



Research



Cite this article: Martin A, Yasin T, Bartlett D, Desmond H, Ferreira P. 2026 Constraining dark matter halo profiles with symbolic regression. *Phil. Trans. R. Soc. A* **384**: 20250090.

<https://doi.org/10.1098/rsta.2025.0090>

Received: 21 July 2025

Accepted: 28 November 2025

One contribution of 15 to a discussion meeting issue ‘Symbolic regression in the physical sciences’.

Subject Areas:

astrophysics, cosmology, galaxies, statistics

Keywords:

dark matter, cosmology, galaxies, galaxy dynamics, machine learning, symbolic regression

Author for correspondence:

Alicia Martin

e-mail: alicia.martin@physics.ox.ac.uk

Constraining dark matter halo profiles with symbolic regression

Alicia Martin¹, Tariq Yasin¹, Deaglan Bartlett², Harry Desmond³ and Pedro Ferreira¹

¹University of Oxford, Oxford, UK

²Institut d’Astrophysique de Paris (IAP), CNRS, Paris, France

³University of Portsmouth Institute of Cosmology and Gravitation, Portsmouth, UK

AM, 0009-0002-4201-7962; TY, 0000-0002-6314-1607; HD, 0000-0003-0685-9791; PF, 0000-0002-3021-2851

Dark matter haloes are typically characterized by radial density profiles with fixed forms motivated by simulations (e.g. Navarro–Frenk–White [NFW]). However, simulation predictions depend on uncertain dark matter physics and baryonic modelling. Here, we present a method to constrain halo density profiles directly from observations using Exhaustive Symbolic Regression (ESR), a technique that searches the space of analytic expressions for the function that best balances accuracy and simplicity for a given dataset. We test the approach on mock weak lensing excess surface density (ESD) data of synthetic clusters with NFW profiles. Motivated by real data, we assign each ESD data point a constant fractional uncertainty and vary this uncertainty and the number of clusters to probe how data precision and sample size affect model selection. For fractional errors around 5%, ESR recovers the NFW profile even from samples as small as approximately 20 clusters. At higher uncertainties representative of current surveys, simpler functions are favoured over NFW, though it remains competitive. This preference arises because weak lensing errors are smallest in the outskirts, causing the fits to be dominated by the outer profile. ESR therefore provides a robust, simulation-independent framework both for testing mass models and determining which features of a halo’s density profile are genuinely constrained by the data.

This article is part of the discussion meeting issue ‘Symbolic regression in the physical sciences’.

1. Introduction

In the Λ CDM cosmological model, structure forms hierarchically: small initial overdensities grow over time through gravitational instability, merging to form larger systems until they reach virial equilibrium. This process primarily affects dark matter, which interacts only through gravity. As a result, dark matter is the first component to form gravitationally bound structures known as dark matter haloes, creating a scaffolding into which baryons, subject also to dissipative processes, later settle. Understanding the properties and dynamics of haloes is important for studying the formation of galaxies and testing the predictions of cosmological models [1,2].

Gravity-only N -body simulations trace the formation and evolution of dark matter haloes, while hydrodynamical simulations also attempt to include baryonic processes. From these simulations, a key result is that haloes seem to follow increasingly steep central density profiles or *cusps*. This behaviour is commonly captured by the Navarro–Frenk–White (NFW) density profile [3]:

$$\rho_{\text{NFW}}(r) = \frac{\rho_0}{r/r_s(1+r/r_s)^2} \quad , \quad (1.1)$$

where r is the radial distance from the centre of the halo, ρ_0 is a characteristic density and r_s a scale radius.

Despite this profile being widely used, relying only on numerical simulations has limitations. These profiles are traditionally derived as empirical fits to simulations rather than predictions from first principles, and therefore lack a fundamental theory explaining their origin. Recently, progress has been made in deriving halo profiles based on the physics of collisionless relaxation, offering an insight on why these profiles appear in cold dark matter simulations [4]. However, the lack of a fully settled theory limits the reliability of these models. More significantly, observational data continue to challenge the universality of cuspy profiles. For instance, galaxy rotation curves often suggest flatter inner density slopes, so-called *cores*, with an approximately constant density near the centre ($\rho \propto r^0$) [5].

Several mechanisms have been proposed to explain the discrepancy between simulations and observations. Most involve baryonic processes, such as supernova-driven feedback expelling gas from the centre, or clumps transferring energy outwards through dynamical friction [6,7]. However, these processes are very complex and remain difficult to model accurately in simulations. Moreover, some of these relevant processes occur below the resolution limit of such simulations, necessitating uncertain, semi-analytic ‘subgrid models’. Beyond baryons, alternative dark matter models, such as fuzzy or self-interacting dark matter [8], can also generate halo profiles that differ from those seen in cold dark matter simulations. The space of theoretical predictions across both varying dark matter and baryonic physics is thus vast and difficult to constrain.

Upcoming surveys will provide higher-precision measurements of dark matter distributions in galaxies and galaxy clusters. For example, data from surveys like Euclid [9], DESI [10] and Rubin [11] will make it possible to combine weak lensing and dynamical data to reconstruct full three-dimensional density profiles of galaxies and clusters. To make full use of these data, and to interpret them without imposing prior assumptions from simulations, we require a model-independent method that can uncover the structure of haloes directly from observations.

To do this, we propose using a machine learning algorithm called Symbolic Regression (SR) [12]. SR searches the space of analytic functions to find a mathematical expression that can describe a dataset well. This makes it particularly suitable for situations like ours where we do not want to assume any specific functional form from the outset. In this way, it can infer dark matter halo density profiles from observations without relying on assumptions from simulations.

One of the main advantages of SR over other machine learning methods is that it produces interpretable mathematical expressions instead of black-box models. While it may be too optimistic to expect that these expressions will directly reveal new physical laws, they can still serve as models for testing correlations between profile parameters or for identifying systematic deviations

from the standard halo profiles. Moreover, because these profiles are more accurate empirically, they can also be dropped into existing analyses (for example, when making predictions using the halo model) as a replacement for the NFW profile, providing an observationally better-motivated alternative.

In this work, we focus on a specific SR algorithm called Exhaustive Symbolic Regression (ESR) [13]. Given a set of mathematical operators, ESR generates all possible combinations to produce functions up to a given level of complexity (defined to be equal to the number of nodes in the tree representation of a function). For each candidate function, it assesses how well it fits the data and how simple it is. This trade-off is formalized using the Minimum Description Length (MDL) principle: essentially, the best model is the one that gives a good fit to the data using the simplest possible formula.

The paper is structured as follows: we begin by introducing the ESR algorithm and describing how it can be used to combine data from multiple datasets (§2). Then we describe how it can be used to constrain halo profiles from weak lensing data. In §3, we apply ESR to mock weak lensing data to illustrate its effectiveness and discuss how data quality affects the resulting functions. Finally, we conclude and outline how the method can be extended to other probes such as galactic rotation curves in §4. Throughout this paper we adopt a concordance Λ CDM cosmology with $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$ and $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ and use \log to denote the natural logarithm.

2. Methods

(a) Symbolic regression framework

SR is a machine learning framework designed to recover mathematical expressions directly from data. Given a dataset and likelihood model, SR aims to find the equations that best describe the underlying relationships [13–15]. The goal is to balance accuracy, ensuring the final expression reproduces the data reliably, and simplicity, to find equations that are easy to interpret and more likely to generalize well. In doing so, SR can produce insight into the physical processes that generated the data.

In recent years, many SR algorithms have been developed. Most begin by specifying a set of mathematical operators (e.g. +, −, ×, exp, etc.) and proceed by combining these to generate candidate expressions. The complexity of an expression is defined as the number of operators, parameters and variables needed to construct the function. The most common approaches to generating functions at different complexities rely on genetic programming, which evolves the initial population of functions through operations such as mutation and crossover [16–18]. These methods aim to be computationally efficient by only studying a subset of the function space. The hope is that this is the optimal subset for the data in question, however this exploration often misses important expressions even at relatively low complexities [13,19].

This limitation motivated the development of ESR, an algorithm that, given a set of operators, generates all possible functions up to a specified complexity [13,14,20]. This brute-force approach constructs and tests every valid expression within the allowed function space, ensuring no potentially good functions are missed. This makes ESR the best algorithm to find expressions at low complexities and a valuable benchmark for guiding stochastic methods such as genetic programming at higher complexity.

Although ESR is described in detail in [13], in this section we briefly describe its two main stages. First, it systematically generates all possible expressions at a given complexity and optimizes their free parameters. Second, it ranks the resulting expressions using an information theory-motivated metric that accounts for both goodness-of-fit and model complexity.

(i) Generating and optimizing functions

To generate mathematical expressions, ESR begins by the user specifying a set of operators. These can be classified into three types: binary (e.g. +, ×), unary (e.g. exp, log) and nullary

(parameters and variables). Mathematical expressions can be represented as trees, where each node corresponds to an operator and the structure of the tree encodes the mathematical composition of the function [21]. The complexity of a function can also be defined as the number of connected nodes in the tree representation. For example, the standard NFW profile can be written as $\theta_0/(r(|\theta_1| + r)^2)$, which is a complexity 9 function (the modulus sign is not counted as an operator).

Given a fixed complexity, ESR systematically generates all possible tree structures with that number of nodes and then populates each node with every possible combination of operators with the correct arity. In this framework, the choice of operators and the maximum complexity are the only degrees of freedom.

Once the complete set of candidate expressions has been generated, ESR simplifies expressions and eliminates duplicates. Since multiple tree structures can generate equivalent mathematical expressions (e.g. $e^{a+x} = e^a \cdot e^x$), it is important to identify these cases, as retaining only unique expressions avoids redundant computations in the next step.

Finally, ESR optimizes the free parameters in each of the candidate models. We consider the best-fit parameters to be those that maximize the likelihood function. Here we use Gaussian likelihood with uncorrelated errors. We consider two optimization algorithms to be particularly useful for parameter optimization. The faster option is to use a gradient-based method such as BFGS [22–25], which uses gradient information to effectively navigate the likelihood surface and locate maxima. However, we find that this algorithm sometimes gets trapped in local maxima, particularly when the likelihood landscape is complex. To mitigate this, we also use a derivative-free method like Nelder–Mead [26], which can be more robust in avoiding local maxima, although it typically takes longer to converge.

To ensure that the parameter estimation has converged to a reliable solution, the optimization procedure is repeated up to N_{iter} times. After each iteration we check whether the same best solution (within some likelihood tolerance) has been recovered N_{conv} times, in which case we terminate the optimization as successful. If the highest maximum has not been reached N_{conv} times after N_{iter} iterations, the algorithm returns the maximum-likelihood solution achieved. The values of N_{conv} and N_{iter} used here are the same as those specified in [13].

This process is repeated for all complexities considered. However, there is a practical upper limit to how high one can go. As the number of nodes is increased, the number of possible expressions grows exponentially, making the computation infeasible beyond a certain complexity. Since the standard NFW profile is a complexity 9 function, we set our maximum search complexity to 9, ensuring NFW is included in our analysis while remaining within these computational constraints.

(ii) Model selection via minimum description length

The result of the previous procedure is a list of all functions up to a given complexity, along with their best-fit parameters. The next step is to rank these functions. A natural choice would be to use the maximum likelihood as a criterion. Plotting each function on the complexity-likelihood plane, the functions that achieve the highest likelihood at each complexity form what is known as the *Pareto front*. Along this front, no function can be made more accurate without also becoming more complex and vice versa.

However, the Pareto front itself only defines a set of optimal trade-offs; it does not provide a principled way of choosing one function over another within that set. In other words, Pareto optimization gives a frontier of good functions in terms of accuracy but not a ranking. To obtain a one-dimensional ordering of the models, an additional criterion is needed that weights complexity against accuracy. To do this, ESR uses an information-theory approach known as the MDL principle [27–29]. MDL states that the best mathematical description of the data is the one that compresses information the most, or in other words that needs the fewest units of information to transmit the data. Here we will perform all calculations with natural logarithms rather than base-2, and thus our unit of information is nats rather than bits.

This idea is represented by the *Description Length*, $L(D)$, which is defined as $L(D) = L(D|H) + L(H)$. Here, $L(D|H)$ quantifies how well the hypothesis H fits the data D . Under the Shannon–Fano coding scheme, the optimal encoding for the residual has $L(D|H) = -\log \mathcal{L}(D|\hat{\theta})$ [30], where \mathcal{L} is the likelihood and $\hat{\theta}$ are the maximum-likelihood parameters. The second term, $L(H)$, measures the complexity of the hypothesis itself. It penalizes models with more operators, more parameters and parameters that need to be specified to higher precision to achieve a high likelihood.

The full expression for $L(D)$ is derived in [13] and is given by

$$L(D) = -\log(\hat{\mathcal{L}}) + k \log(n) + p \log(2) + \sum_{i=1}^p \log(|\theta_i|/\Delta_i) + \sum_j \log(c_j), \quad (2.1)$$

where k is the number of nodes to represent the function, n is the number of unique operators, p is the number of free parameters, $1/\Delta_i$ is the precision to which the parameters can be specified and c_j are constant natural numbers. The hat notation indicates evaluation at the maximum likelihood point.

In some cases, it might happen that $|\theta_i|/\Delta_i < 1$, meaning that the maximum likelihood estimate of a parameter is indistinguishable from 0 within the uncertainty tolerance. In these situations, we set these parameters to 0, recalculate the likelihood, and reduce the parameter count p by one. Since this occurs only when a parameter is already poorly constrained, the overall effect on $L(D)$ is minimal.

Equation (2.1) can be interpreted as an approximation to the Bayesian evidence under a certain choice of parameter prior and function prior, where the latter is given by $L(H)$. More sophisticated choices for the prior can be expressed in terms of language models [31]; however, for simplicity, in this work we use the formulation of the MDL as given by [13].

This expression balances model simplicity and goodness of fit. Simple functions are penalized if they fail to fit the data well (high $L(D|H)$), while highly complex functions that overfit are penalized for their complexity (high $L(H)$). This is naturally influenced by the quality of the data. When the data are not particularly constraining (few data points or large uncertainties), the likelihood term has less influence, and simpler functions with higher prior probabilities (and hence low $L(H)$) are favoured.

(iii) Global versus local parameters

We aim to find dark matter density profiles that perform well across different galaxies or galaxy clusters. This requires combining results from individual systems to evaluate overall model performance. There are different ways to do this depending on how the model parameters are treated.

One option is to consider that the parameters are independent for each dataset, such that they are optimized separately for each galaxy or cluster. This means applying ESR independently on each system and afterwards combining the resulting fits. To compute the overall description length in this case, we sum the likelihood and parameter precision terms across galaxies, but count complexity terms (i.e. those depending on the number of nodes k and operators n) only once. This results in a total description length given by

$$L(D) = -\sum_{a=1}^N \log \mathcal{L}^{(a)} + k \log(n) + N \cdot p \log(2) + \sum_{a=1}^N \sum_{i=1}^p \log(|\theta_i^{(a)}|/\Delta_i^{(a)}) + \sum_j \log(c_j), \quad (2.2)$$

for N independent galaxies/galaxy clusters labelled by the index a .

Alternatively, we can consider the possibility that a subset of parameters is shared across all galaxies; we refer to these as *global* parameters. This allows for the possibility that some of the parameters are universal constants that appear in the mathematical description of dark matter haloes, for example the square in the denominator of the NFW profile. In this case, a single set of shared parameters is optimized using combined data from all systems. This results in a lower

total likelihood, since the model has fewer degrees of freedom to fit each galaxy. However, it also reduces the model complexity by significantly decreasing the number of free parameters. This trade-off can lead the data to prefer global over local parameters.

Ideally, one would like to combine both approaches, allowing some parameters to be global while others remain galaxy-specific in the attempt to find the overall lowest description length models. This allows the data itself to determine which model structure is appropriate. To implement this in ESR, we use a two-step procedure. First, we perform fully local fits, where all parameters are independently fitted to each galaxy. This gives the best possible likelihood that each function can have. Then, we evaluate whether $L(D)$ can be improved by ‘globalizing’ some of the parameters. The aim is to determine whether making some parameters global could potentially improve the $L(D)$ enough to improve the function’s ranking beyond that of the best-performing function with all local parameters. To estimate this, we consider the maximum theoretical possible gain in $L(D)$, which occurs when the likelihood remains unchanged on globalization. This yields (setting the precision cost of new global parameters to zero)

$$\max(\Delta L) = \max(L_{\text{local}} - L_{\text{global}}) = \sum_{a=1}^N \sum_{i=1}^{p_{\text{global}}} \log \left(|\theta_i^{(a)}| / \Delta_i^{(a)} \right) + p_{\text{global}}(N-1) \log 2, \quad (2.3)$$

where p_{global} denotes the number of parameters being considered for globalization. It is necessary to consider all possible combinations of local and global parameters for each function.

Using this, we identify good candidates for globalization as those for which $\max(\Delta L)$ is large enough to put the function in the top 10 ranked models. For any such function, we proceed to globalize the corresponding parameter and re-evaluate the description length. This involves re-fitting the function setting the desired parameters to global.

One straightforward way to do this is to perform a single optimization over all $N_{\text{params}} = p_{\text{global}} + N \cdot p_{\text{local}}$ parameters, where N is the number of galaxies or clusters. However, this requires searching in a very high-dimensional space, and we find that the optimizer often struggles to explore the likelihood surface effectively.

We therefore use instead a nested optimization. In this scheme, an outer optimizer searches in the space of global parameters. For each set of global parameters it tries, an inner optimizer fits the best corresponding local parameters for each cluster. This process is repeated until the global optimizer converges. We define convergence as N_{conv} iterations showing an absolute improvement in the likelihood function of less than 0.1.

While computationally more expensive, this method has the advantage that each optimization, both for the local and global parameters, occurs in a low-dimensional space. This improves the convergence and stability of the fitting procedure. The rest of the terms for the description length are calculated as in [equation \(2.2\)](#), where the sum is performed over local parameters and global parameters are only considered once.

(b) Application to weak lensing

(i) Weak lensing basics

Weak gravitational lensing refers to the magnification and distortion of background galaxy shapes caused by the deflection of light by intervening mass along the line of sight. It offers a powerful and direct way of mapping the matter distribution in galaxies and clusters without relying on assumptions about their dynamical state. Galaxy clusters, being among the most massive gravitationally bound systems, produce strong gravitational wells that make them particularly effective lenses. As such, they provide unique environments for studying the distribution of matter, which is dominated by dark matter.

The effects of weak gravitational lensing are described by the convergence κ , which describes the isotropic magnification, and the complex shear γ that describes the anisotropic distortion of

galaxy shapes. In practice, the observable quantity is not the shear γ but the reduced shear g , which accounts for the magnification effect from the convergence

$$g = \frac{\gamma}{1 - \kappa}. \quad (2.4)$$

The shear γ can be decomposed into two components: the tangential component γ_+ and the 45°-rotated component γ_\times . At the radii we consider, $g_+ \simeq \gamma_+$. The tangential shear γ_+ is directly related to the azimuthally averaged surface mass density $\Sigma(R)$ through [32]

$$\gamma_+ = \frac{\langle \Sigma(< R) \rangle - \Sigma(R)}{\Sigma_{\text{cr}}(z_l, z_s)} \equiv \frac{\Delta \Sigma(R)}{\Sigma_{\text{cr}}(z_l, z_s)}, \quad (2.5)$$

where R is the projected radius from the lens centre, $\Delta \Sigma(R)$ is the Excess Surface Density (ESD) and $\langle \Sigma(< R) \rangle$ is the average surface density within R , given by

$$\langle \Sigma(< R) \rangle = \frac{2}{R^2} \int_0^R \Sigma(R') R' dR'. \quad (2.6)$$

The critical surface density Σ_{cr} is

$$\Sigma_{\text{cr}}(z_l, z_s) = \frac{c^2}{4\pi G} \frac{D_A(z_s)}{(1 + z_l)^2 D_A(z_l) D_A(z_l, z_s)}, \quad (2.7)$$

where c is the speed of light, G is the gravitational constant, z_l is the lens redshift, z_s is the source redshift and D_A is the angular diameter distance. The factor $(1 + z_l)^2$ converts to co-moving surface mass density.

The projected surface density, $\Sigma(R)$, can also be related to the three-dimensional mass density $\rho(r)$ by

$$\Sigma(R) = \int_{-\infty}^{\infty} \rho(\sqrt{R^2 + z^2}) dz, \quad (2.8)$$

where z is the line-of-sight coordinate and r is the three-dimensional distance, such that $r^2 = R^2 + z^2$.

By combining these equations, we can derive the predicted $\Delta \Sigma(R)$ profile for any candidate three-dimensional density function $\rho(r)$. ESR can use this transformation to fit ESD data and rank the performance of each function. To simplify the numerical calculation of $\Delta \Sigma(R)$, we use an alternative expression that avoids double integrals [33]

$$\Delta \Sigma(R) = \frac{4}{R^2} \int_0^R dr r^2 \rho(r) - 4R \int_0^{\pi/2} d\theta \frac{\rho(R \sec \theta)}{4 \sin \theta + 3 - \cos(2\theta)}. \quad (2.9)$$

In principle, the total excess surface density $\Delta \Sigma(R)$ has contributions from several components: the one-halo dark matter term, baryonic components and the two-halo term.

The baryonic contribution itself has two main parts: the stellar mass of the brightest cluster galaxy (BCG), which is important only at very small scales ($R \lesssim 0.1 h^{-1}$ Mpc), and the hot, extended Intracluster Medium (ICM) or gas component. The two-halo term arises from correlated matter in surrounding large-scale structure and becomes important at scales approaching and beyond the virial radius [34].

We analyse data in the intermediate radial range ($0.3 - 3 h^{-1}$ Mpc), where the one-halo term dominates in Λ CDM for the cluster masses under consideration. We therefore neglect the two-halo term in our analysis, providing a detailed quantitative justification in §2b(ii). We do not model the baryonic components separately from the dark matter, with the hot ICM gas expected to contribute around one-sixth of the total mass. Therefore, by neglecting the gas component, this inference procedure applied to real data effectively constrains the total mass profile rather than the purely dark-matter contribution. We discuss how the method could be developed to explicitly model both the two-halo term and the subdominant baryonic component in §4.

An important consideration is the quality of the data itself. Since more precise data tend to prefer more complex models, the quality of the data is important to understand if profiles like NFW can actually be recovered from observations. In the next section, we present a worked example using mock weak lensing data. We focus on the one-halo term only and investigate how data quality affects the performance of ESR, and the extent to which the function generating the data can be expected to be recovered by ESR.

(ii) Mock data generation

To illustrate the method and test the ability of ESR to recover known physical profiles, we generate a synthetic sample of 150 galaxy clusters with an NFW profile, that is designed to closely resemble the dark matter halo properties inferred for the XXL X-ray selected cluster sample in the analysis of [34]. The halo masses were drawn from a Gaussian probability density function in $Z = \log(M_{200}/h^{-1}M_{\odot})$ with a mean $\mu_Z = 14$ and a dispersion $\sigma_Z = 0.5 \log 10$. The range of true M_{200} masses for the simulated sample is $4.0 \times 10^{13} \lesssim M_{200}/(h^{-1}M_{\odot}) \lesssim 5.0 \times 10^{14}$. Corresponding concentrations were drawn from the scattered c_{200} - M_{200} relation of [35], assuming a log-normal intrinsic dispersion of $\sigma(\log c_{200}) = 0.15 \log 10$ and a fixed redshift of $z = 0.3$, the median redshift of the XXL sample.

The true ESD, $\Delta\Sigma(R)$, profile for each cluster was calculated in 10 logarithmically spaced bins of co-moving radius from $R_{\min} = 0.3h^{-1}$ Mpc to $R_{\max} = 3h^{-1}$, again following [34], avoiding the inner regions ($R < 0.3h^{-1}$ Mpc). The lower limit is to ensure our measurements are not significantly affected by masking, imperfect deblending, or photometric biases caused by the BCGs. Furthermore, this R_{\min} is much larger than the typical offsets between the BCG and the X-ray peak for these clusters, meaning smoothing of the lensing signal due to miscentring is also not expected to be a significant effect. The upper limit is to avoid the region where the two-halo term becomes significant.

To model observational uncertainties, we add realistic noise to the simulated measurements. For current surveys, the uncertainty in measured ESD profiles is dominated by the intrinsic dispersion of unlensed galaxy shapes. This *shape noise* scales with the number of source galaxies in a radial bin (n_i) as $\sigma_i \propto 1/\sqrt{n_i}$. Because we use logarithmic radial bins, the annular area—and hence n_i —grows approximately as R^2 , implying $\sigma_i \propto 1/R$. In real cluster data the ESD declines as approximately $1/R$ in the radial range we consider, yielding an approximately constant fractional uncertainty. Therefore, we adopt the following error model for our synthetic data:

$$\Delta\Sigma(R_i) = \Delta\Sigma_{\text{true}}(R_i) + \epsilon_i, \quad (2.10)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i), \quad \sigma_i = f \Delta\Sigma_{\text{true}}(R_i), \quad (2.11)$$

with $i = 1, \dots, 10$ radial bins and \mathcal{N} a normal distribution with mean zero and dispersion σ_i . To test our method's performance under different data quality scenarios, we generated six mock datasets with

$$f \in \{0.01, 0.05, 0.20, 0.40, 0.60, 0.80\}.$$

We can compare these f values with typical fractional uncertainties of current lensing surveys. For instance, the HSC-XXL cluster sample in [34] has a median fractional uncertainty of approximately $f \approx 0.60$, while a higher signal-to-noise survey like CLASH has an uncertainty closer to $f \approx 0.38$ [36].

Our chosen range, $f \in \{0.01, \dots, 0.80\}$, therefore spans the full spectrum of data quality. The low-noise values ($f = 0.01, f = 0.05$) represent highly idealized precision, while the upper end of our range ($f = 0.40, 0.60, 0.80$) is representative of, or even more challenging than, the data from current-generation surveys. This allows us to test our method's performance in both idealized and realistic observational scenarios.

Our analysis is restricted to modelling the one-halo term of the cluster mass profile. One might wonder whether the two-halo term contribution is significant, particularly at the outermost datapoints. In the Λ CDM framework, this question was addressed in detail by [37] for the HSC cluster sample. Since our dataset is designed to be comparable with the XXL-HSC sample, their findings are directly relevant. For their cluster sample, which has a lensing-weighted mean mass (the effective average mass of a cluster sample, obtained by weighting each cluster by the strength or precision of its lensing signal) of $M_{WL} = 8.3 \times 10^{13} h^{-1} M_{\odot}$, they found the two-halo term's contribution to the stacked ESD profile to be negligible across their entire radial range (see fig. 1, [37]). Even at their outermost datapoint ($R \approx 3 h^{-1} \text{Mpc}$), the two-halo term was still an order of magnitude smaller than the fitted one-halo term.

We have confirmed this for our own sample by calculating the two-halo term (following the method in [37]) for our lowest-mass clusters, where this term is expected to be relatively most significant. For the lowest-mass cluster in our sample—where the two-halo contribution is most significant—it remains about 4.5 times smaller than the one-halo term at the outermost radial bin, and roughly 10 times smaller for the second-lowest-mass system.

Given that the two-halo term is so small, one route would be to simply include the standard Λ CDM 2-halo term as an additional source of uncertainty in the covariance. However, the above implies this would not noticeably alter our conclusions on the quality of data required to recover NFW. For a future application to real data, there is a deeper methodological issue: given the aim of the study is to recover the halo profile, assuming NFW to calculate the two-halo term is unsatisfactory. In principle, one could use each halo profile under consideration to calculate its own two-halo term, but this would be costly. In the near-term, the most practical option when applying this method to real data is to test the stability of the preferred profiles against the radial range of the data (e.g. by re-fitting with the outermost bins removed).

3. Results

We applied our ESR pipeline to the mock cluster datasets, optimizing all candidate functions up to complexity 9 for each fractional uncertainty level. To illustrate the fitting procedure and its results, we will first focus on the low-noise regime ($f = 0.05$) as a representative example.

Figure 1 shows the ESR-derived fits for a single mock cluster from this $f = 0.05$ dataset. The left panel compares the mock ESD data (black points) with the best-fit NFW profile and the next two top-ranked functions discovered by ESR for this cluster, which are $|\theta_0|^{|\theta_1|} r$ and $(|\theta_0|/r)^{|\theta_1|}$. The right panel shows the corresponding three-dimensional density profiles (ρ) for these same models.

We then ranked the functions based on their combined performance across all clusters. As explained in §2a(iii), our procedure first optimizes all functions assuming purely local parameters, and then identifies which models benefit from ‘globalizing’ one or more parameters. For this $f = 0.05$ uncertainty level, this method identified 734 functions as candidates for the globalizing step out of the 36 217 total unique functions evaluated up to complexity 9 (2% of the functions). This results in a total of 2432 unique function and parameter-type (local/global) combinations that were then optimized.

Table 1 summarizes the top-ranked expressions across all global and local combinations. The total description length for each function is broken down into three components: the residuals between the predicted and observed data, the structural complexity of the functional form, and the penalty due to the number of free parameters, as detailed in the table footnote.

The best-ranked function is the NFW profile. This is a validation of our mock-data simulation: in the low-noise regime, ESR successfully recovers the true, underlying model that was used to generate the data. The generalized NFW (gNFW) profile, which introduces an extra parameter θ_2 per halo by allowing the inner slope of the profile to vary, achieves a better likelihood but is penalized for its extra complexity, falling to rank 40 in total $L(D)$. We find a median and standard deviation for the slope θ_2 of 2.01 ± 0.31 across all clusters, close to the expected NFW value of 2.

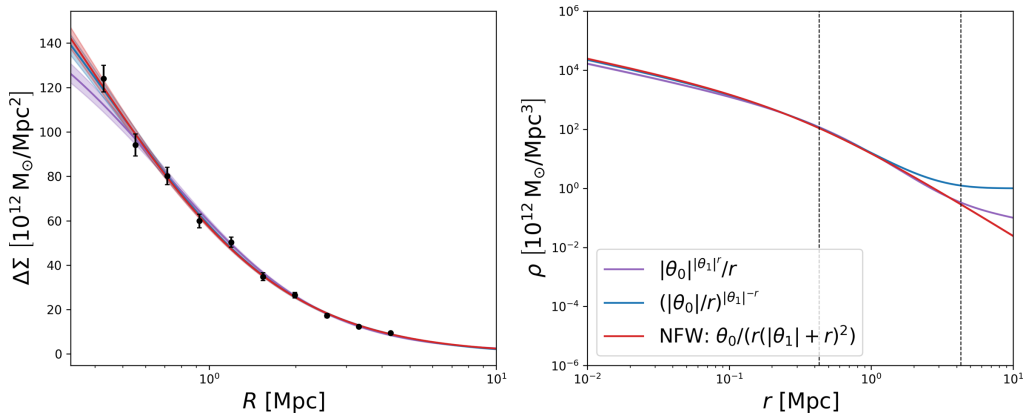


Figure 1. An example result of the ESR fitting procedure for one mock galaxy cluster, generated with a fractional uncertainty of $f = 0.05$. *Left panel:* the mock weak lensing ESD data (black points) are compared with the best-fit NFW profile (red) and the next two best-fitting functions discovered by ESR, $|\theta_0|^{|\theta_1|}r/r$ (purple) and $(|\theta_0|/r)^{|\theta_1|-r}$ (blue). *Right panel:* the corresponding three-dimensional density profiles (ρ) for the three models as a function of the intrinsic three-dimensional radius. The vertical dotted lines mark the radial range of the ESD data from the left panel.

The second-ranked function is the gNFW profile with its outer slope θ_2 treated as a global parameter shared across all clusters. Its best-fit value of $\theta_2 = 2.024 \pm 0.008$ is extremely close to the canonical NFW value of 2. This model achieves a slightly better likelihood (Residual = 617.79) than the standard NFW (618.40) but is penalized for having one extra parameter just enough by the MDL to fall to second place.

We also note that despite the global search, the final ranking is dominated by models with purely local parameters. The gNFW is the only function in this top list that successfully incorporates a global parameter. This makes sense, local parameters are necessary to capture cluster-to-cluster variations in mass and concentration, while global parameters are only appropriate for universal properties like an outer slope.

Interestingly, most of the highest-ranked functions have two free parameters, same as the standard NFW profile. The MDL metric heavily penalizes additional parameters, and our results show that models with three or more parameters do not provide a sufficient improvement in fit to overcome this penalty.

The trade-off between model fit (residuals) and complexity is shown in [figure 2](#). The red axis (left, $\Delta L(D)$) shows the total Description Length. The blue axis (right, $|\Delta \log \mathcal{L}|$) shows the goodness-of-fit. Both metrics are shown relative to the best-performing model in that metric, so a value of 0 represents the minimum (best) score. Therefore, a higher point on the plot indicates a worse-performing model.

The NFW profile is marked with a star. It has the best description length, even when it does not have the best likelihood. The function that achieves the best likelihood ($\mathcal{L} = 481.98$ nats) is

$$\left| \frac{\theta_0}{\theta_2 - r^{\theta_1}} \right|^{\theta_3},$$

but it is penalized for its higher functional complexity, finally ending at rank 545 in overall $L(D)$.

[Figure 2](#) also shows that the total $L(D)$ is still decreasing at complexity 9. This suggests that we have not necessarily found the absolute best model, as better-fitting functions may still exist at higher complexities. In fact, all of the top-ranked functions in [table 1](#) are already at this complexity 9 limit. While extending the search to complexities beyond 9 is a natural next step, it is computationally expensive. We, therefore, present our results as the best-performing functions discovered within the search space up to complexity 9.

Table 1. Highest ranked density models ordered by their total description length $L(D) = \text{Residual} + \text{Parameter length} + \text{Function length}$. The ‘global params’ column specifies which parameters are set to global for each function.

rank	$\rho(r)/10^{12}M_{\odot} \text{Mpc}^{-3}$	complexity	global params	description length			
				residual ^a	parameter ^b	function ^c	total
1	NFW: $\theta_0/(r(\theta_1 + r)^2)$	9	—	618.40	1043.06	24.04	1685.51
2	gNFW: $\theta_0/(r(\theta_1 + r)^{\theta_2})$	9	θ_2	617.79	1053.98	25.08	1696.85
3	$ \theta_0/r - r ^{ \theta_1 ^r}$	9	—	647.02	1115.00	14.48	1776.50
4	$(\theta_0 /r)^{ \theta_1 ^{-r}}$	8	—	650.62	1115.01	12.88	1778.51
5	$\theta_0/(r(\theta_1 - r^2))$	9	—	889.52	945.90	14.48	1849.90
6	$ \theta_0 ^{1/(\theta_1-1/r)^r}$	9	—	902.78	1002.83	14.48	1920.09
7	$ \theta_0 ^{ \theta_1 ^{-r}}/r$	8	—	834.67	1076.40	12.88	1923.94
8	$ \theta_0 - r ^{ \theta_1 ^r}/r$	9	—	834.69	1076.15	14.48	1925.32
9	$ \theta_0 + r ^{ \theta_1 ^r}/r$	9	—	834.69	1076.18	14.48	1925.35
10	$\theta_0/(\theta_1 - r^{1-\theta_2})$	9	—	570.25	1383.21	14.48	1967.94
.
.
.
40	gNFW: $\theta_0/(r(\theta_1 + r)^{\theta_2})$	9	—	545.85	1597.07	25.08	2167.15

$$- \sum_{a=1}^N \log \hat{\mathcal{L}}^{(a)}$$

$$^a - \sum_{a=1}^N \log \hat{\mathcal{L}}^{(a)}$$

$$^b p \log(2) + \sum_{a=1}^N \sum_{i=1}^p \log \left(\left| \hat{\theta}_i^{(a)} \right| / \Delta_i^{(a)} \right)$$

$$^c k \log n + \sum_j \log c_j$$

We repeat this analysis for all fractional uncertainties. Figure 3 shows how the different components of the description length change with different fractional uncertainties (f) for the NFW profile. The residual component (red line), which represents the goodness-of-fit ($-\log \hat{\mathcal{L}}$), is remarkably flat for all uncertainty levels. This indicates that the model is achieving an optimal fit at each noise level. As the noise σ increases, the absolute difference between the data and model also increases, but the normalized residual remains constant. By contrast, the parameter length (blue line) drops significantly as the fractional uncertainty f increases. At low noise, the data are highly constraining, and the parameters are determined with high precision (i.e. narrow posteriors). This high precision represents a large amount of information, which has a high ‘cost’ to encode.

As noise increases, the data’s constraining power drops, leading to less precise, low-information (and thus low-cost) parameters. In this high-noise regime, it becomes more common for parameters to be ‘snapped’ to zero (i.e. $|\hat{\theta}_i|/\Delta_i < 1$), which reduces the effective parameter count p for that cluster. For any parameters that are not snapped but are still poorly constrained, the information-dependent part of the parameter cost, $\sum_{a=1}^N \sum_{i=1}^p \log(|\hat{\theta}_i^{(a)}|/\Delta_i^{(a)})$, tends to zero. Therefore, in the high-noise limit, the parameter length converges to its minimum ‘structural’

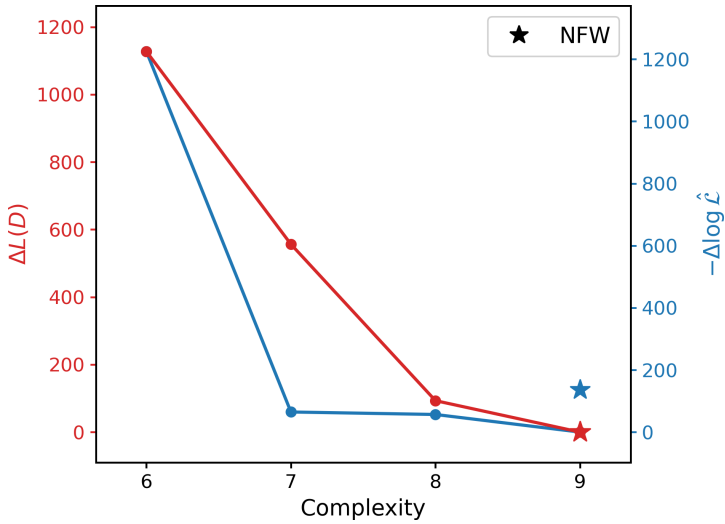


Figure 2. Optimal values of the objectives at each complexity for a fractional uncertainty of $f = 0.05$. The red curve (left axis) shows the per-complexity minimum description length relative to the global best description-length (NFW in this case). The blue curve (right axis) shows the per-complexity best likelihood relative to the global best likelihood across all models. Stars mark NFW in both metrics. For both axes, lower values on the plot indicate better-performing models.

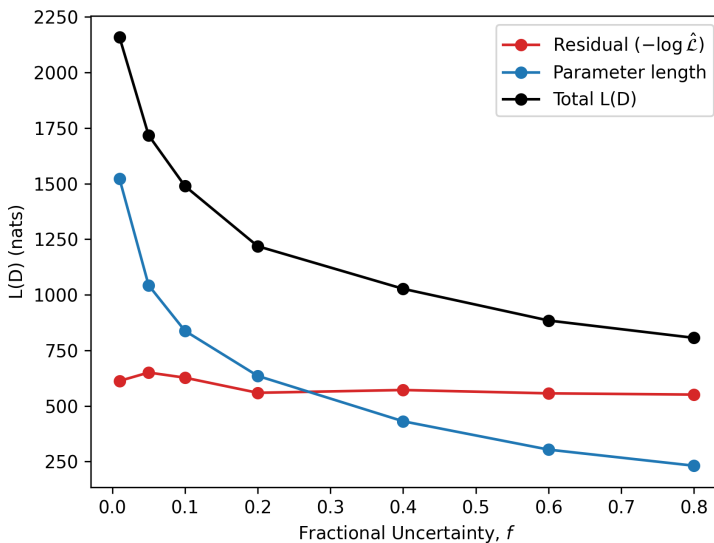


Figure 3. Breakdown of the total Description Length ($L(D)$) for NFW as a function of the fractional uncertainty (f) in the mock data. The Total $L(D)$ (black) is the sum of the Residual component ($-\log \hat{\mathcal{L}}$) (red), the Parameter length (blue) and the function complexity. The plot shows that the parameter cost (blue line) dominates the total cost at low noise and drops significantly as the data become less constraining. The residual term, representing the goodness-of-fit, remains relatively flat at all noise levels.

cost, which is given by the $p \log 2$ term in equation (2.1). For a two-parameter model ($p_{\text{cluster}} = 2$) like NFW and our sample of 150 clusters, this floor is:

$$L(\theta)_{\min} = 150 \times 2 \log 2 \approx 207.94 \text{ nats},$$

which is presumably the value that the blue line is approaching in figure 3.

We also studied how the NFW profile's performance compares with the ESR-discovered functions across all data quality levels. Figure 4 summarizes this analysis. The bottom panel shows

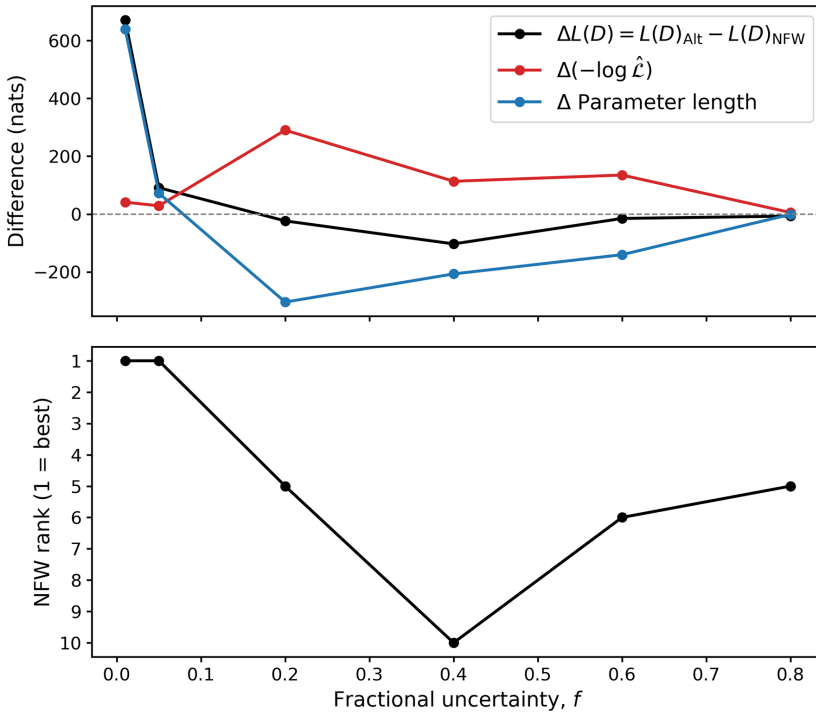


Figure 4. The performance of the NFW profile relative to the other ESR-discovered functions, shown as a function of the fractional uncertainty (f) in the mock data. *Top panel:* the difference in total Description Length, $\Delta L(D) = L(D)_{\text{Alt}} - L(D)_{\text{NFW}}$, where $L(D)_{\text{Alt}}$ is the description length of the best function other than NFW (black). The equivalent difference for likelihood (red) and parameter length (blue) also shown. Positive values indicate that NFW is the best-fitting model function. *Bottom panel:* the absolute rank of the NFW profile in the full list of functions, where a rank of 1 is the best. The plot shows that NFW is the $L(D)$ -preferred model at low uncertainty, but it is overtaken by other functions at intermediate f . In the high-noise regime ($f \gtrsim 0.6$), it once again becomes one of the top-ranked models.

NFW's absolute rank (where 1 is best) among all ESR functions up to complexity 9. The top panel shows the $L(D)$ gap, which we define as $\Delta L(D) = L(D)_{\text{alt}} - L(D)_{\text{NFW}}$, where $L(D)_{\text{alt}}$ is the $L(D)$ of the best model found by ESR at each fractional uncertainty (which may differ with f). Since a lower $L(D)$ is better, positive values on the black line mean NFW is preferred. The red line shows the difference in fit quality and the blue line shows the difference in parameter cost.

Three regimes appear. At very low noise ($f \leq 0.05$), NFW is clearly preferred (rank 1). Both the red and blue curves are positive, showing that NFW achieves a better likelihood and a lower parameter cost than its competitors. As figure 3 highlights, the parameter-length term dominates the $L(D)$ difference here: some functions can approach NFW's likelihood, but only at the price of substantially finer parameter tuning.

However, in the intermediate-noise regime ($f \approx 0.2 - 0.4$), NFW's performance gets worse. Here, the red line is still positive (NFW fits better), but the blue line is strongly negative. This indicates that an alternative function has a lower parameter cost. Actually, the preferred model for $f = 0.2, 0.4, 0.6$ is $\theta_0/(r^2(r+1))$. The MDL framework concludes that NFW's slightly better fit is not worth its much higher parameter 'fine-tuning' cost, so it prefers the alternative.

At very large f , both likelihood differences and parameter costs compress, so NFW may no longer be best but typically remains competitive. This behaviour is likely driven by the fractional-uncertainty model used in our mock. We set $\sigma(r) = f \times \rho(r)$, which mirrors weak lensing-like errors that scale with the signal amplitude. This choice makes the absolute uncertainties at small radii bigger than those in the outskirts, so the outer data points carry more weight in the likelihood. At large r , NFW asymptotes to r^{-3} . Many alternative profiles, like $\theta_0/(r^2(r+1))$, can reproduce

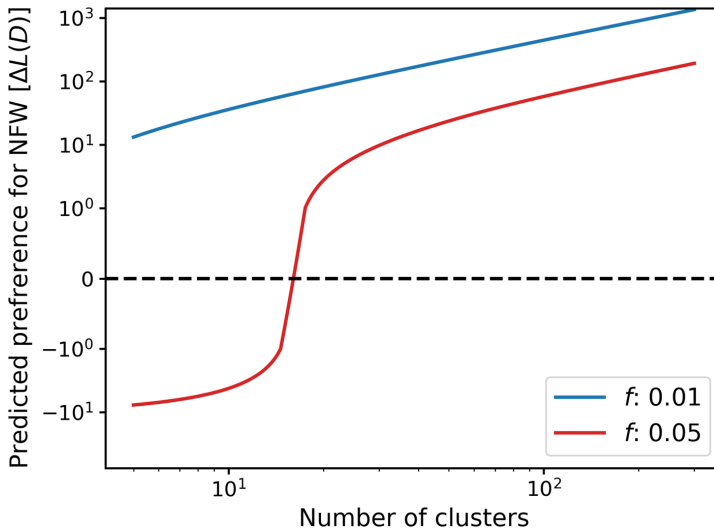


Figure 5. Change in description length ($\Delta L(D)$) between NFW and the best alternative dark matter halo model. The curves show results for two different levels of Gaussian noise in the ESD data: $f \in \{0.01, 0.05\}$ as a function of the number of clusters. Any point higher than 0 shows a preference for NFW.

this outer slope while differing in the centre. As a result, the fractional model effectively down-weights the inner cusp (r^{-1}) that distinguishes NFW, making alternative functions competitive in likelihood. Therefore, more simple functions that can reproduce the outer behaviour but have less parameters are preferred, which is exactly what we are seeing in figure 4.

This is just a property of weak lensing weighting. In practice, one could combine weak lensing with probes that better constrain the inner regions (e.g. strong lensing or stellar kinematics). This also means model recoverability depends as much on the data as on the fitting criterion.

It is interesting to note, however, that even when NFW is not the $L(D)$ -winner, it always remains one of the top models. As the bottom panel of figure 4 shows, NFW remains within the top 10 models across all uncertainty levels. This suggests that if NFW is the best description of the data, ESR should be able to recover it as one of the best-fitting models.

In addition to noise, another important factor that influences model selection is the amount of available data. To study this effect, we assess how the description length changes under variations in the number of clusters N . After running the analysis for $N = 150$ clusters, we extrapolate our results to arbitrary sample sizes. The only terms that depend on N in equation (2.2) are the total likelihood $\sum^N \log \mathcal{L}$ and the parameter information term $\sum^N \sum^p (\frac{1}{2} \log(I_{ii}) + \log(|\theta_i|))$. Both of these terms scale linearly with N , allowing for a trivial transformation.

We can now use this framework to quantify how the statistical preference for the NFW profile scales with data quantity (i.e. the number of clusters, N). Since the extrapolation to a different number of clusters is linear, we can only study regimes for which NFW is identified already as the best model. So we focus on the two low-noise regimes, $f = 0.01$ and $f = 0.05$, which are the cases where figure 4 showed NFW is the best-performing model for the full sample.

For these two noise levels, we compute the description length difference, defined as $\Delta L(D) = L(D)_{\text{alternative}} - L(D)_{\text{NFW}}$, where $L(D)_{\text{alternative}}$ is the $L(D)$ of the next-best function. This difference is plotted as a function of N in figure 5. With this definition, positive values of $\Delta L(D)$ indicate a preference for NFW.

The results show that for the very low-noise case ($f = 0.01$), the preference for NFW is strong even with a very small number of clusters. As the uncertainty increases to $f = 0.05$, a larger dataset is required. While NFW is still preferred, a sample of at least $N \approx 20$ clusters is needed to accumulate significant statistical evidence.

At the next noise level we studied, $f = 0.20$, NFW was no longer the top-ranked model (falling to rank 5) for our $N = 150$ sample. It is possible that NFW could still be recovered with $N = 150$ clusters at an intermediate noise level (e.g. $f \approx 0.10$), or that recovering NFW at $f = 0.20$ would just require a larger sample size ($N > 150$). We leave a detailed exploration of this for future work.

4. Conclusions and future work

(a) Overview

We have demonstrated how ESR can be applied to constrain dark matter density profiles, specifically focusing on recovering halo structures from weak lensing ESD measurements.

To illustrate the method, we applied ESR to a mock sample of galaxy clusters with ESD profiles generated from the standard NFW model. We modelled the observational uncertainty as a constant fractional error, $\sigma_i = f \Delta \Sigma_{true}(R_i)$, and tested the pipeline across six distinct noise levels: $f \in \{0.01, 0.05, 0.2, 0.4, 0.6, 0.8\}$. By running ESR on this set of mocks, we studied the ranking of best-fitting functions to determine the conditions under which the input NFW profile is successfully recovered.

In the low-noise cases ($f = 0.01$ and $f = 0.05$), ESR identifies NFW as the optimal model. In this regime, the data are precise enough to penalize deviations from the true profile shape, and very few alternative profiles can reproduce the signal without requiring finely tuned parameters.

In the intermediate and high-noise regimes ($f \geq 0.2$), NFW is typically not recovered as the rank-1 model, though it consistently remains within the top 10. This result is probably a consequence of the fractional uncertainty modelling. Because the error scales with the signal ($\sigma_i \propto \Delta \Sigma(R_i)$), the absolute uncertainty is smallest at large radii. Consequently, the likelihood calculation is dominated by the cluster outskirts, where the profile asymptotes to r^{-3} . Simpler functions that can mimic this outer slope can achieve a likelihood comparable with NFW but with a lower parameter cost, thus outranking NFW in $L(D)$.

Nevertheless, the fact that NFW remains a leading candidate (top 10) even at higher noise levels is encouraging. It indicates that for current surveys with precisions similar to HSC ($f \approx 0.6$), ESR should successfully recover NFW as one of the top models if the underlying physics is indeed NFW-like.

We also investigated the impact of sample size by extrapolating our results from $N = 150$ clusters. For the low-noise cases ($f \leq 0.05$), we found that the preference for NFW is statistically preferred even with a very small number of clusters. At higher noise levels, larger samples would be required to overcome the preference for simpler functions driven by the outer-slope weighting.

Future work could extend this analysis in several directions. One could probe higher model complexities (beyond complexity 9) to assess the stability of these rankings. Additionally, more realistic physical modelling could be incorporated.

For instance, by not explicitly modelling the ICM gas, our inference effectively constrains the total mass profile rather than the purely dark-matter contribution. A more complete treatment would require marginalizing over the gas distribution (using observed or empirical ICM profiles) to isolate the dark matter profile, or using mock data to quantify any bias introduced by its neglect.

Similarly, we have assumed the dark matter halos are spherically symmetric. However, simulations show that haloes are typically triaxial, with moderate triaxiality also observed in data [38]. Studies have found for a sufficiently large sample of randomly oriented samples the bias in parameters when fitting an NFW halo is small. Hence it seems to first order it should be possible to model this triaxiality as an additional noise term. This could be studied in detail in mocks using a suite of functions with triaxial scale parameters. With future developments to ESR algorithm, and larger datasets, in the future it may be possible to constrain three-dimensional halo parameterizations.

As data quality improves in future surveys, reaching the $\approx 5\%$ level we find it necessary to recover NFW with less than 150 clusters, properly treating systematic effects will certainly be vital

to infer the correct halo profile in data, and a detailed study of potential model misspecification in mocks in this regime should be conducted in the future.

More broadly, the framework we have developed is applicable to any dataset that constrains the radial distribution of dark matter in galaxies or galaxy clusters. The core idea is straightforward: use symbolic regression to generate candidate expressions for the density profile $\rho(r)$, compute the corresponding observables and rank the models using the MDL principle. This methodology could be applied, for instance, to galactic rotation curves.

(b) Application to rotation curves

The gold standard for recovering the dark matter distribution around individual galaxies is rotation curves, the rotational velocity of an observable tracer (stars or gas) as a function of radius. Spiral galaxies are particularly good systems to study this as they are approximately axisymmetric and their stars and gas clouds follow nearly circular orbits hence directly tracing the gravitational potential. Due to the linearity of Poisson's equation, the total circular velocity can be expressed as the quadrature sum of the circular velocity contributed by each mass component

$$v_c^2(r) = v_{\text{baryonic}}^2(r) + v_{\text{DM}}^2(r), \quad (4.1)$$

where $v_{\text{DM}}(r)$ is the contribution of dark matter, and $v_{\text{baryonic}}(r)$ the baryonic component, given by

$$v_{\text{baryonic}}^2(r) = Y_{\text{disc}} v_{\text{disc}}^2(r) + Y_{\text{bulge}} v_{\text{bulge}}^2(r) + Y_{\text{gas}} v_{\text{gas}}^2(r), \quad (4.2)$$

which includes contributions from the disk, bulge and gas, each weighted by their respective mass-to-light ratios Y_x .

We can use the functions generated by ESR as candidate dark matter density profiles $\rho_{\text{DM}}(r)$. These can be integrated to give a mass and a velocity profile

$$M_{\text{DM}}(r) = 4\pi \int_0^r r'^2 \rho_{\text{DM}}(r') dr', \quad v_{\text{DM}}^2(r) = \frac{GM_{\text{DM}}(r)}{r}. \quad (4.3)$$

To fit to observational data, one needs to combine the dark matter model with the baryonic contribution to match the total circular velocity [equation \(4.1\)](#). One can infer the baryonic matter distribution (and thus its contribution to the circular velocity) from measurements of the stellar and gas velocities, typically based on photometry and HI observations [39]. The remaining, unaccounted for contribution to the observed v_c is due to the dark matter, allowing one to infer its density profile. These measurements in turn depend on galaxy-specific parameters such as mass-to-light ratios Y , the galaxy distance D and the disk inclination i . This introduces additional complexity, as fitting the combined model requires simultaneous optimization of both the function parameters describing the dark matter profile and galaxy-specific quantities associated with the baryonic component. As a result, the optimization becomes a significantly more challenging task, involving exploration of a higher-dimensional parameter space.

Still, rotation curves provide a complementary probe to lensing, especially at smaller galactic radii where lensing is less sensitive. It would be interesting to test how closely dark matter density profiles obtained from weak lensing match those obtained from rotation curves. It should also be straightforward to analyse both datasets combined. ESR naturally accounts for differences in the number of data points and measurement precision, one simply needs to combine description lengths carefully: model complexity should only be penalized once, while the likelihood contributions from each dataset are summed, in a similar way to what was done here for multiple galaxy clusters. Together, these two methods can provide a data-driven approach to constraining dark matter profiles. This might help reveal the underlying physical properties of dark matter and guide the development of alternative dark matter models as the quantity and quality of available data continue to grow.

Data accessibility. The data underlying this article will be shared on reasonable request to the corresponding author.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. A.M.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; T.Y.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—original draft, writing—review and editing; D.B.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, writing—original draft, writing—review and editing; H.D.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, writing—original draft, writing—review and editing; P.F.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. T.Y. acknowledges the support of a UKRI Frontiers Research Grant (EP/X026639/1), which was selected by the European Research Council. D.J.B. was supported by the Simons Collaboration on 'Learning the Universe' and acknowledges that support was provided by Schmidt Sciences, LLC. H.D. is supported by a Royal Society University Research Fellowship (grant no. 211046). P.G.F. acknowledges support from STFC and the Beecroft Trust. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

Acknowledgements. We thank the referees for their careful reading of the manuscript and for their valuable suggestions.

References

1. Frenk CS, White SDM, Davis M, Efstathiou G. 1988 The formation of dark halos in a universe dominated by cold dark matter. *Astrophys. J.* **327**, 507–525. (doi:10.1086/166213)
2. Blumenthal GR, Faber SM, Primack JR, Rees MJ. 1984 Formation of galaxies and large-scale structure with cold dark matter. *Nature* **311**, 517–525. (doi:10.1038/311517a0)
3. Navarro JF, Frenk CS, White SD. 1997 A universal density profile from hierarchical clustering. *Astrophys. J.* **490**, 493.
4. Banik U, Bhattacharjee A. 2025 Collisionless relaxation to quasi-steady state attractors in cold dark matter halos: origin of the universal NFW profile. *arXiv Preprint arXiv:2506.02104* (doi:10.1103/jds7-gm27)
5. de Blok WJG. 2010 The core-cusp problem. *Adv. Astron.* **2010**, 789293. (doi:10.1155/2010/789293)
6. Duffy AR, Schaye J, Kay ST, Vecchia CD, Battye RA, Booth CM. 2010 Impact of baryon physics on dark matter structures: a detailed simulation study of halo density profiles. *Mon. Not. R. Astron. Soc.* **405**, 2161–2178. (doi:10.1111/j.1365-2966.2010.16613.x)
7. Schaye J *et al.* 2015 The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Mon. Not. R. Astron. Soc.* **446**, 521–554. (doi:10.1093/mnras/stu2058)
8. Vogelsberger M *et al.* 2014 Properties of galaxies reproduced by a hydrodynamic simulation. *Nature* **509**, 177–182. (doi:10.1038/nature13316)
9. Laureijs R, Amiaux J, Arduini S. 2011 Euclid definition study report. *arXiv e-Prints arXiv:1110.3193*.
10. DESI Collaboration. 2016 The DESI experiment part I: science, targeting, and survey design. *arXiv e-Prints arXiv:1611.00036*. (doi:10.48550/arXiv.1611.00036)
11. LSST Science Collaboration. 2009 LSST science book, version 2.0. *arXiv e-Prints arXiv:0912.0201*. (doi:10.48550/arXiv.0912.0201)
12. Kronberger G, Burlacu B, Kommenda M, Winkler SM, Affenzeller M. 2024 *Symbolic regression*. Boca Raton, FL: Chapman & Hall / CRC Press.

13. Bartlett DJ, Desmond H, Ferreira PG. 2024 Exhaustive symbolic regression. *IEEE Trans. Evol. Computat.* **28**, 950–964. (doi:10.1109/TEVC.2023.3280250)
14. Desmond H, Bartlett DJ, Ferreira PG. 2023 On the functional form of the radial acceleration relation. *Mon. Not. R. Astron. Soc.* **521**, 1817–1831. (doi:10.1093/mnras/stad597)
15. Tenachi W, Ibata R, François TL, Diakogiannis FI. 2024 Class symbolic regression: gotta Fit 'Em all. *Astrophys. J. Lett.* **969**, L26. (doi:10.3847/2041-8213/ad5970)
16. Turing AM. 2009 *Computing machinery and intelligence*. Boca Raton, FL: Springer.
17. Goldberg DE. 1994 Genetic and evolutionary algorithms come of age. *Commun. ACM* **37**, 113–120. (doi:10.1145/175247.175259)
18. Cranmer M. 2020 Pysr: fast & parallelized symbolic regression in Python/Julia. (doi:10.5281/zenodo.4041459)
19. Kronberger G, de Franca FO, Desmond H, Bartlett DJ, Kammerer L. 2024 The inefficiency of genetic programming for symbolic regression—extended version. *arXiv Preprint arXiv:2404.17292*
20. Sousa T, Bartlett DJ, Desmond H, Ferreira PG. 2024 Optimal inflationary potentials. *Phys. Rev. D* **109**, 083524. (doi:10.1103/PhysRevD.109.083524)
21. Petersen BK, Landajuela M. 2021 Deep symbolic regression: recovering mathematical expressions from data via risk-seeking policy gradients. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*.
22. Broyden CG. 1970 The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA J. Appl. Math.* **6**, 76–90. (doi:10.1093/imamat/6.1.76)
23. Fletcher R. 1970 A new approach to variable metric algorithms. *Comput. J.* **13**, 317–322. (doi:10.1093/comjnl/13.3.317)
24. Goldfarb D. 1970 A family of variable-metric methods derived by variational means. *Math. Comput.* **24**, 23–26. (doi:10.1090/S0025-5718-1970-0258249-6)
25. Shanno DF. 1970 Conditioning of quasi-Newton methods for function minimization. *Math. Comput.* **24**, 647–656. (doi:10.1090/S0025-5718-1970-0274029-X)
26. Nelder JA, Mead R. 1965 A simplex method for function minimization. *Comput. J.* **7**, 308–313. (doi:10.1093/comjnl/7.4.308)
27. Rissanen J. 1978 Modeling by shortest data description. *Automatica* **14**, 465–471. (doi:10.1016/0005-1098(78)90005-5)
28. Grünwald PD. 2007 *The minimum description length principle*. Cambridge, MA: MIT Press.
29. Grünwald PD, Roos T. 2019 Minimum description length revisited. *Int. J. Math. Ind* **11**, 1930001. (doi:10.1142/S2661335219300018)
30. Cover TM. 1999 *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.
31. Bartlett D, Desmond H, Ferreira P. 2023 Priors for symbolic regression. In *Proc. Companion Conf. Genetic and Evolutionary Computation*, Lisbon, Portugal, pp. 2402–2411. New York, NY: ACM. (doi:10.1145/3583133.3596327)
32. Bartelmann M, Schneider P. 2001 Weak gravitational lensing. *Phys. Rep.* **340**, 291–472. (doi:10.1016/S0370-1573(00)00082-X)
33. Cromer D, Battaglia N, Miyatake H, Simet M. 2022 Towards 1% accurate galaxy cluster masses: including baryons in weak-lensing mass inference. *J. Cosmol. Astropart. Phys.* **2022**, 034. (doi:10.1088/1475-7516/2022/10/034)
34. Umetsu K *et al.* 2020 Weak-lensing analysis of X-Ray-selected XXL galaxy groups and clusters with Subaru HSC data. *Astrophys. J.* **890**, 148. (doi:10.3847/1538-4357/ab6bca)
35. Bhattacharya S, Habib S, Heitmann K, Vikhlinin A. 2013 Dark matter halo profiles of massive clusters: theory versus observations. *Astrophys. J.* **766**, 32. (doi:10.1088/0004-637X/766/1/32)
36. Umetsu K *et al.* 2014 CLASH: weak-lensing shear-and-magnification analysis of 20 galaxy clusters. *Astrophys. J.* **795**, 163. (doi:10.1088/0004-637X/795/2/163)
37. Umetsu K. 2020 Cluster–galaxy weak lensing. *Astron. Astrophys. Rev.* **28**, 7. (doi:10.1007/s00159-020-00129-w)
38. Chiu IN, Umetsu K, Sereno M, Ettori S, Meneghetti M, Merten J, Sayers J, Zitrin A. 2018 CLUMP-3D: three-dimensional shape and structure of 20 CLASH galaxy clusters from combined weak and strong lensing. *Astrophys. J.* **860**, 126. (doi:10.3847/1538-4357/aac4a0)
39. Lelli F, McGaugh SS, Schombert JM. 2016 SPARC: mass models for 175 disk galaxies with spitzer photometry and accurate rotation curves. *Astron. J.* **152**, 157. (doi:10.3847/0004-6256/152/6/157)