

Genetic and Evolutionary Computation

Leonardo Trujillo

Stephan M. Winkler

Sara Silva

Wolfgang Banzhaf *Editors*

# Genetic Programming Theory and Practice XIX

 Springer

# Genetic and Evolutionary Computation

## Series Editors

Wolfgang Banzhaf , Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

Kalyanmoy Deb , Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

The area of Genetic and Evolutionary Computation has seen an explosion of interest in recent years. Methods based on the variation-selection loop of Darwinian natural evolution have been successfully applied to a whole range of research areas.

The Genetic and Evolutionary Computation Book Series publishes research monographs, edited collections, and graduate-level texts in one of the most exciting areas of Computer Science. As researchers and practitioners alike turn increasingly to search, optimization, and machine-learning methods based on mimicking natural evolution to solve problems across the spectrum of the human endeavor, this growing field will continue to surprise with novel applications and results. Recent award-winning PhD theses, special topics books, workshops and conference proceedings in the areas of EC and Artificial Life Studies are of interest.

Areas of coverage include applications, theoretical foundations, technique extensions and implementation issues of all areas of genetic and evolutionary computation. Topics may include, but are not limited to:

Optimization (multi-objective, multi-level) Design, control, classification, and system identification Data mining and data analytics Pattern recognition and deep learning Evolution in machine learning Evolvable systems of all types Automatic programming and genetic improvement.

Proposals in related fields such as:

Artificial life, artificial chemistries Adaptive behavior and evolutionary robotics Artificial immune systems Agent-based systems Deep neural networks Quantum computing will be considered for publication in this series as long as GEVO techniques are part of or inspiration for the system being described. Manuscripts describing GEVO applications in all areas of engineering, commerce, the sciences, the arts and the humanities are encouraged.

Prospective Authors or Editors:

*If you have an idea for a book, we would welcome the opportunity to review your proposal. Should you wish to discuss any potential project further or receive specific information regarding our book proposal requirements, please contact Wolfgang Banzhaf, Kalyan Deb or Mio Sugino:*

*Areas: Genetic Programming/other Evolutionary Computation Methods, Machine Learning, Artificial Life*

*Wolfgang Banzhaf Consulting Editor* BEACON Center for Evolution in Action Michigan State University, East Lansing, MI 48824 USA [banzhafw@msu.edu](mailto:banzhafw@msu.edu)

*Areas: Genetic Algorithms, Optimization, Meta-Heuristics, Engineering*

*Kalyanmoy Deb Consulting Editor* BEACON Center for Evolution in Action Michigan State University, East Lansing, MI 48824 USA [kdeb@msu.edu](mailto:kdeb@msu.edu)


*Mio Sugino* [mio.sugino@springer.com](mailto:mio.sugino@springer.com)

The GEVO book series is the result of a merger the two former book series: Genetic Algorithms and Evolutionary Computation <https://www.springer.com/series/6008> and Genetic Programming <https://www.springer.com/series/6016>.


Leonardo Trujillo · Stephan M. Winkler ·  
Sara Silva · Wolfgang Banzhaf  
Editors


# Genetic Programming Theory and Practice XIX

*Editors*

Leonardo Trujillo   
Engineering Sciences Graduate Program  
Tecnológico Nacional de México/Instituto  
Tecnológico de Tijuana  
Tijuana, Baja California, Mexico

Sara Silva   
Department of Informatics  
Faculty of Sciences  
University of Lisbon  
Lisbon, Portugal

Stephan M. Winkler   
Heuristic and Evolutionary  
Algorithms Laboratory  
University of Applied Sciences  
Upper Austria  
Wels, Austria

Wolfgang Banzhaf   
Department of Computer Science  
and Engineering  
Michigan State University  
East Lansing, MI, USA

ISSN 1932-0167

ISSN 1932-0175 (electronic)

Genetic and Evolutionary Computation

ISBN 978-981-19-8459-4

ISBN 978-981-19-8460-0 (eBook)

<https://doi.org/10.1007/978-981-19-8460-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

The COVID-19 pandemic has marked our world indefinitely. Its devastating effects on peoples' lives and livelihoods are, unfortunately for many, incalculable. For the academic and scientific world, it meant a sudden halt to in-person meetings, workshops and conferences, among many other consequences. People had to manage a much greater hardship than having to discuss the latest in science and technology by way of a digital screen, but while most stayed positive and made use of the online opportunities, science, learning and teaching were not the same. After canceling the Genetic Programming Theory and Practice (GPTP) workshop in 2020, we decided to organize the 18th edition of GPTP as an online event in 2021, which allowed us to gather our community and become an active group again. It was a success by any measure, but we still hoped to avoid having to repeat that format again.

For the 19th edition of GPTP in 2022, a mix of hopefulness and nervousness were part of the early organizing meetings, as many institutions around the world were gradually returning, at least partially, to in-person activities. New strains of the SARS-COV-2 virus kept appearing and affected the well-being of many. However, as the first few months of 2022 passed, our hopefulness began to change into joy as we began to realize that holding an in-person event was not impossible, and gradually became a realistic scenario.

Our community was engaged, as people started accepting invitations to the workshop, and booking flights and hotel rooms. It was a strange feeling, since such gatherings are the lifeblood of scientific discourse and engagement. Finally, the time had come to get back to normal, or at least as close to normal as possible: GPTP 2022 was held from June 2 to June 4 at Weiser Hall in the University of Michigan in Ann Arbor, hosted by the Center for the Study of Complex Systems. When Prof. Carl Simon graciously greeted us early on the first day of the event, he reminded us that it was wonderful to host GPTP once again, especially since it was the first in-person event held there in almost two years. While this sounded completely reasonable given what transpired over the past two years, it was still shocking to think about such a statement, and motivated us to make the event as productive and engaging as ever. As Carl mentioned, GPTP is characterized by open discussions, thought-provoking talks and a dynamic format, small in number but big in spirit.

Among the many highlights of GPTP over the years have been the amazing keynote talks, and this year was no exception. The opening keynote was delivered by Dr. Frank Crary from the University of Colorado, Boulder, giving the audience a bird's-eye view, or more appropriately a satellite's-eye view, of how our solar system has been, and is being, explored by robotic spacecraft, particularly by always captivating NASA missions. He also provided a perspective on the role that software systems play in such missions, discussing their unique constraints and stringent requirements, along with the possible opportunities that may lie on the horizon as cost of space travel continues to go down and as more countries reach the frontiers of space, especially for those interested in taking machine learning and genetic programming along for the ride.

On the second day, the keynote was given by Prof. Susan Stepney from the University of York, providing a fresh new look at how life could, or should, be studied, and how to get there by way of an engineering program. At the intersection of biology, computing and physics, cyber-bio-physical systems, or Zoetic systems, were succinctly presented by Susan, capturing the imagination of the audience and discussing how Zoetic science could come to be, by thinking of life as a process, and by taking inspiration of how thermodynamics developed from work in engineering before becoming a full-fledged science. The breadth and scope of the talk was inspiring, and the implications and opportunities for the genetic programming community were discussed in a lively Q&A. These topics are explored in greater detail in the chapter "Life as a Cyber-Bio-physical System" contributed by Susan Stepney to the present collection.

On the final day, the keynote was delivered by Craig Reynolds, who talked about some of his most recent work on the evolution of camouflage using genetic programming and co-evolutionary dynamics. The presentation was extremely engaging, as he showed us how his artificial evolutionary system was able to progressively discover the ability to generate camouflage patterns to trick a learning neural net predator, generating a variety of intriguing and aesthetic visual patterns. From conceptualization to design and implementation, Craig's system presented a perfect example of how researchers in our field continue to discover new and amazing ways to leverage the adaptability of evolution in the systems we develop and study, taking care to provide evolution with the necessary elements to construct the building blocks required to solve a given task.

Besides the keynotes, many invited speakers presented some of their most recent findings and ideas concerning genetic programming, artificial evolution and machine learning, covering topics that included auto-machine learning, interpretable machine learning, adversarial learning, symbolic regression and complexity. The book you hold in your hands contains a collection of 12 chapters derived from all those talks given at the workshop, each chapter having been authored, read, reviewed and discussed at the 19th edition of GPTP in Ann Arbor, Michigan, by participants of the workshop. We also had a fantastic in-person gathering at Bill Worzel's home, hosted by one of the founders of GPTP and his amazing spouse; it was easy to see where the gracious, generous and affable spirit of the event comes from, thanks, Bill!

We are very honored and grateful that we could once again organize another GPTP workshop in person, and the accompanying book, after two years of uncertainty. It is our intention that GPTP continues to be a core event for genetic programming research, bringing together academics, practitioners and theorists from diverse fields of science that intersect in our community, providing for a constructive, thoughtful, inspired and open interchange of ideas, and to do so, whenever possible, in-person, with a coffee during breaks or a beer at dinner.

Tijuana, Baja California, Mexico  
East Lansing, MI, USA  
Hagenberg, Wels, Austria  
Lisbon, Portugal  
September 2022

Leonardo Trujillo  
Wolfgang Banzhaf  
Stephan M. Winkler  
Sara Silva



# Acknowledgements

We would like to thank all of the participants for making GP Theory and Practice a successful IN-PERSON workshop once again in 2022. It was a breath of fresh air, literally, to talk, interact and discuss with all of the attendees. Our community deserved it after such prolonged isolation. Special thanks to our three wonderful keynote speakers, Frank, Susan and Craig: Your talks were amazing!

We would also like to thank our financial supporters for making the existence of GP Theory and Practice possible for 19 great editions. For 2022, we are grateful to the following sponsors:

- Michael Affenzeller from HEAL and the University of Applied Sciences Upper Austria (FH Oberösterreich)
- Stuart Card
- Michael Korn
- Mark Kotanchek at Evolved Analytics
- John Koza
- Jason H. Moore at the Department of Computational Biomedicine in Cedars-Sinai.

A number of people made key contributions to the organization of the workshop. Foremost among them is Linda Wood, who helped behind the scenes before, during and after the workshop. Special thanks to Carl Simon at the Center for the Study of Complex Systems at the University of Michigan for hosting GPTP once again. We are particularly grateful for contractual assistance by Mio Sugino and Nobuko Kamikawa, Springer-Nature Tokyo, and editorial assistance by Sivananth S. Siva Chandran, Springer-Nature Chennai. We would also like to express our gratitude to

Erik Goodman and Charles Ofria at the BEACON Center for the Study of Evolution in Action at Michigan State University for their continued support.

Tijuana, Baja California, Mexico  
East Lansing, MI, USA  
Hagenberg, Wels, Austria  
Lisbon, Portugal  
September 2022

Leonardo Trujillo  
Wolfgang Banzhaf  
Stephan M. Winkler  
Sara Silva

# Contents

<b>Symbolic Regression in Materials Science: Discovering Interatomic Potentials from Data</b> .....	1
Bogdan Burlacu, Michael Kommenda, Gabriel Kronberger, Stephan M. Winkler, and Michael Affenzeller	
<b>Correlation Versus RMSE Loss Functions in Symbolic Regression Tasks</b> .....	31
Nathan Haut, Wolfgang Banzhaf, and Bill Punch	
<b>GUI-Based, Efficient Genetic Programming and AI Planning for Unity3D</b> .....	57
Robert Gold, Andrew Haydn Grant, Erik Hemberg, Chathika Gunaratne, and Una-May O’Reilly	
<b>Genetic Programming for Interpretable and Explainable Machine Learning</b> .....	81
Ting Hu	
<b>Biological Strategies ParetoGP Enables Analysis of Wide and Ill-Conditioned Data from Nonlinear Systems</b> .....	91
Mark Kotanchek, Theresa Kotanchek, and Kelvin Kotanchek	
<b>GP-Based Generative Adversarial Models</b> .....	117
Penousal Machado, Francisco Baeta, Tiago Martins, and João Correia	
<b>Modeling Hierarchical Architectures with Genetic Programming and Neuroscience Knowledge for Image Classification Through Inferential Knowledge</b> .....	141
Gustavo Olague, Matthieu Olague, Gerardo Ibarra-Vazquez, Isnardo Reducindo, Aaron Barrera, Axel Martinez, and Jose Luis Briseño	
<b>Life as a Cyber-Bio-Physical System</b> .....	167
Susan Stepney	

**STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison** ..... 201  
Ryan Urbanowicz, Robert Zhang, Yuhao Cui, and Pranshu Suri

**Evolving Complexity is Hard** ..... 233  
Alden H. Wright and Cheyenne L. Laue

**ESSAY: Computers Are Useless ... They Only Give Us Answers** ..... 255  
Bill Worzel

**Index** ..... 261

# Contributors

**Michael Affenzeller** Heuristic and Evolutionary Algorithms Laboratory, University of Applied Sciences Upper Austria, Hagenberg, Austria

**Francisco Baeta** CISUC and LASI, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

**Wolfgang Banzhaf** Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

**Aaron Barrera** EvoVisión Laboratory, CICESE, Ensenada, B.C., Mexico

**Jose Luis Briseño** EvoVisión Laboratory, CICESE, Ensenada, B.C., Mexico

**Bogdan Burlacu** Josef Ressel Center for Symbolic Regression and Heuristic and Evolutionary Algorithms Laboratory, University of Applied Sciences Upper Austria, Hagenberg, Austria

**João Correia** CISUC and LASI, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

**Yuhan Cui** University of Pennsylvania, Philadelphia, PA, USA

**Robert Gold** ALFA, MIT CSAIL, Cambridge, MA, USA

**Andrew Haydn Grant** ALFA, MIT CSAIL, Cambridge, MA, USA

**Chathika Gunaratne** ALFA, MIT CSAIL, Cambridge, MA, USA

**Nathan Haut** Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA

**Erik Hemberg** ALFA, MIT CSAIL, Cambridge, MA, USA

**Ting Hu** School of Computing, Queen's University, Kingston, ON, Canada; Department of Computer Science, Memorial University, St. John's, NL, Canada

**Gerardo Ibarra-Vazquez** ITESM, Institute for Future of Education, Monterrey, N.L., Mexico

**Michael Kommenda** Josef Ressel Center for Symbolic Regression and Heuristic and Evolutionary Algorithms Laboratory, University of Applied Sciences Upper Austria, Hagenberg, Austria

**Kelvin Kotanchek** Evolved Analytics LLC, Rancho Santa Fe, CA, USA

**Mark Kotanchek** Evolved Analytics LLC, Rancho Santa Fe, CA, USA

**Theresa Kotanchek** Evolved Analytics LLC, Rancho Santa Fe, CA, USA

**Gabriel Kronberger** Josef Ressel Center for Symbolic Regression and Heuristic and Evolutionary Algorithms Laboratory, University of Applied Sciences Upper Austria, Hagenberg, Austria

**Cheyenne L. Laue** Computer Science Department, University of Montana, Missoula, USA

**Penousal Machado** CISUC and LASI, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

**Axel Martinez** EvoVisión Laboratory, CICESE, Ensenada, B.C., Mexico

**Tiago Martins** CISUC and LASI, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

**Gustavo Olague** EvoVisión Laboratory, CICESE, Ensenada, B.C., Mexico

**Matthieu Olague** Anahuac University Queretaro, El Márques, Querétaro, Mexico

**Una-May O'Reilly** ALFA, MIT CSAIL, Cambridge, MA, USA

**Bill Punch** Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA

**Isnardo Reducindo** Autonomous University of San Luis Potosí, Information Sciences Faculty, Fracc. Talleres, San Luis Potosí, Mexico

**Susan Stepney** Department of Computer Science, University of York, York, UK

**Pranshu Suri** University of Pennsylvania, Philadelphia, PA, USA

**Ryan Urbanowicz** Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA, USA

**Stephan M. Winkler** Heuristic and Evolutionary Algorithms Laboratory, University of Applied Sciences Upper Austria, Hagenberg, Austria

**Bill Worzel** Evolution Enterprise, Ann Arbor, MI, USA

**Alden H. Wright** Computer Science Department, University of Montana, Missoula, USA

**Robert Zhang** University of Pennsylvania, Philadelphia, PA, USA

# Symbolic Regression in Materials Science: Discovering Interatomic Potentials from Data



Bogdan Burlacu, Michael Kommenda, Gabriel Kronberger,  
Stephan M. Winkler, and Michael Affenzeller

**Abstract** Particle-based modeling of materials at atomic scale plays an important role in the development of new materials and the understanding of their properties. The accuracy of particle simulations is determined by *interatomic potentials*, which allow calculating the potential energy of an atomic system as a function of atomic coordinates and potentially other properties. First-principles-based *ab initio* potentials can reach arbitrary levels of accuracy, however, their applicability is limited by their high computational cost. Machine learning (ML) has recently emerged as an effective way to offset the high computational costs of *ab initio* atomic potentials by replacing expensive models with highly efficient surrogates trained on electronic structure data. Among a plethora of current methods, symbolic regression (SR) is gaining traction as a powerful “white-box” approach for discovering functional forms of interatomic potentials. This contribution discusses the role of symbolic regression in Materials Science (MS) and offers a comprehensive overview of current methodological challenges and state-of-the-art results. A genetic programming-based approach for modeling atomic potentials from raw data (consisting of snapshots of atomic positions and associated potential energy) is presented and empirically validated on *ab initio* electronic structure data.

---

B. Burlacu (✉) · M. Kommenda · G. Kronberger  
Josef Ressel Center for Symbolic Regression and Heuristic and Evolutionary Algorithms  
Laboratory, University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg,  
Austria  
e-mail: [bogdan.burlacu@fh-ooe.at](mailto:bogdan.burlacu@fh-ooe.at)

S. M. Winkler · M. Affenzeller  
Heuristic and Evolutionary Algorithms Laboratory, University of Applied Sciences Upper  
Austria, Softwarepark 11, 4232 Hagenberg, Austria

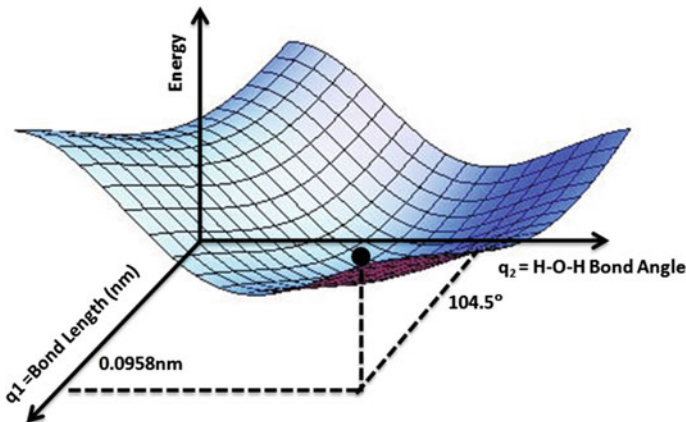
© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
L. Trujillo et al. (eds.), *Genetic Programming Theory and Practice XIX*,  
Genetic and Evolutionary Computation, [https://doi.org/10.1007/978-981-19-8460-0\\_1](https://doi.org/10.1007/978-981-19-8460-0_1)

# 1 Introduction

Materials Science (MS) is a highly interdisciplinary field incorporating elements of physics, chemistry, engineering and more recently, machine learning, in order to design and discover new materials. The rapid increase in processing power over the last decades has made computational modeling and simulation the main tool for studying new materials and determining their properties and behavior. Computational approaches can deliver accurate quantitative results without the need to set up and execute highly complex and costly physical experiments.

Potential energy surfaces (PES), describing the relationship between an atomic system's potential energy and the geometry of its atoms, are a central concept in computational chemistry and play a pivotal role in particle simulations. An example PES for the water molecule is shown in Fig. 1. The mathematical function used to calculate the potential energy of a system of atoms with given positions in space and generate the PES is called an *interatomic potential* function. The form of this function, its physical fidelity as well as its complexity and efficiency are critical components in simulations used to predict material properties.

The ability to simulate large particle systems over long time scales depends critically on the accuracy and computational efficiency of the interatomic potential. Broadly speaking, the more accurate the method, the lower its computational efficiency and the more limited its applicability. For example, first-principles modeling methods such as *density functional theory* (DFT) [33] provide highly accurate results by considering quantum-chemical effects but are not efficient enough to simulate large systems containing thousands of atoms over long time scales of nanoseconds [44].



**Fig. 1** PES for water molecule: the energy minimum corresponding to optimized molecular structure for water-O-H bond length of 0.0958nm and H-O-H bond angle of 104.5°. Image from Wikipedia ©AimNature



Molecular dynamics (MD) simulations treat materials as systems consisting of many microscopic particles (atoms) which interact with each according to the laws of statistical thermodynamics. These interactions are modeled by interatomic potentials depending mainly on particle positions. Macroscopic properties of materials are obtained as time and/or ensemble averages of processes emerging at the microscopic scale [27].

Empirical and semi-empirical methods treat atomic interactions in a more coarse-grained manner via parameterized analytic functional forms and trade-off accuracy for execution speed in order to enable simulations at a larger scale. Although they are computationally undemanding, they are only able to provide a qualitatively reasonable description of chemical interactions [53].

Machine learning (ML) interatomic potentials aim to bridge the gap between quantum and empirical methods in order to deliver the best of both worlds: functional forms that are as efficient as empirical potentials and as accurate as quantum-chemical approaches.

## 1.1 *Materials Informatics and Data-Driven Potentials*

Building upon the three established paradigms of science that have led to many technological advances over time, experimental, theoretical and simulation-based, a fourth “data-driven” paradigm of science is emerging today using machine learning and the large amounts of experimental and simulation data available [1]. “Big-data” science unifies the first three paradigms and opens up new avenues in materials science under the umbrella term of *materials informatics*. The field of material informatics is very new, and many unsolved questions still remain open and wait for proper answers [26].

Machine learning interaction models are generated on the basis of quantum-chemical reference data consisting of a series of snapshots of atomic coordinates, associated potential energy of the system and optionally other properties.

In molecular dynamics simulations, the system’s potential energy is typically decomposed into a set of independent  $m$ -body interactions that are a function of each particle’s position,  $\mathbf{r}$ . For a two-body or pair potential, it is assumed that the energy contributions from each pair of interacting particles are independent of other pairs and therefore:

$$E = \sum_{(i,j)} g(\mathbf{r}_i, \mathbf{r}_j) \quad (1)$$

For a three-body potential, triplets of atoms are also considered:

$$E = \sum_{(i,j)} g(\mathbf{r}_i, \mathbf{r}_j) + \sum_{(i,j,k)} h(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) \quad (2)$$

Traditionally, the functions  $g$  and  $h$  are represented by all kinds of empirical or semi-empirical analytic functions. With the advent of machine learning and data-based modeling, it becomes possible to automatically search for these functional forms with the help of ab initio training data. Substantial effort has already been put into this direction, and many machine learning models have been successful in discovering interatomic potentials for a variety of chemical configurations [42].

## 1.2 Current Challenges

Despite their success in representing atomic interactions, ML methods are not without their own challenges. Deriving highly accurate and tractable analytic functional forms for high-dimensional PESs is a very active field of research. The most important requirements for ML-based PESs are

- general applicability and absence of ad-hoc approximations (transferability);
- accuracy close to first-principles methods (including high-order many-body effects);
- very high efficiency to enable large simulations;
- the ability to describe chemical reactions and arbitrary atomic configurations;
- the ability to be automatically constructed and systematically improved.

Currently available potentials are far from satisfying all the needs [6], mainly due to the following difficulties and shortcomings.

### *Physical plausibility*

Closed physical systems are governed by various conservation laws that describe invariant properties. These fundamental principles of nature provide strong constraints that can be used to guide the search toward physically plausible ML models [53]. In molecular systems, each conserved quantity is associated with a differentiable symmetry of the action of a physical system. Typical conserved quantities include temporal and roto-translational invariance (i.e. total energy, linear and angular momentum). Forces must be the negative gradient of the potential energy  $E$  with respect to atomic positions  $r_i$ :

$$F_i = -\nabla_{r_i} E$$

When atoms move, they always acquire the same amount of kinetic energy as they lose in potential energy, and vice versa—the total energy is conserved. The potential energy of a molecule only depends on the relative positions of atoms and does not change with rigid rotations or translations.

Another aspect of invariance is permutational invariance resulting from the fact that from the perspective of the electrons, atoms with the same nuclear charge appear identical to each other and can thus be exchanged without affecting the energy or the forces. To ensure physically meaningful predictions, ML-based models must exhibit the same invariant behavior as the true potential energy surface.

### *Accuracy*

Accuracy is one of the most important requirements of ML potentials. The predicted energies and forces should be as close as possible to the underlying *ab initio* data. Numerical accuracy of the ML models is restricted by the intrinsic limitations of their functional form and descriptors (input variables) used. For example, conceptual problems related to incorporating rotational, translational and permutational invariance into descriptors are of primary relevance [6, 21, 46, 47] as well as their optimal design [20].

### *Transferability*

Ideally, potentials should be generally applicable and should not be restricted to specific types of atomic configurations. Due to their mathematical unbiased form, ML methods are promising candidates to reach this goal. However in practice, developed potentials often perform very well in applications they have been designed for, but are too system-specific and thus cannot be easily transferred from one system to another. The issues of extensibility, generality and transferability of the ML potentials need to be explicitly addressed [6].

### *Complexity and data requirements*

Another issue worth mentioning here is the mathematical complexity of ML potentials. For example, the most popular ML methods used to represent many-body PESs, ANNs, require complex architectures with many adjustable parameters (weights of neural synapses and neuron biases) to yield sufficiently flexible and invariant PES representations. For this, large amounts of training data (often dozens or even hundreds of thousands of points) are needed. On the other hand, the number of training data should be kept as low as possible since they are calculated via demanding quantum-chemical methods. It means that as simple as possible analytic representations of PESs are needed.

### *Integration of physical knowledge and interpretability*

Related to the mathematical complexity issue, it is also important to note that most of the ML methods (e.g. ANN and SVM) are of a “black-box” nature, and may be less amenable to including physical information in the functional forms, relying at least partially on physics-inspired features considered in atomic descriptors. This often leads to the increased mathematical and computational complexity of resulting interaction models. One of the main directions of the current development in ML-based computational MS is the shift from “black-box” methods toward “white-box” methods which often offer better interpretability.

## **2 State of the Art**

A plethora of machine learning approaches have recently emerged as a powerful alternative for finding a functional relation between an atomic configuration and corresponding energy [6, 17, 23]. Several ML techniques such as polynomial fitting [10],

Gaussian processes [5], spectral neighbor analysis [52], modified Shepard interpolation [29], moment tensor potentials [47], interpolating moving least squares [32], support vector machines [4], random forests [31], artificial neural networks (ANNs) [15, 25, 46, 55] or symbolic regression (SR) [40] have been successfully employed for a variety of systems.

More detailed reviews of current ML potentials can be found for example in [23, 37, 42] or [53]. Particularly, ANNs have received considerable attention and are probably the most popular form of ML potentials used in MS [55]. However, methods based on symbolic regression are gaining in popularity due to the advantages they bring in solving aspects of physical knowledge integration, efficiency and interpretability [7, 8, 11, 12, 24, 38, 41, 45, 48]. In the following, we refer to symbolic regression in its canonical incarnation that employs genetic programming to perform a search over the space of mathematical expressions. Symbolic regression approaches have succeeded in rediscovering simple forms of potentials that deliver qualitatively good results in a series of specific applications, some of which are described below.

## 2.1 Directed Search

The goal of the directed search is to improve search efficiency by limiting the hypothesis space to a functional form known to deliver qualitatively good results, instead of searching for a brand new potential.

Makarov and Metiu [38] use the Morse potential as a functional template for modeling diatomic molecules (see Sect. 5, Eq. 17). They rewrite it in the form:

$$M(D(r), R(r)) = D(r)(1 - \exp(R(r)))^2 \quad (3)$$

and use genetic programming to find the best  $D(r)$  and  $R(r)$ .

The directed search approach is augmented with an error metric that better reflects the physical characteristics of the problem. A standard error metric such as the MSE has the disadvantage of overemphasizing high-energy points which are rarely used during simulation. For this reason, the authors found it advantageous to introduce a scaling factor:

$$F(a) = \sum_i \frac{(E(r_i) - f(r_i; a))^2}{E^2(r_i) + \delta^2} \quad (4)$$

where the constant  $\delta$  is added to prevent division by zero.

For each function  $f_\alpha$  in the population of individuals, the fitness function is then defined as

$$p_\alpha = \exp(-\beta F_\alpha) \quad (5)$$

where parameter  $\beta$  controls how discriminating the function is and is adaptively updated during the run. The search starts with a small value for  $\beta$  which is gradually increased as the search improves.

The authors note the importance of including the derivative of the energy in the training data:

$$F'(a) = \sum_i \frac{|\nabla E(r_i) - \nabla f(r_i; a)|^2}{|\nabla E(r_i)|^2 + \delta^2} \quad (6)$$

leading to an expanded fitness function  $p_\alpha$ :

$$p_\alpha = \exp(-\beta(F + F')) \quad (7)$$

The recombination pool is filled using a proportional selection scheme. An additional “natural selection” operator employs a “badness list”  $b_\alpha = \exp(\beta F_\alpha)$  whose elements are the inverse of the fitness. Old individuals are replaced with a probability proportional to badness.

### Results

The directed search approach is shown to perform better than an undirected search over the search space, on training data generated using the Lippincott potential (Sect. 5, Eq. 19). A population size of 500 individuals is evolved over 150 generations (75,000 evaluations) using the primitive set  $\mathcal{P} = \{+, -, \div, \times, \exp\}$ . Furthermore, a search directed by a Lennard-Jones potential gives accuracy comparable to that directed by a Morse function, suggesting that restricting the hypothesis space with an appropriate functional template is a powerful and general approach in the search for interatomic potentials. In the case of the Lennard-Jones potential (Sect. 5, Eq. 18), the functional template was defined as

$$f(r) = 4D(r) \left[ \frac{1}{4} + \left( \frac{1}{R(r)} \right)^{12} - \left( \frac{1}{R(r)} \right)^6 \right] \quad (8)$$

The authors additionally note that some of the returned models, although accurate, exhibited unphysical behavior and did not extrapolate well. For example, one of the returned models based on the Lennard-Jones functional form had very good accuracy but contained a singularity at  $r = 12 \text{ \AA}$ , a point outside the interpolation range. The authors address overfitting by fitting the parameters of both the energy function and its derivative in the local search phase. This reduces the chance of obtaining pathological curves in the model extrapolation response.

Finally, Makarov and Metiu also model the potential of a triatomic molecule on *ab initio* data consisting of 60 nuclear configurations, showing that directed search maintains high levels of accuracy and scales favorably with dimensionality.

## 2.2 Directed Search with Parallel Multilevel Genetic Program

Belluci and Coker [7, 8] employ symbolic regression to discover empirical valence bond (EVB) models using directed search augmented with a multilevel genetic programming approach: the lower level (LLGP) optimizes co-evolving populations of models, while the higher level (HLGP) optimizes genetic operator probabilities of the lower level populations. The approach entitled Parallel Multilevel Genetic Program (PMLGP) found accurate EVB models for proton transfer in 3-hydroxy-grammapyrone (3-HGP) in the gas phase and protic solvent as well as ultrafast enolketo isomerization in the lowest singlet excited state of 3-hydroxyflavone (3-HF).

At the lower level (LLGP), the authors use the same error metric and fitness as in [38], namely Eqs. 4 and 5. LLGP individuals represent the  $R(r)$  functional part of the Morse potential (see Eq. 3). Remarkably, PMLGP does not use crossover but instead uses six different mutation operators:

- *Point mutation* randomly replaces a subtree with a randomly generated one.
- *Branch mutation* replaces a binary operator with one of its arguments at random.
- *Leaf mutation* replaces a leaf node with another randomly selected leaf.
- *New tree mutation* replaces an entire tree with a newly generated tree.
- *Parameter change* replaces each parameter value  $a_i$  with  $a_i + (R - 0.5)\gamma$ , where  $R$  is a uniform random number on the unit interval and  $\gamma$  is a scaling constant.
- *Parameter scaling* replaces each parameter value  $a_i$  with  $a_i R \gamma$ , where  $R$  is a uniform random number on the unit interval and  $\gamma$  is a scaling constant.

Of the last two types of mutation, parameter change is designed to make small local moves in parameter space, while parameter scaling is designed to make large moves in parameter space to escape the basins of attraction of local optima. Selection is performed using stochastic universal sampling [3].

At the higher level (HLGP), a real vector encoding is used to represent genetic operator probabilities. The population is initialized with  $k$  random vectors  $P_k = (p_1^{(k)}, \dots, p_6^{(k)})$ , with  $\sum_i p_i^{(k)} = 1$ , where  $k$  ranges from 1 to the total number  $N_p$  of processors, such that each vector corresponds to one of the LLGP populations whose operator probabilities it dynamically adapts.

The fitness of each vector  $P_k$  is evaluated based on the maximum fitness delta in the corresponding LLGP population over a specified time interval  $\Delta t$ :

$$F_k^{\text{HLGP}} = \frac{\Delta F_{\max}^{\text{LLGP}}}{\Delta t} \quad (9)$$

This is based on the idea that the larger the magnitude of  $F_k^{\text{HLGP}}$ , the more successful the set of probabilities  $P_k$  at improving the fitness of the population.

Two genetic operators are used to modify the probability vectors  $P_k$ :

- *Mutation* changes each component of the vector by a random amount with the constraint that all components sum up to one. This operator kicks in when the fitness of a vector  $P_k$  drops below a given threshold.

- *Adaptation* attempts to improve the probability distribution given by  $P_k$  by using feedback from the LLGP. Each LLGP builds a histogram of the number of times each mutation produced the most fit member of the population. Then the success frequency of the mutation operator is given by

$$s_i = \frac{w_i m_i}{n}, \quad w_i = \frac{1}{p_i}, \quad n = \sum_i m_i$$

Here,  $w_i$  is a weight,  $m_i$  is the number of successful mutations for the  $i$ th operator (component of  $P_k$ ) and  $n$  is the total number of successful mutations (for all operators). Based on the success frequencies, adaptation shifts a random amount of probability from the least successful operator to the most successful operator.

The number of LLGP populations (and HLGP individuals, respectively) is set to the number of available processors. Initially, all LLGP populations are identical but diverge during evolution as each corresponding fitness function is parameterized with a different value of  $\beta$  evenly sampled over a specified range. In effect, this applies different selection pressures on each LLGP population. Migrations are performed after the last adaptation step in HLGP. At this point, copies of the fittest individual in each LLGP population are sent to all the other populations, where they replace the least fit individual.

### Results

Training data for five different diatomic molecules (CO, H<sub>2</sub>, HCl, N<sub>2</sub> and O<sub>2</sub>) was generated using differently parameterized Morse functions, Gaussian functions and double well functions. The corresponding directed search spaces are given by

$$F_M = D(1 - \exp(-R(r; a)))^2 + c \quad \text{Morse} \quad (10)$$

$$F_G = A \exp(R(r; a)^2) \quad \text{Gaussian} \quad (11)$$

$$F_D = D_1(1 - \exp(-R_1(r; a)))^2 + D_2(1 - \exp(-R_2(r; a)))^2 \quad \text{Double well} \quad (12)$$

Parameters  $D$ ,  $c$ ,  $A$ ,  $D_1$  and  $D_2$  are optimized by including them as leaves in the trees.

The PMLGP approach was compared against a standard parallel genetic programming implementation (SPGP). In both cases, populations of 500 individuals were evolved in parallel on 8 processors for 20,000 generations. The function set  $\mathcal{F} = \{+, -, \times, \div, \exp\}$  was used for internal nodes and the terminal set  $\mathcal{T} = \{r, a_1, \dots, a_{10}\}$  was used for the leaf nodes.

PMLGP was shown to converge faster and achieve higher accuracy than SPGP. The obtained model of the EVB surface accurately reproduced global features of the *ab initio* data. The approach provides a basis for high-quality many-body potentials for studying gas and solution phase photon reactions.

### 2.3 Parallel Tempering

Slepoy et al. [48] use a hybrid approach consisting of genetic programming, Monte Carlo sampling and parallel tempering to discover the functional form of the Lennard-Jones pair potential.

Parallel tempering is an approach for parallel genetic programming where several islands (or *replicas*) evolve at a different effective temperature. High effective temperatures favor exploration by accepting new trees even if their fitness is poor, and low effective temperatures favor exploitation by being sensitive to small changes in fitness. By using replicas at different temperatures, the approach simultaneously performs both exploitation and exploration.

The remarkable aspect of this approach is that it marks the first large-scale application of genetic programming in materials science with interesting extensions to the canonical Koza-style algorithm and without restrictions of the hypothesis space.

The training data used consists of 10 nuclear configurations of 10 particles placed in 3D space. The Lennard-Jones potential describes the interactions between pairs of particles, therefore a nuclear configuration's energy is given by the sum of pairwise potentials:

$$E_{\text{conf}} = \sum_{\langle i, j \rangle} V_{\text{LJ}}(r_{ij}) \quad (13)$$

where  $r_{ij} = \|\mathbf{r}_i, \mathbf{r}_j\|$  is the distance between particles  $i$  and  $j$ . Fitness is defined as the negative mean squared error.

The evolutionary search is organized as a three-stage process consisting of generation, mutation and testing. Offspring individuals are tested for acceptance into the new population. A new tree is unconditionally accepted if its fitness exceeds the old one at the same index. Otherwise, it is accepted with the Boltzmann probability:

$$P_{\text{accept}} = \min \left\{ 1, \exp \left( \frac{F_{\text{new}} - F_{\text{old}}}{T} \right) \right\}$$

where  $F_{\text{old}}$  and  $F_{\text{new}}$  are the old and new fitness values, and  $T$  is the effective temperature.

After each generation, each sub-population exchanges one tree with its left neighbor in temperature space and one tree with its right neighbor. The trees to be swapped are selected with equal probability from their respective populations. The tree swap is accepted with a probability based on the relative Boltzmann weights of the two trees:

$$P_{\text{acc}} = \min \left\{ 1, \exp \left[ \left( \frac{1}{T_i} - \frac{1}{T_{i+1}} \right) (F_{i+1} - F_i) \right] \right\}$$

#### Results

A large-scale experiment was performed on a cluster made of 100 AMD Opteron 2.2 GHz processors. The trees were restricted to minimum depth 3 and maximum depth 4.



200 replicas with temperatures distributed logarithmically from 0.1 to 10 were used. The replica size was chosen to be either  $N = 10,000$  or  $N = 50,000$  individuals. The primitive set consisted of elementary operations  $\mathcal{P} = \{+, -, \times, \div, \exp, |\cdot|\}$ .

The proposed approach successfully discovered the Lennard-Jones potential or arithmetic equivalents within 100 generations. Interestingly, the expended effort was estimated to be somewhere in the range of  $10^9$  evaluated trees, which represents only a small fraction of the possible trees with depth 4 (around  $2.9 \times 10^{36}$ ) [48].

A number of ideas for improving the physical fidelity of the developed functional forms and their generality and transferability are suggested:

- Inclusion of additional properties and forces on individual atoms in the training set.
- Primitive set extension to include three-body interactions.
- Integration of physical knowledge (inclusion of symmetries, invariances).

## 2.4 Symbolically Regressed Table KMC

In order to increase the time scale of simulations, molecular dynamics can be combined with kinetic (dynamic) Monte Carlo (KMC) techniques [9] that coarse-grain the state space, for example via discretization (e.g. assign an atom to a lattice site). The main assumption is that multiscale modeling requires only relevant information at the appropriate length or time scale.

KMC constructs a lookup table consisting of an *a priori* list of events such as atomic jumps or off-lattice jumps. This yields several orders of magnitude increases in simulated time and allows to directly model many processes unapproachable by MD alone. However, identifying barrier energies from a list of events is difficult and restricts the applicability of the method.

Here, symbolic regression is proposed to identify the functional form of the potential energy surface at barrier energy points from a limited set of *ab initio* training data. The method entitled Symbolically Regressed Table KMC (sr-KMC) [45] provides a machine learning replacement for the lookup table in KMC, thus removing the need for explicit calculation of all activation barriers.

Sastry [45] showed that symbolic regression allows atomic-scale information (diffusion barriers on the potential energy surface) to be included in a long-time kinetic simulation without maintaining a detailed description of all atomistic physics, as done within molecular dynamics.

In this approach, fitness is computed as a weighted mean absolute error between the predicted and calculated barriers, for  $N$  random configurations:

$$F = \frac{1}{N} \sum_{i=1}^N w_i |\Delta E_{\text{pred}}(\mathbf{x}_i) - \Delta E_{\text{calc}}(\mathbf{x}_i)| \quad (14)$$

Setting  $w_i = |\Delta E_{\text{calc}}|^{-1}$  gives preference to predicting accurately lower energy (most significant) events over higher energy events.

The algorithm uses the *ramped-half-and-half* tree creation method, tournament selection and Koza-style subtree crossover, subtree mutation and point mutation [34].

### Results

sr-KMC is applied to the problem of vacancy-assisted migration on the surface of phase-separating  $\text{Cu}_x\text{Co}_{1-x}$  at a concentrated alloy composition ( $x = 0.5$ ). Two types of potentials (Morse and TB-SMA) are used to generate the training data via molecular dynamics. The number of active configurations is limited knowing that only atoms in the environment locally around vacancy and migrating atoms significantly influence the barrier energies.

The inline barrier function is represented from the primitive set  $\mathcal{P} = \mathcal{F} \cup \mathcal{T}$ , with  $\mathcal{F} = \{+, -, \times, \div, \text{pow}, \text{exp}, \text{sin}\}$  and  $\mathcal{T} = \{\mathbf{x}, \mathcal{R}\}$ . Here,  $\mathbf{x}$  represents the currently active configuration and  $\mathcal{R}$  is an ephemeral random constant.

The results show that GP predicts all barriers within 0.1% error while using less than 3% of the active configurations for training. This leads to a significant scale-up in real simulation time and a significant reduction in the CPU time needed for KMC. sr-KMC is also compared against the basic KMC approach (using a table look up) where it was shown to perform orders of magnitude faster.

The authors note that standard basis-set regression methods are generally not competitive to GP due to the inherent difficulty in choosing appropriate basis functions and show that quadratic and cubic polynomials perform worse in terms of accuracy (within 2.5% error) while requiring energies for  $\sim 6\%$  of the active configurations.

They also note that GP is robust to changes in the configuration set, the order in which configurations are used or the labeling scheme used to convert the configuration into a vector of inputs.

## 2.5 Hierarchical Fair Competition

Brown, Thompson and Schultz [11, 12] are able to rediscover the functional forms of known two- and three-body interatomic potentials using a parallel approach to genetic programming with extensions toward better generalization. Their implementation is based on Hierarchical Fair Competition (HFC) by Jianjun et al. [28].

The HFC framework [28] is designed toward maintaining a continuous supply of fresh genetic diversity in the population and protecting intermediate individuals who have not reached their evolutionary potential from being driven to extinction by unfair competition. It implements these goals with the help of a hierarchical population structure where individuals only compete with other individuals of similar fitness.

Brown et al. note that a correlation-based fitness measure would increase the efficiency of the search and propose the following formula using the Pearson correlation coefficient:

$$F = \frac{N}{N + 100 - 100 \left| \sum_{i=1}^N \frac{(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{p_i \sigma_y \cdot p_i \sigma_{\hat{y}}} \right|} \quad (15)$$

Here,  $N$  is the number of configurations and  $p_i$  is the number of terms in the summation over  $g$  (see Eq. 1). Ordinary least squares is then used to fit the prediction  $\hat{y}$  to the data by introducing scale and intercept terms to the functions  $g$  and  $h$ :

$$E = \sum_{(i,j)} (a \cdot g(\mathbf{r}_i, \mathbf{r}_j) + b) + \sum_{(i,j,k)} (c \cdot h(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + d) \quad (16)$$

The approach is implemented in PM- DREAMER, an open-source software package developed on top of the Open Beagle library for evolutionary computation [19], using its available genetic operators. These include several mutations (standard, shrink, swap, constant), subtree-swapping crossover, tournament selection and elitism:

- *Standard* mutation replaces a node in the tree with a randomly generated subtree.
- *Swap* mutation swaps two nodes in the tree.
- *Shrink* mutation replaces a subtree with one of its arguments.
- *Swap subtree* mutation swaps a subtree’s arguments.
- *Ephemeral* mutation changes the value of a constant in the tree.

Additionally, PM- DREAMER implements support for distributed evolution using the MPI standard and introduces migration operators that exchange individuals between sub-populations at fixed intervals.

Bloat reduction strategies are implemented to prevent the expression trees from becoming increasingly large, a tendency observed especially in the case of three-body modeling. Two strategies are tested:

- Using a simplification operator which replaces subtrees that evaluate to a constant value with the constant value: this operator is applied generationally at a fixed interval.
- Using penalty terms to the fitness function: in this case, the fitness is decreased based on a threshold penalty size value  $s_b$  and a maximum penalty size  $s_e$ , such that trees with length  $< s_b$  are not penalized at all, and trees with length  $> s_e$  are penalized fully (fitness is set to zero).

*Local search.* Local search based on the derivative-free Nelder-Mead simplex algorithm is employed with a set probability, optimizing either a single constant or all the constants in the expression.

*HFC Extension.* Brown et al. implement HFC in a parallel manner by allowing populations with different fitness thresholds to evolve in parallel, with periodic migrations between them. After migrations, populations that grow too large are “decimated” by the removal of the least fit individuals, while populations that grow too small are supplemented with new randomly generated individuals.