Available online at www.eccomasproceedia.org
Eccomas Proceedia COMPDYN (2023) 1249-1254

ECCOMAS
Proceedia

COMPDYN 2023
9th ECCOMAS Thematic Conference on
Computational Methods in Structural Dynamics and Earthquake Engineering
M. Papadrakakis, M. Fragiadakis (eds.)
Athens, Greece, 12-14 June 2023

DERIVING COARSE-GRAINED MODELS OF MOLECULAR SYSTEMS BY APPROXIMATING THE FREE ENERGY SURFACE WITH MACHINE LEARNING ALGORITHMS

Nikolaos P. Bakas[1], Antonis Chazirakis[2], Eleftherios Christofi[1], and Vangelis Harmandaris[1]

[1]Computation-based Science and Technology Research Center, The Cyprus Institute
20 Konstantinou Kavafi Street, 2121, Aglantzia Nicosia, Cyprus.
e-mail: {n.bakas,e.christofi,v.harmandaris}@cyi.ac.cy

[2] Department of Applied Mathematics, University of Crete
Heraklion GR-71409, Crete, Greece

Abstract. *Atomistic molecular simulations are capable to predict the properties of systems, and materials, starting from their atomic microstructure. To enhance the range of spatiotemporal scales accessible by simulations coarse-grained (CG) models, that reduce the dimensionality of the underlying system, are used. In such models CG particles, representing group of atoms, are introduced. The derivation of accurate effective interactions between the CG particles (CG force field) is a grand challenge in order to study realistic complex systems associatted with important technological applications. The problem of coarse-grained simplification of the representation of the underlying atomistic microstructure (or molecular potential) is a hard task in the generic case. Significant information, such as the orientation of atoms, might be lost when using the Cartesian coordinates of the participating molecules to develop a model predicting the potential. Furthermore, the actual relationships among the input coordinates and the target potential stem from vastly complex interactions among the neighbouring molecules with the central one, as well as the corresponding atoms. Accordingly, even if we use vastly big datasets and supercomputing for training the models, it is not known whether an approximation algorithm yielding low accuracy is a weak learner or important features have been committed during the coarse-grained representation. In this work, we provide a detailed investigation of several algorithms that are used to approximate accurately the effective CG interaction (free energy surface) and present empirical evidence for their efficiency applied in a pool of databases for various materials. Particularly, we investigated tree-based models (gradient boosting and random forests), artificial neural networks' architectures, and higher-order polynomial regression with heuristic feature selection. The CG force field is automatically obtained from the differentiation of the ML model trained for Energy without training for the Forces as a target variable. Finally, we suggest a generic framework for approximating such systems.*

Keywords: coarse-grained, molecular dynamics, feature selection, polynomial approximation

ISSN:2623-3347 © 2023 The Authors. Published by Eccomas Proceedia.
Peer-review under responsibility of the organizing committee of COMPDYN 2023.
doi: 10.7712/120123.10473.21195

1 Introduction

In this paper we present how Machine Learning (ML) Algorithms may approximate the Energy Distribution for the Methane CH_4 case. Particularly, we use only a few molecules samples in the train set, and we accurately predict 132.800 in the test set. Interestingly, by computing the derivatives of the trained model, we directly compute the Forces' distribution without training on the Forces, using only the Energies' model. Our model involves the utilization of inverted distances $\frac{1}{r_1}$ as features for the Energy and later for Forces.

In Figure 1 an indicative three-dimensional representation of the atoms is depicted. The atoms are coloured with respect to their distance from the reference atom in the center. Furthermore, in Figure 2, we present the histograms of the first feature $\frac{1}{r_1}$ as well as Energy. We may see that the feature's energy is close to a Gaussian shape, while Energy's is skewed. In Table 1, we present the corresponding statistical properties for the five first features.

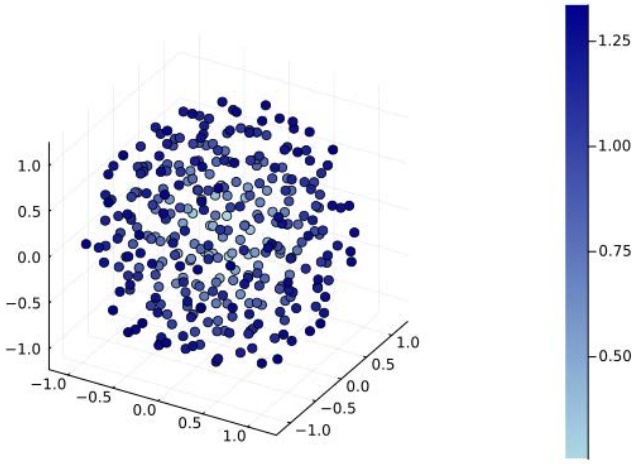
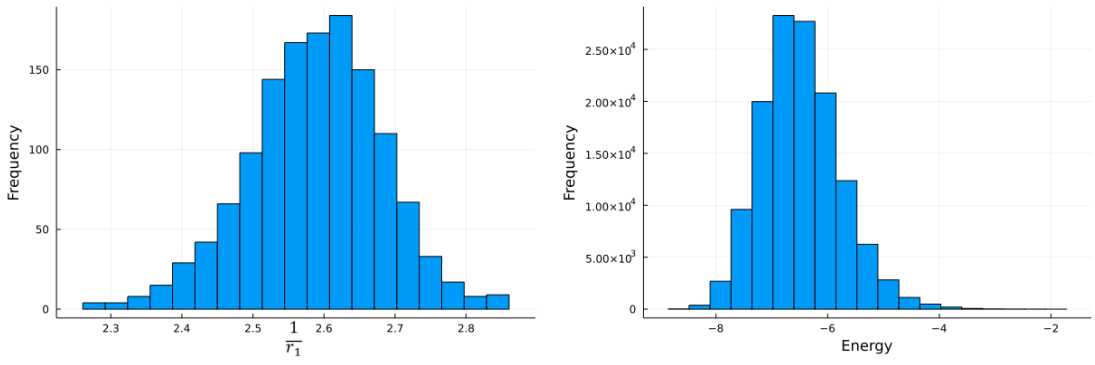


Figure 1: Colored distances of neighbours to the central molecule for CH_4



(a) (b)

Figure 2: Histogramms of first feature $\frac{1}{r_1}$ and Energy

Table 1: Statistical Measurements for 5 first features

	$\frac{1}{r_1}$	$\frac{1}{r_2}$	$\frac{1}{r_3}$	$\frac{1}{r_4}$	$\frac{1}{r_5}$
Average	2.58623296	2.49124579	2.41860704	2.35226661	2.28641703
Maximum	2.95144524	2.87002243	2.75256192	2.69820288	2.63408836
95% Quantile	2.73942366	2.63204125	2.56163643	2.50067215	2.44320657
75% Quantile	2.65102813	2.55409326	2.4828045	2.42021519	2.35841022
Median	2.58851559	2.49547326	2.42361757	2.35826757	2.29316787
25% Quantile	2.52395721	2.43315146	2.35973786	2.29094022	2.22179436
5% Quantile	2.42590571	2.33598572	2.25885619	2.18273488	2.106839
Minimum	2.086301	2.0042334	1.86254944	1.82117039	1.74705074

2 Mathematical Formulation

Let $[i] = \{1, 2, \ldots, m\}$ the iterator for m molecules and $[j] = \{1, 2, \ldots, n\}$ the corresponding for neighbors n, and $r_{i\in[i]j\in[j]}$ the distances of each neighbor j from the cetral atom of each sample i.

We define the input matrix

$$\mathbf{rrr} = \mathbf{rr}_{i\in[i]} = rr_{i\in[i]j\in[j]} = \frac{1}{r_{i\in[i]j\in[j]}}, \tag{1}$$

and assume that a function f exists, mapping the distances r_{ij} to the Energy E_i, that is

$$f(\mathbf{rr}_i) = E_i + e_i. \tag{2}$$

The process is deterministic; hence the errors of an adequate model should be of $\mathcal{O}(eps)$, with eps the round of accuracy of the machine, thus

$$f(\mathbf{rr}_i) \approx E_i = \mathbf{E}. \tag{3}$$

2.1 Conjecture

We assume that f is a polynomial of arbitrary order n_p, and the polynomial exponents are real numbers, represented by a vector $[p] \in \mathcal{R}^{n_p}$, and all distances of neighbours $[j]$, are raised to the same order $p \in [p]$ and hence, participate with the same weight in f. We do not use interactions among features, as the complexity of the model increases vastly fast, and in some preliminary results we tried exhibited lower accuracy. We keep $n = 75$ neighbours of the central atom for the model.

2.2 Representation of f

Let p_1 be the maximum polynomial order, p_{n_p} the minimum, and n_p the number of polynomial terms. Hence, we define $[p] = \{p_1, p_2 \ldots, p_{n_p}\}$, a particular combination of the polynomial orders, within the bounds p_1, p_{n_p} and length n_p. Accordingly, we may write

$$f(\mathbf{rr}_i) = \sum_{p\in[p]} w_p \times \sum_{j\in[j]} rr_{ij}^p \tag{4}$$

Hence, for a given $[p]$ the nonlinear problem can be written as a linear system, in the form of

$$\mathbf{X} \times \mathbf{w} = \mathbf{E}, \tag{5}$$

where $\mathbf{w} = w_p$, and

$$\mathbf{X} = X_{ip} = \sum_{j \in [j]} rr_{ij}^{p} \tag{6}$$

However, we want to select the best subset $[\hat{o}] \subset [p]$ exhibiting the best possible accuracy. All the potential combinations of the polynomial orders are given by the binomial coefficient (n_p choose k), by

$$\binom{n_p}{k} = \frac{n_p!}{k!(n_p - k)!}, \tag{7}$$

resulting in an exponentially increasing size for high orders and a number of polynomial terms. For example, if we select from $n_p = 100$, the best combination with $k = 10$ terms, we have to solve $> 17 \times 10^{12}$ linear systems and select the best. Hence, instead of an exhaustive search, we use the Algorithm 1 [1–3].

2.2.1 Algorithmic Solution

For low-dimensional problems, we may use an exhaustive search, however, because of the exponential character of the problem, we solve it with the following Algorithm 1. We have to optimize simultaneously:

1. the weights w_p,
2. the set of polynomial orders $[p]$,

For a particular set $[p]$, we can easily optimise the weights w_p by using least square regression. We do this in each step of the following Algorithm 1, for the simultaneous selection of the number of polynomial terms n_p, as well as the set of polynomial orders $[p]$. We use the mean absolute error $\overline{|e|}$ as a convergence criterion, and tol as the desired threshold for accuracy.

Algorithm 1: Polynomial Feature selection Algorithm

Data: $\mathbf{rrr}, \mathbf{E}$
Result: Initialize $[o]$ with a random subset of $[p]$
Solve Linear System $\mathbf{X} \times \mathbf{w} = \mathbf{y}$, where $\mathbf{X}' \subset \mathbf{X}$, with $[o]$ columns.
Compute regression errors $\mathbf{e} = \mathbf{X}' \times \mathbf{w} - \mathbf{E}$.
Set as optimal error $\hat{e} \leftarrow \overline{|\mathbf{e}|}$.
Set as optimal indices $[\hat{o}] \leftarrow [o]$.
while $\hat{e} > tol$ **do**
  select an index $d \in [o]$ randomly.
  substitute o_d with another index $o_{d'} \in [p], o_{d'} \notin [o]$.
  Solve Linear System $\mathbf{X}' \times \mathbf{w} = \mathbf{y}$.
  Compute regression error $\overline{|\mathbf{e}|}$.
  if $\hat{e} > \overline{|\mathbf{e}|}$ **then**
    $\hat{e} \leftarrow \overline{|\mathbf{e}|}$
    $[\hat{o}] \leftarrow [o]$
  else
    $[o] \leftarrow [\hat{o}]$
  end
end

2.2.2 Derivation of Forces

We may write the Energy E_i by

$$E_i = f(\mathbf{x}_i) = \sum_{j\in[j]}\sum_{p\in[p]} w_p \times r r_{ij}^p = \sum_{j\in[j]}\sum_{p\in[p]} w_p \times \hat{E}_{ijp}, \tag{8}$$

for each polynomial order $p \in [p]$.
We know that

$$r_{ij} = \sqrt{(x_{ij} + y_{ij} + z_{ij})}, \tag{9}$$

thus

$$\hat{E}_{ijp} = r_{ij}^{-p}. \tag{10}$$

Hence, for each molecule i and neighbor j, the part of the energy corresponding to p order is

$$\hat{E}_p = r^{-p} \tag{11}$$

Accordingly, for the corresponding force in the x direction $\hat{F}_{xp}$

$$\hat{F}_{xp} = \frac{\partial \hat{E}_p}{\partial x} = \frac{\partial r^{-p}}{\partial x} = -p \times r^{-p-1} \times \frac{1}{2} \times (x^2 + y^2 + z^2)^{-1/2} \times 2x, \tag{12}$$

and hence

$$\hat{F}_{xp} = -p \times r^{-p-2} \times x. \tag{13}$$

Similarly, for the other 2 dimensions, y, z.

In the following Figure 3, we demonstrate the predicted vs given values of Energy and Forces in x-dimension. We should underline that the train of the model was for Energies only, while the Forces are derived by the above-mentioned Equations.

3 Conclusions

This work presented a generic methodology for approximating atoms' Energies with arbitrary order polynomials. We do this for CH_4 with great accuracy, and the applied methodology is generic. We foresee in the future to try new datasets for other materials, as well as other methods such as artificial neural networks [4].

...

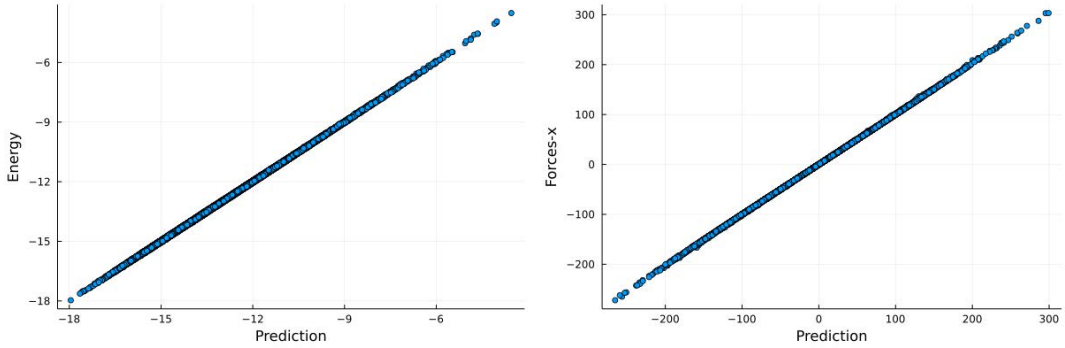


Figure 3: Prediction of Energies E_i in the test set (left) with polynomial model with heuristic features' selection, and automatic derivation of Forces F_{x_i} from the ML model.

REFERENCES

[1] N. P. Bakas, V. Plevris, A. Langousis, and S. A. Chatzichristofis, "Itso: A novel inverse transform sampling-based optimization algorithm for stochastic search," *Stochastic Environmental Research and Risk Assessment*, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00477-021-02025-w

[2] N. Bakas, G. Markou, A. Langousis, S. Lavdas, and S. Chatzichristofis, "nbml: Computer software for data analysis and predictive modelling with artificial intelligence algorithms," 2023. [Online]. Available: https://github.com/nbakas/nbml/blob/main/docs/__nbml__.pdf

[3] V. Plevris, N. P. Bakas, and G. Solorzano, "Pure random orthogonal search (pros): A plain and elegant parameterless algorithm for global optimization," *Applied Sciences*, vol. 11, no. 11, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/11/5053

[4] N. Bakas, A. Langousis, M. Nicolaou, and S. Chatzichristofis, "Gradient free stochastic training of anns, with local approximation in partitions," *Stochastic Environmental Research and Risk Assessment*, pp. 1–15, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00477-023-02407-2