graphs are the fundamental data structure for characterizing atomistic systems, where individual atoms are denoted as nodes, and the localized interactions between these atoms are depicted as edges. I use node attributes, such as the atomic species and position, and edge attributes, such as relative positions, as well as intermediate values.

The construction of graphs in this study uses a fixed-radius near-neighbor approach. Each atom within the atomistic system is a node and any pair of atoms within a predefined cutoff radius is an edge. This approach allows for the creation of a graph that succinctly captures the

local interactions and relationships between atoms within the specified radius while accommodating the periodicity of the system. It is a crucial step in encoding the atomic environment and facilitating the subsequent analysis of the molecular system's behavior and properties.

2.1.2 Group Theory

Group theory, a branch of abstract algebra, plays a fundamental role in understanding the symmetries and transformations inherent in molecular systems. In the realm of molecular simulations and computational chemistry, group theory provides a rigorous mathematical framework to study and categorize the symmetries exhibited by atomic and molecular structures. Group theory's significance lies in its ability to elucidate the underlying symmetries of molecular systems, enabling researchers to predict and interpret a wide range of molecular behaviors, from spectroscopic outcomes to reaction mechanisms. This mathematical discipline is instrumental in characterizing and harnessing the symmetrical properties of molecules, facilitating a deeper comprehension of their physical and chemical attributes.

Mathematically, a group is a pair $(G, \cdot)$ consisting of a group G and a binary operation on G with the following axioms:

1. Associativity: For all a, b, c in G, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
2. Identity element: There exists an element e in G such that, for all a in G, we have $e \cdot a = a \cdot e = a$.
3. Inverse element: For each a in G, there exists an element b in G such that $a \cdot b = b \cdot a = e$.

In this work, the groups I am working with are the group SO(3)—the set of real 3×3 orthogonal matrices with unit determinant, SE(3)—the set of translations and proper rotations in Euclidean space, and GL(n)—the set of $n \times n$ invertible matrices.

2.1.3 Equivariance Theory

Equivariance theory, within the context of molecular simulations and computational chemistry, provides a powerful framework rooted in mathematical symmetries to capture and reproduce the physical behaviors of atomic systems. At its core, equivariance recognizes that the laws of nature governing molecular interactions are consistent under different transformations, such as rotations or translations. These transformations reflect the inherent symmetries of molecular systems, allowing us to study how atoms and molecules behave irrespective of their spatial orientations. It allows us to describe complex molecular behaviors while maintaining the fundamental symmetries that govern the atomic world.

To have a clear definition of equivariance, I first build from simpler mathematical constructs from groups. First, allow the group to act on other groups through group actions and encapsulate those actions in G-sets.

If G is a group and X is a set, then a group action α of G on X is a function α: $G \times X \rightarrow X$ that satisfies the following axioms:

1. Identity: $\alpha(e, x) = x$ where e is the identity element of G and x is in X.

where g, h are in G and x is in X.

Expressions such as $\alpha(g, x)$ are shortened to $g \cdot x$ when the action is clear from context. A set X together with an action α of G is a G-set. The concept of a G-set allows for the following

compact definition of equivariance: If X and Y are both G-sets for the same group G, then a function f: $X \to Y$ is said to be equivariant if $f(g \cdot x) = g \cdot f(x)$ for all g in G and x in X.

2.1.4 Representation Theory

Representation theory is a fundamental mathematical framework that plays a pivotal role in the study of equivariance, particularly in the context of machine learning and deep learning. It provides a formalism for understanding mathematical objects, such as linear transformations or symmetries, using matrices or linear operators.

In the context of equivariance, representation theory describes a group action as linear transformations on vector spaces. This is crucial because real-world data, including images, molecular structures, or physical systems, share the same mathematical properties as vectors and tensors.

A representation of a group G on a vector space V is a function ρ: $G \to \mathrm{GL}(n)$ that is a homomorphism, i.e., it satisfies the following homomorphism property: $\rho(g \cdot h) = \rho(g)\,\rho(h)$ for all g, h in G. An irreducible representation is a nonzero representation with no proper nontrivial subrepresentation (a vector subspace that is itself a representation).

2.1.5 Tensor Theory

Tensors serve as a fundamental mathematical construct in the context of equivariant neural networks, enabling the representation and manipulation of data that have various symmetries and transformations. Understanding tensors is crucial for developing machine learning models that respect and use these symmetries effectively.

There are a variety of equivalent ways to define tensors, as multidimensional arrays, as multilinear maps, using tensor products of vector spaces, and using (tensor) representations of an arbitrary group G. The latter approach is the most natural for my purposes, as equivariant neural networks are rooted in representation theory and this definition most clearly shows that the objects used in equivariant neural networks are indeed tensors.

A tensor T: $\boldsymbol{B} \rightarrow W$ is a mapping from the set of ordered bases of a vector space V to another vector space W (also over the field F) to the vector space W. Here, $B = \{(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \mid \boldsymbol{v}_i \in V\}$ is a set of ordered bases $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ on V. The mapping T is equivariant under the action of a group G, meaning that there is a representation $\rho: G \rightarrow GL(W)$ such that

$$T(\rho(g)(v)) = \rho(g)\big(T(v)\big) \qquad (M1)$$

for all $g \in G$ and basis $v \in B$. In practice, inference with an equivariant neural network works in a fixed basis so we efficiently store tensors as arrays in computer memory. However, equivariant neural networks have restrictions so intermediate features transform properly as tensors under basis transformation. Thus, the symmetry constraint transfers from the tensors to the neural network itself as the new equivariant condition:

$$\mathrm{NN}\big(\rho(g)(\boldsymbol{T})\big) = \rho(\boldsymbol{R})(\mathrm{NN}(\boldsymbol{T})) \qquad (M2)$$

for all invertible matrices $g \in G$ where $\boldsymbol{T}$ is a tensor and $\mathrm{NN}: V \rightarrow W$ is the neural network mapping from positional space to an arbitrary vector/tensor space W.

A fundamental choice of the group G in the context of linear algebra and differential geometry is $\mathrm{GL}(W)$. In the context of equivariant neural networks based in 3-dimensional Euclidean space, the group $\mathrm{GL}(\mathbb{R}^3)$ includes transformations such as shearing and scaling which do not correspond to conserved physical quantities in molecules. Therefore, I restrict the

equivariance condition to rotational symmetries. While it is possible to include the symmetry of improper rotations (those including a reflection), for this work I restrict to the symmetries of proper rotations, those described by $\mathrm{SO}(\mathbb{R}^3)$. Since those are the same symmetries of a sphere, these tensors are "spherical tensors."

The properties of $\mathrm{SO}(\mathbb{R}^3)$ are well documented due to its usage in quantum mechanics, rigid body mechanics, and computer graphics and vision. Any representation ρ of $\mathrm{SO}(\mathbb{R}^3)$ can be linearly transformed into a basis so that ρ is the direct sum of Wigner D-matrices, the irreducible representations of $\mathrm{SO}(\mathbb{R}^3)$. Therefore, it is convenient for equivariant neural networks to use irreducible vector spaces that transform by the action of Wigner-D matrices. These are the same transformation laws that govern spherical harmonics, so vectors of spherical harmonics $\left(Y_l^{-l}, \dots, Y_l^l\right)$ are irreducible spherical tensors and use the same language as spherical harmonics to describe irreducible spherical tensors by degree *l* and order *m*.

New tensors can be made by combining other tensors with tensor products. The new tensors are often reducible, meaning they can be decomposed into the direct sum of other tensors, and the components of the irreducible tensors can be calculated using Clebsch-Gordan coefficients $C^{J\,M}_{j_1 m_1 j_2 m_2}$ according to

$$z_M^J = \sum_{m_1=-j_1}^{j_1} \sum_{m_2=-j_2}^{j_2} C^{J\,M}_{j_1 m_1 j_2 m_2} x_{m_1}^{j_1} y_{m_2}^{j_2} \,. \qquad (M3)$$

Where $x_{m_1}^{j_1}$ and $y_{m_2}^{j_2}$ are input tensors of degree $j_1$and j_2 respectively, and z_M^J is the output tensors of degree J.

This ability to combine information from different tensors together to produce a new tensor is the foundation of equivariant neural networks, generating tensors out of spherical harmonics and other tensors.

Notably, the Clebsch-Gordan coefficients have symmetries such as "conservation of angular momentum", where $m_1 + m_2 = M$ and the degrees must satisfy the triangle inequalities, where $|j_1 - j_2| \leq J \leq j_1 + j_2$, $|j_2 - J| \leq j_1 \leq j_2 + J$, and $|J - j_1| \leq j_2 \leq J - j_1$.

2.2 Deep Learning Foundations

Currently, message-passing graph neural networks are often used in machine learning potentials. In this framework, a molecule is a graph $G = (V, E)$, with atoms abstracted as a set of nodes *V* and local interactions as a set of edges *E*, where each edge is a pair of nodes. Graph neural networks pass information between their nodes using message-passing layers, which can be written as:

$$\boldsymbol{f}_u^{i+1} = \alpha^i \left(\boldsymbol{f}_u^i, \bigoplus_{v \in N_u}^i \beta^i \left(\boldsymbol{f}_u^i, \boldsymbol{f}_v^i, \boldsymbol{e}_{uv}^i \right) \right) \quad (1)$$

where *i* is the layer index, α^i and β^i are differentiable functions, $\bigoplus_{v \in N_u}^i$ is a permutation invariant aggregation operation over the "source" nodes *v* in the neighborhood N_u—the set of nodes that share an edge—of the destination node *u*, $\boldsymbol{x}_u^i$ is the initial vector on the "destination" node, different versions of $\boldsymbol{f}$ are the node vectors, and $\boldsymbol{e}_{uv}^i$ is the feature vector on the edge from *v* to *u*. This design enables information to propagate through different atoms

across multiple interaction layers, fostering the exchange of knowledge and enabling a machine learning algorithm to learn chemical environments and interactions.

An early example of an invariant approach was the SchNet potential, which uses continuous convolutions of the form:

$$\boldsymbol{f}_u^{i+1} = \boldsymbol{f}_u^i + \mathrm{MLP}^i\left(\sum_{v \in N_u} \boldsymbol{W}^{\boldsymbol{i}}(\|\vec{\boldsymbol{r}}_{vu}\|)\, \mathbf{Aff}^i(f_v)\right) \tag{2}$$

Where $\mathbf{Aff}^i$ is a learnable affine transformation, the relative position of node u relative to node v is

$$\vec{\boldsymbol{r}}_{vu} = \vec{\boldsymbol{r}}_u - \vec{\boldsymbol{r}}_v, \tag{3}$$

where $\boldsymbol{W}^{\boldsymbol{i}}$ is a multilayer perceptron serving as a radial profile function, and MLP^i is a multilayer perceptron to calculate the overall interaction.

Equivariant neural networks take a more general approach with spherical tensors[35]. An example of a fundamental equivariant message-passing layer that I use in this work comes from Tensor Field Networks (TFN)[36], where the features are now described as irreducible spherical tensors that update according to:

$$\boldsymbol{f}_u^{i+1,l} = w^{i,l}\boldsymbol{f}_u^{i,l} + \sum_{k \geq 0} \sum_{v \in N_u} \boldsymbol{W}^{i,lk}(\vec{\boldsymbol{r}}_{vu})\boldsymbol{f}_v^{i,k} \tag{4}$$

Where $w^{i,l}$ is a "self-interaction" scalar and $\boldsymbol{W}^{lk,i}(\vec{\boldsymbol{r}})$: $\mathbb{R}^3 \rightarrow \mathbb{R}^{(2l+1)\times(2k+1)}$ are learnable weight kernels defined as

$$\boldsymbol{W}^{lk}(\vec{\boldsymbol{r}}) = \sum_{j=|k-l|}^{k+l} \rho^{lkj}(\|\vec{\boldsymbol{r}}\|)\, \boldsymbol{W}^{lkj} \tag{5}$$

using radial profile functions ρ^{lkj}: $\mathbb{R}^3 \rightarrow \mathbb{R}$ and similarly sized basis kernel $\boldsymbol{W}^{lkj}$: $\mathbb{R}^3 \rightarrow \mathbb{R}^{(2l+1)\times(2k+1)}$ defined component-wise as

$$\left(\boldsymbol{W}^{lkj}\right)_{m_l m_k}(\vec{\boldsymbol{r}}) = \sum_{m_J=-j}^{j} Y_j^{m_j}\left(\frac{\vec{\boldsymbol{r}}}{\|\vec{\boldsymbol{r}}\|}\right) C_{jm_j km_k}^{lm_l}. \tag{6}$$

Where $Y_{m_j}^{j}: \mathbb{R}^3 \to \mathbb{R}$ are spherical harmonics of degree j and order m_j evaluated in the direction of the destination node in terms of the source node and $C_{jm_j km_k}^{lm_l}$ are the Clebsch-Gordan coefficients. The radial basis functions are MLPS with the following layers: Linear(1, 32), LayerNorm(32), SiLU, Linear (32, 32), LayerNorm(32), SiLU, Linear(32, number of components of ρ^{lkj}).

My approach also uses SE(3)-Transformer message-passing layers, which incorporate learnable edge-wise attention values α_{uv} into the TFN layer described in Eq 3 according to:

$$\boldsymbol{f}_u^{i+1,l} = w^{i,l}\boldsymbol{f}_u^{i,l} + \sum_{k\geq 0}\sum_{v\in N_u} \alpha_{uv}^{i} \boldsymbol{W}_V^{i,lk}(\vec{\boldsymbol{r}}_{vu})\boldsymbol{f}_v^{i,k} \tag{7}$$

Where the attention weight α_{uv} and its constituents are

$$\alpha_{uv} = \frac{\exp\left(\boldsymbol{q}_u^{\intercal}\boldsymbol{k}_{uv}\right)}{\sum_{w\in N_u}\exp(\boldsymbol{q}_u^{\intercal}\boldsymbol{k}_{uw})}, \tag{8}$$

$$\boldsymbol{q}_u = \bigoplus_l \boldsymbol{W}_Q^l \boldsymbol{f}_u^l, \tag{9}$$

$$\boldsymbol{k}_{uv} = \bigoplus_l \sum_k \boldsymbol{W}_K^{lk}(\vec{\boldsymbol{r}}_{vu})\, \boldsymbol{f}_v^k, \tag{10}$$

where $\boldsymbol{W}_K^{lk}$ and $\boldsymbol{W}_V^{lk}$ are taken from Eq 6 and $\boldsymbol{W}_Q^l$ is a matrix of learnable parameters.

Analogous to conventional transformers, an SE(3)-Transformer multi-head attention layer has multiple attention layers stacked together. The channels of the incoming feature vector are split, allowing each attention head (SE(3)-Transformer attention layer) to attend to different subspaces. The output of the multi-head attention layer is the concatenation of the outputs of the heads.

The SE(3)-Transformers also use norm nonlinearity layers, defined as:

$$\boldsymbol{f}_u^{i+1,l} = \mathrm{ReLU}(\mathrm{LN}^{i,l}(\|\boldsymbol{f}_u^{i,l}\|)\frac{\boldsymbol{f}_u^{i,l}}{\|\boldsymbol{f}_u^{i,l}\|} \tag{11}$$

Where ReLU is the rectified nonlinear unit, $\mathrm{LN}^{i,l}$ is a layer norm applied over all features of degree *l*, and

$$\|\boldsymbol{f}^l\| \coloneqq \sqrt{\sum_{\boldsymbol{m}=-\boldsymbol{l}}^{\boldsymbol{l}} \left(f_m^l\right)^{\mathbf{2}}}. \tag{12}$$

The default architecture of an SE(3)-Transformer as implemented by NVIDIA consists of seven multi-head attention layers, each followed by the equivariant norm layer above, and a final Tensor Field Network layer. The multilayer perceptrons in the radial profile functions use ReLU as an activation function.

Chapter 3

Design of TrIP

In this chapter, we embark on an exploration of the design choices behind the Transformer Interatomic Potential, or TrIP. TrIP is a fusion of machine learning techniques and an understanding of the fundamental physics governing interatomic interactions. This chapter unveils TrIP's architecture, shedding light on its core components, informed adaptations, and design choices that have resulted in a potent tool for simulating and understanding molecular systems. From the mathematical underpinnings to practical implementation, I delve into TrIP's construction, exploring how it expands upon SE(3)-Transformer networks to achieve physical realism.

3.1 Network

Transformer Interatomic Potential (TrIP) is a deep learning model that uses atomic species and positions to predict the energy of atomistic systems. In the landscape of existing methodologies, the core architecture of TrIP closely resembles NequIP[28] and MACE[29] due to their shared reliance on Tensor Field Network[36] (TFN) graph convolutions. However, TrIP takes a distinct route by mainly incorporating SE(3)-Transformer Multi-Headed Attention layers[37], which subsume the role of the TFN layers.

In my approach, I introduce two strategies that infuse physical realism into near- and long-range interactions. I employ Ziegler-Biersack-Littmark (ZBL) screened Coulombic interactions[38] to model nuclear repulsions. I augment the training set with lone atoms[39] to enforce the expected long-range energy and a modified bump function as a cutoff function to enforce gradual and exponentially asymptotic behavior.

My method also follows a "species-agnostic" strategy, wherein all parameters in the network are shared across atomic species, except those in the embedding layer. This idea draws inspiration from other deep learning algorithms, such as convolutional neural networks[40], where parameter-sharing enhances the model's ability to grasp underlying patterns within the data. This approach encourages the network to capture fundamental physical principles universally valid across diverse atomic environments. This strategy also augments the method's scalability to accommodate systems that include dozens of different atomic species by preventing an explosion in the number of parameters. Additionally, it contributes to regularizing the model, mitigating the risk of overfitting. Therefore, TrIP is a universal potential.

My network architecture is intentionally designed to exhibit mathematical smoothness, ensuring that the derivatives of all orders remain continuous[41]. Although the potential only needs to be differentiable to first order for force evaluations, the smoothness of the network reduces the likelihood of introducing anomalies or abrupt changes in the predicted energy and force profiles, which could potentially lead to unrealistic behaviors during simulations. I used a smooth cutoff function, a novel equivariant norm layer, smooth activation functions, and appropriate modifications to the attention mechanism.

3.2 Implementation

The atoms within the system are abstracted as nodes in a graph, and distinct nodes are connected via edges if the atoms lie within a $r_c = 4.6$ Å cutoff distance. To ensure a smooth transition over the cutoff boundary, TrIP uses a smooth cutoff function $\varphi: \mathbb{R} \to \mathbb{R}$ based on the bump function:

$$\varphi^{\text{cutoff}}(x) = \begin{cases} \exp\left(k - \frac{k}{1-x^2}\right), & x < 1 \\ 0, & x \geq 1 \end{cases} \tag{13}$$

where $k = 3$ is a hyperparameter that controls the shape of the cutoff function. TrIP uses the smooth cutoff function φ in conjunction with the cutoff distance r_c to generate "cutoff scalars" that scale messages passed between nearby nodes.

The nodes include atomic number and position information, while the edges include the relative position vectors between connected atoms. These relative position vectors are used in the creation of equivariant matrices described in Eq 5.

The neural network of TrIP is built upon NVIDIA's implementation of SE(3)-Transformers[37, 42] with modifications noted in the following section. A trainable embedding layer looks up a scalar-valued vector corresponding to each atom's species and is the initial features vector for the corresponding node. Next, the graph and data are passed through seven successive SE(3)-Transformer multi-headed attention layers—each followed by an equivariant norm layer—then through a Tensor Field Network layer[36]. The outputs of each SE(3)-Transformer layer and equivariant norm layer are tensors of order l=0,1,2[35] while the Tensor Field Network convolutional layer outputs a scalar-valued vector for each atom. These vectors

are then used in a multi-layer perceptrons (MLPs) that predict raw atomic energies. This MLP consists of the following layers: Linear(48, 48), SiLU, Linear (48, 48), SiLU, Linear(48, 1).

The other contribution to the total energy comes from the Coulombic interactions between nuclei that become significant in the near-range regime (< 1 Å). The standard Coulomb potential between the two nuclei multiplied by the ZBL screening function:

$$V_{uv}^{ZBL}(r) = \frac{1}{4\pi\epsilon_0}\frac{Z_u Z_v e^2}{r}\varphi^{\mathrm{ZBL}}\left(\frac{r}{a}\right) \tag{14}$$

Is the ZBL-screened Coulomb energy, where *a* is the "universal screening length":

$$a = \frac{0.8854 \cdot \left(0.529\ \text{Å}\right)}{(Z_u)^{0.23} + (Z_v)^{0.23}} \tag{15}$$

And φ^{ZBL}: $\mathbb{R} \to \mathbb{R}$ is the empirical ZBL screening function:

$$\varphi^{\mathrm{ZBL}}(x) = 0.1818\, e^{-3.2x} + 0.5099\, e^{-0.9423x} + 0.2802\, e^{-0.4028x} + 0.02817\, e^{-0.2016x}. \tag{16}$$

This potential is written edge-wise and does not use cutoff scaling, so to work with my node-wise total energy sum pooling layer I transfer this contribution to each node by the following:

$$V_u^{ZBL} = \frac{1}{2}\sum\nolimits_{v \in N_u} V_{uv}^{ZBL}(\|\vec{\boldsymbol{r}}_{vu}\|)\, \varphi^{\mathrm{cutoff}}\left(\frac{\|\vec{\boldsymbol{r}}_{vu}\|}{r_c}\right). \tag{17}$$

The sum of this value, along with the raw atomic energies predicted by the neural network, constitutes the system's total energy. TrIP's architecture is shown in Figure 1. Automatic differentiation calculates the forces on the atoms as the negative gradients of the energy with respect to the atomic positions. The end-to-end differentiability of the total energy from atomic positions ensures forces that are properly linked to the energy.

3.3 Changes to NVIDIA's SE(3)-Transformer

I adjusted NVIDIA's implementation of SE(3)-Transformers to work appropriately as an interatomic potential. Their implementation expects a predetermined graph structure, does not use a cutoff function to scale messages by distance, and uses either sum or mean pooling to generate a global quantity. Due to the lack of continuity, graph construction, and sum aggregation, their implementation is not suitable as an interatomic potential without substantial modifications.

I incorporate smooth cutoff boundary transitions by inserting the smooth cutoff function defined in Eq 12 in the appropriate places. I use it to change the basis kernels $\boldsymbol{W}^{lkj}$ from Eq 5 to:

$$\left(\boldsymbol{W}^{lkj}\right)_{m_l m_k}(\vec{\boldsymbol{r}}) = \varphi^{\text{cutoff}}\left(\frac{\|\vec{\boldsymbol{r}}\|}{r_c}\right)\sum\nolimits_{m_J=-j}^{j} Y_{m_j}^{j}\left(\frac{\vec{\boldsymbol{r}}}{\|\vec{\boldsymbol{r}}\|}\right) C_{jm_jkm_k}^{lm_l}. \tag{18}$$

And the attention values from Eq 7 to:

$$\alpha_{uv} = \frac{\varphi^{\text{cutoff}}\left(\frac{\|\vec{\boldsymbol{r}}_{vu}\|}{r_c}\right)\exp\left(\boldsymbol{q}_u^{\intercal}\boldsymbol{k}_{vu}\right)}{\sum_{w\in N_u}\varphi^{\text{cutoff}}\left(\frac{\|\vec{\boldsymbol{r}}_{wu}\|}{r_c}\right)\exp(\boldsymbol{q}_u^{\intercal}\boldsymbol{k}_{wu})}. \tag{19}$$

I incorporated smoothness by changing the activation functions from ReLU to SiLU in the radial profile function. I also use a custom norm layer that normalizes all features, removing most discontinuities in the derivatives. While there is still a discontinuity at the origin, it is inconsequential for practical purposes due to the exceptionally high dimensionality of the feature space. For a feature vector $\mathbf{f} \in \mathbb{R}^{C\times M}$ composed of C channels and M indices per feature, my norm is:

$$\boldsymbol{f}^{i+1} = \text{Norm}\left(\boldsymbol{f}^{i}\right) = \gamma \cdot \frac{\boldsymbol{f}^{i}}{\|\boldsymbol{f}^{i}\|_2/C} \tag{20}$$

where γ is a learnable parameter and $\| \quad \|_2$ is the Frobenius matrix norm, effectively generalizing equation 11 to work over multiple channels and degrees of spherical tensors by effectively "norming the norms".

I also changed the radial profile functions in Eq 4 to use learned invariant edge features instead i.e., $\rho^{lkj}(\boldsymbol{e}_{uv}^{0})$. This enables the radial profile function to use the atomic identities while reasoning about the geometry. I set the initial, invariant edge feature vector $\boldsymbol{e}_{uv}^{0,0}$ to have the distance in the first index and zeros everywhere else. The edge feature vector updates after each multi-headed attention layer using a ResNet-like connection:

$$\boldsymbol{e}_{uv}^{i+1,0} = \boldsymbol{e}_{uv}^{i,0} + \boldsymbol{f}_{u}^{i,0}. \tag{21}$$

Finally, I changed the pooling layer to allow sum pooling, which aligns with the observation that the total energy scales with the number of atoms and naturally identifies the output of the final MLP as an atomic energy.

A diagram of how the weight kernel, attention mechanism, TFN layer, SE(3)-Transformer Attention layer, and the SE(3)-Transformer Multi-Head Attention is shown in Fig 1. A diagram of TrIP itself is shown in Fig 2.

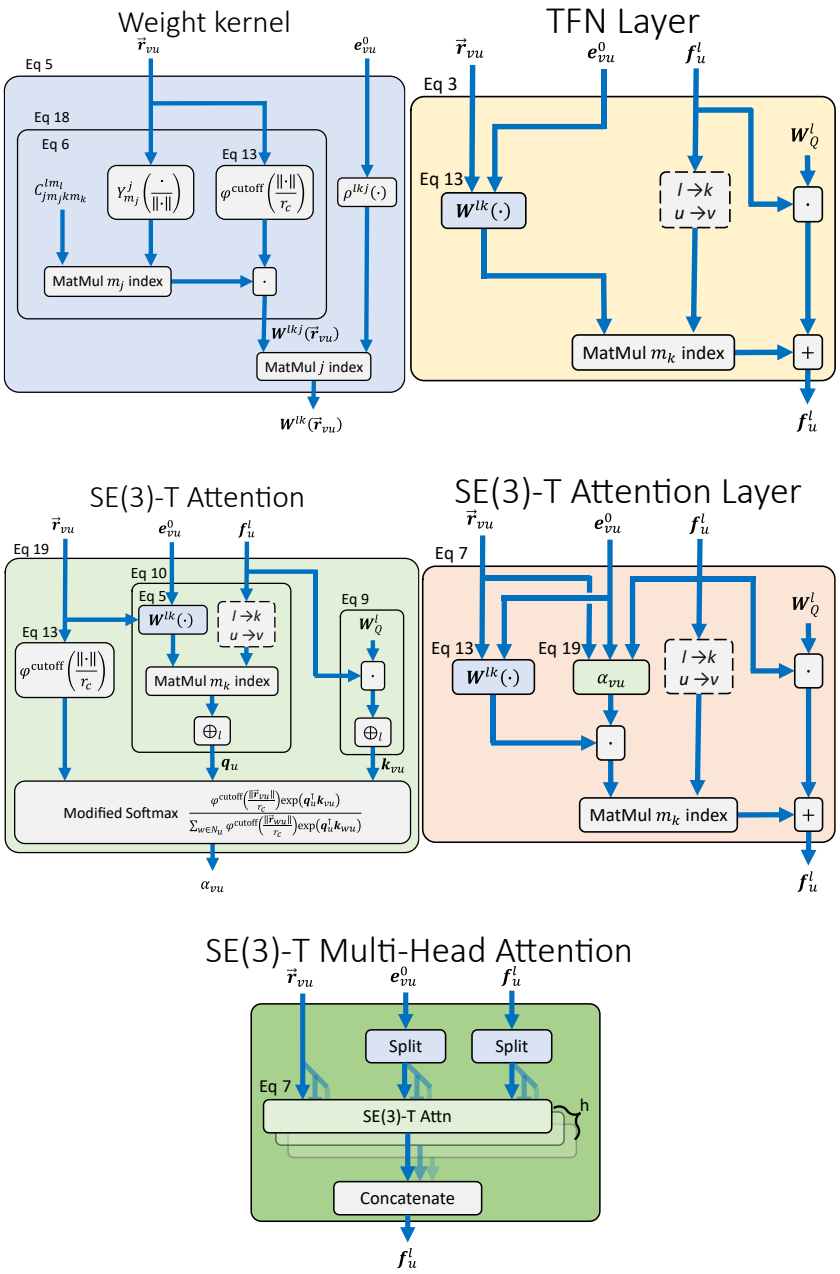

Figure 1: *Diagrams of the main layer types used in TrIP. The weight kernel learns equivariant matrices. The TFN layer uses the weight kernel to calculate a fused linear-tensor product of the node features. The SE(3)-Transformer (SE(3)-T) Attention uses learned spherical tensors to calculate attention values. The SE(3)-T Attention layer incorporates the SE(3)-T Attention into the TFN layer to scale messages. The SE(3) Multi-Head Attention uses multiple SE(3)-T Attention Layers in parallel that attend to different feature subspaces. These diagrams are for illustrating mathematics and are equivalent to the implementation.*

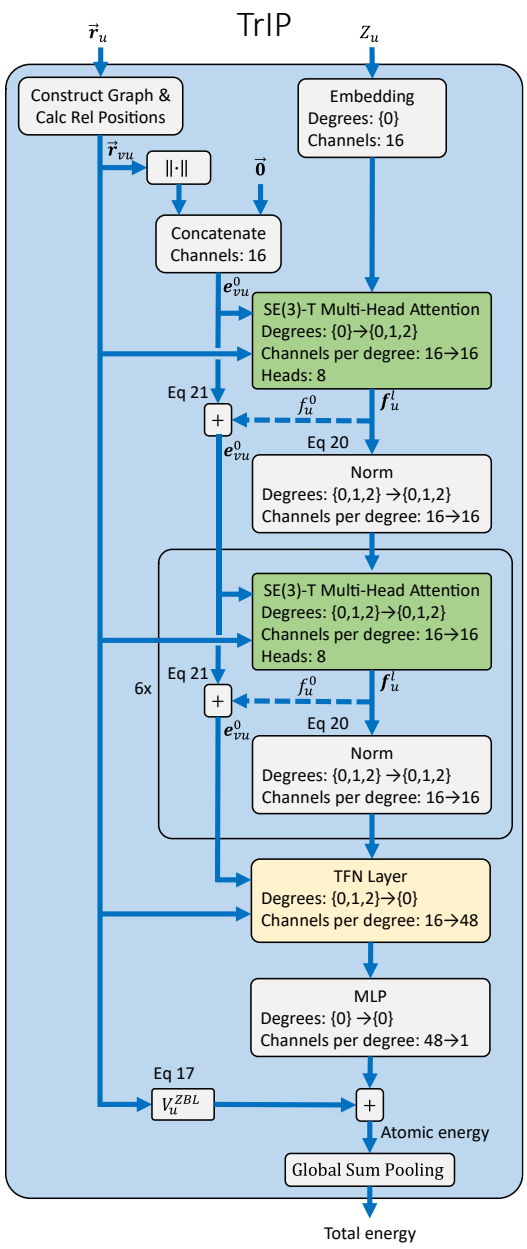

Figure 2: *Architecture of TrIP. This figure shows the connectivity of the layers and how the data changes. The atom species pass through an embedding layer, resulting in a scalar-valued vector for each atom. TrIP uses the graph and relative positions to generate higher degree spherical tensors from these features via SE(3)-T Multi-Head Attention layers. A final TFN layer reduces all the features to scalars, which an MLP uses to generate a scalar for each atom. The sum of these scalars with the ZBL-screened Coulombic energy makes up the total energy. Automatic differentiation calculates the forces on atoms as the negative gradient of the total energy with respect to the atomic coordinates.*

Chapter 4

Experimental Design

This chapter serves as a guide to the methods and procedures employed throughout the research. It starts with the preparation of both the training dataset and the benchmark set, The later sections delve into the intricate process of training the TrIP model and how I designed the experiments.

4.1 Training, Validation, and Test Datasets Preparation

I used the ANI-1x dataset to train TrIP. I performed a 95%-5% training-validation split on the ANI-1x dataset such that all conformations of a given molecular formula are in the same subset. The ANI-1x dataset consists of molecules from the GDB-11 and ChEMBL databases and generated amino acids and 2-amino acid peptides generated in RDKit. The conformations are sampled from molecular dynamics trajectory sampling, normal mode sampling, dimer sampling, and torsion sampling using active learning. The energies and forces were calculated in Gaussian 09 using ωB97x/6-31*.

I use the Comprehensive Machine-learning Potential (COMP6) benchmark[23] to test TrIP. COMP6 was designed to assess models trained on the ANI-1x dataset for generalizability,

especially to larger molecules. The mean molecule size in the COMP6 subsets ranges from seventeen atoms to 75 atoms with the largest molecule having 312 atoms. COMP6 encompasses over 38,000 conformations originating from a diverse range of molecules, including those sourced from DrugBank[43], random Tripeptides, GDB13[44], and S66x8[45]. Like the ANI-1x dataset, the energies were calculated in Gaussian 09 using ωB97x/6-31*. I did not perform any filtering on COMP6 so that it would be more stringent and so the results would be directly comparable to other models. Statistics of the training, validation, and test sets are found in Table 3.

Purpose	Dataset (Source)	Molecules (Conformations)	Atoms/molecule Mean (STD)	Heavy atoms/molecule Mean (STD)
Training	ANI-1x (GDB-11, ChEMBL)	~60,000 (4,404,973)	15 (6)	8 (3)
Validation	ANI-1x (GDB-11, ChEMBL)	~3,000 (212,173)	15 (6)	8 (3)
Testing	S66x8 (S66x8)	66 (528)	20 (7)	9 (3)
Testing	ANI-MD (PDB)	14 (1,791)	75 (72)	37 (38)
Testing	GDB7to9 (GDB-11)	1,500 (36,000)	17 (3)	9 (1)
Testing	GDB10to13 (GDB-13)	2,996 (47,670)	25 (4)	13 (1)
Testing	Tripeptides (RDKit)	248 (1,984)	53 (7)	27 (4)
Testing	DrugBank (DrugBank)	837 (13,379)	44 (20)	22 (10)

Table 3*: Statistics about training, validation, and testing sets. This proves that the training and validation sets have comparatively small molecules while the testing set has a wide range of molecular sizes. Information on the exact number of molecules is unavailable because the ANI-1x dataset groups molecules with the same formula together so the values are estimated based on the facts that the full ANI-1x dataset has 64,865 atoms, I use a 95%-5% split, and that some molecules were excluded due to not having any usable energy nor force values.*

As in the ANI models, I standardized the energies using so-called self-interaction energies and the remaining standard deviation, which were calculated on the training set using multiple linear regression. The independent variables for the multiple linear regression were the number of each atomic species and the dependent variable was the quantum mechanical energy for each conformation. The resulting fitting parameters describe the average energy per atomic species that best describes the dataset and the error from linear regression is the dataset's effective standard deviation. For a given point in the dataset, TrIP standardizes the energy by subtracting the self-interaction energies for each atomic species and then dividing by the standard deviation. During inference, TrIP reverses the standardization process by scaling the prediction of the model by the standard deviation and adding the self-interaction energies.

Additionally, I included experimental values for electron binding energies for the four lone atomic species with the rest of the ANI-1x dataset. TrIP standardizes these lone energies in the same way it standardizes the conformational energies. The training protocol includes the lone energies in each minibatch during training to constrain the model in the limiting case.

4.2 Training

My loss function was based on the Pseudo-Huber loss function[46]:

$$L_\delta(r) = \delta^2 \left(\sqrt{1 + (r/\delta)^2} - 1\right). \tag{22}$$

Where *r* is the residual—the difference between a target value and predicted value—and $\delta = 0.2$ is a hyperparameter that sets the maximum slope of the loss function. I use this loss function due to its smoothness and that it approximates the L2 squared loss when $r \ll \delta$ and

the L1 absolute loss when $r \ll \delta$, and is positive except when *r = 0* when the loss is zero. Like the loss function for the ANI-2 model, my loss function has terms for both energy and force loss:

$$L = \frac{1}{N}\sum_{i}^{N} L_{\delta}\left(E_i - \widehat{E_i}\right) + \frac{l_o}{\sum_{i=1}^{N} M_i}\sum_{i=1}^{N}\sum_{j=1}^{M_i} L_{\delta}\left(\left\|\boldsymbol{f}_{ij} - \hat{\boldsymbol{f}}_{ij}\right\|\right) \tag{23}$$

where $\widehat{E_i}$ and $\hat{\boldsymbol{f}}_{ij}$ are the respective standardized energies and forces predicted for a given molecule, E_i and $\boldsymbol{f}_{ij}$ are the respective standardized QM energies and forces, *i* indexes *N* molecules within a minibatch, *j* index the M_i atoms within molecule *i*, and $l_o = 0.1$ weights the force loss relative to the energy loss as was used in ANI2.

I trained TrIP on my training subset of the ANI-1x dataset. I trained different versions of TrIP networks on 8 GPUs for 10 epochs with automatic mixed precision using minibatch sizes of 25 conformations and the four lone atoms. I used the AdamW optimizer with a weight decay of 0.1, a learning rate of 1e-3 that decayed by a factor of 0.5 each epoch, and a gradient norm clipping of 10.0. The cutoff radius r_c was set to 4.6 Å and the force weight l_o was set to 0.1; the same values as used in the ANI networks. The hyperparameter k was set to 3.0 in the cutoff function Eq 12 and δ was set to 0.2 in the Pseudo-Huber loss function Eq 21. I experimented with various hyperparameters, but the validation errors did not significantly improve for any model I evaluated beyond the defaults for NVIDIA's implementation of the SE(3)-Transformer and my initial guess for hyperparameters. Training took about 20 hours to complete on the full TrIP model.

For the ablation studies, I changed the network architecture while using the same training settings. In the “Equivariant & one-head attention” model, I only used a single attention head per multi-head attention layer, which eliminated splitting the features into groups and later concatenation. In the “Equivariant” model, multi-headed attention layers were replaced with Tensor Field Network layers that use the same number of channels and types of spherical tensors—equivalent to using a single attention head setting the attention value α to one—and in the “Invariant model”, only scalar values were used, removing the equivariant components of the spherical tensors and the network.

4.3 Experimental Protocols

After training, I benchmarked TrIP and the ablation models on the COMP6 dataset. I calculated the energy and force predictions mean absolute error (MAE) and root mean square error (RMSE) for using the same method as the ANI networks.

I performed a rigid scan of an H_2O molecule by fixing the H-O-H angle at 104.5° degrees and varying the O-H distances between 0.5 Â and 4.5 Â in increments of 0.05 Â. I calculated the single point energies using TrIP, ANI-2x, and singlet, triplet, and quintet ωB97x/6-31* Density Functional Theory (DFT) method in Gaussian 16. I calculated the ground-state energy by taking the minimum of the singlet, triplet, and quintet states of each conformation.

I created and optimized an initial structure for ephedrine in Avogadro[47]. I rotated the O-C-C-N torsion angle by increments of 10 degrees, resulting in 36 initial positions. Then I optimized the structure by minimizing TrIP and ANI-2x potentials using conjugate gradient

descent while using a torsion error function to constrain the angle. I then calculated the energy and angle for the resulting conformation.

I use a similar protocol to run a torsion benchmark on QC Torsion Drive structures[48]. First, I filtered the QC Torsion Drive dataset by removing all structures with atomic species not contained in the dataset. I optimized the QC initial structures with conjugate gradient descent on the TrIP and ANI-2x potentials while constraining the torsion angle. Then I compared the ML minimized structures against the reference structures using the average RMSDs.

I conducted molecular dynamics (MD) simulations using OpenMM[49]. I ran simulations using TrIP, ANI-2x, and Amber14[50]. I incorporated TrIP into OpenMM as an external force, updating the energies and forces at each step of a simulation. I modeled the system using explicit water: ANI-2x and TrIP directly predicted the solvent interactions while Amber14 used TIP3P-FB[51]. I confined the system to a cubic box with periodic boundary conditions. This was implemented in TrIP by subjecting the graph construction to toroidal boundary conditions[52] and using modified relative positions and distances between atoms and their virtual neighbors. I calculated the modified relative positions by

$$\vec{\boldsymbol{r}}_{vu} = \left[\left(\vec{\boldsymbol{r}}_u - \vec{\boldsymbol{r}}_v + \frac{\vec{\boldsymbol{b}}}{2}\right) \bmod \vec{\boldsymbol{b}}\right] - \frac{\vec{\boldsymbol{b}}}{2} \tag{24}$$

Where $\vec{\boldsymbol{b}}$ is the box diagonal vector.

First, I minimized the energy of the system, and performed an NVT (constant number of particles, volume, and temperature) equilibration for 100 ps followed by an NPT equilibration for 100 ps. A Monte Carlo barostat was employed to keep a constant pressure for the NPT

equilibration and the production simulation. I ran 20 ns production simulations with step sizes of 0.5 fs for TrIP and ANI and 2 fs for Amber.

I employed the MDAnalysis toolkit[53,54] for the analysis of the molecular dynamics (MD) simulations, We. I extracted the phi (φ) and psi (ψ) torsion angles for each residue sampled at fixed intervals from the trajectory. I assigned the secondary structure state of each residue based on the location of the torsion angles in the Ramachandran plot. I define the secondary structures in terms of rectangular sections of the Ramachandran domain according to Table 4. The statistics for these angles, including the average and standard deviation for each method, were computed.

Secondary Structure	ϕ lower bound (degrees)	ϕ upper bound (degrees)	ψ lower bound (degrees)	ψ upper bound (degrees)
β	130	-90	50	-110
P_{II}	-90	20	50	-40
α_R	130	20	-110	50
α_L	20	130	-90	120
P	20	130	120	-90

Table 4*: Definition of secondary structures in terms of rectangular subsets of the Ramachandran domain.*

Chapter 5

Results

This chapter stands as a testament to the rigorous training and design of the Transformer Interatomic Potential (TrIP) model. With the foundation firmly laid, this chapter embarks on a journey of empirical validation and practical benchmarking. The central aim is to assess the capabilities and performance of TrIP in the realm of molecular simulations and interatomic energy predictions. I highlight TrIP's prowess in accurately predicting interatomic energies and forces, surpassing the benchmarks set by its predecessors.

5.1 Potential Benchmark

I present the accuracy of TrIP in predicting interatomic energies and forces on the COMP6 benchmark in Table 5. Trip outperforms all other models in all regards, with about half the error compared to the ANI models. Notably, all my ablation models that use equivariant perform as well as ANI-1x in all metrics, except the “Equivariant” model’s forces RMSE.

Potential	Energy Error (kcal/mol)	Force (kcal/mol/A)
ANI-1[55]	5.01/16.9	3.70/7.13
ANI-1x[56]	1.93/3.37	3.09/5.29
NewtonNet*[30]	1.45/…	1.79/…
HIP-NN-TS (n_{int}=2)[57]	1.25/…	1.44/…
GM-NN[58]	2.03/...	1.85/…
TrIP (Equivariant & multi-head attention)	**1.04/**1.71	**1.41/2.80**
TrIP (Equivariant & one-head attention)	1.07/**1.68**	1.50/2.94
TrIP (Equivariant)	1.27/2.86	1.75/6.65
TrIP (Invariant)	5.30/8.05	7.56/11.6

Table 5*: Comparison of energy and force prediction accuracies between ANI models, other interatomic potentials, TrIP, and various ablation models on the COMP6 benchmark. Unreported values are marked by "…". The errors are reported in mean absolute error (MAE)/root mean square error (RMSE) for each model. The energy errors are per molecular conformation and the force values are evaluated per component. TrIP has significantly lower errors in both energy and force predictions compared to the other two models.*

** For NewtonNet, the energies outside of 100 kcal/mol were excluded*

5.2 Water Scan Experiments

The next experiment was a rigid energy scan of a water (H2O) molecule using four different methods: singlet-state DFT, ground-state DFT, TrIP, and ANI 1-x. Here, the DFT method is ωB97x/6-31G*, the same used in the generation of the dataset. The results shown in Figure 6 show the effects of the physical bias used in the near- and far-region when comparing TrIP against ANI-2x. I attribute the physical bias caused TrIP to have the correct physical tendencies in all parts of the geometry.

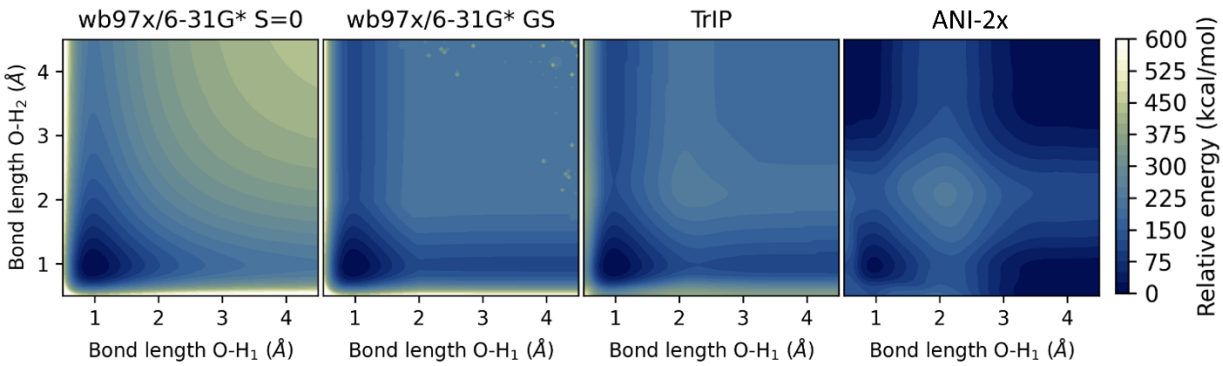

Figure 6*: Rigid energy scan on an H_2O molecule using four different methods: ωB97x/6-31G* S=0, ωB97x/6-31G* GS, TrIP, and ANI1-x. All four methods used the same molecule conformations to evaluate the point energies. The minimum evaluated energy was subtracted from each model's outputs to obtain the contour plots above. These plots demonstrate the physical realism incorporated into TrIP by adding the ZBL-screened Coulombic interactions to the potential and the lone atoms to the training set. Notably, there is a peak in TrIP and ANI-2x around 2 Å, thought to be due to the ANI-1x dataset using singlet electron multiplicity rather than the ground-state.*

5.3 Torsion Scans

I conducted relaxed torsion scans of the complex molecule ephedrine using ωB97x/6-31G*, TrIP, and ANI-2x. I scanned the dihedral angle of the O-C-C-N atom sequence from 0 to 360 degrees in 10-degree increments. For each setting of the dihedral angle, an initial conformation with a dihedral constraint using either TrIP or ANI-2x. Furthermore, I cross-validated the results with the ωB97x/6-31G* S=0 method. The error between TrIP and ωB97x/6-31G* was 0.54 kcal/mol RMSE and the error between ANI-2x and ωB97x/6-31G* was 0.66 kcal/mol. The torsion profile is shown in Figure 7.

Torsion benchmarking on the QC Torsion Drive structures[48] was performed by comparing QC minimized structures against TrIP and ANI-2x minimized structures. Of the 942 molecules, I removed 460 for containing species beyond those in the ANI-1x dataset, leaving 482 for

scanning. The average RMSD between TrIP and QC is 0.0368 Å and the average RMSD between ANI-2x and QC is 0.123 Å.

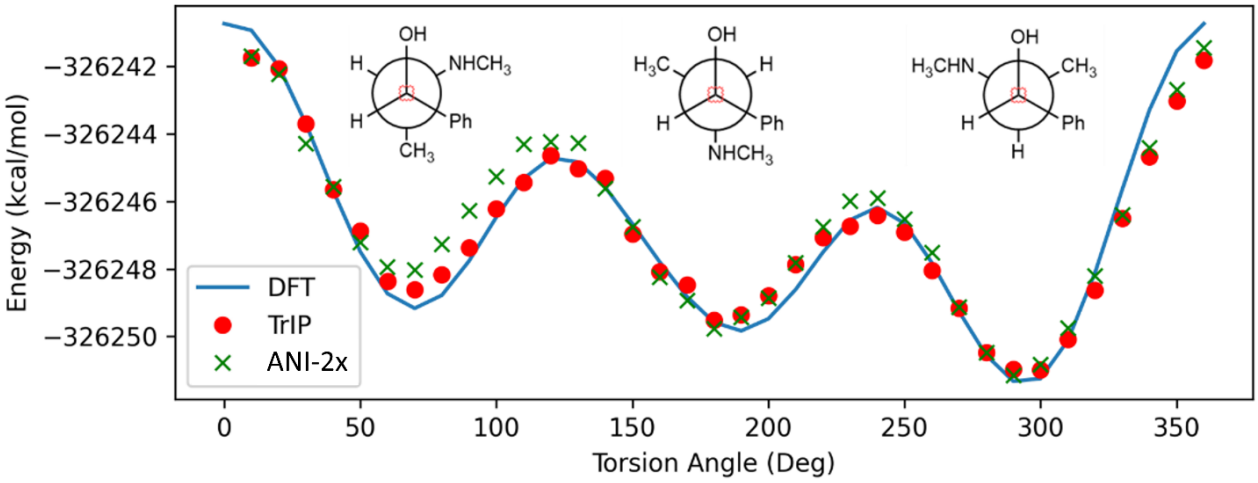

Figure 7: *Results of the Ephedrine torsion scan. The plot shows the calculated energies as a function of the dihedral angle between O-C-C-N. Results obtained with the TrIP model (red dots) are compared with those from the ANI-2x (green x), and the ωB97x/6-31G* S=0 method (blue line).*

5.4 Trialanine Simulations

To comprehensively assess TrIP's performance, I conducted extensive 20 ns simulations on Trialanine, employing three distinct methodologies: Amber14/TIP3P-FB, ANI-2x, and TrIP. The results of these simulations offer valuable insights into the behavior and structural dynamics of Trialanine molecules. Figure 8 provides a visual representation in the form of surface Ramachandran plots, showcasing the conformations sampled during the molecular dynamics trajectories, similar to one found elsewhere.[59] These plots reveal how the secondary structures evolve, allowing for insightful comparisons between TrIP, ANI-2x, and Amber14 models.

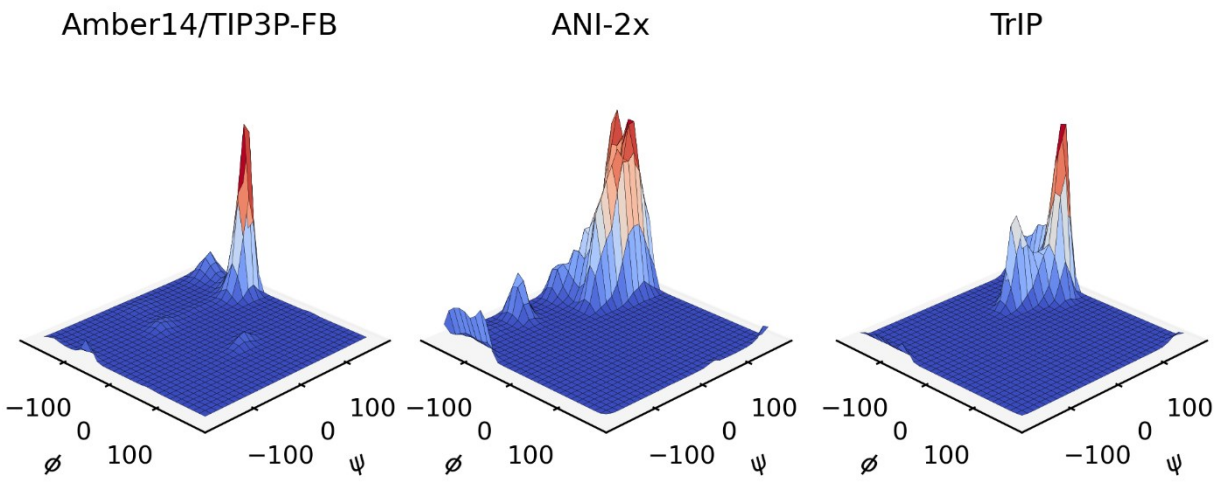

Figure 8*: Surface Ramachandran Analysis of Trialanine Dynamics. This figure offers a surface Ramachandran analysis, unveiling the dynamic behavior of Trialanine molecules throughout molecular dynamics simulations. This plot provides an insightful visual exploration of conformational changes as Trialanine evolves, highlighting the significant differences between the TrIP, ANI-2x, and Amber14 models.*

Model	State	Φ (deg)	Ψ (deg)	ΔΦ (deg)	ΔΨ (deg)	P (%)
Amber	β	-132.0	155.4	24.6	19.9	17.5
	P_{II}	-67.5	152.0	10.4	17.1	72.3
	α_R	-95.5	-12.5	33.8	24.9	6.1
	α_L	56.1	23.1	8.4	22.5	3.7
ANI-2x	B	-124.4	133.6	23.1	36.5	67.7
	P_{II}	-78.3	125.1	8.4	32.6	26.3
	α_R	-141.7	-13.5	21.6	44.1	6.0
TrIP	B	-122.5	145.0	27.2	34.6	35.4
	P_{II}	-70.3	133.9	12.3	32.2	63.7
	α_R	-86.6	43.4	8.9	5.9	0.9

Table 9*: Torsional Angle Statistics in Trialanine Molecular Dynamics Simulations. This table presents statistics characterizing the torsional angles in Trialanine molecules during molecular dynamics simulations. These statistics encompass the angles Φ and Ψ, their standard deviations ΔΦ and ΔΨ, and the corresponding population percentages (P) for different standard der states. The models evaluated include Amber, ANI-2x, and TrIP, shedding light on the distinctive conformational dynamics exhibited by Trialanine under the influence of these computational approaches.*

Chapter 6

Discussion

The preceding chapter presented a detailed examination of the Transformer Interatomic Potential (TrIP), highlighting its capabilities in accurately predicting interatomic energies and forces across various molecular systems. This chapter delves deeper, aiming to dissect and interpret the significance of these findings, placing them within the broader context of computational molecular modeling and interatomic potential research. The core goal is to understand not just the 'how' but the 'why' behind TrIP's performance, drawing connections between my results and the underlying principles of molecular dynamics and machine learning.

6.1 Analysis of TrIP's Performance in COMP6

In the COMP6 benchmark, a renowned testing ground for molecular modeling methods, the Transformer Interatomic Potential (TrIP) showed a significant advancement in predicting interatomic energies and forces. This performance is a testament to the sophisticated integration of machine learning techniques and a deep understanding of atomic interactions, as meticulously outlined in the design of TrIP.

The benchmark results, particularly the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), place TrIP ahead of the ANI-1 and ANI-1x models. This improved accuracy is

not merely a numerical achievement but reflects the efficacy of TrIP's core components – the Tensor Field Network (TFN) layers and the SE(3)-Transformer. These elements enable TrIP to grasp the complexities of molecular systems, ensuring that its predictions are not only accurate but also physically realistic.

One of the key features contributing to TrIP's success is the incorporation of equivariant layers, which maintains consistency in network responses under transformations like rotations and translations. This aspect is crucial in preserving the physical symmetries inherent in molecular systems, a factor often overlooked in traditional approaches. Additionally, the SE(3)-Transformer's attention mechanism allows TrIP to consider multi-particle interactions dynamically, enhancing the model's ability to scale and adapt its predictions according to the surrounding atomic environment.

The physical realism in TrIP's design, particularly its novel adaptations like the smooth cutoff function and the custom equivariant norm, plays a pivotal role in its performance. These design choices ensure that the model not only captures the intricate interplay of atoms within molecules but also aligns closely with the fundamental principles governing these interactions. This alignment is evident in TrIP's superior performance in the benchmark, especially in its ability to accurately model both near- and far-range interactions, a challenge for existing models.

Furthermore, TrIP's superior performance in the COMP6 benchmark carries significant implications for the field of molecular simulations. Its ability to predict atomic energies and forces with high precision paves the way for more accurate molecular dynamics simulations, potential energy surface mappings, and geometry optimizations in diverse applications ranging from drug discovery to materials science.

TrIP’s performance not only highlights the potential of machine learning in scientific applications but also opens new avenues for future research, where further refinements and enhancements to TrIP could lead to even more groundbreaking advancements in the field of computational chemistry.

6.2 Evaluation of Physical Bias in Water Molecule Energy Scans

The water molecule rigid scan results provide a compelling demonstration of TrIP's capabilities in capturing bond formation and breaking behavior accurately. It shows that the inclusion of the lone atoms in the training set and the addition of the ZBL potential improve near-range and long-range interactions compared to the ANI-2x model.

Interestingly, this experiment shows that my approach matches ground-state DFT calculations better than the other methods. I attribute this alignment to the intrinsic properties of graph neural networks, particularly the linearity in the global sum aggregation step. Consequently, in the non-interactive domain, where the influence of interatomic interactions diminishes, the energies predicted by the TrIP model approach those of isolated atoms. This is consistent with the behavior of ground-state energies, which are fundamentally linear combinations of atomic energies in the non-interactive limit. However, it is crucial to differentiate between the ground-state and the singlet-state. The singlet-state is not always the ground-state and does not inherently have the same linearity as the ground-state.

I interpret that as a consideration when generating data for machine learning interatomic potentials. If intersystem crossing (nonradiative transitioning between different electron states with different spin multiplicities) is readily permissible—e.g., through interactions with heavy atoms with large spin-orbit couplings—then it is advisable to calculate the dataset in the ground-state. On the other hand, if the system is constrained and cannot easily switch between electron states, selecting the appropriate electronic configuration becomes more critical and should be guided by the specific characteristics of the systems one aims to study. For instance, if the system is usually in a singlet-state and rarely undergoes intersystem crossing, it may be more relevant to use singlet-state configurations. Then, if the inherent nonlinearity that will appear in the long-range region is important for fitting or inference, carefully consider what sort of physical constraint is appropriate and how to construct the dataset.

There is a high likelihood that training a linear machine learning algorithm (those predicting a per atom energy and using a global sum aggregation for total energy) on non-linear data (such as non-ground-state singlet electron state energies) could put significant pressure on the algorithm to be steep where those assumptions disagree. It could also lead to poor training convergence, poor prediction of physical properties, or instability during molecular dynamics. It is vital to carefully assess the dynamics and behavior of the systems in question to choose the electronic configurations that most accurately represent the relevant physical phenomena, and thereby design the algorithm and dataset.

Another interesting feature in the TrIP and ANI potential energy scans of water is that they both show peaks near 2 Å. This is due to all energies in the ANI-1x dataset being calculated in singlet electronic state while some conformations may have ground-states in triplet electron state.

TrIP predictions align closely with the reference data during progressive elongation and eventual breaking of the O-H bond in a water molecule. Importantly, it managed to capture the steep increase in potential energy typically associated with bond-breaking events and an approximate plateau value due to the inclusion of the loan atoms during training and the exponentially decaying cutoff function, showing that it accurately models the substantial energy input needed for chemical transformations.

In contrast, traditional force fields cannot accurately describe bond-breaking and bond-forming events due to their intrinsic limitation of typically considering bonds as fixed connections between atoms. This experiment's results hence underscore the advantage of these approaches in capturing the complex nature of chemical bonding.

6.3 Assessment of Ephedrine Torsion Scan

Torsion scans play an integral role in pharmaceutical research, particularly in finding the conformational preferences of molecular structures and thus influencing drug design strategies. The torsion scan of the ephedrine molecule exemplifies TrIP's applicability in such important pharmaceutical contexts. My model was able to accurately reproduce the torsional energy profiles, highlighting its potential in guiding the design of novel molecules with desirable properties.

While my results are promising, they also expose an area for improvement - the need for more robust training sets incorporating a wider variety of atom types. By incorporating more varied and complex atom types in the training sets, TrIP can be further optimized for application

in pharmaceutical research, potentially opening new avenues for drug discovery and design. This approach will enable my model to provide even more accurate and informative torsion scans.

6.4 Interpretation of Trialanine Molecular Dynamics Simulations

I found that the ANI-2x trajectory seemed to explore more of the Ramachandran space than the other potentials, even regions that are not occupied by reference conformations in the PDB. On the other hand, TrIP did not explore the landscape as much but hardly ever occupied any secondary structure other than β and P_{II}. This might also be a solvation result, given that AMBER was optimized against Ramachandran probabilities with TIP3P water models.

One of the most compelling outcomes was the demonstration of the stability of MD simulations using TrIP. Even in complex, demanding computational scenarios, TrIP showed remarkable consistency and stability. The generated trajectories did not display erratic behavior, and the energy conservation was excellent. These results provide compelling evidence that TrIP can accurately simulate the intricate dynamics of chemical systems, including the behavior of individual atoms and the entire system.

A key feature of my work with TrIP is its reusability and adaptability. While the model itself presents a significant advancement in ML potentials, the true capability of this research lies in its promise for iterative improvements and refinements. The code base, available via my repository, is specifically designed to enable retraining of the model on different datasets, other than ANI-2x. Additionally, the code base provides a framework for conducting various

experiments. Researchers can use the available scripts to perform their own torsion scans, potential energy surface mappings, and molecular dynamics simulations, among others. The availability of these tools should encourage widespread engagement with my model and lead to collective enhancements in the field of machine learning potentials.

6.5 Broader implications

There are dozens of published machine learning potentials. The most important considerations when selecting a model for practical use are performance, accuracy, generality, and physicality. TrIP is based on a highly performant SE(3)-Transformer implementation and is fast enough for simple simulations as demonstrated in the 20 ns simulation of solvated Trialanine. However, equivariant machine learning potentials are still much slower than force fields, and much more research needs to be done to reduce the number of required floating point operations and better parallelizing calculations.

Comparing models tested on COMP6, TrIP's accuracy on energy and force predictions surpasses all others. The ablation studies show that the most important contributions to the accuracy were using an equivariant architecture and attention. Similarly, the highest accuracy models on COMP6 are equivariant. However, multi-headed attention marginally increased accuracy. This suggests that the model did not need to pay attention to many things at a time or was not able to effectively leverage multiple heads. This could have been due to the dataset being simple (molecules containing first- and second-row elements with near equilibrium geometry), too few parameters, or the architecture not providing good enough information for the heads. The last option seems the most likely, considering that the inputs for the radial basis

functions come from learned scalar features and interatomic distances. While equivariant neural networks can learn molecular geometry from relative positions alone and do not require handcrafted features to achieve high accuracy, it is likely that better geometric descriptors—such as angles, torsions angles, and cross-products of radial positions—would improve accuracy.

Like other equivariant architectures, TrIP's species-agnostic architecture encourages it to learn more general features that are shared between atomic species. It also did not prevent TrIP from still achieving high accuracy. This property will be useful looking forward to universal potentials managing dozens of species by not requiring parameters to scale specific to the number of parameters. However, a completely species-agnostic architecture may require a deeper architecture and may become a hindrance for models featuring dozens of species. Future models will benefit from a better balance between species-specific and general parameters while being data efficient.

TrIP achieves the best physicality of any published potential. Adding limiting cases to the training set and the ZBL-screened potential to the output of the neural network improved the behavior in the short- and long-range interactions respectively, exemplified by the water scans. While other potentials have used ZBL-screening terms, including However, there is likely more that can be gleaned from physics that would improve the predictions before requiring machine learning. For example, it may be possible to predict partial charges to predict a coulombic interaction beyond those in the non-shielding regime. These contributions would need to be computationally efficient—with at most linear scaling—to be practical for molecular dynamics simulations. However, they could reduce what is required for the network to learn, reducing the variance of the effective training set, and allowing the network to focus on finer details.

In this rapidly changing field, TrIP will not remain state-of-the-art for long. However, TrIP has demonstrated that the SE(3)-Transformer architecture has useful machinery, a powerful mathematical framework, and that attention improves on equivariant architectures. TrIP also has introduced improved physicality in the long range. Finally, it has hopefully elucidated the current challenges and considerations for the field for future research. For these reasons, TrIP is a useful stepping stone towards an accurate, performant, universal potential.

Chapter 7

Conclusions

In this study, I describe my machine learning potential, Transformer Interatomic Potential (TrIP), and evaluate it across a range of critical computational chemistry tasks, showing its robustness and versatility. Testing on the COMP6 dataset reveals that my method achieves state-of-the-art results in all predictions, approaching the accuracy of Density Functional Theory on molecules outside of the training set. The water energy scan shows the physical bias I incorporated into TrIP improve its behavior energy predictions in the near- and far-fields, demonstrating its potential for high-quality molecular simulations and potential energy surface mapping and overcoming nonphysical behavior where training data is not present. The effectiveness of TrIP was further evident in my torsion scans, where the network delivered precise energy and conformation predictions, revealing its capacity to be a powerful tool in pharmacological studies. Finally, through molecular dynamics simulations, TrIP showed excellent stability and predicted a similar Ramachandran plot as a classical force field designed for that sort of system, showing its potential for free energy calculations. The success of TrIP is a milestone towards the development of accurate universal machine learning potentials, opening the door to more detailed and accurate explorations of complex molecular phenomena. The code and TrIP parameters for the experiments are available at https://github.com/dellacortelab/trip.

Bibliography

(1) Cai, X. Exact Stochastic Simulation of Coupled Chemical Reactions with Delays. *The Journal of chemical physics* **2007**, *126*, 124108.

(2) Bash, P. A.; Field, M. J.; Davenport, R.; Petsko, G. A.; Ringe, D.; Karplus, M. Computer Simulation and Analysis of the Reaction Pathway of Triosephosphate Isomerase. *Biochemistry* **1991**, *30*, 5826-5832.

(3) Ess, D. H.; Wheeler, S. E.; Iafe, R. G.; Xu, L.; Celebi-Oelcuem, N.; Houk, K. N. Bifurcations on Potential Energy Surfaces of Organic Reactions. *Angewandte Chemie International Edition* **2008**, *47*, 7592-7601.

(4) Coates, T. L.; Young, N.; Jarrett, A. J.; Morris, C. J.; Moody, J. D.; Corte, D. D. Current Computational Methods for Enzyme Design. *Modern Physics Letters B* **2021**, *35*, 2150155.

(5) Stevens, J. A.; Grunewald, F.; van Tilburg, M.; Thornburg, Z. R.; Gilbert, B. R.; Brier, T. A.; Luthey-Schulten, Z.; Marrink, S. J. The Onset of Whole-Cell Modeling Using the Martini Force Field. *Biophysical journal* **2023**, *122*, 421a.

(6) Morris, C. J.; Corte, D. D. Using Molecular Docking and Molecular Dynamics to Investigate Protein-Ligand Interactions. *Modern Physics Letters B* **2021**, *35*, 2130002.

(7) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M. The Reaxff Reactive Force-Field: Development, Applications and Future Directions. *npj Computational Materials* **2016**, *2*, 1-14.

(8) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular Mechanics Methods for Predicting Protein–Ligand Binding. *Physical Chemistry Chemical Physics* **2006**, *8*, 5166-5177.
(9) Betzler, B. R.; Powers, J. J.; Worrall, A. Molten Salt Reactor Neutronics and Fuel Cycle Modeling and Simulation with Scale. *Annals of Nuclear Energy* **2017**, *101*, 489-503.
(10) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Advances in protein chemistry* **2003**, *66*, 27-85.
(11) Lin, X.; Wang, W.; He, B.; Li, J.; Wang, Q.; Ma, J. Transport Parameters for Combustion Species Based on Camoeba Polarizable Force Field. *Journal of Chemical Theory and Computation* **2023**.
(12) Ghahremanpour, M. M.; van Maaren, P. J.; Caleman, C.; Hutchison, G. R.; Van der Spoel, D. Polarizable Drude Model with S-Type Gaussian or Slater Charge Density for General Molecular Mechanics Force Fields. *Journal of chemical theory and computation* **2018**, *14*, 5553-5566.
(13) Andersson, D.; Beeler, B. W. Ab Initio Molecular Dynamics (Aimd) Simulations of Nacl, Ucl3 and Nacl-Ucl3 Molten Salts. *Journal of Nuclear Materials* **2022**, *568*, 153836.
(14) Dushanan, R.; Weerasinghe, S.; Dissanayake, D. P.; Senthilinithy, R. Implication of Ab Initio, Qm/Mm, and Molecular Dynamics Calculations on the Prediction of the Therapeutic Potential of Some Selected Hdac Inhibitors. *Molecular Simulation* **2022**, *48*, 1464-1475.
(15) Hu, D.; Gu, X.; Lyu, L.; Wang, G.; Cui, B. Unraveling Oxidative Aging Behavior of Asphaltenes Using Ab Initio Molecular Dynamics and Static Density Functional Theory. *Construction and Building Materials* **2022**, *318*, 126032.

(16) Molani, F.; Webb, S.; Cho, A. E. Combining Qm/Mm Calculations with Classical Mining Minima to Predict Protein–Ligand Binding Free Energy. *Journal of Chemical Information and Modeling* **2023**.
(17) Nwanosike, E. M.; Conway, B. R.; Merchant, H. A.; Hasan, S. S. Potential Applications and Performance of Machine Learning Techniques and Algorithms in Clinical Practice: A Systematic Review. *International Journal of Medical Informatics* **2022**, *159*, 104679.
(18) Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. *International Journal of Quantum Chemistry* **2015**, *115*, 1074-1083.
(19) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical review letters* **2007**, *98*, 146401.
(20) Wang, H.; Zhang, L.; Han, J.; Weinan, E. Deepmd-Kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Computer Physics Communications* **2018**, *228*, 178-184.
(21) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The Ani-1ccx and Ani-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Scientific data* **2020**, *7*, 134.
(22) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the Ani Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation* **2020**, *16*, 4192-4202.

(23) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *The Journal of chemical physics* **2018**, *148*, 241733.
(24) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Advances in neural information processing systems* **2017**, *30*.
(25) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet–a Deep Learning Architecture for Molecules and Materials. *The Journal of Chemical Physics* **2018**, *148*, 241722.
(26) Esteves, C. Theoretical Aspects of Group Equivariant Neural Networks. *arXiv preprint arXiv:2004.05154* **2020**.
(27) Jeevanjee, N. *An Introduction to Tensors and Group Theory for Physicists*; Springer, 2011.
(28) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nature communications* **2022**, *13*, 2453.
(29) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csányi, G. Mace: Higher Order Equivariant Message-passing Neural Networks for Fast and Accurate Force Fields. *Advances in Neural Information Processing Systems* **2022**, *35*, 11423-11436.
(30) Haghighatlari, M.; Li, J.; Guan, X.; Zhang, O.; Das, A.; Stein, C. J.; Heidar-Zadeh, F.; Liu, M.; Head-Gordon, M.; Bertels, L. Newtonnet: A Newtonian Message-passing Network for Deep Learning of Interatomic Potentials and Forces. *Digital Discovery* **2022**, *1*, 333-343.

(31) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A. Machine-Learned Potentials for Next-Generation Matter Simulations. *Nature Materials* **2021**, *20*, 750-761.
(32) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific data* **2014**, *1*, 1-7.
(33) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Science advances* **2017**, *3*, e1603015.
(34) Porter, T.; Vaka, M. M.; Steenblik, P.; Della Corte, D. Computational Methods to Simulate Molten Salt Thermophysical Properties. *Communications Chemistry* **2022**, *5*, 69.
(35) Landsberg, J. M. Tensors: Geometry and Applications. *Representation theory* **2012**, *381*, 3.
(36) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor Field Networks: Rotation-and Translation-Equivariant Neural Networks for 3d Point Clouds. *arXiv preprint arXiv:1802.08219* **2018**.
(37) Fuchs, F.; Worrall, D.; Fischer, V.; Welling, M. Se (3)-Transformers: 3d Roto-Translation Equivariant Attention Networks. *Advances in Neural Information Processing Systems* **2020**, *33*, 1970-1981.
(38) Ziegler, J. F.; Biersack, J. P. *The Stopping and Range of Ions in Matter*; Springer, 1985.
(39) A.~Kramida; and, J. R. a. 2022.
(40) Savarese, P.; Maire, M. Learning Implicitly Recurrent Cnns through Parameter Sharing. *arXiv preprint arXiv:1902.09701* **2019**.
(41) Lee, J. M.; Lee, J. M. *Smooth Manifolds*; Springer, 2012.

(42) Milesi, A. Accelerating Se(3)-Transformers Training Using an Nvidia Open-Source Model Implementation. NVIDIA Developer, 2021.
(43) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z. Drugbank 5.0: A Major Update to the Drugbank Database for 2018. *Nucleic acids research* **2018**, *46*, D1074-D1082.
(44) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database Gdb-13. *Journal of the American Chemical Society* **2009**, *131*, 8732-8733.
(45) Rezác, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *Journal of Chemical Theory and Computation* **2011**, *7*, 3466-3470.
(46) Lange, K. Convergence of Em Image Reconstruction Algorithms with Gibbs Smoothing. *IEEE transactions on medical imaging* **1990**, *9*, 439-446.
(47) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *Journal of cheminformatics* **2012**, *4*, 1-17.
(48) Horton, J. T.; Boothroyd, S.; Wagner, J.; Mitchell, J. A.; Gokey, T.; Dotson, D. L.; Behara, P. K.; Ramaswamy, V. K.; Mackey, M.; Chodera, J. D. Open Force Field Bespokefit: Automating Bespoke Torsion Parametrization at Scale. *Journal of Chemical Information and Modeling* **2022**, *62*, 5622-5633.
(49) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D. Openmm 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS computational biology* **2017**, *13*, e1005659.
(50) Case, D.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham III, T.; Darden, T.; Duke, R.; Gohlke, H. Amber 14; University of California: San

Francisco, 2014. *Google Scholar There is no corresponding record for this reference* **2014**, 1-826.
(51) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The journal of physical chemistry letters* **2014**, *5*, 1885-1891.
(52) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature methods* **2020**, *17*, 261-272.
(53) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. Mdanalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of computational chemistry* **2011**, *32*, 2319-2327.
(54) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M. Mdanalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th python in science conference*, 2016; SciPy Austin, TX: Vol. 98, p 105.
(55) Smith, J. S.; Isayev, O.; Roitberg, A. E. Ani-1: An Extensible Neural Network Potential with Dft Accuracy at Force Field Computational Cost. *Chemical science* **2017**, *8*, 3192-3203.
(56) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *The Journal of chemical physics* **2018**, *148*.
(57) Chigaev, M.; Smith, J. S.; Anaya, S.; Nebgen, B.; Bettencourt, M.; Barros, K.; Lubbers, N. Lightweight and Effective Tensor Sensitivity for Atomistic Neural Networks. *The Journal of Chemical Physics* **2023**, *158*.

(58) Zaverkin, V.; Holzmüller, D.; Bonfirraro, L.; Kästner, J. Transfer Learning for Chemically Accurate Interatomic Neural Network Potentials. *Physical Chemistry Chemical Physics* **2023**, *25*, 5383-5396.
(59) Mu, Y.; Kosov, D. S.; Stock, G. Conformational Dynamics of Trialanine in Water. 2. Comparison of Amber, Charmm, Gromos, and Opls Force Fields to Nmr and Infrared Experiments. *The Journal of Physical Chemistry B* **2003**, *107*, 5064-5073.