School of Information Technology and Engineering at the ADA University

School of Engineering and Applied Science at the George Washington University

REGRESSION ON INTERATOMIC DESCRIPTOR DATA: DIRECT SOLUTION STRATEGIES FOR LINEAR REGRESSION IN CPU AND MEMORY-CONSTRAINED ENVIRONMENTS

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

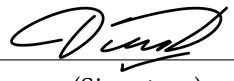
In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Dursun Dashdamirov

April, 2024

ACADEMIC INTEGRITY STATEMENT

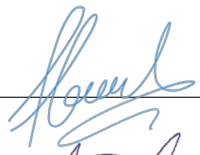
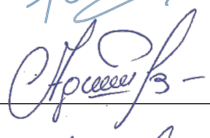
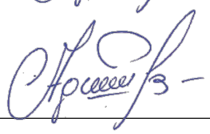
"I affirm that this is my own work, I attributed where I used the work of others, I did not facilitate academic dishonesty for myself or others, and I used only authorized resources for my Thesis, per the ADA University Academic Integrity requirements. If I failed to comply with this statement, I understand consequences will follow my actions. Consequences may range from failing the course to expulsion from the program/university and may include a transcript notation."

Dursun Dashdamirov		28.04.2025
(Full Name)	(Signature)	(Date: DD.MM.YY)

This Thesis by: Dursun Dashdamirov
Entitled: *Regression on Interatomic Descriptor Data: Direct Solution Strategies for Linear Regression in CPU and Memory-Constrained Environments*

has been approved as meeting the requirements for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

Dr. Jamaladdin Hasanov		08.08.2025
(Advisor)		(Date)
Dr. Abzatdin Adamov		08.08.2025
(Program Director)		(Date)
Dr. Abzatdin Adamov		08.08.2025
(Dean)		(Date)

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to my supervisor, Prof. Jamaladdin Hasanov, for providing invaluable guidance and feedback.

I am also extremely thankful to my co-supervisor, Ilgar Baghishov, who guided me from both the domain and technical perspectives with weekly meetings throughout the last seven months. His knowledge and experience were instrumental in the successful completion of this thesis.

In addition, I wish to acknowledge the Henkelman Research Group from the University of Texas at Austin for providing access to their data and High-Performance Computing System, which enabled the execution of scientifically valuable experiments.

List of Figures

1 Neural network structure for a three-atom system. Cartesian coordinates $\mathbf{R}_i^\alpha$ are transformed into symmetry function values G_i^μ, capturing the local environment of atom i based on all atomic positions. These values feed into subnet S_i, yielding atom i's energy contribution E_i to the system's total energy. [1] . . xiii

2 The outline of the followed methodology. All subsections of the proposed methodology have been completed. The testing of the developed solution on existing benchmark datasets is recommended for future work. xx

3 When using linear models or the Random Forest algorithm, each input feature to the model is a sum of the descriptor value for all atoms in a cell. This can create a correlation between descriptors since they carry information that is too general and possibly redundant. Switching from this atomic cell to the next one can yield a similar response in all descriptors. xxii

4 In this configuration, a descriptor for each atom is part of the learning process. Instead of producing one row for each cell, the result is N rows for each cell creating a dataset of Nx20000 rows. xxii

5 Removing the columns with the highest VIF value xxiii

6 The first row and second row represent the dependency plots before and after the removal of multicollinearity, respectively. The plots are colored by the effect of the feature that they are affected by the most. For instance, feature 15 affects the behavior of feature 2 in the second row. After the removal, clearer relationships can be seen. xxv

7 Energy and force RMSE values plotted against training time for different fractions of training data. xxviii

8 The applied feature reduction scheme is visualized. The first split of the data is based on the Pearson Correlation coefficient estimate between features and target labels. As a result, the number of columns was halved. xxix

9 Error and computation time of feature reduction techniques. xxx

10 Computational graph of QR solution with 2 nodes and 2 cores on each node. . . xxxi

11 The scaling performance of SVD. The x-axis is the relative speed-up compared to 1 core. Three cases with different chunk sizes have been tested. The dotted line illustrates ideal scaling. xxxiii
12 Comparison of standard SVD versus Dask implementation. Dotted lines indicate the ideal scaling, while bar plots show the actual performance. xxxiv
13 Energy RMSE is plotted against training time. Green and blue colored markers indicate methods including data reduction and parallelization, respectively. . . xxxvi
14 Force RMSE is plotted against training time. Green and blue colored markers indicate methods including data reduction and parallelization, respectively. . . xxxvii
15 Code Snippet demonstrating QR factorization with Dask. xxxviii
16 Radar chart illustrating the mean absolute error across different physical property evaluations as we vary the α parameter. xl
17 Correlation matrix for the used dataset, showing high inter-feature correlations. xli

List of Tables

1 Structure of the dataset xxi
2 Testing RMSE Errors Before and After Reduction xxx
3 Comparison of methods xxxv
4 TSQR Execution Times for Different Configurations xxxviii

ABBREVIATIONS

Abbreviation	Explanation
ACE	Atomic Cluster Expansion
ASE	Atomic Simulation Environment
DFT	Density Functional Theory
FitSNAP	Fitting framework based on SNAP
GA	Genetic Algorithm
GBDT	Gradient Boosted Decision Trees
HDNNP	High-Dimensional Neural Network Potentials
HPC	High-Performance Computing
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
MAE	Mean Absolute Error
MD	Molecular Dynamics
ML	Machine Learning
MLIP	Machine Learning Interatomic Potential
MPNN	Message Passing Neural Network
MSE	Mean Squared Error
PCA	Principal Component Analysis
PES	Potential Energy Surface
RAM	Random Access Memory
RMSE	Root Mean Square Error
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SLURM	Simple Linux Utility for Resource Management
SNAP	Spectral Neighbor Analysis Potential
SOAP	Smooth Overlap of Atomic Positions
SSWRP	Stability-Weighted Square Sequence Product Aggregate Strategy
SVD	Singular Value Decomposition
VIF	Variance Inflation Factor

ABSTRACT

Molecular-dynamics (MD) simulations convert Newton's laws into atom-by-atom trajectories, from which pressure, temperature, and free energy are extracted for various experiments. While simulating any atomic environment, the most crucial feature of the atomic system is the energy. Also, it is the most challenging feature to estimate. The Machine Learning Interatomic Potentials (MLIP) solve this by fitting models to the system's definitive functions, which are utilized to make fast inferences regarding the energy of a system. The accuracy of these models depends on the domain-specific hyperparameter optimization, which is quite slow due to the use of complex deep neural networks. With this work, a new interatomic descriptor called quadratic ACE (qACE) is proposed, which surpasses the neural network's accuracy. Then, we explore and benchmark possible ways to fit a linear regression model to this computationally demanding solution in CPU and memory-bound environments. To solve the regression problem, several strategies are explored, including data reduction techniques and parallel processing. By leveraging Dask's comprehensive task scheduling infrastructure, we compute the direct least-squares regression on a 27 GB dataset in under five minutes, demonstrating both scalability and computational efficiency.

Overall, this work demonstrates that efficient feature engineering, combined with lightweight parallel regression strategies, can substitute for deep models without sacrificing accuracy or scalability.