Recent Advances in Machine Learning for Charge Density Prediction

A Review of State-of-the-Art Approaches

Review

May 27, 2025

Outline

What is Charge Density?

- Charge density $\rho(\vec{r})$ is a fundamental quantity in quantum mechanics
- Represents the probability distribution of electrons in a system
- Key properties:
  - Non-negative: $\rho(\vec{r}) \geq 0$
  - Normalized: $\int \rho(\vec{r}) d\vec{r} = N$ (number of electrons)
  - E(3) equivariant: $\rho(R\vec{r} + \vec{t}) = \rho(\vec{r})$ for any rotation R and translation $\vec{t}$

Why is Charge Density Important?

- Foundation of Density Functional Theory (DFT)
- Determines many physical properties:
  - Total energy
  - Forces on atoms
  - Electronic structure
  - Chemical bonding
- Computational bottleneck in materials science
- Key for materials discovery and design

Challenges in Charge Density Prediction

- High-dimensional output space
- Need for physical constraints:
  - E(3) equivariance
  - Electron conservation
  - Non-negativity
- Computational efficiency
- Accuracy requirements for downstream tasks

E(3) Equivariance

- E(3) = Euclidean group in 3D
- Includes:
  - Rotations (SO(3))
  - Translations
  - Reflections
- For charge density:

$$\rho(R\vec{r} + \vec{t}) = \rho(\vec{r})$$

where $R \in \mathrm{SO}(3)$ and $\vec{t} \in \mathbb{R}^3$

Higher-Order Tensor Representations

- Irreducible representations (irreps) of SO(3)
- Tensor features $V_{cm}^{(\ell,p)}$ where:
  - ℓ: rotation order ($\ell \in \{0, 1, 2, ...\}$)
  - p: parity ($p \in \{-1, 1\}$)
  - c: channel index
  - m: index in $[-\ell, \ell]$
- Size at each order: $\mathbb{R}^{N_{\text{channels}} \times (2\ell+1)}$

Tensor Product Operations

- Combines representations using Clebsch-Gordan coefficients

$$(U^{(\ell_1,p_1)} \otimes V^{(\ell_2,p_2)})^{(\ell_o,p_o)}_{cm_o} = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} C^{(\ell_o,m_o)}_{(\ell_1,m_1)(\ell_2,m_2)} U^{(\ell_1,p_1)}_{cm_1} V^{(\ell_2,p_2)}_{cm_2}$$

where:

- $|\ell_1 - \ell_2| \leq \ell_o \leq |\ell_1 + \ell_2|$
- $p_o = p_1 p_2$

Impact of Higher-Order Features

- Higher ℓ values capture more complex angular dependencies
- Performance improvement:
  - 44.6% median improvement for materials with non-metals/metalloids
  - 23.0% median improvement for materials with only metals
- Particularly important for:
  - Covalent bonding
  - High angular variance systems
  - Complex electronic structures

ChargE3Net Architecture Overview

- E(3)-equivariant graph neural network
- Key components:
  - Graph construction with atoms and probe points
  - Higher-order equivariant features (up to L=4)
  - Message passing between atoms and probes
  - Equivariant convolution operations
- Input: Atomic species and positions
- Output: Charge density at probe points

Graph Construction

- Vertices:
  - Atoms: One-hot encoding of atomic number
  - Probe points: Initialized as zero scalar
- Edges:
  - Atom-atom: Unidirectional, cutoff 4Å
  - Atom-probe: Directed from atoms to probes
- Periodic boundary conditions supported

Message Passing Architecture

- Two types of convolutions:
  - $\text{Conv}_{\text{atom}}$: Bidirectional between atoms
  - $\text{Conv}_{\text{probe}}$: From atoms to probes
- Each layer:
  - Updates atom representations
  - Updates probe representations
  - Uses tensor product operations
- Final layer: Regression to predict charge density

Performance on Benchmark Datasets

- Materials Project (MP):
  - ChargE3Net: 0.523% mae
  - DeepDFT: 0.799% mae
  - invDeepDFT: 0.859% mae
- QM9:
  - ChargE3Net: 0.206% mae
  - OrbNet-Equi: 0.284% mae
  - DeepDFT: 0.357% mae
- NMC:
  - ChargE3Net: 0.061% mae
  - DeepDFT: 0.060% mae

Impact on DFT Calculations

- SCF step reduction:
  - MP materials: 26.7% reduction
  - GNoME materials: 28.6% reduction
- Non-self-consistent property prediction:
  - 40% of materials: energy errors < 1 meV/atom
  - 70% of materials: forces < 0.03 eV/Å
  - 76% of materials: band gaps within chemical accuracy
- Linear scaling O(N) with system size

Higher-Order Features Analysis

- Performance vs. rotation order:
  - L=0: Basic scalar features
  - L=1: Vector features
  - L=2,3,4: Higher-order tensor features
- Channel distribution:
  - $N_{channels} = \lfloor 500/(L+1) \rfloor$
  - Equal representation size across orders
- Consistent improvement with increasing L

Angular Variance Analysis

- High angular variance materials:
  - Example: Cs(H2PO4)
  - Strong covalent bonding
  - Significant L=4 improvement
- Low angular variance materials:
  - Example: Rb2Sn6
  - Primarily ionic interactions
  - Similar L=0 and L=4 performance
- Metric ζ for angular variance:

$$\zeta(G) = 1 - \frac{\sum_{\vec{g}_k \in G} |\nabla \rho(\vec{g}_k) \cdot \hat{r}_{ki}|}{\sum_{\vec{g}_k \in G} ||\nabla \rho(\vec{g}_k)||}$$

SCDP Architecture Overview

- Orbital-based approach
- Key components:
  - Spherical-harmonics-based atomic orbitals
  - Learnable basis sets
  - Efficient evaluation of spherical fields
- Input: Atomic species and positions
- Output: Charge density through orbital coefficients

Orbital Representation

- Charge density as sum of spherical fields:

$$\rho(\vec{r}) = \sum_i \sum_{nlm} c^i_{nlm} \phi_{nlm}(\vec{r} - \vec{r}_i)$$

where:

  - ϕ_{nlm}: Atomic orbital basis functions
  - c^i_{nlm}: Learnable coefficients
  - $\vec{r}_i$: Atomic positions

- Basis functions:
  - Radial part: Gaussian-type orbitals
  - Angular part: Spherical harmonics

Learnable Basis Sets

- Reference basis: def2-QZVPPD
- Learnable parameters:
  - Orbital exponents (α)
  - Contraction coefficients
  - Radial scaling factors
- Optimization:
  - End-to-end training
  - Physical constraints preserved
  - Adaptive to different elements

Efficiency Optimizations

- Spherical channel representation:
  - Efficient evaluation of spherical harmonics
  - Pre-computed Clebsch-Gordan coefficients
  - Optimized tensor operations
- Computational improvements:
  - >10x faster than existing methods
  - Linear scaling with system size
  - GPU-optimized implementation

Performance Analysis

- Accuracy:
  - Competitive with grid-based methods
  - Better for systems with strong atomic character
  - Flexible accuracy-efficiency trade-off
- Efficiency:
  - Faster inference than grid-based methods
  - Lower memory requirements
  - Scalable to large systems

Downstream Applications

- Property prediction:
  - Total energy
  - Forces
  - Electronic structure
- Materials discovery:
  - High-throughput screening
  - Property optimization
  - Structure prediction
- Molecular dynamics:
  - Force field generation
  - Trajectory simulation
  - Property evolution

Uni-3DAR Overview

- Tokenization-based approach
- Key components:
  - Hierarchical octree compression
  - Fine-grained structural tokenization
  - Masked next-token prediction
- Unifies:
  - 3D structure generation
  - Property prediction
  - Multi-modal tasks

Hierarchical Tokenization

- Octree-based compression:
  - Coarse-to-fine subdivision
  - Non-empty cell detection
  - Level-wise tokenization
- Fine-grained tokenization:
  - Atom types and positions
  - In-cell coordinate discretization
  - Structural details
- 2-level subtree compression:
  - 8 subcells $\rightarrow$ 1 token
  - 256 possible states
  - 8x reduction in tokens

Masked Next-Token Prediction

- Challenge: Dynamic token positions
- Solution:
  - Token duplication
  - Masked token replacement
  - Position-aware prediction
- Benefits:
  - Handles varying token positions
  - Maintains causal sampling
  - Improves prediction accuracy

Unified Framework

- Single-frame generation:
  - Unconditional generation
  - Property-conditioned generation
  - Text-guided generation
- Multi-frame generation:
  - Molecular dynamics
  - Pocket-based generation
  - Frame-by-frame prediction
- Understanding tasks:
  - Token-level properties
  - Structure-level properties
  - Cross-modal tasks

Performance Analysis

- Generation tasks:
  - Up to 256% relative improvement
  - 21.8x faster inference
  - Better quality and diversity
- Understanding tasks:
  - Competitive with specialized models
  - Effective transfer learning
  - Multi-task learning benefits

Efficiency Optimizations

- Training:
  - FlashAttention with bfloat16
  - Sequence packing
  - Efficient memory usage
- Inference:
  - KV-cache acceleration
  - Paired token generation
  - GPU utilization optimization

Cross-Modal Applications

- Protein folding:
  - Sequence to structure
  - Multi-frame generation
  - Property prediction
- Crystal structure prediction:
  - PXRD-guided generation
  - NMR signal conditioning
  - Property optimization

Approach Comparison

- ChargE3Net:
  - E(3)-equivariant GNN
  - Higher-order tensor features
  - Grid-based representation
- SCDP:
  - Orbital-based approach
  - Learnable basis sets
  - Spherical channel representation
- Uni-3DAR:
  - Tokenization-based
  - Octree compression
  - Unified framework

Performance Comparison

- Accuracy:
  - ChargE3Net: Best on MP and QM9
  - SCDP: Competitive with faster inference
  - Uni-3DAR: Best for generation tasks
- Efficiency:
  - ChargE3Net: O(N) scaling
  - SCDP: >10x faster than grid-based
  - Uni-3DAR: 21.8x faster inference
- Memory usage:
  - ChargE3Net: Moderate
  - SCDP: Low
  - Uni-3DAR: High (transformer-based)

Equivariance Handling

- ChargE3Net:
  - Explicit E(3) equivariance
  - Higher-order tensor representations
  - Clebsch-Gordan coefficients
- SCDP:
  - Implicit through basis functions
  - Spherical harmonics
  - Physical constraints
- Uni-3DAR:
  - Data augmentation
  - Tokenization structure
  - Position-aware prediction

DFT Acceleration

- SCF step reduction:
  - ChargE3Net: 26.7-28.6%
  - SCDP: Efficient initialization
  - Uni-3DAR: Multi-frame prediction
- Property prediction:
  - Energy and forces
  - Electronic structure
  - Chemical properties

Materials Discovery

- High-throughput screening:
  - Property prediction
  - Structure optimization
  - Composition design
- Property optimization:
  - Inverse design
  - Multi-objective optimization
  - Constraint satisfaction

Molecular Dynamics

- Force field generation:
  - Energy and force prediction
  - Trajectory simulation
  - Property evolution
- Long-time dynamics:
  - Rare event sampling
  - Phase transitions
  - Chemical reactions

Key Takeaways

- Higher-order features:
  - Significant performance improvement
  - Better for complex systems
  - Important for covalent bonding
- Tokenization approach:
  - Efficient compression
  - Unified framework
  - Cross-modal capabilities

Future Directions

- Model improvements:
  - Higher-order features beyond L=4
  - Better equivariance handling
  - More efficient architectures
- Applications:
  - Larger systems
  - More complex properties
  - Real-time prediction
- Integration:
  - Multi-modal learning
  - Transfer learning
  - Active learning