A Survey of Advanced Computational Techniques in Machine Learning and Artificial Intelligence

www.surveyx.cn

Abstract

The survey paper explores a spectrum of advanced computational techniques pivotal in the evolution of machine learning and artificial intelligence. It highlights symbolic regression's role in deriving mathematical expressions that accurately represent datasets, bridging traditional regression with complex models. Sparse regression focuses on identifying pertinent features in high-dimensional data, enhancing model efficiency. Equation discovery automates deriving mathematical equations, leveraging evolutionary algorithms for complex behaviors. Genetic algorithms, inspired by natural selection, offer robust optimization across domains. Reinforcement learning trains agents to make decisions, rewarding desired behaviors, while large language models transform natural language processing through extensive text corpora. Automated model discovery identifies optimal predictive models, emphasizing interpretability. Evolutionary computation mimics biological evolution to solve complex problems. The paper delves into these techniques' methodologies, applications, and recent advancements, underscoring their transformative potential in enhancing model interpretability, scalability, and efficiency. The integration of domain knowledge, hybrid approaches, and adaptive learning mechanisms further augments these methodologies, fostering innovation and addressing complex challenges in diverse fields. The survey concludes with future research directions, emphasizing the need for standardized benchmarks, enhanced function libraries, and interdisciplinary applications to advance the capabilities of AI and machine learning systems.

1 Introduction

1.1 Overview of Advanced Computational Techniques

The rapid advancement of machine learning and artificial intelligence is driven by a range of sophisticated computational techniques that have transformed data analysis and model generation. Symbolic regression is a key method that aims to discover mathematical expressions that accurately represent datasets, merging traditional linear regression with complex models like neural networks to achieve a balance of interpretability and sophistication [1]. Recent innovations, such as the Neural-Enhanced Monte-Carlo Tree Search (NEMoTS), combine Monte-Carlo Tree Search with neural networks to efficiently derive analytical expressions from time series data. Additionally, evolutionary symbolic regression-based classification algorithms address the limitations of conventional methods like logistic regression and decision trees.

Sparse regression plays a critical role in high-dimensional data contexts, focusing on identifying the most relevant features for predictive modeling. Advanced methodologies, including LLM-SR and LLM4ED, automate the extraction of mathematical equations from data, leveraging evolutionary algorithms and Large Language Models (LLMs) to generate diverse candidate equations while incorporating domain-specific knowledge. These systems treat equations as programmable structures, iteratively refining hypotheses based on data fitting and physical understanding, ultimately leading to

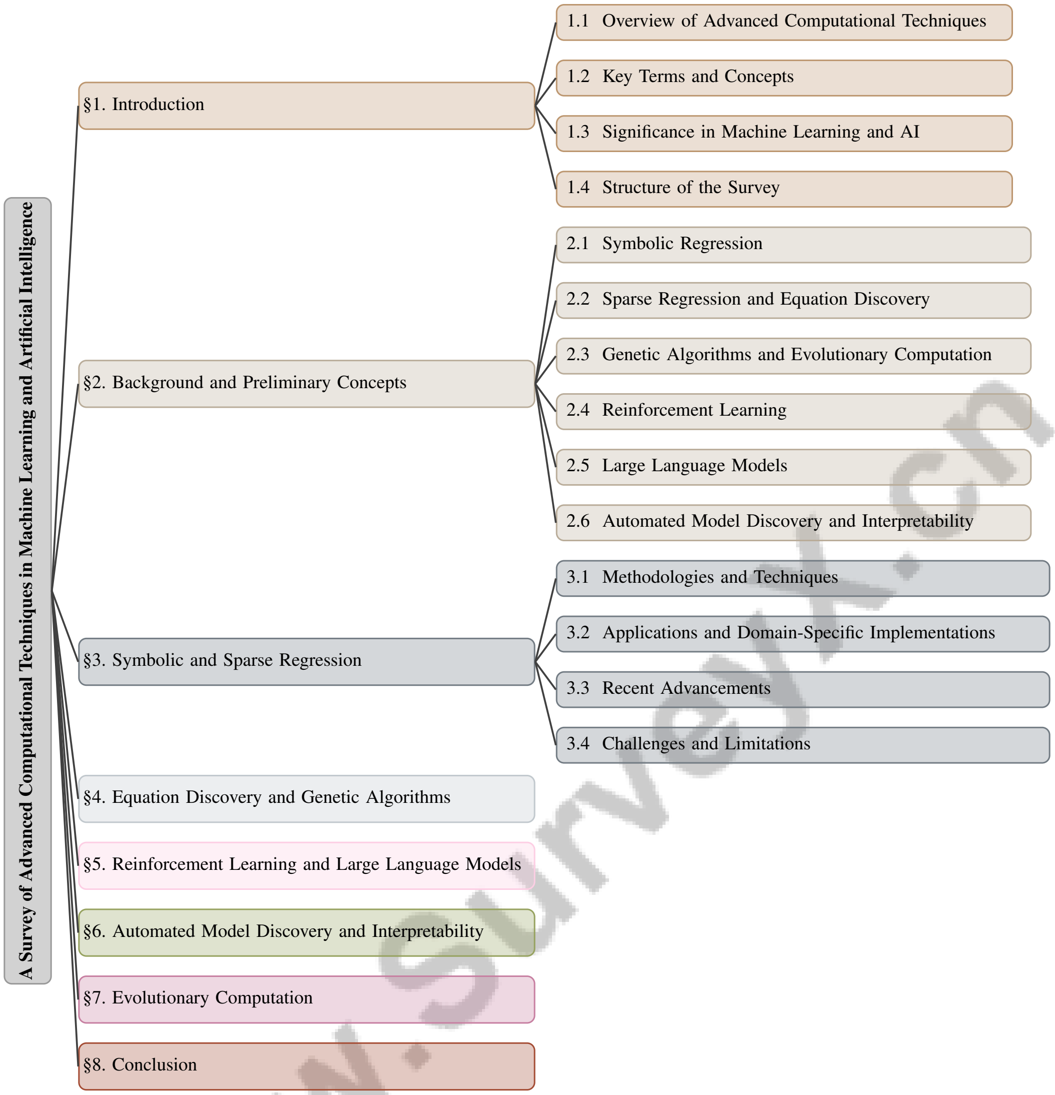

A Survey of Advanced Computational Techniques in Machine Learning and Artificial Intelligence
§1. Introduction
1.1 Overview of Advanced Computational Techniques
1.2 Key Terms and Concepts
1.3 Significance in Machine Learning and AI
1.4 Structure of the Survey
§2. Background and Preliminary Concepts
2.1 Symbolic Regression
2.2 Sparse Regression and Equation Discovery
2.3 Genetic Algorithms and Evolutionary Computation
2.4 Reinforcement Learning
2.5 Large Language Models
2.6 Automated Model Discovery and Interpretability
§3. Symbolic and Sparse Regression
3.1 Methodologies and Techniques
3.2 Applications and Domain-Specific Implementations
3.3 Recent Advancements
3.4 Challenges and Limitations
§4. Equation Discovery and Genetic Algorithms
§5. Reinforcement Learning and Large Language Models
§6. Automated Model Discovery and Interpretability
§7. Evolutionary Computation
§8. Conclusion

Figure 1: chapter structure

the identification of governing equations that describe complex natural phenomena across scientific disciplines [2, 3, 4]. The incorporation of active learning and physical constraints further enhances the rediscovery of mathematical equations from data.

Genetic algorithms and evolutionary computation, inspired by natural selection, provide robust optimization techniques that evolve potential solutions across generations, proving effective in various domains. Reinforcement learning, which trains agents to make decisions through rewards, has become a fundamental approach in developing autonomous systems capable of intelligent decision-making [5].

Large language models, trained on extensive text corpora, have revolutionized natural language processing, enabling machines to understand and generate human language with remarkable accuracy. This shift includes the application of LLMs in symbolic regression, aiming to identify mathematical expressions that describe numerical datasets. Frameworks such as In-Context Symbolic Regression (ICSR) iteratively refine functional forms using LLMs while determining coefficients through external optimization, resulting in symbolic equations that often outperform traditional methods on benchmark datasets. Additionally, novel Transformer architectures specifically designed for symbolic regression enhance flexibility and efficiency in scientific discovery without incurring extra computational costs [6, 7]. Automated model discovery is crucial in identifying optimal predictive models, while

AI-generated, for reference only.

interpretable machine learning ensures that these models' decisions remain comprehensible to humans.

These advanced computational techniques collectively offer powerful tools for addressing complex challenges in machine learning and artificial intelligence, facilitating innovative applications across diverse domains. For instance, the development of a flexible, data-driven system identification approach that adapts to multibody dynamics complexities without prior assumptions exemplifies the versatility of these techniques [8].

1.2 Key Terms and Concepts

Symbolic regression is a sophisticated machine learning technique aimed at discovering mathematical expressions that accurately represent the relationship between input and response variables [9]. This process emphasizes deriving interpretable symbolic expressions from data [10]. Techniques such as gene expression programming (GEP) and sequential threshold ridge regression (STRidge) are commonly employed to model complex relationships within datasets [11]. The integration of symbolic regression with neural networks, exemplified by Physics-informed Neural Networks (PINNs) and Differentiable Program Architecture (DPA), enhances the modeling of partial differential equations (PDEs) with improved precision [12].

Sparse regression, closely related to symbolic regression, identifies a minimal set of relevant features for predictive modeling, often employing hybrid methodologies to enhance efficiency and accuracy [13]. Symbolic integration and hybrid methods further refine the extraction of key features from complex datasets.

Equation discovery, synonymous with symbolic regression, aims to derive concise and interpretable symbolic mathematical expressions that approximate unknown underlying equations [2]. This methodology is crucial in areas requiring interpretable modeling, such as constitutive modeling in hyperelasticity [14], and in converting neural network outputs into interpretable mathematical expressions to enhance model usability [15].

The advent of Transformer models has revolutionized symbolic regression by enabling the simultaneous prediction of entire mathematical expressions, including both structure and numerical constants [16]. This innovation significantly enhances the efficiency of symbolic regression tasks.

Interpretable machine learning is essential in these techniques, focusing on developing models with decisions that are easily understandable by humans. This involves addressing interpretability and uncertainty quantification challenges, especially in noisy data environments [17]. Information fusion in symbolic regression is also critical for developing interpretable models [18].

The MLLM-SR method exemplifies advancements in symbolic regression by allowing users to specify their requirements in natural language, facilitating the generation of expressions that align with observed data [19]. These key terms and concepts underscore the importance of advancing computational techniques that enhance predictive accuracy while ensuring model interpretability and usability across diverse domains.

1.3 Significance in Machine Learning and AI

The integration of advanced computational techniques in machine learning and artificial intelligence signifies a paradigm shift towards improved model interpretability, scalability, and efficiency. Symbolic regression is pivotal, constructing dynamic process models that yield parsimonious analytic expressions, surpassing traditional black-box models in accuracy and interpretability [20]. Innovations such as Neural Symbolic Regression with Hypotheses (NSRwH) enhance algorithmic performance by incorporating user-defined prior knowledge, boosting precision and interpretability.

In domains where decision-making carries significant consequences, the need for interpretable machine learning models is critical [21]. Symbolic regression advances model interpretability and generalization, essential for scientific comprehension [10]. The recursive-LASSO-based symbolic (RLS) regression method exemplifies the construction of interpretable mathematical models capable of revealing new natural laws from noisy datasets, highlighting interpretability's role in uncovering underlying relationships [22]. The QLattice approach addresses the complexities of modeling intricate data relationships through symbolic regression, overcoming the interpretative limitations of traditional methodologies [23].

Large language models (LLMs) aim to streamline the extraction of governing equations from data, facilitating the discovery of hidden relationships via machine learning, particularly neuro-symbolic approaches. Additionally, the IGEP method fosters a more informed initial population, expediting the optimization process in symbolic regression compared to conventional techniques [24].

The significance of these computational techniques is highlighted by their role in automated scientific discovery, where systems autonomously generate hypotheses, conduct experiments, and interpret outcomes [25]. The benchmark for evaluating AutoML systems emphasizes the importance of synthesizing pipelines efficiently, advancing automated machine learning [26]. Evolutionary symbolic regression methods demonstrate efficacy in classifying binary and multinomial datasets, outperforming traditional classification techniques [27].

Collectively, these advancements illustrate the transformative potential of advanced computational techniques in fostering innovation and enhancing the interpretability and functionality of machine learning and AI systems. They provide robust tools for addressing complex challenges, enabling the development of models that are both accurate and comprehensible across various fields. The necessity for novel methods, such as SymFormer, arises from existing techniques' limitations in efficiently predicting formulas, further underscoring the need for continuous innovation [28]. The exploration of hidden semantics in neural networks through approaches like SRNet enhances interpretability, facilitating a deeper understanding of model behavior [29]. Moreover, integrating active learning and physical constraints in symbolic regression improves rediscovery rates of equations with fewer data points [30], while the GSR method innovatively modifies the symbolic regression optimization problem to focus on transformations of the target variable [31]. These innovations exemplify the potential of machine learning techniques to bridge knowledge gaps and advance scientific discovery [32]. The primary innovation of SymPDE, leveraging deep reinforcement learning to derive closed-form symbolic solutions, contrasts with existing methods reliant on approximations and numerical fitting, further demonstrating the significance of these computational techniques in advancing AI capabilities [8].

1.4 Structure of the Survey

This survey is organized to provide a comprehensive exploration of advanced computational techniques in machine learning and artificial intelligence. It begins with an introductory section outlining the primary focus of the survey, presenting an overview of key computational techniques and their significance in the field. Following the introduction, the survey delves into background and preliminary concepts, offering detailed explanations of core methodologies such as symbolic regression, sparse regression, equation discovery, genetic algorithms, reinforcement learning, large language models, automated model discovery, interpretable machine learning, and evolutionary computation.

The survey is structured into thematic sections, each dedicated to a specific set of techniques. The first major section focuses on symbolic and sparse regression, examining their methodologies, applications, recent advancements, and challenges. This is followed by a detailed discussion on equation discovery and genetic algorithms, highlighting multi-objective optimization, genetic programming dynamics, innovative techniques, and hybrid approaches.

Subsequent sections explore the principles and applications of reinforcement learning, alongside the development and impact of large language models, including an analysis of their integration and recent breakthroughs. The survey further investigates automated model discovery and interpretability, emphasizing the balance between model complexity and interpretability, supported by case studies and applications.

The penultimate section provides an in-depth overview of evolutionary computation, its enhancements, and applications in machine learning program design, highlighting adaptive learning with feedback mechanisms. The survey concludes with a reflection on the current state of these advanced techniques, discussing potential future directions and areas for further research, thereby offering a roadmap for ongoing and future advancements in the field.The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Symbolic Regression

Symbolic regression is a pivotal computational tool in machine learning and AI, tasked with uncovering mathematical expressions that accurately depict dataset relationships. Unlike traditional regression, it navigates a vast hypothesis space without relying on predefined model structures, offering flexibility essential for creating interpretable models that reveal underlying mechanisms across scientific domains [33]. The technique's ill-posed nature poses challenges, as similar expression structures may vary in coefficients, complicating supervision. Feature space transformations are crucial for modeling specific functions effectively. Recent progress has enhanced expression skeleton recovery, surpassing state-of-the-art methods in R^2 performance across benchmarks [33].

Despite promising advancements, the absence of standardized benchmarks hinders comparison across symbolic regression methods. Establishing benchmarks is vital for evaluating these methods' ability to model complex relationships while maintaining simplicity [33]. Innovative frameworks like SymPDE leverage deep reinforcement learning to derive closed-form solutions for partial differential equations, showcasing symbolic regression's utility in complex systems modeling [8]. By approximating true data-generating processes, symbolic regression enhances model interpretability, contributing to scientific discovery across diverse fields. Continuous improvements in scalability and efficiency broaden its applicability, solidifying symbolic regression as an indispensable tool in exploring complex scientific domains.

2.2 Sparse Regression and Equation Discovery

Sparse regression and equation discovery are critical methodologies in AI and machine learning, focusing on deriving mathematical models that encapsulate data relationships. Sparse regression is particularly effective in high-dimensional datasets, identifying minimal feature subsets to enhance interpretability and efficiency, crucial in domains with sparse data or limited resources [34]. Traditional methods like LASSO often struggle with correlated features, impacting their effectiveness in uncovering dynamical system equations [35].

Equation discovery, akin to symbolic regression, seeks to uncover scientific laws from data [36]. It navigates a vast hypothesis space to identify symbolic expressions accurately describing variable relationships [37]. However, the non-linear, high-dimensional nature of data and the NP-hardness of symbolic regression present challenges in identifying governing equations from extensive datasets [38]. Scalability and computational efficiency issues often result in slow performance and difficulties in obtaining optimal expressions [39].

The lack of systematic comparison among symbolic regression algorithms underscores the need for benchmark studies to evaluate genetic programming symbolic regression (GPSR) methods against traditional machine learning approaches, particularly in recovering historical empirical equations [40]. Such benchmarks are crucial for advancing the field by providing insights into various methodologies' strengths and limitations.

A primary challenge in symbolic regression is extracting complex mathematical expressions from sparse datasets, essential for scientific research [41]. High computational costs and overfitting risks complicate accurate model discovery [34]. Moreover, the inability of symbolic regression and Generalized Additive Models (GAMs) to capture complex variable interactions poses a significant challenge in scientific applications.

Advancements in machine learning enhance sparse regression and equation discovery's applicability and effectiveness, facilitating precise and interpretable models through symbolic regression and GAMs. These developments enable researchers to navigate complex data, derive meaningful insights, and pave the way for models that improve accuracy while maintaining interpretability [42, 32, 43].

2.3 Genetic Algorithms and Evolutionary Computation

Genetic algorithms (GAs) and evolutionary computation are foundational optimization techniques in AI, inspired by natural selection. These methodologies simulate evolutionary processes to evolve solutions iteratively, effectively addressing complex optimization challenges. Genetic programming (GP), a key subset, excels in symbolic regression by optimizing function structures and coefficients without predefined assumptions [44]. This adaptability is crucial for navigating symbolic regression's vast search spaces, where solutions are represented as expression trees of varying complexities [45].

Recent advancements have enhanced GP capabilities. The Physically Inspired Neural Dynamics Symbolic Regression (PI-NDSR) combines neural networks for denoising with GP for symbolic regression, improving accuracy and robustness [46]. Neuro-encoded expression programming (NEEP) offers a continuous representation, enhancing efficiency in deriving explicit functions for data simulation [47].

Innovative frameworks like SymbolNet address the computational intensity of exploring potential expressions, enabling dynamic pruning of model weights, input features, and mathematical operators within a single training process [48]. This approach alleviates existing symbolic regression methods' rigidity, which often require extensive dataset training [6]. Furthermore, TPSR, a Transformer-based Planning strategy for Symbolic Regression, integrates a Monte Carlo Tree Search (MCTS) algorithm into the transformer decoding process to optimize equation generation [49].

Hybrid approaches further enhance genetic programming's flexibility and robustness. The GP-RVM method combines GP and Relevance Vector Machine to evolve a linear combination of basis functions [50]. The GSR method employs GP with a matrix-based encoding scheme to identify and optimize expressions [31].

The integration of evolutionary algorithms to explore large model spaces converges on models that accurately represent network structures [51]. Evolutionary symbolic sparse regression methods utilize GP to discover symbolic function expressions and estimate system parameters from data, showcasing these techniques' potential in data-driven discovery [52].

Ensuring genetic symbolic regression expressions adhere to domain-specific constraints is a critical challenge, particularly in Gene Expression Programming (GEP) [53]. Active learning methods select experiments that maximize information gain while incorporating physical constraints to ensure meaningful rediscovered equations [30]. Integrating asymptotic constraints into symbolic regression, using neural networks to generate production rules and guiding a Monte Carlo Tree Search, illustrates innovative approaches in this domain [1].

Genetic algorithms and evolutionary computation's transformative potential is evident across diverse fields, providing robust tools for solving complex optimization problems. By enhancing interpretability and optimizing performance, these techniques significantly contribute to developing machine learning systems capable of addressing intricate challenges, as demonstrated in deriving control laws for dynamical systems [54]. Ongoing innovations continue to expand these techniques' applicability and effectiveness, fostering advancements across various scientific and engineering disciplines.

2.4 Reinforcement Learning

Reinforcement learning (RL) is a dynamic field within AI, focusing on training agents to make sequential decisions through environmental interaction. Agents learn to achieve goals by receiving rewards or penalties, refining strategies to maximize cumulative rewards over time. This trial-and-error process enables autonomous systems to perform complex tasks without explicit programming [55].

A significant RL challenge is modeling human-like decision-making, often oversimplified by traditional models. Conventional RL approaches typically use linear updating rules for reward expectations, failing to capture human reward processing's nuanced nature [55]. More sophisticated models are needed to accurately reflect human behavior and decision-making dynamics.

Incorporating RL into symbolic regression enhances model discovery. The Symbolic Q-network (Sym-Q) redefines symbolic regression as a sequential decision-making task, using RL to optimize symbolic expression searches. This approach leverages RL's strengths in exploring vast search spaces, improving symbolic regression tasks' efficiency and accuracy [56].

RL also plays a crucial role in identifying differential equations from empirical data, where minimal assumptions about physical processes are vital. This application highlights RL's potential in scientific discovery, enabling mathematical models that accurately represent complex phenomena [57].

Integrating RL with symbolic regression techniques, such as Neural-Enhanced Monte-Carlo Tree Search (NEMoTS), exemplifies the synergy between these methodologies. NEMoTS enhances computational efficiency and generalization capabilities, addressing challenges posed by larger datasets and the need for scalable solutions [58].

Despite its potential, RL faces challenges related to processing time and resource consumption, particularly with large datasets. Existing algorithms' slow processing speeds and high computational demands necessitate more efficient RL frameworks capable of handling extensive data efficiently [59]. Such advancements are crucial for expanding RL's applicability across diverse domains, from automated machine learning to dynamic network modeling, addressing complexities like node additions and temporal evolution [60].

2.5 Large Language Models

Large language models (LLMs) have emerged as transformative tools in natural language processing, characterized by their ability to understand, generate, and manipulate human language with remarkable accuracy. Typically based on deep learning architectures such as transformers, these models are trained on extensive corpora, enabling them to learn complex linguistic patterns and relationships. The capabilities of LLMs extend beyond basic text generation; they are increasingly applied in complex domains such as machine translation, sentiment analysis, and conversational agents, as well as in advanced tasks like symbolic regression for scientific equation discovery. This versatility is exemplified by their ability to analyze data and generate mathematical expressions reflecting underlying relationships, leveraging both natural language instructions and domain-specific prior knowledge. Furthermore, LLMs have demonstrated effectiveness in optimizing solutions and discovering physically accurate equations across various scientific fields, highlighting their significant impact on enhancing research methodologies and facilitating deeper insights into complex phenomena [2, 32, 61, 19].

Advancements in LLMs have been propelled by innovative developments in model architectures and training methodologies, resulting in substantial improvements in performance and scalability. These enhancements enable LLMs to tackle complex tasks, such as scientific equation discovery, by leveraging domain-specific knowledge and sophisticated optimization techniques, thereby outperforming traditional methods across various scientific disciplines [2, 4, 61, 62]. The transformer architecture has revolutionized the field by introducing mechanisms like self-attention, allowing models to dynamically weigh the importance of different words in a sentence, facilitating the processing of long-range dependencies in text, which is critical for understanding nuanced language structures.

One notable application of LLMs is in code generation tasks, where they exhibit the ability to generate syntactically and semantically correct code snippets. The Self-Taught Optimizer for Programming (STOP) exemplifies this capability by leveraging LLMs to recursively improve code generation outputs. STOP utilizes the language model to generate initial solutions and iteratively refines them, demonstrating the potential for self-improvement and optimization in programming tasks [63].

Beyond technical applications, LLMs play a crucial role in advancing scientific discovery by facilitating the extraction of knowledge from unstructured data. Their capacity to formulate hypotheses and integrate insights from extensive textual datasets positions machine learning technologies as essential assets in advancing research and fostering innovation across diverse scientific fields. This capability enhances the understanding of complex natural processes and facilitates the discovery of new scientific laws and relationships, bridging theoretical knowledge with empirical data to drive meaningful advancements in disciplines such as social science, physics, and cognitive psychology [32, 64, 65, 66, 67]. The integration of LLMs with other computational techniques, such as symbolic regression and reinforcement learning, further enhances their utility, enabling the development of models that are both interpretable and efficient.

Despite their notable achievements in various applications, LLMs encounter significant challenges related to interpretability, making it difficult for users to understand their decision-making processes, and computational demands, which can hinder their efficiency and scalability in complex scientific tasks [32, 4, 42, 2, 61]. The complexity of these models complicates the understanding of their underlying decision-making processes, raising concerns about transparency and bias. Additionally, the resource-intensive nature of training and deploying large-scale models necessitates ongoing efforts to optimize efficiency and accessibility.

AI-generated, for reference only.

2.6 Automated Model Discovery and Interpretability

Automated model discovery represents a significant advancement in machine learning and artificial intelligence, emphasizing the development of predictive models with minimal human intervention while ensuring transparency and interpretability. This process enhances the efficiency and accessibility of machine learning systems, particularly through the integration of symbolic regression frameworks that facilitate the conversion of complex neural network-based optimization rules into interpretable symbolic expressions, thereby reducing computational overhead and improving model transparency [62].

The importance of interpretability in symbolic regression is underscored by the challenges of achieving a balance between model complexity and interpretability, essential for deriving insights into complex systems. This balance is demonstrated in space-related applications, where complex models can lead to opacity in understanding variable relationships [68]. The Generalized Symbolic Regression (GSR) method exemplifies this by outperforming traditional methods in delivering interpretable mathematical models, thus enhancing the discovery of models that are both transparent and robust.

Large language models (LLMs) play a pivotal role in automating the extraction of governing equations from data, showcasing the synergy between machine learning techniques and symbolic regression. This integration leverages the capabilities of LLMs to facilitate the discovery of interpretable models that accurately capture the dynamics of complex systems. Recent approaches have focused on simplifying complex multi-dimensional symbolic regression problems into simpler one-dimensional tasks, thereby enhancing both interpretability and expressivity [68].

A significant challenge in symbolic regression is determining unknown numerical constants, which limits the performance of existing techniques. Addressing this issue is crucial for improving the accuracy and reliability of symbolic regression models [68]. The introduction of constrained optimization methods has further contributed to developing interpretable models, particularly in discovering partial differential equations (PDEs) from data.

The integration of feature selection and transformation techniques effectively manages high-dimensional data, providing a new hybrid approach that enhances interpretability. Additionally, the development of symbolic metamodels addresses the issue of interpreting black-box machine learning models by proposing a method for finding interpretable metamodels using innovative mathematical frameworks. This approach overcomes the limitations of traditional methods that depend on fixed building blocks, which can restrict the model's expressiveness and often overlook complex interactions among multiple features. By integrating Lp regularization with neural networks, this method enhances the ability to discover interpretable and predictive models while allowing for a nuanced understanding of feature interactions, thereby facilitating the automatic discovery of material models and advancing scientific knowledge in various domains [32, 43, 64, 42, 69].

Automated model discovery and interpretability increasingly focus on developing methodologies that prioritize transparency and robustness. By integrating symbolic regression with other computational techniques, researchers are advancing machine learning systems' capabilities to produce models that are both accurate and comprehensible, facilitating their application across diverse scientific and engineering domains. The ongoing development of self-improvement strategies in language models further underscores the potential for future AI advancements, enhancing model discovery processes [62].

In recent years, the exploration of symbolic and sparse regression has gained significant attention within the academic community, particularly due to its diverse applications and the innovative methodologies that have emerged. To elucidate this complex landscape, Figure 2 presents a comprehensive overview of the hierarchical structure of symbolic and sparse regression. This figure illustrates not only the methodologies and applications but also the recent advancements and challenges faced in the field. Notably, it highlights key techniques such as genetic programming and sparse data-driven discovery, while also addressing their applications in engineering and scientific discovery. Furthermore, the figure showcases recent advancements like the SPRINT algorithm, alongside the challenges posed by expansive search spaces and computational costs. This visual representation serves to enhance our understanding of the intricate relationships and dynamics that characterize this area of research.

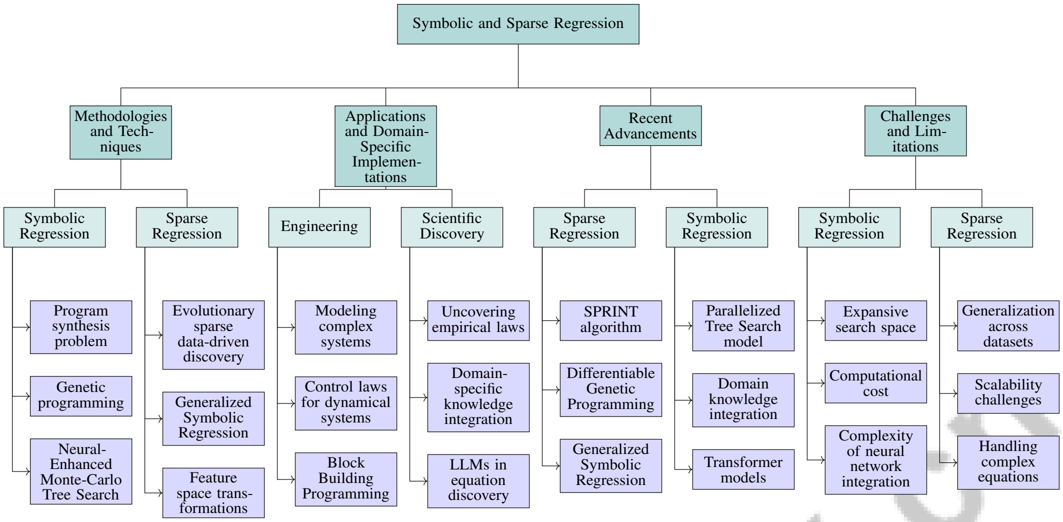

Symbolic and Sparse Regression
Methodologies and Techniques
Applications and Domain-Specific Implementations
Recent Advancements
Challenges and Limitations
Symbolic Regression
Sparse Regression
Engineering
Scientific Discovery
Sparse Regression
Symbolic Regression
Symbolic Regression
Sparse Regression
Program synthesis problem
Genetic programming
Neural-Enhanced Monte-Carlo Tree Search
Evolutionary sparse data-driven discovery
Generalized Symbolic Regression
Feature space transformations
Modeling complex systems
Control laws for dynamical systems
Block Building Programming
Uncovering empirical laws
Domain-specific knowledge integration
LLMs in equation discovery
SPRINT algorithm
Differentiable Genetic Programming
Generalized Symbolic Regression
Parallelized Tree Search model
Domain knowledge integration
Transformer models
Expansive search space
Computational cost
Complexity of neural network integration
Generalization across datasets
Scalability challenges
Handling complex equations

Figure 2: This figure illustrates the hierarchical structure of symbolic and sparse regression, detailing methodologies, applications, recent advancements, and challenges. It highlights key techniques like genetic programming and sparse data-driven discovery, applications in engineering and scientific discovery, recent advancements like the SPRINT algorithm, and challenges such as expansive search spaces and computational costs.

3 Symbolic and Sparse Regression

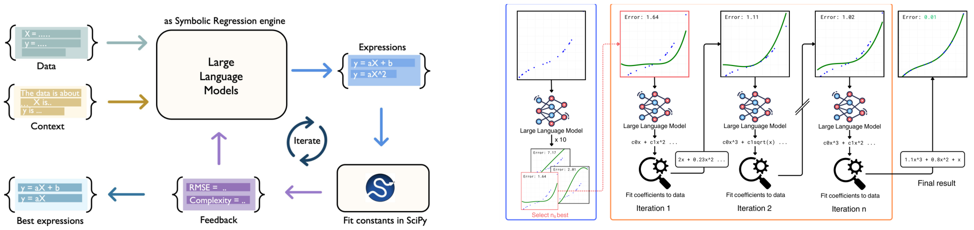
3.1 Methodologies and Techniques

Symbolic and sparse regression are pivotal in deriving interpretable mathematical models from complex datasets, advancing machine learning and AI. These methodologies focus on optimizing model discovery for accuracy and interpretability. Symbolic regression, framed as a program synthesis problem, explores extensive hypothesis spaces without predefined structures, balancing model complexity and accuracy through genetic programming [70]. Neural-Enhanced Monte-Carlo Tree Search (NG-MCTS) exemplifies this by using neural networks to generate production rules and Monte Carlo Tree Search to identify fitting expressions, enhancing efficiency [1].

Sparse regression techniques, such as the evolutionary sparse data-driven discovery method, integrate evolutionary algorithms with symbolic regression to extract equations of motion from high-dimensional datasets, emphasizing minimal feature subsets [52]. The Generalized Symbolic Regression (GSR) method discovers relationships between input features and transformed target variables using expressions like $g(y) = f(x)$, combining genetic programming with matrix-based encoding [31]. Feature space transformations, such as LCF nodes in expression trees, refine symbolic regression by enhancing data representation [68]. The Parallelized Tree Search (PTS) model further enhances symbolic regression by distilling expressions from data, improving evaluation efficiency [71].

These methodologies integrate advanced techniques like deep learning and transformers, facilitating the extraction of interpretable expressions from complex datasets. Comparative analyses of techniques such as Bayesian optimization, reinforcement learning, and evolutionary algorithms demonstrate their varying effectiveness across different automated ML categories [62]. This progress supports the development of models that are accurate and insightful, proving valuable across scientific and engineering applications, including equation recovery and expression optimization for improved generalization and efficiency [72, 10, 7, 73].

Symbolic and sparse regression enable the extraction of meaningful expressions from complex datasets. Figure 3 illustrates methodologies like using LLMs to derive symbolic expressions and training deep learning models, highlighting the integration of advanced ML techniques with symbolic and sparse regression to enhance accuracy and interpretability [61, 6].



(a) Large Language Models as Symbolic Regression Engine[61]

(b) Deep Learning Model Training Process[6]

Figure 3: Examples of Methodologies and Techniques

3.2 Applications and Domain-Specific Implementations

Symbolic and sparse regression techniques are increasingly utilized across fields due to their capacity to derive clear mathematical models from intricate datasets. Symbolic regression facilitates the discovery of succinct expressions that capture variable relationships, bridging simple linear models and complex neural networks. Recent advancements, including deep learning integration and novel algorithms like Generalized Symbolic Regression and Transformer models, have enhanced their applicability and efficiency, proving valuable in scientific discovery and problem-solving [10, 7, 72].

In engineering, symbolic regression aids in modeling and optimizing complex systems, such as control laws for dynamical systems, enhancing control strategies [70]. The Block Building Programming (BBP) approach optimizes target functions by dividing them into blocks, streamlining the modeling process [74].

In scientific discovery, symbolic regression uncovers empirical laws from experimental data. Researchers derive equations characterizing natural phenomena by integrating domain-specific knowledge with data-driven modeling, enhancing understanding across disciplines like physics, chemistry, and biology. LLMs aid in discovering equations that fit data while aligning with scientific principles, transforming inquiry and accelerating knowledge advancement [2, 32, 25, 66].

Sparse regression is impactful in high-dimensional data environments, crucial in genomics and bioinformatics for identifying genetic markers vital for understanding diseases. Machine learning techniques enhance predictive capabilities, uncovering hidden relationships and advancing knowledge of genetic influences on health [75, 32]. Sparse regression enhances model interpretability and accuracy by focusing on minimal feature subsets, aiding in targeted therapies.

In finance and economics, symbolic and sparse regression methodologies model market behaviors and economic trends. Techniques like Neural-Enhanced Monte-Carlo Tree Search (NEMoTS) improve efficiency and generalizability, resulting in accurate models for understanding complex financial dynamics. Algorithms combining symbolic regression with linear regression techniques enhance model accuracy and efficiency, providing robust frameworks for analyzing datasets [58, 76]. Interpretability is crucial for decision-makers requiring transparent insights.

The diverse applications of symbolic and sparse regression highlight their versatility in addressing complex challenges across fields. By developing interpretable and accurate models, these techniques drive progress in scientific research, engineering optimization, and data-driven decision-making. AI-assisted frameworks enhance the formulation of quantitative models, bridging parametric and non-parametric approaches, underscoring the role of ML and AI in advancing discovery and optimizing processes [42, 32, 65].

3.3 Recent Advancements

Recent advancements in symbolic and sparse regression have enhanced methodologies and broadened applications, leading to more interpretable and precise models. The SPRINT algorithm is a breakthrough in sparse regression, reducing computational complexity and making it feasible for large datasets [34]. Differentiable Genetic Programming (DGP) has revolutionized symbolic regression by introducing a differentiable symbolic tree structure, optimizing genetic programming in continuous space and enhancing performance on high-dimensional tasks [44]. The Generalized

AI-generated, for reference only.

Method Name	Methodological Innovations	Performance Enhancements	Application Scope
SPRINT[34]	Sprint Algorithm	Computational Efficiency	Large Symbolic Libraries
DGP[44]	Differentiable Symbolic Tree	Gradient-based Optimization	High-dimensional Datasets
GSR[31]	Matrix-based Encoding	High Recovery Rates	Complex Relationships
PTS[71]	Parallelized Tree Search	Accuracy Improvement	Interdisciplinary Domains
LCF[77]	Lcf Nodes	Improve Performance	Complex Functions
DSRA[41]	Context-free Grammar	Deterministic Algorithm	Synthetic Benchmark Datasets

Table 1: Overview of recent methodological innovations, performance enhancements, and application scopes in symbolic and sparse regression methods. The table highlights six advanced techniques, detailing their unique contributions to computational efficiency, optimization, accuracy, and applicability across various domains.

Symbolic Regression (GSR) approach demonstrates competitive performance with high recovery rates and low RMSE on benchmarks [31]. Table 1 presents a comprehensive summary of recent advancements in symbolic and sparse regression, showcasing the methodological innovations and performance enhancements of several prominent methods.

The Parallelized Tree Search (PTS) model achieves significant accuracy improvements and speedups compared to existing methods [71]. Incorporating domain knowledge into symbolic regression improves performance, particularly in handling complex data relationships [77]. The deterministic symbolic regression algorithm competes with genetic programming, achieving reliable and interpretable models [41]. Recent advancements include innovations like Generalized Symbolic Regression (GSR), a Transformer model for scientific data analysis, and the PTS model, which improve accuracy and efficiency in equation discovery. These enhancements address computational challenges and facilitate insights from complex datasets, fostering broader utilization of symbolic regression in interdisciplinary research [72, 10, 78, 71, 7]. Integrating innovative techniques like reinforcement learning emphasizes the potential for these methodologies to tackle complex challenges in ML and AI.

3.4 Challenges and Limitations

Method Name	Computational Complexity	Model Interpretability	Generalization Capability
SymTree[45]	Computational Constraints	Simpler, Interpretable Models	Rugged Search Space
NEEP[47]	High Computational Cost	Overly Intricate Solutions	Reduced Performance
LPM[79]	Computational Cost Significant	Improving Interpretability Decision-making	-
GOSR[37]	Exhaustive Searches Impractical	Simplest Mathematical Expression	Rediscovering Known Laws
DN-CL[80]	Substantial Computational Resources	-	Enhanced Robustness TO Noise
TPSR[49]	High Computational Costs	Overly Intricate Solutions	Better Generalization
mSNGP[81]	Population-based Genetic	Expansive Search Space	Constraint Sampling Process
SGPT[82]	Faster Inference Times	Overly Intricate Solutions	Diverse OR Noisy
SR4MDL[78]	Computational Efficiency	Complexity OF Models	Noisy Data
LCF[77]	High Computational Cost	Expansive Search Space	Reduced Performance
AL-SRPC[30]	High Dimensionality	Interpretability And Complexity	Resilient TO Noise
PTS[71]	High Computational Efficiency	Concise And Interpretable	Limited Data Generalization
Racing-CVGP[70]	Extensive Training Time	-	-
NG-MCTS[1]	High Computational Cost	Expansive Search Space	Generalize Better

Table 2: Comparison of symbolic and sparse regression methods based on computational complexity, model interpretability, and generalization capability. The table highlights the diverse challenges faced by each method, such as high computational costs, model intricacy, and varying abilities to generalize across datasets. This comparison provides insights into the trade-offs involved in selecting appropriate regression techniques for different applications.

Table 2 presents a comparative analysis of various symbolic and sparse regression methods, illustrating the challenges and limitations in terms of computational complexity, interpretability, and generalization capability. Symbolic and sparse regression methodologies, while invaluable for deriving interpretable models, face challenges that can limit their effectiveness. A primary issue in symbolic regression is the expansive search space, often resulting in overly complex models that compromise interpretability [45]. The discrete nature of traditional methods exacerbates this, creating sharp fitness landscapes and insufficient neighborhood information, complicating solution space navigation [47]. The computational cost of methods like the likelihood profile approach can be prohibitive for large datasets due to multiple optimization runs [79].

AI-generated, for reference only.

Sparse regression techniques encounter limitations in generalizing across diverse datasets, particularly with high-dimensional data or complex formulations. The reliance on input data quality is a constraint, as excessive noise can hinder accurate model construction, evident in methods like RLS regression [22]. Scalability of approaches like Branch-and-Bound in MINLP is a limitation, struggling with larger datasets [37].

Integrating neural networks with symbolic regression introduces challenges. For instance, the DNC-L method requires substantial resources due to complexity and multiple encoders [80]. Despite improvements, approaches like TPSR may struggle with complex or noisy datasets, leading to overfitting [49]. Ensuring models pass constraint checks across the domain, not just sampled points, is critical in symbolic regression-driven training [81].

Handling complex equations or datasets with irregular distributions remains challenging, even for models like Symbolic GPT leveraging GPT technology [82]. Methods like MDLformer may experience decreased performance with complex data relationships [78]. Configurations using globally synchronized LCFs have shown suboptimal performance, indicating drawbacks in sharing transformations [77]. The effectiveness of physical constraints in symbolic regression can vary; some constraints may hinder optimization [30].

The Parallelized Tree Search model faces challenges related to increasing memory demands and limited heuristic guidance [71]. Sub-optimal experiment schedule selections can hinder regression, increasing training time and decreasing performance [70].

Addressing these challenges requires ongoing research to enhance robustness, scalability, and interpretability of symbolic and sparse regression methods. This involves developing novel techniques leveraging domain knowledge, optimizing computational efficiency, and improving feature library expressiveness. Despite strengths, methods like NG-MCTS may face challenges with complex expressions or significant deviations from training conditions [1]. Contemporary symbolic regression methods may face limitations like overfitting and challenges in generalizing findings across regression types [33].

4 Equation Discovery and Genetic Algorithms

The integration of computational techniques in equation discovery is pivotal for addressing the complexities inherent in symbolic regression. This section explores the significance of multi-objective optimization, a key strategy balancing competing objectives, enhancing model accuracy and interpretability. Understanding multi-objective optimization nuances leads to robust solutions in this field. The following subsection delves into specific applications, methodologies, and implications of multi-objective optimization in equation discovery.

4.1 Multi-Objective Optimization in Equation Discovery

Multi-objective optimization is crucial in symbolic regression, tackling the dual challenges of optimizing model accuracy and interpretability. The complexity of symbolic regression necessitates balancing these objectives to explore parsimonious functional relationships [27]. Advanced methodologies, such as Neural-Enhanced Monte-Carlo Tree Search (NEMoTS), integrate neural networks with Monte-Carlo Tree Search, significantly reducing search space and enhancing fitting capabilities [58].

The Generative Flow Network for Symbolic Regression (GFN-SR) showcases the potential of generative models to produce diverse candidate solutions, particularly in noisy environments where traditional methods may falter [83]. This diversity is essential for exploring the expansive hypothesis spaces characteristic of symbolic regression, ensuring discovered models maintain both accuracy and interpretability.

Innovative approaches like DISCOVER demonstrate the ability to uncover complex partial differential equations (PDEs) without extensive prior knowledge, emphasizing computational efficiency and scalability to high-dimensional problems [84]. This highlights the importance of multi-objective optimization in managing trade-offs between model complexity and computational resources, enhancing symbolic regression's applicability across various scientific domains.

AI-generated, for reference only.

The integration of evolutionary strategies within neural networks, exemplified by SRNet, underscores the role of multi-objective optimization in adaptively searching for mathematical expressions that approximate relationships modeled by neural networks [29]. This synergy between evolutionary computation and symbolic regression facilitates the discovery of interpretable and robust models.

Thus, multi-objective optimization is pivotal in advancing equation discovery, enabling the development of models that balance complexity and interpretability while maintaining high accuracy. By integrating advanced computational techniques and domain-specific knowledge, these methodologies expand the frontiers of symbolic regression, offering robust solutions to the intricate challenges of model discovery [25].

4.2 Genetic Programming Dynamics and Operator Effectiveness

Genetic programming (GP) is a powerful evolutionary computation technique that simulates natural selection to evolve programs or models represented as tree structures. The dynamics of GP involve iterative application of genetic operators such as selection, crossover, and mutation, guiding populations toward optimal solutions. The effectiveness of these operators is crucial for GP's success in symbolic regression and other applications [85].

The Subtree Tracing Methodology (STM) advances understanding of GP dynamics by identifying the origins of genes within genetic programs and quantifying genetic operators' effectiveness. By tracing specific subtrees' contributions to overall fitness, STM provides insights into the most beneficial operators for evolving high-quality models, optimizing GP processes for greater efficiency in discovering accurate and interpretable models [85].

In symbolic regression, multi-objective evolutionary algorithms simultaneously optimize the accuracy and complexity of discovered differential equations. This approach balances trade-offs between model fidelity and interpretability, ensuring resulting equations are precise and comprehensible. Integrating multi-objective optimization into GP frameworks enhances their capability to discover meaningful mathematical expressions capturing complex system dynamics [57].

Innovative methodologies, such as employing analytic continued fractions combined with a memetic algorithm, offer alternative representations to traditional tree-based GP. This approach diverges from conventional methods by utilizing analytic continued fractions, yielding more compact and interpretable solutions. The memetic algorithm further refines these solutions through local search strategies, enhancing GP performance [86].

The Neural Symbolic Regression with Hypotheses (NSRwH) framework exemplifies integrating privileged information into GP processes. By conditioning predictions on desired properties of output expressions, NSRwH enhances control over solution evolution, allowing for models aligning closely with user-defined criteria. This approach underscores the importance of incorporating domain knowledge into GP to improve discovered models' relevance and applicability [87].

Advancements in symbolic regression depend on understanding genetic programming dynamics and operator efficacy, crucial for optimizing search processes and improving generated mathematical models' quality. Recent studies emphasize analyzing evolutionary dynamics, including selection pressure and genetic diversity, and integrating neural network methodologies to enhance symbolic regression solutions' interpretability and performance. Combining GP with quality-diversity strategies shows promise in maintaining exploration and avoiding premature convergence and bloat, highlighting GP dynamics' vital role in symbolic regression's ongoing development [88, 89, 90, 85, 91]. By exploring innovative methodologies and integrating multi-objective optimization and privileged information, researchers enhance GP's capability to discover robust and interpretable models across various scientific and engineering domains.

4.3 Innovative Genetic Programming Techniques

Innovative genetic programming techniques have significantly advanced symbolic regression by enhancing solution diversity, improving complex equation structure handling, and increasing noise resistance during the discovery process [57]. A notable innovation is the subtree tracing approach, providing a detailed view of how genetic operators influence solution evolution, offering insights into genetic operations' effectiveness and enabling refined genetic programming processes for more efficient model discovery [85].

Hybrid approaches, such as integrating genetic programming (GP) with Functional Tree Generators (FTG), represent a promising research direction. This hybridization aims to create algorithms effectively spanning subspaces while optimizing constants, enhancing symbolic regression techniques' robustness and applicability across diverse datasets [92].

The SBP-GEP (Structure-Based Programming Gene Expression Programming) framework exemplifies an innovative approach outperforming traditional Gene Expression Programming (GEP) in accuracy and robustness, particularly in noisy environments. This advancement highlights the potential of integrating structural constraints into genetic programming to improve model reliability and precision [53].

Moreover, the Generalized Symbolic Regression (GSR) method broadens symbolic regression's scope by seeking mathematical relationships in the form of $g(y) = f(x)$, where g is a transformation of the target variable. This approach enhances symbolic models' flexibility and expressiveness, facilitating discovering more complex and meaningful relationships within data [31].

Recent advancements in genetic programming significantly enhance symbolic regression by introducing innovative methodologies integrating mathematical models' interpretability with modern machine learning techniques' predictive power. These developments aim to produce succinct and interpretable mathematical expressions directly from data while addressing traditional genetic programming limitations, such as slow convergence in large-scale problems. New hybrid methods combining deep learning with genetic programming are explored to improve feature extraction and guide the symbolic regression process. Techniques like the divide and conquer method and deterministic algorithms are implemented to optimize target function discovery efficiently. These innovations pave the way for more accurate and interpretable models, bridging the gap between machine learning and scientific discovery [93, 10, 91, 94]. By integrating advanced computational strategies and leveraging domain-specific knowledge, these approaches enhance genetic programming's ability to address complex challenges in machine learning and artificial intelligence.

4.4 Hybrid Approaches in Equation Discovery

Hybrid approaches in equation discovery represent significant advancements, combining various computational techniques to enhance the accuracy, efficiency, and interpretability of discovered models. These methodologies integrate elements from genetic programming, symbolic regression, and expert domain knowledge to tackle equation discovery's inherent challenges, including high-dimensional data structures and model selection optimization. For instance, the LLM-SR approach utilizes Large Language Models to merge scientific priors with evolutionary search techniques, improving accurate equation discovery efficiency. Incorporating soft constraints into symbolic regression frameworks generates meaningful expressions aligned with theoretical expectations while enhancing search effectiveness. Additionally, applying probabilistic grammars in equation discovery introduces a flexible way to encode prior knowledge, facilitating efficient exploration of the equation space. These integrated methodologies produce interpretable and scientifically valid mathematical expressions that better represent complex phenomena across disciplines [43, 10, 36, 2, 95].

Integrating expert knowledge into the optimization process has shown promise in refining candidate selection and improving discovery efficiency. This approach leverages domain-specific insights to guide the search process, ensuring generated models are accurate and relevant to the specific application context [57].

The Taylor Genetic Programming (TaylorGP) framework exemplifies a hybrid approach combining genetic programming principles with Taylor series expansions to approximate complex functions. While TaylorGP offers a novel symbolic regression method, it faces limitations in high-dimensional datasets where low-order Taylor polynomials may inadequately capture global features, highlighting the need for further refinement and technique integration to enhance its effectiveness [96].

Another innovative hybrid methodology involves using constraints to guide the genetic programming process, as seen in the SBP-GEP framework. This approach incorporates structural constraints into genetic program evolution, improving model robustness and precision, particularly in noisy environments. However, enforcing these constraints may increase computational overhead, potentially slowing overall evaluation time [53].

These hybrid approaches underscore the transformative potential of integrating diverse computational strategies, such as Large Language Models (LLMs) and probabilistic grammars, to enhance scientific equation discovery's efficiency and effectiveness, addressing traditional symbolic regression methods' shortcomings that often overlook valuable domain-specific knowledge [2, 36, 4]. By integrating expert knowledge, leveraging advanced mathematical techniques, and incorporating structural constraints, researchers push the boundaries of symbolic regression and equation discovery, paving the way for more sophisticated and interpretable models across scientific and engineering domains.

5 Reinforcement Learning and Large Language Models

5.1 Principles of Reinforcement Learning

Reinforcement learning (RL) is a subfield of machine learning focused on training agents to make decisions by interacting with their environment to maximize cumulative rewards. RL's core components include agents, environments, actions, states, and rewards. Agents perceive environmental states and select actions that transition to new states while receiving feedback in the form of rewards, facilitating the learning of optimal policies through trial and error [97]. A critical aspect of RL is managing the exploration-exploitation trade-off. Exploration involves testing new actions to uncover their effects, while exploitation focuses on leveraging known actions to maximize rewards. Effective exploration strategies are essential for discovering optimal policies in complex environments with extensive state spaces [98]. Techniques such as policy gradient methods and Q-learning are employed to optimize agents' policies, with Q-learning specifically updating the value of action-state pairs to inform decision-making [55].

The integration of Monte Carlo Tree Search (MCTS) within RL frameworks, as demonstrated in the RSRM approach, enhances expression generation and mitigates overfitting via double Q-learning, effectively combining the strengths of MCTS and Q-learning [99]. RL also incorporates predictive modeling principles, exemplified by methods like GPF, which capture complex, non-linear relationships between dataset characteristics and classifier accuracy [64]. RL's versatility is further illustrated by approaches such as Algorithmic Distillation, enabling the training of general RL policies on extensive offline data, as seen with the FormulaGPT model [100]. Recent advancements, including the Quadratic Q-Weighted model, reveal non-linear patterns in reward prediction errors, enhancing our understanding of human learning and decision-making intricacies [98, 55]. The principles of RL, comprising exploration-exploitation balance, policy optimization, and integration with computational techniques, underpin its efficacy in addressing dynamic decision-making and optimal control challenges across various domains.

5.2 Applications of Reinforcement Learning

Reinforcement learning (RL) has diverse applications, demonstrating its effectiveness in complex decision-making scenarios. In autonomous control systems, RL constructs value functions for continuous state and input spaces, essential for developing effective control policies [97]. In robotics, RL enables robots to learn and adapt to dynamic environments, enhancing navigation, manipulation, and human interaction capabilities. In game playing, RL has achieved significant milestones, notably with AlphaGo, which mastered complex board games through self-play and deep learning. Recent innovations in RL have improved strategy game applications, allowing agents to develop adaptive strategies against opponents. The Quadratic Q-Weighted model captures reward prediction complexities and human decision-making, surpassing traditional linear models in predictive accuracy [67, 55].

In finance, RL is applied to algorithmic trading and portfolio management, modeling the financial market as a dynamic system. RL algorithms derive optimal trading strategies that maximize returns while incorporating risk management techniques, enhancing interpretability and efficiency in decision-making [99, 55, 97]. This adaptability allows RL to respond to changing market conditions using historical and real-time data. In healthcare, RL optimizes treatment strategies and personalizes patient care by analyzing patient data and treatment outcomes. Traditional linear RL models may oversimplify the complex relationship between behavior and rewards, while innovative approaches like the Quadratic Q-Weighted model reveal non-linear dynamics in reward prediction, leading to more accurate patient response predictions [64, 55, 22]. This application highlights RL's potential in improving decision-making in medical diagnostics and treatment planning.

Moreover, RL is utilized in transportation for autonomous vehicle navigation and traffic management. By simulating complex traffic environments, RL algorithms optimize routing and scheduling strategies, minimizing congestion and enhancing efficiency. This optimization leverages techniques like symbolic regression and data-driven equation discovery to model dynamic systems accurately [55, 98, 101, 20]. The wide-ranging applications of RL underscore its transformative potential in addressing intricate challenges across various fields, including symbolic regression, scientific discovery, and human behavior modeling. Techniques like FormulaGPT and the Reinforcement Symbolic Regression Machine demonstrate how RL enhances mathematical modeling efficiency and accuracy, paving the way for significant breakthroughs in science and engineering [102, 32, 55, 98, 100]. By refining decision-making processes through dynamic interactions, RL offers robust solutions that enhance efficiency, adaptability, and performance in real-world applications.

5.3 Development and Impact of Large Language Models

The rise of large language models (LLMs), particularly transformer-based architectures like GPT-4, has transformed natural language processing (NLP) by enabling machines to comprehend and generate human language with remarkable precision. LLMs excel in various tasks, including symbolic regression, where they derive accurate mathematical expressions from datasets using natural language prompts and scientific context. Recent studies illustrate that LLMs can rediscover well-known scientific equations and generate expressions that meet specific requirements, effectively integrating theoretical knowledge with empirical data [61, 19]. Their extensive training on vast text corpora has revolutionized applications ranging from machine translation to scientific discovery.

A notable advancement in LLMs is their integration with symbolic regression and equation discovery, exemplified by LLM-SR, which employs GPT-3.5-turbo and Mixtral-8x7B-Instruct as backbone models [2]. This integration enhances the derivation of scientific equations, improving model interpretability and usability in complex domains. Experiments on nonlinear systems described by partial differential equations (PDEs) and ordinary differential equations (ODEs), such as Burgers' equation and the Navier-Stokes equation, demonstrate LLMs' capability to model intricate scientific phenomena.

The impact of LLMs extends to symbolic regression, where methods like In-Context Symbolic Regression (ICSR) leverage large pre-trained models without task-specific training, offering flexibility and efficiency [6]. This capability enhances model discovery processes, resulting in compact and interpretable models that outperform traditional numerical approaches [97]. Additionally, the integration of stochastic context-free grammars (SCFG) into linear genetic programming (LGP) exemplifies the dynamic evolution of probability distributions within LLM frameworks, improving the search for symbolic regression solutions [103]. This adaptability allows LLMs to evolve complex models that generalize well to out-of-sample data, as demonstrated by the superior performance of the QLattice symbolic regressor [104].

LLMs' versatility is further illustrated in predictive modeling, where models like the Quadratic Q-Weighted model have shown superior predictive accuracy across multiple datasets, providing new insights into human learning complexities [55]. These advancements underscore LLMs' transformative potential in enhancing machine learning systems, equipping researchers with robust tools for tackling complex linguistic and computational challenges across various scientific and engineering fields.

5.4 Integration of Reinforcement Learning and Large Language Models

The integration of reinforcement learning (RL) and large language models (LLMs) marks a significant advancement in artificial intelligence, merging RL's decision-making capabilities with LLMs' linguistic prowess to create systems that understand, generate, and respond to complex human language in dynamic environments. This synergy between machine learning (ML) and symbolic reasoning enhances AI systems' capacity to learn and adapt in real-time. By combining data-driven techniques with formal logical frameworks, these systems effectively tackle complex challenges across cognitive psychology, social sciences, and scientific discovery, facilitating interpretable model extraction and the discovery of fundamental laws governing various phenomena [65, 32, 67, 66].

Key innovations in this integration include methods that evolve ordinary differential equations (ODEs) from data, fine-tuning parameters using gradient-based optimization, which is crucial for accurately

AI-generated, for reference only.

modeling dynamics in complex systems [105]. Incorporating RL techniques optimizes the learning process, refining models through iterative feedback and adaptation. The introduction of feedback gates into Markov Brains exemplifies the potential of RL and LLM integration, allowing systems to generate internal feedback based on actions. This mechanism facilitates learning during their lifetime without external signals, enhancing AI models' autonomy and adaptability [106]. Such innovations highlight RL's role in providing continuous learning capabilities, improving LLM performance in dynamic environments.

The development of symbolic regression methods addressing the Bellman equation further underscores the synergy between RL and LLMs, offering novel approaches to finding value functions that enhance model interpretability and accuracy in complex decision-making scenarios [97]. Additionally, the Quadratic Q-Weighted model offers a framework for understanding how individuals learn from rewards, accounting for nonlinear dynamics and biases in reward estimation, thereby enhancing LLM learning processes [55].

The integration of RL and LLMs significantly enhances AI systems' adaptability, interpretability, and learning efficiency by enabling them to leverage existing knowledge, optimize symbolic regression tasks, and incorporate domain-specific insights through natural language instructions. This combination facilitates efficient exploration of combinatorial search spaces, improving performance in tasks such as symbolic optimization and automatic machine learning, ultimately advancing AI capabilities in real-world applications [107, 5, 19, 26, 61]. By harnessing the strengths of both methodologies, researchers are developing innovative solutions to complex challenges across diverse fields, paving the way for more sophisticated and responsive AI technologies.

5.5 Recent Breakthroughs and Ongoing Research

Recent breakthroughs in symbolic regression and reinforcement learning have significantly advanced machine learning systems, offering new methodologies and insights to address longstanding challenges. Notable advancements include the SNIP framework, which outperforms traditional supervised models in symbolic regression tasks, especially in data-scarce scenarios [108]. This innovation highlights the potential of novel approaches to overcome data limitations common in machine learning applications. In genetic programming, the introduction of -lexicase selection with automatic threshold adaptation has markedly improved regression tasks by selecting candidate solutions based on performance across diverse cases, enhancing robustness and accuracy [109].

The evoNSGA-II algorithm represents another significant advancement, addressing evolvability degeneration in multi-objective genetic algorithms. By outperforming traditional NSGA-II and other algorithms, evoNSGA-II facilitates the discovery of larger and more accurate solutions, underscoring the importance of maintaining evolvability in evolutionary computation [110]. In automated machine learning, the AlphaD3M framework has demonstrated competitive performance, significantly outpacing AutoSklearn in speed while achieving comparable results. This advancement emphasizes the potential for automated systems to streamline the model discovery process, providing efficient solutions to complex machine learning tasks [26].

Ongoing research continues to explore challenges associated with off-policy training methods in reinforcement learning, particularly focusing on the intractability of certain benchmarks [90]. Addressing these challenges is crucial for enhancing the scalability and applicability of RL algorithms across diverse domains. Recent advancements and ongoing research in symbolic regression and reinforcement learning signify a significant evolution in machine learning systems. The introduction of the Reinforcement Symbolic Regression Machine (RSRM) showcases a sophisticated approach to automatically deriving complex mathematical equations from limited data. The RSRM integrates three innovative modules: a Monte Carlo tree search agent for optimal exploration of mathematical expression trees, a Double Q-learning block for efficient navigation of the search space, and a modulated sub-tree discovery mechanism to enhance mathematical operation representation. These efforts not only address existing challenges in symbolic regression but also pave the way for more interpretable and computationally efficient models, demonstrating superior performance over traditional methods across various benchmark datasets and application domains [102, 10, 111]. By integrating advanced computational techniques and addressing key challenges, researchers continue to push the boundaries of artificial intelligence, offering new possibilities for innovation and discovery across scientific and engineering fields.

6 Automated Model Discovery and Interpretability

As machine learning progresses, automated model discovery and interpretability have gained prominence. This section delves into methodologies that enhance model interpretability, particularly through symbolic regression, emphasizing its significance and innovative approaches. The subsequent subsection will explore the importance of interpretability and techniques to achieve it, underscoring their critical role in ensuring transparency and reliability in machine learning applications.

6.1 Significance and Techniques for Enhancing Interpretability

Interpretability is crucial in machine learning, especially in sectors like healthcare, finance, and scientific research, where understanding model decisions is vital. Symbolic regression is valued for generating interpretable mathematical expressions that elucidate complex data relationships, thus enhancing model credibility and utility. The framework by [84] incorporates domain knowledge through customized constraints, improving equation quality and prioritizing interpretability as a key aspect of model utility.

Various methodologies have been developed to enhance interpretability by combining symbolic representations with advanced computational techniques. The Multi-objective Memetic Evolutionary Strategy (MOMES) integrates constant learning with evolutionary processes, enabling the discovery of complex data relationships while maintaining interpretability [68]. This approach highlights the need to balance complexity and interpretability, ensuring comprehensibility alongside intricate data pattern capture.

The Parallelized Tree Search (PTS) model excels at discovering accurate, parsimonious expressions from limited data within a short timeframe, demonstrating superior efficiency and recovery rates [71]. This efficiency is vital for applications requiring rapid model deployment and understanding. Additionally, SRNet provides a comprehensive method for explaining all neural network layers using explicit mathematical representations, significantly enhancing prediction interpretability [29].

The Racing Control Variable Genetic Programming (Racing-CVGP) method dynamically selects and prioritizes effective experiment schedules, resulting in faster discovery times and improved expression quality [70]. This dynamic approach enhances interpretability by focusing on relevant data transformations and relationships. Moreover, RBG2-SR generates interpretable, domain-consistent symbolic expressions, outperforming other methods in various benchmarks [5]. Techniques like SymPDE provide accurate closed-form symbolic solutions with high interpretability, addressing limitations faced by existing numerical methods [8].

The integration of advanced computational techniques and domain-specific knowledge continues to bolster the interpretability and utility of machine learning models. By emphasizing transparency and robustness, these advancements collectively underscore interpretability's importance in machine learning, facilitating the development of accurate and comprehensible models across diverse domains [62].

6.2 Balancing Complexity and Interpretability

Achieving a balance between complexity and interpretability in symbolic regression and machine learning models is a significant challenge affecting their effectiveness across various domains. While symbolic regression is recognized for generating interpretable models, it often struggles to match the predictive accuracy of traditional machine learning methods, highlighting the trade-off between model simplicity and predictive power [112].

A primary challenge in maintaining this balance is the limitation of existing regularization techniques, such as L1 and L2 regularization, which can introduce bias or fail to promote sparsity effectively [87]. These methods often compromise interpretability without enhancing accuracy, emphasizing the need for innovative approaches to navigate this trade-off. The QLattice method allows user-defined constraints to guide the search, balancing model complexity and interpretability [23].

Integrating neural networks with symbolic regression, as seen in methods like NeSymReS, enhances expressivity while preserving interpretability, especially with increasing data availability. However, this integration can lead to computational complexity and potential model overcomplication without substantial performance improvements, known as horizontal bloat [28].

AI-generated, for reference only.

Complexity-aware approaches like MOMES effectively discover algebraic expressions that generalize well to unseen data, achieving a balance between interpretability and accuracy [68]. However, some symbolic regression frameworks' pruning strategies may not yield concise expressions, as they may overlook optimal pruning paths due to their greedy nature [45].

The sensitivity of genetic programming-based symbolic regression methods to operator selection and training data volume complicates the complexity-interpretability balance [113]. This sensitivity can hinder the success of evolving correct analytical forms, underscoring the importance of careful operator selection and data preparation. Furthermore, the focus on limited meta-features in benchmark studies may fail to capture all relevant dataset quality aspects, potentially overlooking factors influencing model complexity and interpretability [21]. Future research should examine multivariate models and less restrictive grammars to enhance symbolic regression benchmarks' applicability and findings [95].

Balancing complexity and interpretability remains a significant challenge in developing symbolic regression and machine learning models, particularly as researchers explore advanced methods like deep symbolic regression, which employs transformers and breadth-first search to enhance robustness and interpretability. While traditional genetic programming methods have established a foundation for symbolic regression, recent deep learning advancements highlight the need for effective trade-offs between data fitness and expression complexity, as evidenced by using Bayesian information criterion (BIC) for model optimization. This ongoing evolution emphasizes developing methodologies that improve learning performance while ensuring interpretable outcomes across various application domains [10, 114]. Addressing this challenge requires novel methodologies to manage trade-offs effectively, ensuring models are both accurate and comprehensible.

6.3 Case Studies and Applications

Symbolic regression has demonstrated remarkable potential in delivering interpretable solutions across diverse domains, as highlighted by numerous case studies and applications. A notable application is the development of interpretable elastoplasticity models, where symbolic regression enhances robustness against non-smoothness and refines techniques to improve efficiency and accuracy [15]. This underscores symbolic regression's role in generating models that are both accurate and comprehensible, facilitating their application in complex material modeling.

Another compelling application involves the automatic discovery of network families, where symbolic regression identifies underlying processes governing empirical networks. This study reveals that similar mathematical expressions can describe different empirical networks, showcasing symbolic regression's utility in uncovering fundamental patterns across diverse structures [115].

The development of symbolic metamodels represents a significant advancement in interpreting black-box machine learning models. By providing clear and interpretable mathematical expressions for underlying data relationships, these metamodels outperform existing methods in approximating black-box functions, which is crucial for enhancing transparency and understanding in machine learning applications [116].

Benchmark studies play a vital role in evaluating symbolic regression algorithms, offering fair comparisons that highlight each approach's strengths and weaknesses in controlled settings. Such benchmarks are essential for advancing the field, providing insights into algorithm performance and guiding future research directions [117]. Future research should explore additional optimization methods and datasets to validate and refine these findings, ensuring the robustness and applicability of symbolic regression techniques across various domains [118].

The Vertical Symbolic Regression (VSR) framework has shown promise in enhancing model interpretability by integrating with deep learning methods or adapting to datasets where control experiments are challenging. This integration could significantly broaden VSR's applicability, facilitating its use in more complex and dynamic environments [119].

In the realm of deep generative symbolic regression, future research could investigate meta-learning frameworks to enhance the pre-training process, thereby reducing the search budget across diverse datasets [120]. This approach highlights symbolic regression's potential to evolve and adapt to new challenges, ensuring its continued relevance and effectiveness in addressing complex scientific and engineering problems.

The Multi-objective Memetic Evolutionary Strategy (MOMES) is notable for producing interpretable mathematical expressions that maintain high accuracy, making it suitable for complex applications such as space exploration [68]. Experimental results indicate that the Neural-Enhanced Monte-Carlo Tree Search (NG-MCTS) significantly outperforms existing symbolic regression methods, solving a higher percentage of expressions and achieving superior extrapolation accuracy [1].

The case studies and applications presented illustrate symbolic regression's transformative potential as a machine learning approach, excelling in generating interpretable and accurate mathematical models directly from data. This method addresses interpretability issues associated with traditional artificial neural networks and demonstrates versatility across diverse domains, including fundamental and applied sciences. Recent advancements, such as the development of a Transformer model and the innovative Symbolic Expression Transformer, underscore symbolic regression's efficiency and effectiveness in scientific discovery, achieving state-of-the-art results while mitigating computational costs [10, 7, 73]. By leveraging advanced computational techniques and integrating domain-specific knowledge, symbolic regression continues to enhance the interpretability and utility of machine learning models, paving the way for innovative solutions to complex challenges.

7 Evolutionary Computation

7.1 Overview of Evolutionary Computation

Evolutionary computation is a robust paradigm in artificial intelligence and machine learning, inspired by natural selection and genetic principles. It tackles complex optimization problems by simulating evolutionary processes, refining candidate solutions through selection, crossover, mutation, and reproduction. Techniques such as Genetic Programming and MAP-Elites enhance exploration and diversity, mitigating premature convergence, particularly in symbolic regression tasks. Innovations like Machine Learning-driven Distance-based Selection (DBS) optimize computational efficiency by reducing fitness evaluation costs on large datasets. Furthermore, integrating transfer learning within gene expression programming allows insights from prior optimizations to inform initial solutions, accelerating convergence [24, 89, 121].

Central to evolutionary computation is the population concept, where each individual represents a potential solution evaluated against an objective function. Fitness evaluation guides the evolutionary process, favoring fitter individuals for reproduction, akin to survival of the fittest. The efficiency of genetic variation operators is crucial for transferring beneficial traits and enhancing population fitness. Research shows that a few ancestor individuals significantly contribute to optimal solutions, underscoring selection pressure's role in maintaining genetic diversity and directing evolutionary dynamics [24, 85, 89].

Evolutionary computation excels in navigating vast search spaces, making it suitable for problems with rugged or poorly understood landscapes. Its adaptability is demonstrated in the automatic discovery of network families, where evolutionary algorithms classify network generators based on their performance in aligning with empirical data [115]. This versatility captures the underlying structures and dynamics of complex systems.

The field includes various algorithmic frameworks like genetic algorithms (GAs), genetic programming (GP), evolutionary strategies, and differential evolution, each employing unique methodologies for representation, selection, and variation. Genetic algorithms evolve solutions through structured genetic operations, while genetic programming focuses on evolving program structures. Evolutionary strategies adapt parameters in response to selection pressures, and differential evolution guides searches using candidate differences. Collectively, these frameworks enhance exploration in complex optimization problems, applied across domains from engineering to financial modeling and neural network architecture evolution [110, 24, 85, 122].

The flexibility of evolutionary computation addresses high-dimensional and multimodal optimization challenges. By leveraging natural evolution principles, these algorithms adopt dynamic problem-solving strategies that enhance exploration and maintain diversity. The integration of Genetic Programming, MAP-Elites, and Covariance Matrix Adaptation Evolution Strategy in symbolic regression exemplifies this adaptability, allowing knowledge transfer from previous optimizations to improve initial candidate solutions and reduce computational costs [24, 89].

AI-generated, for reference only.

Ongoing research aims to enhance the efficiency, scalability, and real-world applicability of evolutionary computation. Hybrid approaches merging evolutionary algorithms with complementary computational techniques, such as machine learning and symbolic regression, improve exploration and exploitation in problem-solving tasks. Recent advancements successfully combine genetic programming with transfer learning and natural language processing to optimize symbolic regression, enhancing convergence rates and result interpretability. Innovative frameworks incorporating learned concept libraries and advanced search strategies demonstrate superior performance in discovering new hypotheses and scaling laws, extending these methodologies' impact across fields from computational fluid dynamics to astrophysics [24, 89, 123, 91].

7.2 Genetic Programming in Dynamical Systems

Genetic programming (GP) is a potent tool for modeling and analyzing dynamical systems, offering a flexible approach to uncover governing equations and system behaviors. Leveraging evolutionary algorithms, GP evolves programs or mathematical expressions that describe complex dynamical phenomena without predefined structures. This capability is advantageous for exploring vast search spaces and identifying models capturing intricate temporal dynamics [44].

GP excels in modeling non-linear and high-dimensional systems, which are often challenging for traditional techniques. Using a tree-based representation, GP evolves expressions elucidating relationships between system variables, providing insights into underlying mechanisms. This is exemplified by GP's application in symbolic regression, deriving equations governing physical systems' dynamics, such as fluid flows and mechanical oscillations [46].

Recent advancements focus on enhancing efficiency and scalability in dynamical systems contexts. Integrating neural networks with GP bolsters symbolic regression robustness, enabling more accurate model discovery amidst noise and uncertainty [47]. Frameworks like SymbolNet dynamically prune model components during the GP process, optimizing both structure and parameters of evolved expressions [48].

GP's application extends beyond traditional modeling to areas like control system design and optimization. By evolving control laws and strategies, GP aids in developing adaptive and robust control systems responsive to changing environmental conditions and system parameters [45]. This adaptability is crucial for systems operating in dynamic environments where conventional control techniques may falter.

GP is integral to analyzing and optimizing dynamical systems, providing a robust framework for modeling complex behaviors. Innovations such as Differentiable Genetic Programming (DGP) enhance GP's capacity for high-dimensional symbolic regression through continuous data structures and gradient-based optimization. Methodologies like Control Variable Genetic Programming (CVGP) and Taylor Genetic Programming (TaylorGP) refine the approach, enabling incremental learning and feature extraction via polynomial approximations. These advancements elevate GP's efficiency in handling complex datasets and underscore its potential for interpretable machine learning across diverse applications [96, 85, 44, 124].

7.3 Enhancements in Evolutionary Algorithms

Recent advancements in evolutionary algorithms have broadened their capabilities, enhancing efficiency and applicability across complex optimization problems. The -lexicase selection method introduces a novel candidate solution selection approach in genetic programming, adapting thresholds automatically to improve robustness and accuracy in regression tasks [109]. By focusing on diverse selection criteria, -lexicase selection effectively navigates the search space, ensuring high-quality solution discovery.

The evoNSGA-II algorithm addresses evolvability degeneration in multi-objective genetic algorithms by maintaining a diverse population and encouraging exploration of larger solution spaces. This approach outperforms traditional NSGA-II and other algorithms, facilitating the discovery of more accurate and comprehensive solutions [110]. This enhancement underscores the importance of preserving evolvability in evolutionary computation, ensuring algorithm effectiveness over successive generations.

AI-generated, for reference only.

Incorporating domain knowledge into evolutionary algorithms has proven beneficial, as demonstrated by methods integrating expert insights into the optimization process. This approach refines candidate selection and improves the efficiency and relevance of discovered solutions, emphasizing the synergy between human expertise and computational techniques [57]. Domain-specific knowledge is particularly valuable in applications with complex constraints and objectives.

Hybrid approaches combining evolutionary algorithms with other computational techniques further enhance flexibility and effectiveness. For instance, integrating neural networks with genetic programming improves symbolic regression robustness, enabling accurate model discovery in noisy environments [46]. These hybrid methodologies leverage different paradigms' strengths, offering robust solutions to intricate optimization challenges.

Recent advancements in evolutionary algorithms, particularly through integrating transfer learning and machine learning techniques, reflect efforts to enhance performance, scalability, and adaptability across scientific and engineering fields. Transfer learning in gene expression programming allows reusing insights from previous optimizations, significantly improving convergence rates in symbolic regression tasks. Machine learning-driven distance-based selection algorithms for grammatical evolution reduce computational costs by optimizing test case selection, enhancing fitness evaluation efficiency across large datasets. These innovations address challenges faced by traditional evolutionary algorithms in diverse applications [24, 121]. By integrating innovative selection methods, preserving evolvability, and incorporating domain knowledge, researchers continue to push evolutionary computation boundaries, paving the way for sophisticated and adaptable optimization strategies.

7.4 Applications in Machine Learning Program Design

Evolutionary computation has advanced machine learning program design by optimizing model architectures, parameter tuning, and feature selection. Techniques like Gene Expression Programming (GEP) generate interpretable equations for regression tasks, while integrating transfer learning improves convergence rates by leveraging prior task insights. Innovations like the AI Programmer utilize genetic algorithms to autonomously generate software, showcasing evolutionary strategies' potential in automating complex programming tasks. Combining Grammatical Evolution with regression techniques facilitates feature engineering and symbolic regression, yielding accurate models with minimal human intervention. These advancements illustrate how evolutionary computation streamlines optimization processes and enhances machine learning applications' efficiency and effectiveness across domains [125, 24, 89, 122, 126].

A primary application of evolutionary computation in machine learning is optimizing neural network architectures. Genetic algorithms (GAs) and genetic programming (GP) automate neural network design, optimizing hyperparameters such as layer configurations, activation functions, and learning rates. This approach allows discovering tailored network architectures that enhance performance without manual intervention [44].

Besides architecture optimization, evolutionary algorithms facilitate feature selection and dimensionality reduction, crucial for managing high-dimensional data. By identifying relevant features for specific tasks, these algorithms reduce computational complexity and improve model interpretability. Techniques integrating symbolic regression with evolutionary strategies have successfully derived compact, interpretable models that maintain high predictive accuracy [46].

Moreover, evolutionary computation supports developing ensemble learning methods, where multiple models combine to enhance overall performance. Evolutionary algorithms optimize the selection and weighting of individual models within an ensemble, ensuring robust and accurate collective outputs. This application is valuable in scenarios where diverse data characteristics necessitate combining different modeling approaches [110].

The adaptability of evolutionary algorithms extends to optimizing reinforcement learning policies. By dynamically adjusting policy parameters and employing advanced strategies, these algorithms enhance learning, enabling agents to navigate complex and changing environments effectively. This adaptability is achieved through reinforcement learning methods that mitigate issues like early commitment and initialization bias, alongside integrating symbolic regression techniques that extract meaningful data patterns. Transfer learning further improves convergence rates and solution efficiency

AI-generated, for reference only.

in diverse tasks [24, 98, 106, 65, 67]. This capability is crucial in applications such as autonomous robotics and adaptive control systems, where real-time decision-making is essential.

Applications of evolutionary computation in machine learning program design highlight its adaptability and efficacy in solving intricate optimization problems. Advancements such as integrating Genetic Programming with MAP-Elites and Covariance Matrix Adaptation Evolution Strategy for symbolic regression, incorporating transfer learning to enhance convergence rates in gene expression programming, and utilizing guided evolution with binary discriminators to optimize machine learning architectures efficiently exemplify this [24, 89, 122, 85, 126]. By automating key model development aspects and leveraging natural evolution principles, evolutionary algorithms provide powerful tools for advancing machine learning systems across diverse domains.

7.5 Adaptive Learning with Feedback Mechanisms

Adaptive learning, enhanced by feedback mechanisms, significantly improves optimization processes in evolutionary computation. This integration enables systems, such as Markov Brains, to evolve learning capabilities in response to environmental feedback, enhancing adaptability and performance over time. Incorporating transfer learning techniques in evolutionary algorithms, like gene expression programming, facilitates the reuse of knowledge from previous optimizations, further accelerating convergence towards optimal solutions [64, 24, 89, 106, 63]. This approach involves dynamically adjusting learning parameters and strategies based on feedback, enabling evolutionary algorithms to adapt to changing conditions and improve performance.

A critical component of adaptive learning is integrating feedback loops that continuously inform the algorithm about candidate solution quality. These mechanisms refine search strategies, allowing the algorithm to focus on promising search space regions while avoiding premature convergence to suboptimal solutions [106]. By leveraging feedback, evolutionary algorithms can dynamically adjust parameters such as mutation rates, crossover probabilities, and selection pressures, optimizing search behavior in response to the evolving problem landscape.

Recent advancements emphasize incorporating domain-specific knowledge into feedback mechanisms. This integration enhances the algorithm's ability to navigate complex search spaces by providing additional context, improving the relevance and accuracy of discovered solutions [57]. Expert knowledge in shaping feedback loops ensures the evolutionary process aligns with specific problem domain requirements and constraints.

Moreover, adaptive learning techniques enhance evolutionary algorithms' robustness in noisy environments. By incorporating mechanisms that account for noise and uncertainty in feedback signals, these algorithms maintain effectiveness even in challenging conditions. This capability is crucial for applications where data quality and reliability vary, such as real-time optimization and adaptive control systems [106].

Hybrid approaches combining adaptive learning with other computational techniques further extend evolutionary algorithms' capabilities. For instance, integrating neural networks with adaptive feedback mechanisms allows continuous refinement of model parameters, improving accuracy and interpretability [57]. These hybrid methodologies leverage multiple computational paradigms' strengths, offering robust solutions to complex optimization challenges.

Adaptive learning with feedback mechanisms represents a significant advancement in evolutionary computation, providing a dynamic framework for optimizing complex systems. By continuously refining optimization strategies using real-time feedback, these techniques enhance the adaptability and efficacy of evolutionary algorithms. This evolution enables developing more sophisticated and responsive optimization solutions, particularly in areas like symbolic regression, where methods such as MAP-Elites and Covariance Matrix Adaptation Evolution Strategy improve exploration and diversity while addressing complex mathematical problems. The integration of transfer learning with gene expression programming accelerates convergence rates by leveraging previous optimizations, and the use of language-model-infused scaffolding programs exemplifies innovative self-improvement strategies that further optimize algorithm performance across various scientific and engineering disciplines [63, 24, 89, 57].

8 Conclusion

8.1 Future Directions

The trajectory of advanced computational techniques in machine learning and artificial intelligence is set to experience transformative growth, particularly through the integration of symbolic regression with empirical data from diverse domains. A pivotal element in this evolution will be the refinement of function libraries and the integration of domain-specific insights, which are crucial for enhancing both interpretability and practical application. Establishing standardized benchmarks and improving the efficiency of automated machine learning (AutoML) systems will play a significant role in expanding their utility across various fields.

In symbolic regression, future research will focus on the simultaneous optimization of constants and structural elements, alongside the integration of partial and complete expression searches to enhance the robustness of differentiable genetic programming. Broadening encoding schemes to include more complex operations will be essential for strengthening the generalized symbolic regression framework, thereby enabling the recovery of a more diverse range of expressions.

Advancing the autonomy of discovery systems will involve exploring novel AI methodologies for hypothesis generation, bridging computational approaches with autonomous systems, and fostering innovation in automated scientific discovery. This includes extending frameworks like DISCOVER to address a wider spectrum of scientific challenges, with an emphasis on resilience to noisy data and the exploration of varied equation types.

Interdisciplinary applications in science will benefit from enhanced data acquisition methods and improved model explainability. Optimizing algorithms for diverse data distributions and assessing their applicability in real-time processing contexts will be vital for maximizing the impact of these techniques.

Future investigations will also explore the application of the Generative Flow Network for Symbolic Regression (GFN-SR) to larger datasets and develop more adaptable tree construction methods to improve performance and flexibility. The role of Linear Combination Features (LCFs) in high-dimensional contexts will be scrutinized, focusing on tuning parameter optimization and extending this concept to other node types in symbolic regression.

Furthermore, advancements in evolutionary computation will emphasize enhancements through larger populations and domain-specific variables, refining parameter optimization for specific network architectures. Optimizing the SRNet algorithm for higher-dimensional datasets and integrating it with other interpretability frameworks will further enhance model transparency and utility.

Collectively, these research directions aim to expand the capabilities of AI and machine learning, paving the way for more sophisticated technologies that address complex challenges across various scientific and engineering domains.

AI-generated, for reference only.

References

[1] Li Li, Minjie Fan, Rishabh Singh, and Patrick Riley. Neural-guided symbolic regression with asymptotic constraints, 2019.

[2] Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models, 2024.

[3] Mengge Du, Yuntian Chen, Zhongzheng Wang, Longfeng Nie, and Dongxiao Zhang. Llm4ed: Large language models for automatic equation discovery, 2024.

[4] Mengge Du, Yuntian Chen, Zhongzheng Wang, Longfeng Nie, and Dongxiao Zhang. Llm4ed: Large language models for automatic equation discovery. *arXiv preprint arXiv:2405.07761*, 2024.

[5] Laure Crochepierre, Lydia Boudjeloud-Assala, and Vincent Barbesant. A reinforcement learning approach to domain-knowledge inclusion using grammar guided symbolic regression, 2022.

[6] Matteo Merler, Katsiaryna Haitsiukevich, Nicola Dainese, and Pekka Marttinen. In-context symbolic regression: Leveraging large language models for function discovery, 2024.

[7] Florian Lalande, Yoshitomo Matsubara, Naoya Chiba, Tatsunori Taniai, Ryo Igarashi, and Yoshitaka Ushiku. A transformer model for symbolic regression towards scientific discovery, 2023.

[8] Shu Wei, Yanjie Li, Lina Yu, Weijun Li, Min Wu, Linjun Sun, Jufeng Han, and Yan Pang. Closed-form solutions: A new perspective on solving differential equations, 2025.

[9] Àlex Ferrando De Las Morenas. Symbolic regression using a transformer neural network trained with supervised and reinforcement learning. B.S. thesis, Universitat Politècnica de Catalunya, 2022.

[10] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: A review, 2023.

[11] Harsha Vaddireddy, Adil Rasheed, Anne E Staples, and Omer San. Feature engineering and symbolic regression methods for detecting hidden physics from sparse sensors, 2020.

[12] Ritam Majumdar, Vishal Jadhav, Anirudh Deodhar, Shirish Karande, Lovekesh Vig, and Venkataramana Runkana. Symbolic regression for pdes using pruned differentiable programs, 2023.

[13] Shahriar Iravanian, Carl Julius Martensen, Alessandro Cheli, Shashi Gowda, Anand Jain, Yingbo Ma, and Chris Rackauckas. Symbolic-numeric integration of univariate expressions based on sparse regression, 2022.

[14] Jixin Hou, Xianyan Chen, Taotao Wu, Ellen Kuhl, and Xianqiao Wang. Automated data-driven discovery of material models based on symbolic regression: A case study on human brain cortex, 2024.

[15] Bahador Bahmani, Hyoung Suk Suh, and WaiChing Sun. Discovering interpretable elastoplasticity models via the neural polynomial method enabled symbolic regressions, 2024.

[16] Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers, 2022.

[17] G. F. Bomarito, P. E. Leser, N. C. M Strauss, K. M. Garbrecht, and J. D. Hochhalter. Automated learning of interpretable models with quantified uncertainty, 2022.

[18] Jennifer J. Schnur and Nitesh V. Chawla. Information fusion via symbolic regression: A tutorial in the context of human health, 2023.

[19] Yanjie Li, Weijun Li, Lina Yu, Min Wu, Jingyi Liu, Wenqiang Li, Shu Wei, and Yusong Deng. Mllm-sr: Conversational symbolic regression base multi-modal large language models, 2024.

[20] Erik Derner, Jiří Kubalík, Nicola Ancona, and Robert Babuška. Constructing parsimonious analytic models for dynamic systems via symbolic regression, 2020.

[21] Guilherme Seidyo Imai Aldeia and Fabricio Olivetti de Franca. Interpretability in symbolic regression: a benchmark of explanatory methods using the feynman data set, 2024.

[22] Yuma Iwasaki and Masahiko Ishida. Data-driven formulation of natural laws by recursive-lasso-based symbolic regression, 2021.

[23] Kevin René Broløs, Meera Vieira Machado, Chris Cave, Jaan Kasak, Valdemar Stentoft-Hansen, Victor Galindo Batanero, Tom Jelen, and Casper Wilstrup. An approach to symbolic regression using feyn, 2021.

[24] Maximilian Reissmann, Yuan Fang, Andrew S. H. Ooi, and Richard D. Sandberg. Accelerating evolutionary exploration through language model-based transfer learning, 2025.

[25] Stefan Kramer, Mattia Cerrato, Sašo Džeroski, and Ross King. Automated scientific discovery: From equation discovery to autonomous discovery systems, 2023.

[26] Iddo Drori, Yamuna Krishnamurthy, Raoni Lourenco, Remi Rampin, Kyunghyun Cho, Claudio Silva, and Juliana Freire. Automatic machine learning by pipeline synthesis using model-based reinforcement learning and a grammar, 2019.

[27] Moshe Sipper. Binary and multinomial classification through evolutionary symbolic regression, 2022.

[28] Martin Vastl, Jonáš Kulhánek, Jiří Kubalík, Erik Derner, and Robert Babuška. Symformer: End-to-end symbolic regression using transformer-based architecture, 2022.

[29] Yuanzhen Luo, Qiang Lu, Xilei Hu, Jake Luo, and Zhiguang Wang. Exploring hidden semantics in neural networks with symbolic regression, 2022.

[30] Jorge Medina and Andrew D. White. Active learning in symbolic regression with physical constraints, 2024.

[31] Tony Tohme, Dehong Liu, and Kamal Youcef-Toumi. Gsr: A generalized symbolic regression approach, 2023.

[32] Ricardo Vinuesa, Jean Rabault, Hossein Azizpour, Stefan Bauer, Bingni W Brunton, Arne Elofsson, Elias Jarlebring, Hedvig Kjellstrom, Stefano Markidis, David Marlevi, et al. Opportunities for machine learning in scientific discovery. *arXiv preprint arXiv:2405.04161*, 2024.

[33] William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H. Moore. Contemporary symbolic regression methods and their relative performance, 2021.

[34] Matthew Golden. Scalable sparse regression for model discovery: The fast lane to insight, 2024.

[35] Chinmay S. Kulkarni. Sparse regression and adaptive feature generation for the discovery of dynamical systems, 2019.

[36] Jure Brence, Ljupčo Todorovski, and Sašo Džeroski. Probabilistic grammars for equation discovery, 2021.

[37] Vernon Austel, Sanjeeb Dash, Oktay Gunluk, Lior Horesh, Leo Liberti, Giacomo Nannicini, and Baruch Schieber. Globally optimal symbolic regression, 2017.

[38] Shahab Razavi and Eric R Gamazon. Neural-network-directed genetic programmer for discovery of governing equations. *arXiv preprint arXiv:2203.08808*, 2022.

[39] Vernon Austel, Cristina Cornelio, Sanjeeb Dash, Joao Goncalves, Lior Horesh, Tyler Josephson, and Nimrod Megiddo. Symbolic regression using mixed-integer nonlinear optimization, 2020.

[40] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023.

[41] Lukas Kammerer, Gabriel Kronberger, Bogdan Burlacu, Stephan M. Winkler, Michael Kommenda, and Michael Affenzeller. Symbolic regression by exhaustive search: Reducing the search space using syntactical constraints and efficient semantic structure deduplication, 2021.

[42] Marco Virgolin, Andrea De Lorenzo, Eric Medvet, and Francesca Randone. Learning a formula of interpretability to learn interpretable formulas, 2020.

[43] Krzysztof Kacprzyk and Mihaela van der Schaar. Shape arithmetic expressions: Advancing scientific discovery beyond closed-form equations, 2024.

[44] Peng Zeng, Xiaotian Song, Andrew Lensen, Yuwei Ou, Yanan Sun, Mengjie Zhang, and Jiancheng Lv. Differentiable genetic programming for high-dimensional symbolic regression, 2023.

[45] Fabricio Olivetti de Franca. A greedy search tree heuristic for symbolic regression, 2018.

[46] Haiquan Qiu, Shuzhi Liu, and Quanming Yao. Neural symbolic regression of complex network dynamics, 2024.

[47] Aftab Anjum, Fengyang Sun, Lin Wang, and Jeff Orchard. A novel neural network-based symbolic regression method: Neuro-encoded expression programming, 2021.

[48] Ho Fung Tsoi, Vladimir Loncar, Sridhara Dasu, and Philip Harris. Symbolnet: Neural symbolic regression with adaptive dynamic pruning, 2024.

[49] Parshin Shojaee, Kazem Meidani, Amir Barati Farimani, and Chandan Reddy. Transformer-based planning for symbolic regression. *Advances in Neural Information Processing Systems*, 36:45907–45919, 2023.

[50] Hossein Izadi Rad, Ji Feng, and Hitoshi Iba. Gp-rvm: Genetic programing-based symbolic regression using relevance vector machine, 2018.

[51] Telmo Menezes and Camille Roth. Symbolic regression of generative network models, 2014.

[52] Ehsan Askari and Guillaume Crevecoeur. Evolutionary sparse data-driven discovery of complex multibody system dynamics, 2022.

[53] Maximilian Reissmann, Yuan Fang, Andrew Ooi, and Richard Sandberg. Constraining genetic symbolic regression via semantic backpropagation, 2024.

[54] Markus Quade, Thomas Isele, and Markus Abel. Explainable machine learning control – robust control and stability analysis, 2020.

[55] Kyle LaFollette, Janni Yuval, Roey Schurr, David Melnikoff, and Amit Goldenberg. Data driven equation discovery reveals non-linear reinforcement learning in humans. 2023.

[56] Yuan Tian, Wenqi Zhou, Michele Viscione, Hao Dong, David Kammer, and Olga Fink. Interactive symbolic regression through offline reinforcement learning: A co-design framework, 2025.

[57] Mikhail Maslyaev and Alexander Hvatov. Comparison of single- and multi- objective optimization quality for evolutionary equation discovery, 2023.

[58] Yi Xie, Tianyu Qiu, Yun Xiong, Xiuqi Huang, Xiaofeng Gao, and Chao Chen. An efficient and generalizable symbolic regression method for time series analysis, 2024.

[59] Grant Dick. Interval arithmetic and interval-aware operators for genetic programming, 2017.

[60] Govind Gandhi. Symbolic regression of dynamic network models, 2023.

[61] Samiha Sharlin and Tyler R. Josephson. In context learning and reasoning for symbolic regression with large language models, 2024.

[62] Yi-Wei Chen, Qingquan Song, and Xia Hu. Techniques for automated machine learning, 2019.

[63] Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (stop): Recursively self-improving code generation, 2024.

[64] Olivier Risser-Maroix and Benjamin Chamand. What can we learn by predicting accuracy?, 2022.

[65] Julia Balla, Sihao Huang, Owen Dugan, Rumen Dangovski, and Marin Soljacic. Ai-assisted discovery of quantitative and formal models in social science, 2023.

[66] Cristina Cornelio, Sanjeeb Dash, Vernon Austel, Tyler Josephson, Joao Goncalves, Kenneth Clarkson, Nimrod Megiddo, Bachir El Khadir, and Lior Horesh. Ai descartes: Combining data and theory for derivable scientific discovery, 2023.

[67] Sannyuya Liu, Qing Li, Xiaoxuan Shen, Jianwen Sun, and Zongkai Yang. Automated discovery of symbolic laws governing skill acquisition from naturally occurring data, 2024.

[68] Symbolic regression for space applications: Differentiable cartesian genetic programming powered by multi-objective memetic algorithms.

[69] Jeremy A. McCulloch, Skyler R. St. Pierre, Kevin Linka, and Ellen Kuhl. On sparse regression, lp-regularization, and automated model discovery, 2024.

[70] Nan Jiang and Yexiang Xue. Racing control variable genetic programming for symbolic regression, 2023.

[71] Kai Ruan, Ze-Feng Gao, Yike Guo, Hao Sun, Ji-Rong Wen, and Yang Liu. Discovering symbolic expressions with parallelized tree search, 2024.

[72] Tony Tohme. *Advances in Symbolic Regression: From Generalized Formulation to Density Estimation and Inverse Problem*. PhD thesis, Massachusetts Institute of Technology, 2024.

[73] Jiachen Li, Ye Yuan, and Hong-Bin Shen. Symbolic expression transformer: A computer vision approach for symbolic regression, 2022.

[74] Chen Chen, Changtong Luo, and Zonglin Jiang. Block building programming for symbolic regression, 2017.

[75] Dominic P. Searson. Gptips 2: an open-source software platform for symbolic data mining, 2015.

[76] Jan Žegklitz and Petr Pošík. Symbolic regression algorithms with built-in linear regression, 2017.

[77] Jan Žegklitz and Petr Pošík. Learning linear feature space transformations in symbolic regression, 2017.

[78] Zihan Yu, Jingtao Ding, and Yong Li. Symbolic regression via mdlformer-guided search: from minimizing prediction error to minimizing description length, 2024.

[79] Fabricio Olivetti de Franca and Gabriel Kronberger. Prediction intervals and confidence regions for symbolic regression models based on likelihood profiles, 2022.

[80] Jingyi Liu, Yanjie Li, Lina Yu, Min Wu, Weijun Li, Wenqiang Li, Meilan Hao, Yusong Deng, and Shu Wei. Dn-cl: Deep symbolic regression against noise via contrastive learning, 2024.

[81] J. Kubalík, E. Derner, and R. Babuška. Symbolic regression driven by training data and prior knowledge, 2020.

[82] Mojtaba Valipour, Bowen You, Maysum Panju, and Ali Ghodsi. Symbolicgpt: A generative transformer model for symbolic regression, 2021.

[83] Sida Li, Ioana Marinescu, and Sebastian Musslick. Gfn-sr: Symbolic regression with generative flow networks, 2023.

[84] Mengge Du, Yuntian Chen, and Dongxiao Zhang. Discover: Deep identification of symbolic open-form pdes via enhanced reinforcement-learning. *arXiv preprint arXiv:2210.02181*, 2022.

[85] Bogdan Burlacu, Michael Affenzeller, and Michael Kommenda. On the effectiveness of genetic operations in symbolic regression, 2021.

[86] Pablo Moscato, Haoyuan Sun, and Mohammad Nazmul Haque. Analytic continued fractions for regression: A memetic algorithm approach, 2019.

[87] Tommaso Bendinelli, Luca Biggio, and Pierre-Alexandre Kamienny. Controllable neural symbolic regression. In *International Conference on Machine Learning*, pages 2063–2077. PMLR, 2023.

[88] Gabriel Kronberger, Lukas Kammerer, Bogdan Burlacu, Stephan M. Winkler, Michael Kommenda, and Michael Affenzeller. Cluster analysis of a symbolic regression search space, 2021.

[89] J. P. Bruneton, L. Cazenille, A. Douin, and V. Reverdy. Exploration and exploitation in symbolic regression using quality-diversity and evolutionary strategies algorithms, 2019.

[90] T. Nathan Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P. Santiago, Daniel M. Faissol, and Brenden K. Petersen. Symbolic regression via neural-guided genetic programming population seeding, 2021.

[91] Yousef A. Radwan, Gabriel Kronberger, and Stephan Winkler. A comparison of recent algorithms for symbolic regression to genetic programming, 2024.

[92] Kirill Antonov, Roman Kalkreuth, Kaifeng Yang, Thomas Bäck, Niki van Stein, and Anna V Kononova. A functional analysis approach to symbolic regression, 2024.

[93] Changtong Luo, Chen Chen, and Zonglin Jiang. A divide and conquer method for symbolic regression, 2017.

[94] Baihe He, Qiang Lu, Qingyun Yang, Jake Luo, and Zhiguang Wang. Taylor genetic programming for symbolic regression. In *Proceedings of the genetic and evolutionary computation conference*, pages 946–954, 2022.

[95] Charles Fox, Neil Tran, Nikki Nacion, Samiha Sharlin, and Tyler R. Josephson. Incorporating background knowledge in symbolic regression using a computer algebra system, 2023.

[96] Baihe He, Qiang Lu, Qingyun Yang, Jake Luo, and Zhiguang Wang. Taylor genetic programming for symbolic regression, 2022.

[97] Jiří Kubalík, Erik Derner, Jan Žegklitz, and Robert Babuška. Symbolic regression methods for reinforcement learning, 2021.

[98] Mikel Landajuela, Brenden K. Petersen, Soo K. Kim, Claudio P. Santiago, Ruben Glatt, T. Nathan Mundhenk, Jacob F. Pettit, and Daniel M. Faissol. Improving exploration in policy gradient search: Application to symbolic optimization, 2021.

[99] Yilong Xu, Yang Liu, and Hao Sun. Rsrm: Reinforcement symbolic regression machine, 2023.

[100] Yanjie Li, Weijun Li, Lina Yu, Min Wu, Jingyi Liu, Wenqiang Li, Meilan Hao, Shu Wei, and Yusong Deng. Generative pre-trained transformer for symbolic regression base in-context reinforcement learning, 2024.

[101] Wenqing Zheng, Tianlong Chen, Ting-Kuei Hu, and Zhangyang Wang. Symbolic learning to optimize: Towards interpretability and scalability, 2022.

[102] Yilong Xu, Yang Liu, and Hao Sun. Reinforcement symbolic regression machine. In *The Twelfth International Conference on Learning Representations*, 2024.

[103] Léo Françoso Dal Piccol Sotto and Vinícius Veloso de Melo. A probabilistic linear genetic programming with stochastic context-free grammar for solving symbolic regression problems, 2017.

[104] Casper Wilstrup and Jaan Kasak. Symbolic regression outperforms other models for small data sets, 2021.

[105] Gabriel Kronberger, Lukas Kammerer, and Michael Kommenda. Identification of dynamical systems using symbolic regression, 2021.

[106] Leigh Sheneman and Arend Hintze. Machine learned learning machines, 2017.

[107] Felipe Leno da Silva, Andre Goncalves, Sam Nguyen, Denis Vashchenko, Ruben Glatt, Thomas Desautels, Mikel Landajuela, Daniel Faissol, and Brenden Petersen. Language model-accelerated deep symbolic optimization. *Neural Computing and Applications*, pages 1–17, 2023.

[108] Kazem Meidani, Parshin Shojaee, Chandan K. Reddy, and Amir Barati Farimani. Snip: Bridging mathematical symbolic and numeric realms with unified pre-training, 2024.

[109] William La Cava, Lee Spector, and Kourosh Danai. Epsilon-lexicase selection for regression, 2019.

[110] Dazhuang Liu, Marco Virgolin, Tanja Alderliesten, and Peter A. N. Bosman. Evolvability degeneration in multi-objective genetic programming for symbolic regression, 2022.

[111] Yilong Xu, Yang Liu, and Hao Sun. Rsrm: Reinforcement symbolic regression machine. *arXiv preprint arXiv:2305.14656*, 2023.

[112] Grant Norman, Jacqueline Wentz, Hemanth Kolla, Kurt Maute, and Alireza Doostan. Constrained or unconstrained? neural-network-based equation discovery from data, 2024.

[113] Marco Virgolin, Tanja Alderliesten, Cees Witteveen, and Peter A. N. Bosman. Improving model-based genetic programming for symbolic regression of small expressions, 2021.

[114] Zachary Bastiani, Robert M. Kirby, Jacob Hochhalter, and Shandian Zhe. Complexity-aware deep symbolic regression with robust risk-seeking policy gradients, 2024.

[115] Telmo Menezes and Camille Roth. Automatic discovery of families of network generative processes, 2019.

[116] Mahed Abroshan, Saumitra Mishra, and Mohammad Mahdi Khalili. Symbolic metamodels for interpreting black-boxes using primitive functions, 2023.

[117] Sohrab Towfighi. Symbolic regression by uniform random global search, 2019.

[118] L. G. A dos Reis, V. L. P. S. Caminha, and T. J. P. Penna. Benchmarking symbolic regression constant optimization schemes, 2024.

[119] Nan Jiang, Md Nasim, and Yexiang Xue. Vertical symbolic regression, 2023.

[120] Pierre-Alexandre Kamienny, Guillaume Lample, Sylvain Lamprier, and Marco Virgolin. Deep generative symbolic regression with monte-carlo-tree-search, 2023.

[121] Krishn Kumar Gupt, Meghana Kshirsagar, Douglas Mota Dias, Joseph P. Sullivan, and Conor Ryan. A novel ml-driven test case selection approach for enhancing the performance of grammatical evolution, 2023.

[122] Kory Becker and Justin Gottschlich. Ai programmer: Autonomously creating software programs using genetic algorithms, 2017.

[123] Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. *arXiv preprint arXiv:2409.09359*, 2024.

[124] Nan Jiang and Yexiang Xue. Symbolic regression via control variable genetic programming, 2023.

[125] Patricia Arroba, José L. Risco-Martín, Marina Zapater, José M. Moya, and José L. Ayala. Enhancing regression models for complex systems using evolutionary techniques for feature engineering, 2024.

[126] John D. Co-Reyes, Yingjie Miao, George Tucker, Aleksandra Faust, and Esteban Real. Guided evolution with binary discriminators for ml program search, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.