POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Data-driven discovery of dynamical systems through Symbolic-SINDy

LAUREA MAGISTRALE IN MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: MATTIA GASTOLDI

Advisor: PROF. ANDREA MANZONI

Co-advisor: DR. PAOLO CONTI

Academic year: 2025-2026

1. Introduction

The process of distilling the underlying laws directly from measurements of a system goes under the name of *(data-driven) model discovery*, and plays a fundamental role in those fields of science for which the governing equations are unknown, such as climate science, neuroscience, ecology, finance, and epidemiology. In all these cases, we aim at modeling the evolution of the state of the system through a nonlinear dynamical system under the form:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t); \boldsymbol{\beta}),$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the observed state vector, $\mathbf{f}(\cdot\,; \boldsymbol{\beta})$ denotes the unknown function encoding the system's dynamics, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of parameters on which the system depends.
Symbolic Regression (SR) [5] has proven to be an effective method to determine the complete structure of $\mathbf{f}$, and its parameters $\boldsymbol{\beta}$, entirely from observations. This result is achieved by taking advantage of genetic programming optimization. By construction, this method does not require any prior assumption about the functional structure of the equation. However, SR is extremely expensive, as it requires evaluating a very large number of possible functional configurations. Moreover, it does not scale well to large systems, and it may be prone to overfitting. Based on the formulation of the objective function, in this work, Symbolic Regression will be referred as **SR-T**[5] or **D-CODE** [3]. An alternative approach is the so-called Sparse Identification of Nonlinear Dynamics **(SINDy)** [1], which formulates model discovery as a sparse linear regression problem to identify a small set of active terms and their coefficients:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t); \boldsymbol{\beta}) \approx \boldsymbol{\Theta}\left(\mathbf{x}(t)\right) \boldsymbol{\Xi}.$$

This reformulation substantially reduces the search space and scales more favorably with the size of the problem. However, the exact recovery of $\mathbf{f}$ depends heavily on the nonlinearities pre-specified by the user within the library $\boldsymbol{\Theta}$.
Symbolic-SINDy [4] is a promising framework that bridges these two algorithms, leveraging their complementary strengths and overcoming their limitations. This work consolidates and extends the Symbolic-SINDy framework in various scenarios, including Online Model Discovery and Reduced Order Modeling.

2. The Symbolic-SINDy idea

The Symbolic-SINDy algorithm is conceptually divided into two main stages. Initially, a preliminary execution of Symbolic Regression is run.

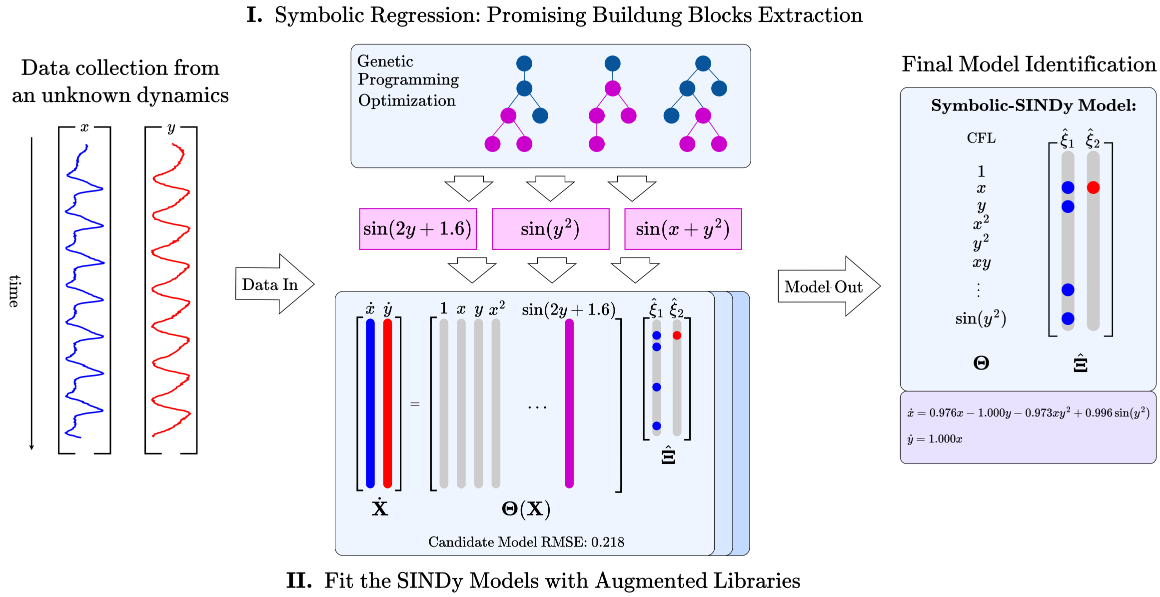

Figure 1: Schematic Symbolic-SINDy algorithm.

The goal of this stage is to extract some suitable building blocks from the most promising expression trees. Subsequently, each of these small mathematical expressions extracted is integrated into a polynomial library, effectively becoming an automatically added nonlinearity. Sparse regression acts as a regularizer, ensuring the identification of a parsimonious model in closed-form. Among the various SINDy models fitted, one for each extracted building block, the best one is then chosen, i.e., the one that minimizes the error with respect to the training data. A schematic representation is provided in Figure 1. To validate the discovered model and the extracted building block, an ensemble-based uncertainty quantification framework is integrated. Using bootstrapping techniques, it is possible to aggregate the outputs of several SINDy models estimated on the final Symbolic-SINDy library; this naturally provides inclusion probabilities and uncertainty estimates for the identified coefficients. In the following section, we demonstrate this idea on a low-dimensional system.

3. The van der Pol oscillator

Consider a revised version of the Pol oscillator:

$$\dot{x} = \mu(1 - y^2)x - y + A\sin(\omega y^2),$$
$$\dot{y} = x,$$

where $A \in \mathbb{R}$, $\mu \in \mathbb{R}^+$, and $\omega \in \mathbb{R}^+$. With this definition, an additional nonlinearity is introduced into the van der Pol oscillator, described by a sinusoidal force composed with a quadratic power of the state. Note that the resulting dynamics is therefore defined by the contribution of two interacting oscillations. To discover the underlying equation directly from data, we compare the performance of *(i)* SINDy, equipped with a polynomial library, *(ii)* Symbolic Regression, and *(iii)* Symbolic-SINDy. As numerically shown, SINDy is computationally efficient. However, the model identification is only partial. Since the library used does not contain the necessary nonlinearity, SINDy correctly identifies only a few terms of the dynamics, and uses additional nonlinearities to approximate the observed trajectories as closely as possible. SR methods also highlight critical issues in model identification. When they do not identify the correct model in closed-form, they often fit dynamics that combine x and y through complex trigonometric expressions. Moreover, they exhibit slow convergence, potentially identifying completely incorrect dynamics, despite the large number of generations employed. Symbolic-SINDy is equipped with a wider functions set and fewer generations, with the aim of maximizing user-independence and reducing the computational costs of genetic programming. The improvement in performance is evident. In both unparametrized and ω-parametrized problems, Symbolic-SINDy proves to be more flexible, outperforming existing methodologies despite a less

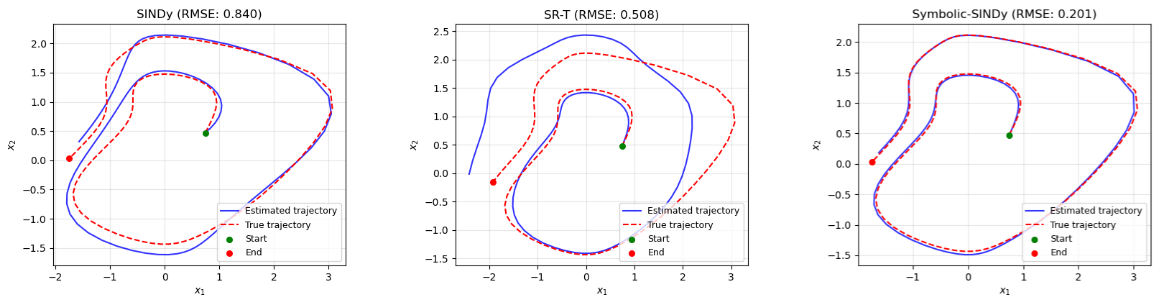

Figure 2: Result on the ω-parametrized revised Van der Pol with $(x_0.y_0, \omega) = (0.7,\, 0.4,\, 1.5)$.

informed functions set and fewer generations. A summary of the experiment is provided in Figure 2. Moreover, the stability of the model identified by Symbolic-SINDy can be assessed in Figure 3.

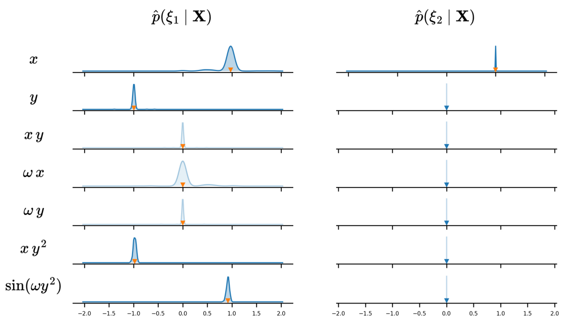

Figure 3: UQ on the Symbolic-SINDy estimated coefficients for the ω-parametrized van der Pol.

In this problem, the key element for the model discovery is the inclusion of the term $\sin(\omega y^2)$ into the library, which is unlikely to be included *a priori*. By automatically identifying and incorporating this building block, Symbolic-SINDy successfully recovers the underlying dynamics, effectively overcoming the limitations of the library construction.

4. Online Symbolic-SINDy

Many existing methods struggle to properly handle situations in which changes in dynamics occur due to variations of the underlying system over time. In particular, critical issues arise when new terms are introduced or existing ones are eliminated. Symbolic-SINDy can potentially be used for this purpose in its online version. Since temporal variations in the system often occur in a nonlinear manner, the flexibility of Symbolic Regression can be exploited to identify promising temporal nonlinearities, which may then be integrated into the function library.

4.1. Online learning framework

Symbolic-SINDy is coupled with the CUSUM algorithm. CUSUM is a statistical technique that is used to detect deviations in the mean level of a process over time. In this context, the mean level of model error is monitored to detect changes in the underlying system. This allows the output model to change according to the ground-truth behavior. Note that building block extraction is not performed at every detected change-point. This avoids unnecessary computational overhead, especially if a search has recently been completed.

4.2. Time-varying Lotka-Volterra

Consider the non-autonomous Lotka-Volterra model:

$$\dot{x} = \alpha(t)x - \beta(t)xy,$$
$$\dot{y} = \delta(t)xy - \gamma(t)y.$$

In the beginning, all the parameters are kept fixed. Subsequently, at $t = 20$, a sinusoidal forcing is introduced, with pulsation $\omega = 5$, that acts on the coefficient α, simulating a seasonal effect on the rate of change of the prey. Finally, at $t = 50$, the interaction coefficient between prey and predator is slightly modified through a switching mechanism, introducing a transient regime into the system.
The first identified model is obtained simply by applying SINDy to the simulated trajectories in the time interval $[0, 10]$. The starting Lotka-Volterra model is hence correctly estimated, since a polynomial library is sufficient to recover the ground-truth dynamics. The first change in system behavior occurs at $t = 20$. The CUSUM detects this change at $t = 30$, and the Symbolic Regression is executed. Among the promising building blocks extracted, a good estimate of the missing nonlinearity is recovered: $\sin(5.007\,t)$. Although this term is already iden-

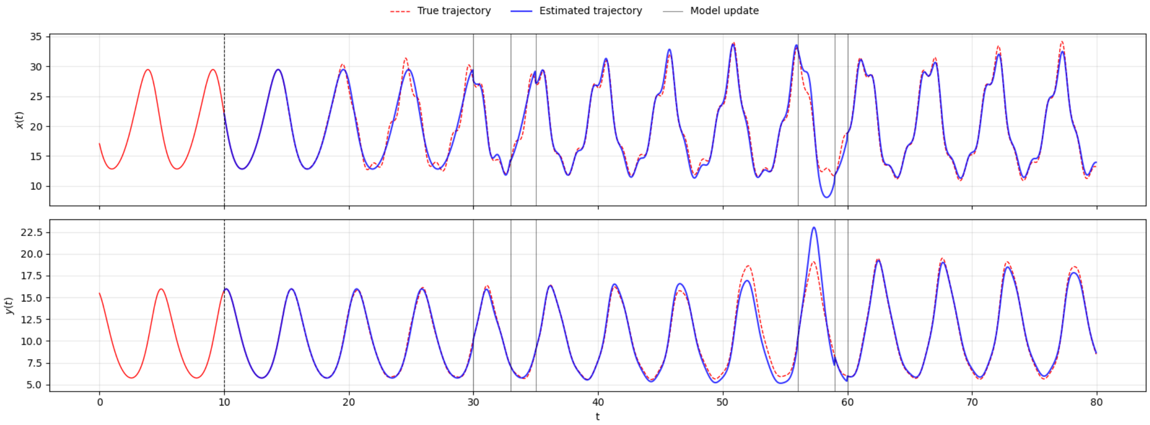

Figure 4: Online model identification of the non-autonomous Lotka-Volterra model with Symbolic-SINDy.

tified, only at $t = 35$ the estimated model is close enough to the ground-truth system. With this update, all the terms of the dynamics are properly identified, even the characteristic nonlinearity $x \sin(5\,t)$. The obtained model is stable, remaining aligned with the observed dynamics and maintaining a steady low error, until the second change occurs at $t = 50$. This variation does not introduce additional nonlinearities, but rather a sudden change in the value of a coefficient. After a transient phase, the execution of the Symbolic Regression is triggered again at $t = 60$, and other promising building blocks are extracted, including $\sin(4.997\,t)$. With this update, the model obtained with a new execution of SINDy adequately describes the dynamics. As before, all terms of the system are recovered with high accuracy. This model is stably aligned with the ground-truth dynamics and maintains a steady low error over time, until the end of the experiment. A summary of the online discovery process is provided in Figure 4. In this experiment, the combination of Symbolic Regression and SINDy proves to be effective in the context of Online Model Discovery. Indeed, the method is capable of detecting changes in dynamics, recovering characteristic nonlinearities, and correctly identifying the underlying system in the long term.

5. Symbolic-SINDy for high-dimensional problems

For physical phenomena that are distributed in space and evolve over time, the direct application of traditional model discovery techniques is often prohibitive. However, the dynamics of these high-dimensional systems is frequently governed by the evolution of a small number of latent variables.

5.1. Latent dynamics identification

To simultaneously identify a coordinate system that enables accurate state reconstruction and a closed-form description of its latent dynamics, coupling SINDy with an autoencoder (AE+SINDy) [2] for the sake of dimensionality reduction provides an effective framework. In this setting, the SINDy regularization terms act as explicit physical constraints during the training process. Although the AE+SINDy framework has demonstrated remarkable performances when trying to identify a reduced order model directly from data, it inherits from the standard SINDy approach the need to define a suitable library *a priori*. An intermediate step with Symbolic-SINDy might be useful to overcome this limitation and automatically build a problem-specific library, thus enhancing the identification of latent dynamics. In the following, we demonstrate this intuition by means of a numerical example.

5.2. Double gyre flow

The double gyre flow is a velocity field $\mathbf{v} = [u, v]^T$ in the domain $[0, 2] \times [0, 1]$ perturbed by periodic external forcing. It is governed by:

$$u(x, y, t) = -\pi A \sin\left(\pi f(x, t)\right) \cos\left(\pi y\right)$$
$$v(x, y, t) = \pi A \cos\left(\pi f(x, t)\right) \sin\left(\pi y\right) \frac{\partial f(x, t)}{\partial x}$$

where the time dependency is introduced by $f(x,t) = a(t)x^2 + b(t)x$, with $a(t) = \epsilon \sin(\omega t)$ and $b(t) = 1 - 2\epsilon \sin(\omega t)$. Here, $A \in \mathbb{R}$ controls the velocity magnitude, while $\epsilon \in \mathbb{R}$ and $\omega \in \mathbb{R}$ respectively determine the amplitude and frequency of the oscillation in the x-direction. The simulated dynamics is shown in Figure 5.

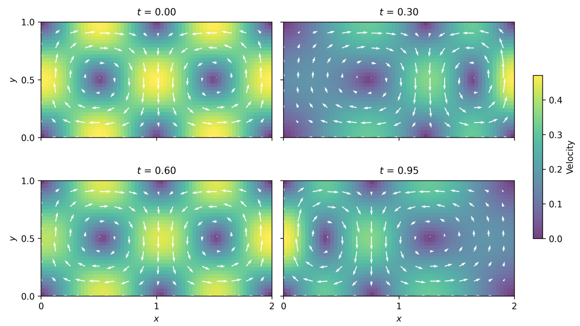

Figure 5: Snapshots of the double gyre dynamics with $A = 0.1$, $\epsilon = 0.25$, and $\omega = 5.0$.

Given the nature of the double gyre, a **Convolutional Autoencoder** (CAE) is employed for the extraction of latent features. This architecture allows for a more effective characterization of the spatial correlations of the field, and the multi-channel structure of enables a better handling of the two components of the velocity, u and v. The latent space is chosen to be one-dimensional, since when considering a larger bottleneck, the high-dimensional projected observations systematically lie on a one-dimensional manifold. The total loss function used in the training is defined as:

$$\mathcal{L} = \underbrace{\|\mathbf{v} - \hat{\mathbf{v}}\|_2^2}_{\text{reconstruction}} + \lambda_1 \underbrace{\|\dot{\mathbf{v}} - \hat{\dot{\mathbf{v}}}\|_2^2}_{\text{SINDy (decoded) in } \dot{\mathbf{x}}} + \lambda_2 \underbrace{\|\dot{z} - \mathbf{\Theta}(z)\mathbf{\Xi}\|_2^2}_{\text{SINDy in } \dot{z}} + \lambda_3 \underbrace{\|\mathbf{\Xi}\|_1}_{L_1 \text{ reg}} \quad (1)$$

following the AE+SINDy approach. We compare the performance of the architecture when a standard library is used in Equation (1), versus when a problem-specific library obtained with Symbolic-SINDy is employed.

Standard CAE+SINDy. Using the standard nonlinear library

$$\mathbf{\Theta}(z) = [1,\ z,\ z^2,\ z^3,\ \sin(z)], \quad (2)$$

the identified latent dynamics are:

$$\dot{z} = 0.422z - 0.425\sin(z). \quad (3)$$

The test set is encoded into the latent space, and its first snapshot is evolved according to the dynamics identified in (3). The resulting trajectory is then compared with the ground-truth behavior. As shown in Figure 6, the inability to learn the dynamics of the system clearly emerges.

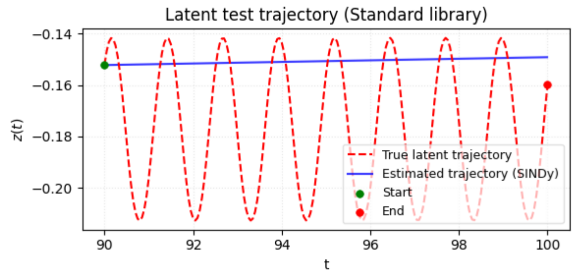

Figure 6: Latent dynamics according to (3).

Employing the decoder to project the latent trajectory from Figure 6 back into the high-dimensional space, a substantial inconsistency can be observed, as shown in Figure 7.

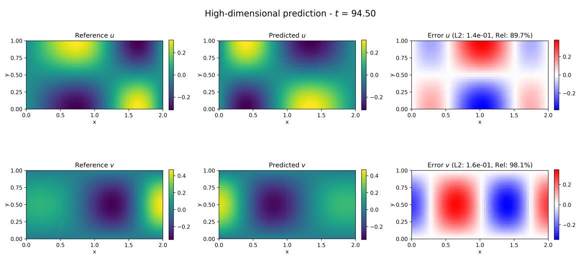

Figure 7: Estimated high-dimensional reconstruction and absolute error at $t = 94.50$ according to (3).

CAE+Symbolic-SINDy. Symbolic-SINDy is executed within the latent space obtained from a partial training of a standard CAE+SINDy architecture. In this phase, the weight of the λ_2 term in the total loss function is relaxed to reduce the bias induced by the preliminary SINDy library. Symbolic-SINDy extracts $b(z,t) = \cos(4.998\,t)$ and the new problem-specific library is defined as follows:

$$\mathbf{\Theta}(z,t) = [1,\ z,\ z^2,\ z^3,\ \cos(4.998\,t)]. \quad (4)$$

The identified latent dynamics are:

$$\dot{z} = -0.052 - 0.311z - 0.639\cos(4.998\,t). \quad (5)$$

The test set is encoded into the latent space, and its first snapshot is evolved according to the dynamics identified in (5). The resulting trajectory is then compared with the ground-truth behavior. As shown in Figure 8, the discovered governing equation is closely aligned with the actual physical behavior of the system.

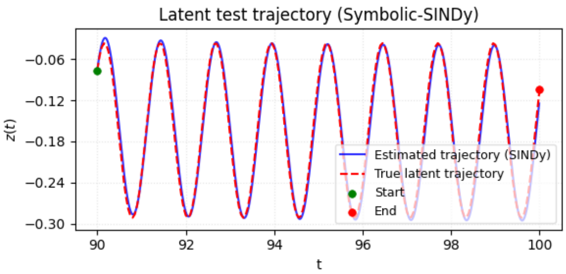

Figure 8: Latent dynamics according to (5).

Employing the decoder to project the latent trajectory (see Figure 8) back onto the high-dimensional space, it is possible to retrieve the original dynamics, as shown in Figure 9.

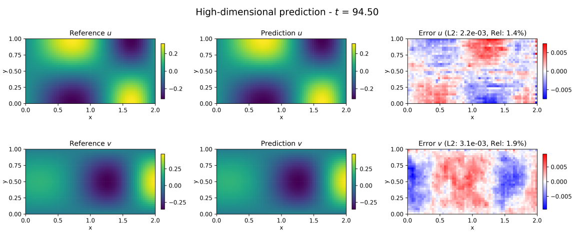

Figure 9: Estimated high-dimensional reconstruction and absolute error at $t = 94.50$ according to (5).

Hence, the integration of Symbolic-SINDy allows us to effectively identify the dynamics, solving a problem that would otherwise be too complex for existing model discovery approaches.

6. Conclusions

In this work, we proved the effectiveness of Symbolic-SINDy as a flexible model discovery method, succeeding in the identification of the correct dynamics from data in different scenarios. First, Symbolic-SINDy has been tested on more realistic and involved low-dimensional problems, where both Symbolic Regression and SINDy fail. In these examples, the current implementation outperformed the existing methodologies. Subsequently, Symbolic-SINDy has addressed Online Model Discovery by extracting local nonlinearities in both state and temporal domains, enabling the model to adapt to sudden changes in the underlying dynamics. Finally, Symbolic-SINDy has been integrated into the latent space of a convolutional autoencoder, proving to be a fundamental tool in defining a new problem-specific library, thus improving the discovery of equations that govern high-dimensional systems.
However, in complex scenarios, Symbolic-SINDy often requires a constrained function set and numerous generations to be effective. This compromises the flexibility of symbolic regression as a tool for discovering arbitrary building blocks with low computational costs. To overcome these limitations, future developments could involve replacing genetic programming-based approaches with emerging deep-symbolic models.
Finally, extremely promising results emerge from the application of Symbolic-SINDy within the context of Reduced Order Modeling. These preliminary and encouraging results justify further investigation of the real potential of the method with respect to more complex high-dimensional dynamical systems.

References

[1] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[2] Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019.

[3] Zhaozhi Qian, Krzysztof Kacprzyk, and Mihaela van der Schaar. D-CODE: Discovering closed-form odes from observed trajectories. In *International Conference on Learning Representations*, 2022.

[4] Giorgio Romano. Model Discovery integrating Symbolic Regression into Sparse Identification of Nonlinear Dynamics. Master's thesis, Politecnico di Milano, 2024.

[5] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.