Journal of Sound and Vibration 506 (2021) 116162

Contents lists available at ScienceDirect

Journal of Sound and Vibration

journal homepage: www.elsevier.com/locate/jsv

Inferring empirical wall pressure spectral models with Gene Expression Programming

Joachim Dominique[a,*], Julien Christophe[a], Christophe Schram[a], Richard D. Sandberg[b]

[a] von Karman Institute, Chaussée de Waterloo 72, Rhode-Saint-Genése 1640, Belgium
[b] University of Melbourne, Parkville, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 31 July 2020
Revised 8 March 2021
Accepted 21 April 2021
Available online 24 April 2021

Keywords:
Gene Expression Programming
Machine learning
Turbulent boundary layer
Wall pressure spectral models

ABSTRACT

This paper presents a new data-driven approach for the establishment of empirical models describing turbulent boundary layer wall-pressure spectra. Unlike other models presented in literature, the new models are not derived by extending previously existing ones, but are directly built from a given dataset through symbolic regression using a machine learning algorithm known as Gene Expression Programming. Two modifications of the GEP algorithm presented in literature are proposed in this work to cope with some issues that are specific to the modelling of wall pressure spectra: a new power terminal and a local optimization loop. The validity of the new approach is first demonstrated using as input a dataset synthesized following the Chase-Howe and Goody models. The method is then applied to experimental data for a flat plate boundary layer. The results indicate that the wall pressure model obtained with the proposed approach remains consistent with previous formulations for zero pressure gradient, while showing a better match with the data and suggesting new ways to predict the influence of moderate pressure gradient

1. Introduction

The modelling of the pressure field induced by a turbulent boundary layer is relevant to a broad range of applications in fields including aeronautics, ground transportation or wind energy. The space and time characteristics of the pressure field beneath the boundary layer developing over an aircraft fuselage or train rooftop, and how they match with the structural response of the wall, dictate the importance of the fluid-structural coupling and noise transmission [1]. In the automotive sector, wind noise is considered to be the dominating sound source within the cabin for speeds above 100 km/h [2]. Also, the noise generated by the interaction of the turbulent boundary layer with an airfoil trailing edge is the minimum sound level emitted by a fan or an airfoil subjected to a clean, laminar inflow [3].

The pressure fluctuations beneath a turbulent boundary layer at low Mach number are given by the divergence of the Navier-Stokes equations, reducing to the Poisson equation in the case of incompressible flows:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} - \frac{\partial^2 p}{\partial x_i^2} = \frac{\partial^2\left(\rho v_i v_j - \langle \rho v_i v_j \rangle\right)}{\partial x_i \partial x_j}, \tag{1}$$

* Corresponding author.
E-mail address: joachim.dominique@vki.ac.be (J. Dominique).

https://doi.org/10.1016/j.jsv.2021.116162
0022-460X/

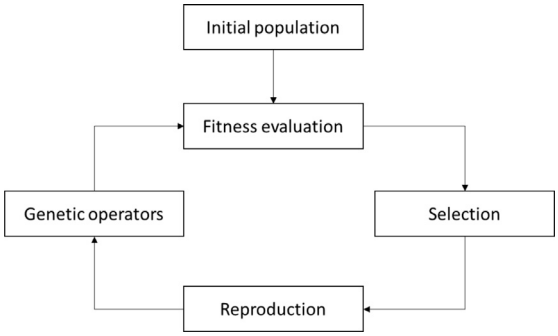


Fig. 1. Gene Expression Programming uses the evolutionary optimization flowchart.

where ρ is the fluid density, c is the speed of sound, v_i is the i-th velocity component and $\langle\cdot\rangle$ denotes an ensemble average. The above equation can be solved through integration [4]. However, such an integral model is complex for three reasons: *i*) the velocity fluctuations in all parts of the boundary layer participate to the wall pressure fluctuations, *ii*) the turbulent velocity fluctuations are advected at different speeds, and *iii*) different scaling approaches can be used to describe the velocity field in the boundary layer [5].

Those issues have motivated sustained research on turbulent wall pressure fluctuations. Following the work of Kraichnan [4], Panton & Linebarger [6] and Blake [7] derived integral models of the Poisson equation in the wavenumber-frequency domain. However, such solutions may present difficulties from the numerical point of view due to the presence of multi-dimensional integrals [8]. Following a different approach, Amiet [9], Chase & Howe [10] proposed empirical models of increasing complexity for engineering applications. The empirical model of Chase-Howe was improved by Goody [11] to account for Reynolds number effects and improve the high frequency behaviour. This model describes accurately boundary layers in absence of pressure gradient, as demonstrated through an extensive review by Hwang *et al.* [12]. The effect of a pressure gradient requires further modifications, such as proposed by Kamruzzaman [13], Rozenberg [14], Hu [15] or Lee [16].

An important issue remains nevertheless: even if each of those models describes fairly well the dataset on which it was based, they rarely provide a satisfactory agreement over the complete range of available data [16]. Many of the available models are based on the mathematical expression proposed by Goody, designed to yield the desired frequency dependence that is expected from first principle or observed in experiments (e.g. the ω^2 growth at low frequencies resulting from the Kraichnan-Phillips theorem [5], the ω^{-1} decay predicted by Bradshaw [17] for the inertial range, and the ω^{-5} decay at high frequencies predicted by Blake [7]). In this methodology, referred to as the *physics-based* approach in this paper, the only remaining degrees of freedom are the model constants, tuned to best fit the amplitudes and transition frequencies between the low, intermediate and high frequency regimes that are observed in a given experiment.

However, the growing number of available experimental datasets and the recent improvements of data analysis techniques have paved the way towards a new approach, where the underlying mathematical expression is no longer obtained from the Goody model, but directly built from data using symbolic regression. Such *data-based* approach will be investigated in this work using a symbolic regression algorithm named Gene Expression Programming (GEP). Alternative *data-based* regression tools such as collocation methods, Bayesian-based methods (e.g. Kriging) or machine learning (e.g. neural networks and random forests) can be mentioned for the sake of completeness, but will not be considered in this paper.

This paper presents an implementation of GEP for the symbolic regression of wall spectral pressure fluctuations models using the open source Python library Geppy [18]. We start by describing the GEP algorithm and its extension for an effective determination of the decimal power. We use this algorithm on synthetic noisy data to retrieve the previously proposed models of Chase-Howe and Goody. Finally, we will use the results to discuss the applicability of GEP as a data-based method to develop new wall pressure spectral models.

2. Traditional GEP

Gene Expression Programming is an evolutionary algorithm that yields mathematical models from data [19]. Over the past decade, GEP has been used with success in different fields such as time series regression, data mining or knowledge discovery, but it will be mostly employed in this work as a symbolic regression tool [20]. The reader is referred to Refs. [21–24] for previous applications of the method in the field of fluid dynamics, for the generation of turbulence models from high fidelity numerical data.

As with other evolutionary algorithms, GEP optimizes solutions by iteratively altering, in a non-deterministic way, a population of individuals following a process that mimics natural selection and survival of the fittest (Fig. 1). The best individuals among an initial population of candidate solutions are selected based on their fitness, while the least fitting ones are rejected. Those least-fit individuals are later replaced by new candidates that are build from random genetic variations.

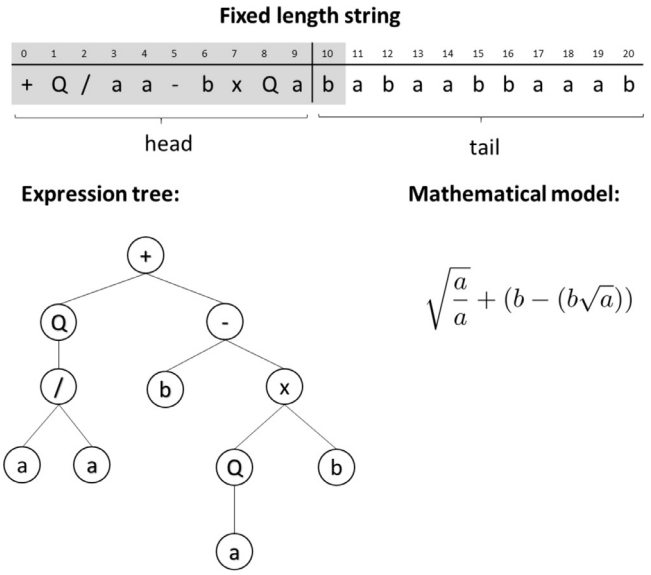


Fig. 2. Example of an individual representation.

The different steps of the traditional GEP optimization will be detailed below, starting with the construction of the individuals and the mean used to evaluate their fitness. Then, the procedure used to select and reproduce the best fitting candidates will be discussed along with the operators used to introduce genetic variations into the remaining pool of candidates. We will conclude this section with specific issues such as the determination of numerical constants.

2.1. Individuals

2.1.1. Single-gene individuals

In GEP, the individuals are encoded as linear strings of fixed length. Each element of the string is either a function or a terminal selected among predefined sets. As an illustration, the string in Fig. 2 has been constructed from the function set $F : [+, -, /, \times, Q]$ and the terminal set $T : [a, b]$, where Q is the square root operator and $[a, b]$ are numerical variables.

The linear string translates into an expression tree that corresponds to a specific valid equation. In this case, the linear string has a fixed total length of 21 elements, but the corresponding length of the expression tree may vary. Indeed, Fig. 2 shows an expression tree that only uses the first 11 elements of the linear string shown in grey. This is refereed by Ferreira as the concept of Open Reading Frame [19]. Such a formulation allows to create solutions of different lengths while conserving a fixed length string representation. For example, when the element 10 of the individual is changed from b to + in Fig. 3, the length of the expression tree increases by two.

To ensure that every expression tree results in a correct mathematical model, every branch of the expression tree must finish by a terminal. This is achieved by dividing the linear string into two parts called *the head* and *the tail*. The first h elements of the string are the head and consist of mathematical operators or terminal functions. The tail of length $t = h(n_{\max} - 1) + 1$, where $n_{\max}$ is the maximum arity of the functional set, is only composed of terminals. Therefore, regardless of the number of operators used within the head region, there always will be enough terminals inside the tail to provide a valid mathematical expression. In the previous examples, the head length is $h = 10$ and the tail length is $t = h(2 - 1) + 1 = 11$.

2.1.2. Multigenic individuals

Individuals are usually composed of multiple genes. When you combine multiple genes of similar length, you compose a solution of multiple sub-expression trees which are able to evolve separately from each other as independent entities. This facilitates the finding of relevant shorter solutions. To connect the different genes together and form a valid mathematical model, the individual must be provided with a linking function. Three different types of linking functions are discussed here:

- Prescribed linking functions : all sub expression trees are linked together by a prescribed mathematical function that remains unchanged during the evolution process as illustrated in Fig. 4. It could be argued that prescribed linking functions can artificially drive the algorithm towards specific solutions. This pitfall can be avoided by using only relatively simple rational prescribed linking functions. It can be recommended to start with simple single gene solutions, and progressively increase the complexity by adding additional genes [19] if needed.

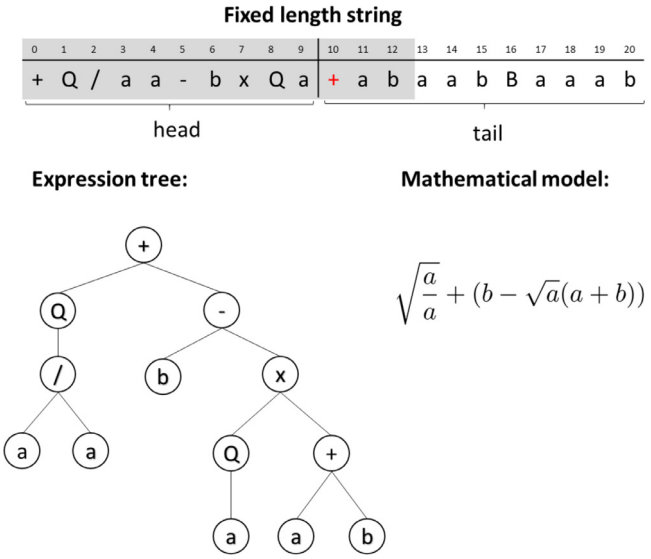


Fig. 3. Modification of element 11 from the previous individual example.

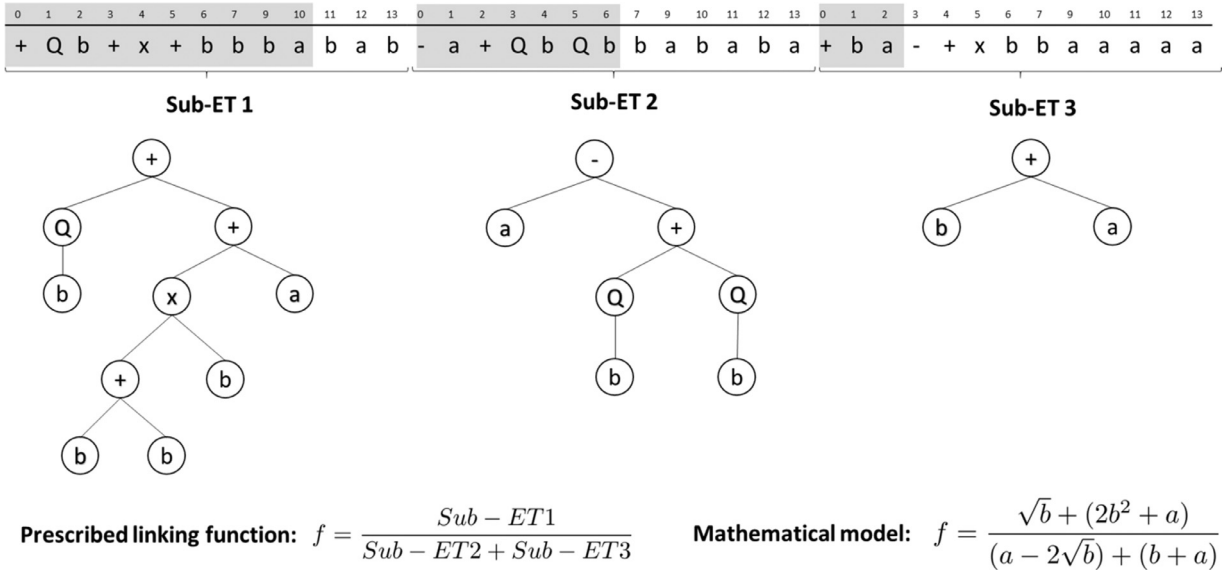


Fig. 4. Construction of multi-genic individual using prescribed linking functions.

- Random linking functions : the linking functions are chosen randomly among a set of operators. This allows the shape of the individuals to change during the evolution as the same gene can interact in a variety of ways and produce different solutions. However, since the connection between the genes is always a linear combination of the operators, a non-linear gene structure can never be created. Indeed, $f = G_1/G_2 + G_3$ is not equivalent to $f = \frac{G_1}{G_2+G_3}$. This also restricts the search space of solutions that individuals can represent.
- Automatically Defined Functions (ADF): another additional gene is added at the end of the multi-genic structure that describes how the sub expression trees are linked together [25]. This solution allows both the global shape to evolve during the evolution and the creation of non-linear entities. It is also the most difficult method to implement.

2.2. Fitness

The fitness of an individual is defined as its capability to match a set of points, for example the training dataset. It is also considered as an indicator of its quality and is the quantity to minimize through the optimization process. Assuming a set of n training points (x_n, y_n) with objective value f_n and an individual function f_{ind}, the fitness is here defined as the

mean square error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}(f_{ind}(x_k, y_k) - f_k)^2 \tag{2}$$

A distinctive feature of GEP, compared with other evolutionary algorithms, is that the selection/mutation processes and fitness computation are performed on different entities. Here, the selection and mutation processes are performed on the linear string formulation which is linear, compact and easy to manipulate genetically, while the fitness is computed from the expression tree, which is a complex non-linear expression. The compactness of the representation of the linear string entities, combined with the complexity of the expression tree formulation, is the strength of the GEP algorithm.

2.3. Selection and reproduction

At each generation, the best members of the population are selected based on their fitness (Eq. 2) to breed a new population of individuals through evolutionary processes. Such selection is accomplished using what is called tournament selection. This is a competition between individuals where n randomly selected individuals compete with each other k times and the less fitting ones of each tournament are eliminated. This selection is used to remove badly adapted solutions from the pool of candidates. The k winners of the tournament are passed to the population of the next generation. To fill the empty spots in this new population, the winners will reproduce sometimes with modifications due to mutations.

Another possible method is called roulette wheel selection. In this method each individual has a probability proportional to their fitness to be selected for the next generation. This enables good candidates to have a higher chance to be selected and conserved to the next generation. The wheel is turned as many time as necessary to fill the next generation. Although this method seems more elaborate, it has been shown to converge less rapidly than the tournament selection [26]. Therefore, the tournament selection was preferred in this work.

2.4. Mutation operators

Using selection only, the pool of candidates remains unchanged and the algorithm cannot create new and better fitting solutions. To trigger those changes, GEP uses a process called mutation where each individual of the selected population has a chance to modify part of its solution using mutation operators. Those mutation operators describe how evolution is introduced in the algorithm. They are usually separated into three groups: Mutation, transposition and recombination.

2.4.1. Mutation

Mutation is the most basic operator where a random element of different genes is changed. This allows to create diversity into the selected pool of candidates. It is described by two probabilities: The first (p_m^1) gives the probability of a given individual to be selected for mutation. The second (p_m^2) describes the probability of each gene to be randomly changed into another function or terminal. For instance, a large p_m^1 and a small p_m^2 implies that many individuals will be slightly modified. It should also be noted that this operator must respect the structure of the individual. In the head a gene can be transformed into a function or a terminal but in the tail only a terminal can be selected.

2.4.2. Transposition

A part of a gene is moved to another location within the same gene. The number of elements moved is variable, it can be a single element or multiple functions and terminals. When the sequence is moved at the root of the individual this process is called Root Insertion Sequence (RIS) transposition and when it is moved at another position it is called Insertion Sequence (IS) transposition. They have respectively transposition rate of p_T^1 and p_T^2.

2.4.3. Recombination

Recombination, the last type of mutation operator, is when a single element or a sequence of elements are exchanged between different genes. There are three types of recombination:

- one-point recombination (p_r^1) : a position is randomly selected in two genes and all the information past this point is exchanged.
- two-points recombination (p_r^2) : the information exchange between genes is everything in between two randomly selected positions.
- Gene recombination (p_r^3) : this only occurs in multigenic individuals, all the information of a gene is exchanged between two individuals.

All those mutation operators are not necessarily applied to every individual simultaneously due to their respective rates. As a result, some individuals might be affected by several mutation operators and others by none. In fact, their probabilistic rate affects strongly the convergence of the GEP algorithm. The best convergence of the algorithm is obtained by balancing selection and mutation. When there is not enough mutation, the algorithm does not generate many new individuals and tends to converge toward local minima. On the other hand, when there are too many mutations, the selection and evolutionary process cannot take place and the GEP reduces to a random search algorithm.

2.5. Numerical constants

The determination of numerical constants has always been a bottleneck for non-gradient based methods such as genetic algorithms [27]. GEP suffers from the same limitations. In the past, numerical constants were introduced within the GEP using mostly two methods: Random Numerical Constants (RNC) or Ephemeral Numerical Constants (ENC).

RNC works as a particular type of terminal because it also introduces a new array of genes that comes after the tails named the DC domain [28]. As illustrated in Fig. 5 this domain is constituted of numerical constants that can evolve (using the same selection-mutation process as other genes) during the optimization. Every time the RNC terminal ? is encountered, the numerical constants from the DC array are inserted from left to right into the expression tree. To ensure that there are always enough RNC variables available, the length of the DC array should always be as long as the tail of the individual.

ENC works in a similar fashion, except that the pool of numerical constants is not selected from an additional array but randomly selected within a user-defined range. This implies that the search region of the algorithm is reduced, but also that the numerical constants do not have the chance to evolve as the optimization carries on.

2.6. Modifications of the existing GEP algorithm

As it will be shown in Section 3, most wall pressure spectral models available in literature include multiple variables with high fractional exponent values. To approach such exponent using the traditional function set $[+, -, /, \times, Q]$, the GEP must create long tree structures as illustrated in Fig. 6. This makes the convergences of the algorithm weak as a very long linear string must be considered. As a result, the algorithm provides very long and complicated mathematical expressions which are often not similar to the existing wall pressure models.

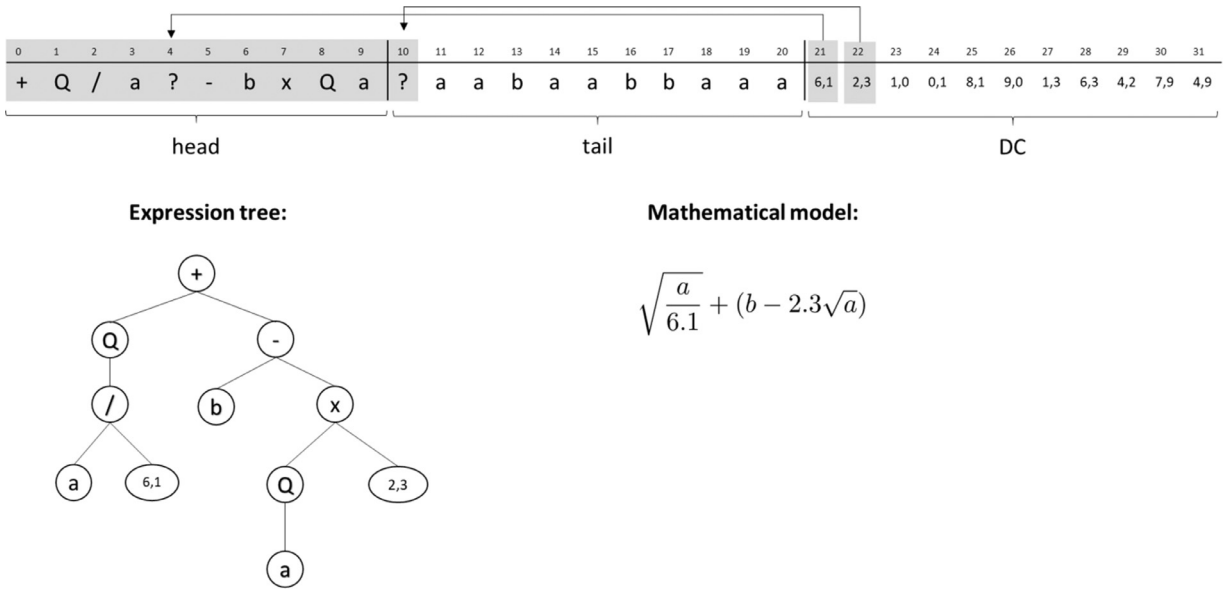


Fig. 5. Construction of an individual including RNC terminals and an DC array.

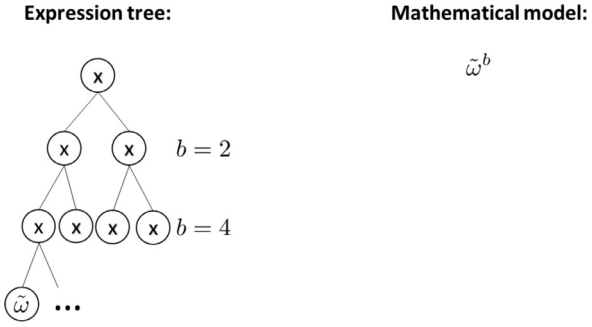


Fig. 6. Illustration of the problem due to the approximation of high exponent variables.

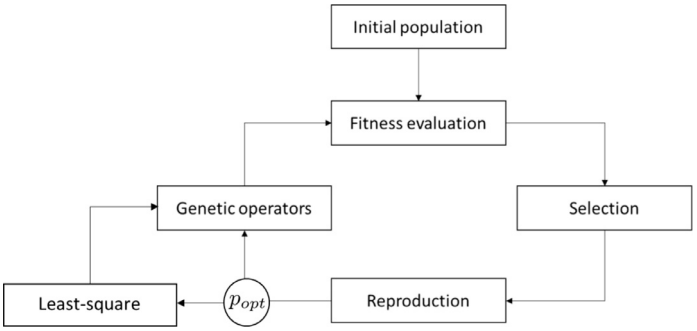


Fig. 7. Local optimization in the GEP flowchart.

A new type of terminal, named the *pow* terminal, was introduced in this work as a solution. The said *pow* terminal works similarly to an RNC terminal, except that the numerical constants selected act as a numerical exponent to a variable. As an example, the pow terminal of the variable named a can be written as $a^?$ where ? is a numerical constant selected among the DC terminals. This formulation permits an easy creation of variables with large fractional exponent and relatively short linear string.

As discussed in the previous section, evolutionary algorithms are non-gradient based optimization techniques. Therefore, they have trouble to correctly approach numerical constants. This is normally not an issue, but the introduction of the *pow* terminal that uses numerical constants from the DC array as exponent value for some variables has put additional strain on the ability of the GEP to accurately approach numerical constants. For this reason, an additional local optimization loop was added to the traditional evolutionary algorithm as shown in Fig. 7. This local optimizer has the effect that every time that the condition is satisfied the numerical constants in the DC array of some selected individuals are changed to their best fitting values using a least-square regression. The least square regression is far slower than the reproduction/mutation optimization performed within the evolutionary process. However, by controlling its occurrence rate through the probability p_{opt}, the optimization process can be fed with well-fitting numerical constants while not penalizing too much the convergence time.

3. Reproducing existing wall pressure spectral model

As a first consistency check, the proposed new *data-based* approach has been applied to datasets synthesized using *physics-based* models: the Chase-Howe [10] model and the Goody model [11]. The GEP solutions will be assessed regarding 1) their capability to match their training dataset (the fitness of the model); and 2) the similarity between the proposed GEP mathematical expression and the mathematical model that served to generate the datasets in the first place. In particular, specific attention will be given to the ability of the GEP approach to retrieve meaningful trends of the models such as the Reynolds number dependency or the low and high frequency trends.

3.1. Validation on Chase-Howe model

Using the theoretical development of Chase [29], Howe proposed a semi-empirical model that describes the behavior of the wall-pressure field at low Mach number based on mixed variables [10]:

$$\frac{\Phi_{pp}U_e}{\tau_w^2\delta^*} = \frac{2\tilde{\omega}^2}{(\tilde{\omega}^2 + 0.0144)^{3/2}} + N, \tag{3}$$

where Φ_{pp} is the surface pressure spectrum, U_e is the external boundary layer velocity, τ_w is the wall shear stress, δ^* is the displacement thickness and $\tilde{\omega} = \omega\delta^*/U_e$. This model scales as $\tilde{\omega}^2$ at low frequency and as $\tilde{\omega}^{-1}$ at high frequency, but does not provide the expected ω^{-5}[7] slope at higher frequencies and does not account for Reynolds number effects. Despite those shortcomings, its simple mathematical expression makes it a good first test for the GEP algorithm.

The training dataset used within the GEP algorithm will be constituted of 1000 points logarithmically spaced obtained from Eq. 3 with additional Gaussian noise N that is used to represent the uncertainty that can be found in experimental data. In clear, the training data was generated as $\Phi_{pp}U_e/(\tau_w^2\delta^*) = f(\tilde{\omega})(1 + N_1[0, 0.05]) + |N_2[0, 0.01]|$. The first noise source has a standard deviation set to 5% of the amplitude of $\Phi_{pp}U_e/(\tau_w^2\delta^*)$ corresponding to the stochastic uncertainties of the measurements while the second has a constant standard deviation set to 0.01 to model systematic uncertainties caused by the sensitivity of the microphones and the background noise of the wind tunnel.

The function set is F : $[+, \times, /, Q]$ and the terminal set is T : $[\tilde{\omega}, 1.0]$. A RNC terminal and a DC array composed of strictly positive constants is adopted to determine the numerical constant of Eq. 3. When defining the above sets, it must be ensured that any combination of operators and terminals cannot result in an inconsistent solution, e.g. due to a division by zero or root of a negative number. Since $\tilde{\omega}$ is strictly positive and non-null, such mathematical operations will never happen in this

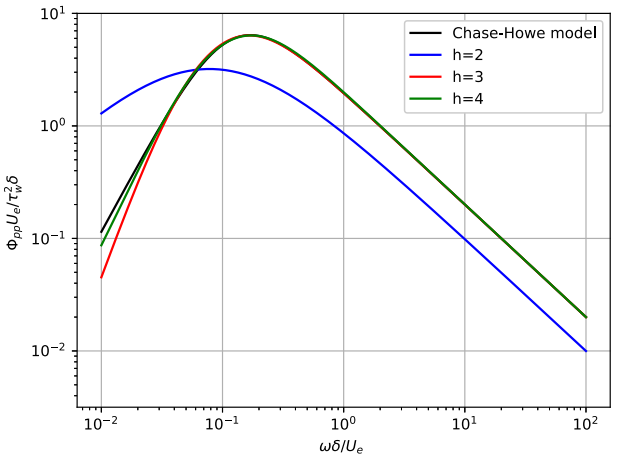


Fig. 8. Representation of the minimum length of the linear string.

Table 1
Representation of 4 different solutions proposed by the GEP to approach the Chase-Howe model.

N	Expression	Simplified	Fitness	$\tilde{\omega}\to 0$	$\tilde{\omega}\to\infty$
Solution 1	$\frac{2\tilde{\omega}}{\left(\tilde{\omega}+\frac{0.021}{\tilde{\omega}}\right)\left(\tilde{\omega}+\frac{0.002}{\tilde{\omega}}\right)}$	$\frac{2\tilde{\omega}^3}{(\tilde{\omega}^2+0.021)(\tilde{\omega}^2+0.002)}$	0.024	$\tilde{\omega}^3$	$\tilde{\omega}^{-1}$
Solution 2	$\frac{(1.06+\tilde{\omega})}{\left(\sqrt{\tilde{\omega}}+\frac{0.05}{\tilde{\omega}}\right)\left(\tilde{\omega}+\frac{0.017}{\tilde{\omega}}\right)}$	$\frac{(1.06+\tilde{\omega})\tilde{\omega}^2}{(\tilde{\omega}^{3/2}+0.05)(\tilde{\omega}^2+0.017)}$	0.026	$\tilde{\omega}^2$	$\tilde{\omega}^{-1/2}$
Solution 3	$\frac{\tilde{\omega}}{(0.038+\tilde{\omega}^2)(0.037)^{0.25}}$	$\frac{\tilde{\omega}}{(0.038+\tilde{\omega}^2)(0.037)^{0.25}}$	0.157	$\tilde{\omega}$	$\tilde{\omega}^{-1}$
Solution 4	$\frac{2.076}{\left(\tilde{\omega}+\frac{0.017}{\tilde{\omega}}\right)\left(1+\frac{0.065}{\tilde{\omega}}\right)}$	$\frac{2.076\tilde{\omega}^2}{(\tilde{\omega}^2+0.017)(\tilde{\omega}+0.065)}$	0.04	$\tilde{\omega}^2$	$\tilde{\omega}^{-1}$

simulation. Otherwise, some conditions should be implemented within the optimization loop to avoid those operations. For this simple model, we are using the traditional GEP, not including the improvements introduced in Section 2.6.

The number of genes and the head length that describes the shape of the solution need to be defined. The number of genes usually depends on the type of linking functions. When a theory is available to infer the functional shape of the quantity of interest, a prescribed linking function of similar shape can be adopted to facilitate the convergence. Three genes connected using the linking function $f = G_1/(G_2\,G_3)$ are used for this first application case.

The head length sets the maximum length of the individuals. If the head is too short, the solutions proposed by the algorithm will not be complex enough to fit the dataset. If the head is too long, the mathematical solution proposed will tend to overfit the dataset by proposing unnecessarily complex, and often physically meaningless, formulations. The head length h should thus be as small as possible while capturing the complexity of the training dataset. Fig. 8 shows solutions for three different head lengths. It can be seen that for $h \leq 2$ the algorithm is not able to fit the Chase-Howe model. The smallest head length for which it is possible to capture the complexity of the problem is thus $h = 3$.

Four different solutions obtained by the GEP algorithm with these parameters are given in Table 1 and represented in Fig. 9. It should first be noted that the fit quality is calculated on a linear scale, while wall pressure spectra are normally represented in log-log scale. This explains how different solutions having similar fitness (0.024 and 0.026 in this instance) can exhibit fairly different behaviours at low and high frequencies in a log-log graph, where the data are one or two decades below their maximum values and do therefore not contribute significantly to the fitness value. In this instance, the solution 4 provides the most meaningful solution, despite not showing the best fitness value, with asymptotic slopes tending to 2 at low frequencies and -1 at high frequencies, as expected from the original model. Otherwise, the solutions obtained by the GEP algorithm share a functional shape that is similar to the model of Chase and Howe. The main difference stands in the numerical constants, including the exponent values, which are more difficult to converge to with a GEP algorithm.

3.2. Validation on Goody model

Goody [11] proposed an extension of the Chase-Howe model (3) to include a third regime with a different frequency dependence ω^{-5} at high frequencies, with a ω^{-1} decay in an intermediate frequency range whose extent grows for increasing

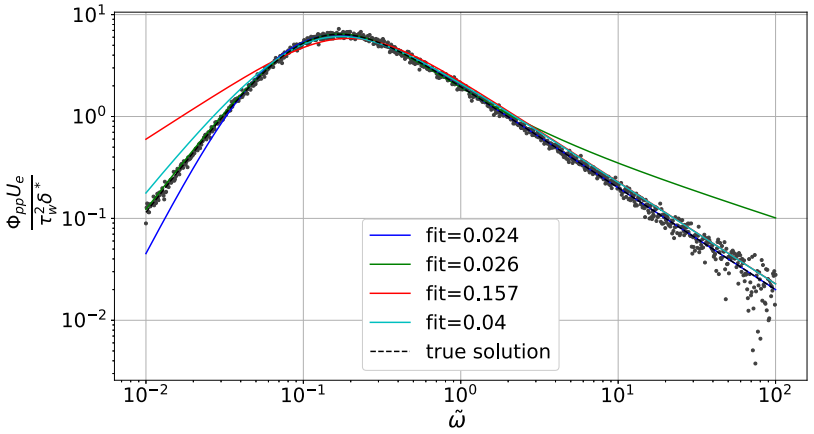


Fig. 9. Comparison between the GEP solutions and the Chase-Howe model. The grey dots represent the training set and the black dotted line is the true Chase-Howe model. The four different GEP models expressed in Table 1 are represented in coloured lines.

Table 2
Representation of 4 different solutions proposed by the GEP to approach the Goody model.

N	Simplified	Fitness	Reynold dependency	$\tilde{\omega} \to 0$	$\tilde{\omega} \to \infty$
Solution 1	$\frac{\tilde{\omega}^{0.9}}{\tilde{\omega}^{5.5}R_T^{-3.8} + \sqrt{1+\tilde{\omega}^{2.7}}}$	0.001	$R_T^{3.8}$	$\tilde{\omega}^{0.9}$	$\tilde{\omega}^{-4.6}$
Solution 2	$\frac{\tilde{\omega}^{3/2}}{0.4 + \tilde{\omega}^2 + \tilde{\omega}^{6.3}R_T^{-4.1}}$	0.0016	$R_T^{4.1}$	$\tilde{\omega}^{1.5}$	$\tilde{\omega}^{-4.8}$
Solution 3	$\frac{\tilde{\omega}}{0.6 + \tilde{\omega}^{1.4} + \tilde{\omega}^{6.3}R_T^{-3.9}}$	0.0016	$R_T^{3.9}$	$\tilde{\omega}$	$\tilde{\omega}^{-5.3}$
Solution 4	$\frac{2\tilde{\omega}}{1.7 + \tilde{\omega}^{1.8} + \tilde{\omega}^{5.8}R_T^{-3.7}}$	0.0006	$R_T^{3.7}$	$\tilde{\omega}$	$\tilde{\omega}^{-4.8}$

Reynolds numbers:

$$\frac{\Phi_{pp}U_e}{\tau_w^2\delta} = \frac{C_2(\omega\delta/U_e)^2}{\left((\omega\delta/U_e)^{0.75} + C_1\right)^{3.7} + \left(C_3(\omega\delta/U_e)\right)^7} + N, \tag{4}$$

where $C_1 = 0.5,\ C_2 = 3.0,$ and $C_3 = 1.1R_T^{-0.57}$ describes the effect of Reynolds number, with $R_T = (u_\tau^2\delta)/(U_e\nu)$. The original scaling of the Goody model uses the boundary layer thickness δ instead of the displacement thickness δ^* in Eq. 3. The frequency scaling $\tilde{\omega} = \frac{\omega\delta}{U_e}$ remains unchanged. This model depends on two variables: $\tilde{\omega}$ and R_T and has a longer mathematical expression with several numerical constants, which is likely to make the convergence of GEP more difficult. However, the Goody expression shows a good agreement with experimental data obtained in different laboratories and is used by many authors as the baseline for more elaborate models, e.g. accounting for pressure gradient effects [13–16] and constitutes therefore an interesting application case for the proposed approach.

The dataset used for this model comprises again 1000 points, for 5 different Reynolds numbers $R_T = [1, 3.16, 10, 31.62, 100]$, which amounts to a training set of 5000 points. The noise N is modelled as in the Chase-Howe model with two Gaussian noise sources. The functional and terminal set are defined as $\mathrm{F} : [+, \times, /, \mathrm{Q}]$ and $\mathrm{T} : [\tilde{\omega}, 1.0, \tilde{\omega}^?, R_T^?]$. The terminal set is composed of the modified power terminal introduced in Section 2.6. The local optimization loop is not used in a first step. Three genes connected using the linking function $f = G_1/(G_2 + G_3)$ were used. The head length was set to $h = 4$.

Four different solutions proposed by the GEP approach are shown in Table 2 and represented in Fig. 10. As for the Chase-Howe model, the expressions returned by the algorithm show clear similarities with the Goody model that was used to synthesize the dataset. The power terminal allowed to find trends for the different variables very close to the ones expected from the model, with a Reynolds dependency as R_T^4 and a decay with $\tilde{\omega}^{-5}$ at high frequencies. Only the low-frequency behaviour with $\tilde{\omega}^2$ is not observed for every solution in Table 2. This is likely due to the Gaussian noise that impact the trend at low frequency. Without the power terminal, the GEP approach is unable to propose models that fits the high frequency decay with $\tilde{\omega}^{-5}$. As in the Chase-Howe model, below a given threshold, the fitness gives little indication on

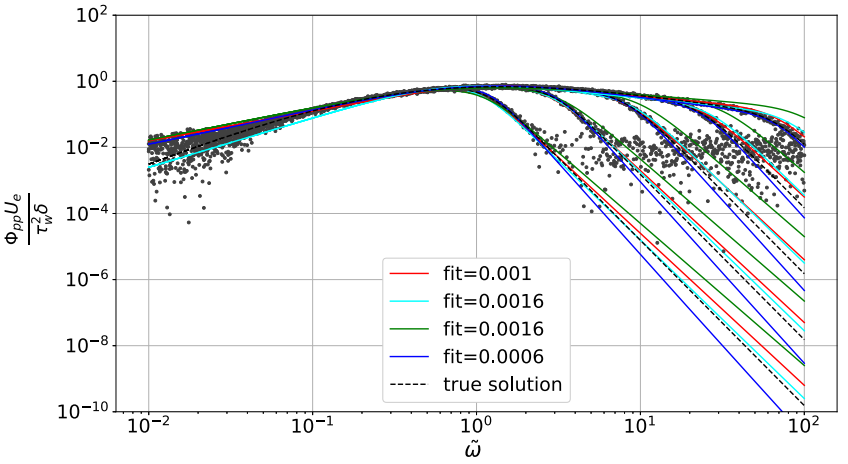


Fig. 10. Comparison between the GEP solutions and the Goody model. The grey dots represent the training set and the black dotted line is the true Goody model. The four different GEP models expressed in Table 2 are represented in coloured lines.

Table 3
Representation of 4 GEP solutions for the Goody model using the local optimization loop for the determination of constants.

N	Simplified	Fitness	Reynolds dependency	$\tilde{\omega} \to 0$	$\tilde{\omega} \to \infty$
Solution 1	$\frac{\tilde{\omega}^{1.72}/(\tilde{\omega}+0.18)}{3.93+\frac{4.8\tilde{\omega}^{5.79}}{R_T^{4.02}}+5.78\tilde{\omega}+\tilde{\omega}^{1.72}}$	3.6×10^{-4}	$R_T^{4.02}$	$\tilde{\omega}^{1.72}$	$\tilde{\omega}^{-5.07}$
Solution 2	$\frac{1.65\tilde{\omega}^{0.63}}{0.24(\tilde{\omega}^{0.76}+\tilde{\omega}^{0.82})+\frac{\tilde{\omega}^{5.68}}{R_T^{4.01}}+\tilde{\omega}^{1.24}+1.37}$	3.6×10^{-4}	$R_T^{4.01}$	$\tilde{\omega}^{0.63}$	$\tilde{\omega}^{-5.05}$
Solution 3	$\frac{1.46\tilde{\omega}^{1.64}}{0.15+\tilde{\omega}+\frac{\tilde{\omega}^{6.51}}{R_T^{3.9}}+\tilde{\omega}^{2.21}}$	3.6×10^{-4}	$R_T^{3.9}$	$\tilde{\omega}^{1.64}$	$\tilde{\omega}^{-4.87}$
Solution 4	$\frac{1.46\tilde{\omega}^{1.34}}{0.25\tilde{\omega}^{2.34}+1.46\tilde{\omega}^{1.34}+0.47+\frac{\tilde{\omega}^{6.26}}{R_T^{3.97}}}$	3.6×10^{-4}	$R_T^{3.97}$	$\tilde{\omega}^{1.34}$	$\tilde{\omega}^{-4.92}$

the quality of the results. Again, the model that is the closest to the Goody model (solution 2) is not the one with the best fitness.

The benefit of the least square local optimization loop can now be assessed, using the same parameters as before. The local optimization is enabled for each individual, i.e. $p_{opt} = 1$: all individuals have their numerical constant in the DC array transformed through a least-square regression.

The solutions for this optimization are shown in Table 3 and represented in Fig. 11. It appears that the local optimization not only allows improving the prediction of the numerical constant and the fitness of the individuals, it also improves the accuracy of the model trends at low and high frequencies. This comes nevertheless at the expense of a considerably higher numerical cost. Using a Intel Core i7-4800MQ processor, without local optimization, it takes only 30 minutes for the GEP to generate a model, while using the local least-square regression on every individual can last up to 24 hours. This is not an optimal strategy as most of the optimized individuals are later removed from the population in the following selection tournament. A better strategy to be verified in a future study would be to set p_{opt} such that only a couple of individuals are optimized at each generation.

It can be seen that the fitness values are identical for all solutions although there are four different equations. Once again, it can be observed that below a certain threshold, the fitness defined as a MSE is not a good criterion to characterize the quality of candidate solutions.

This is related to the fact that the training dataset may gather wall pressure spectra obtained from different flow conditions (e.g. freestream velocity or streamwise pressure gradient), therefore with different orders of magnitudes. And even within a set of data obtained for similar conditions, the large dynamic range of pressure fluctuations across the frequencies of interest makes it difficult to compare the fitness MSEs at low, medium or high frequencies. Nevertheless, the asymptotic

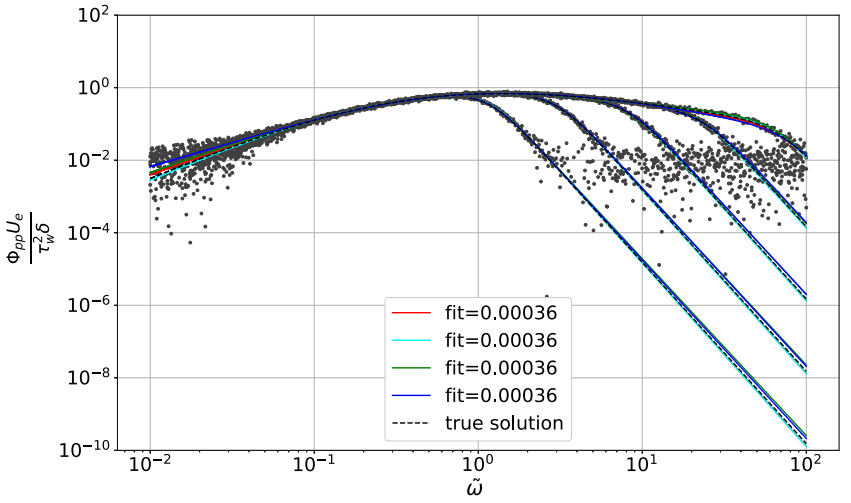


Fig. 11. Comparison between the GEP solutions and the Goody model using the local optimization loop. The grey dots represent the training set and the black dotted line is the true Chase-Howe model. The four different GEP models expressed in Table 3 are represented in coloured lines.

decays at low and high frequencies bear as much importance as the amplitude of the mid-frequency hump to the modeller. This optimisation problem investigated should therefore be seen as a multi-objectives problem where the first objective is to predict the amplitude of the spectrum using MSE and the second is to predict the trends at low and high frequencies using log means square error (logMSE).

$$\text{logMSE} = \frac{1}{n}\sum_{k=1}^{n}(10\log_{10}(f_{ind}(x_k, y_k)) - 10\log_{10}(f_k))^2 \tag{5}$$

A new individual fitness is proposed for this multi-objectives optimisation as a weighted sum of two contributions:

$$\text{Fit} = \sqrt{\alpha\frac{logMSE}{A_{log}^2} + (1-\alpha)\frac{MSE}{A_{lin}^2}} \tag{6}$$

where A_{lin} and $A_{\log}$ are the maximum amplitude of the training points in linear and logarithmic scale. A value of $\alpha = 0.5$ was used in this work, giving the same importance to both contributions. For the remaining of this work, individuals will be evaluated using the fitness in Eq. (6). The code used to obtain those results is available at https://github.com/DominiqueVKI/VKI_researchWPS.

4. A model for turbulent boundary layer wall pressure spectra

The modified GEP algorithm is finally tested on an experimental dataset constituted of 10 separate measurements of boundary layers velocity profiles under adverse, favorable and neutral pressure gradients. The measurements were performed at Ecole Centrale de Lyon (ECL) by Salze and Bailly [30]. They conducted an experimental campaign in the main subsonic wind tunnel at the Centre Acoustique of ECL using a rectangular section with sloped ceiling to control the mean pressure gradient inside the channel. The velocity profiles in the boundary layer were measured using hot-wire anemometry operating in constant voltage mode and the wall pressure spectra were recorded using a rotating antenna composed of 63 flush mounted pinhole microphones. More details about this experimental campaign can be found in Ref. [30].

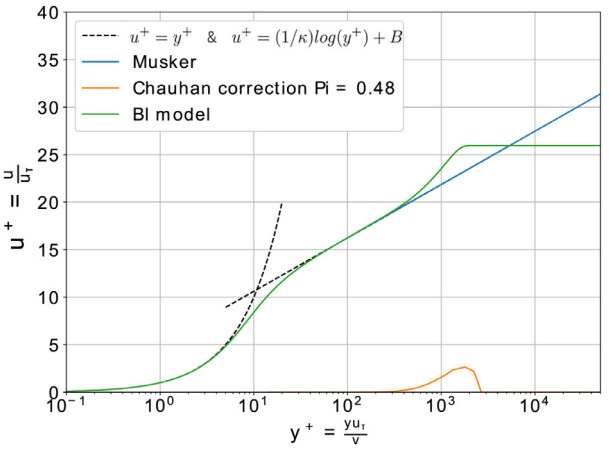
4.1. Boundary layer parameters

The boundary layer parameters can be computed directly from the hot-wire measurements using the definition for the displacement and momentum thickness for incompressible boundary layers:

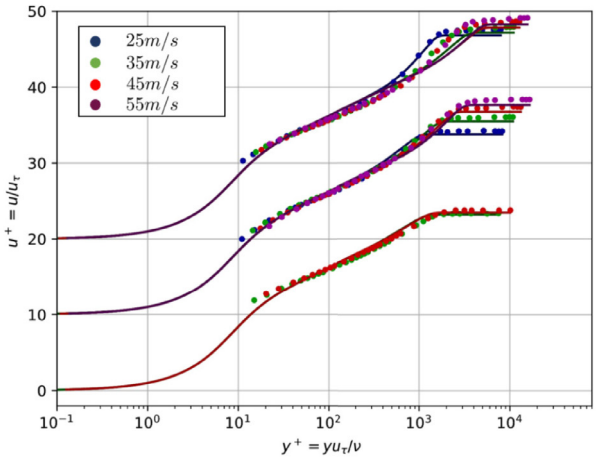
$$\delta^* = \int_0^{\delta}\left(1 - \frac{u(y)}{U_e}\right)dy, \tag{7}$$

$$\theta = \int_0^{\delta}\frac{u(y)}{U_e}\left(1 - \frac{u(y)}{U_e}\right)dy, \tag{8}$$

where δ is the boundary layer thickness that corresponds to 0.99 of the maximum velocity of the velocity profile U_e.



(a)



(b)

Fig. 12. a: Boundary layer velocity profile model. b: Experimental measurements scaled with boundary layer parameters. The dashed lines correspond to the boundary-layer model and the dotted points correspond to the experimental points. For clarity, the measurements have been shifted by $u^+ = 10$ for ZPG and by $u^+ = 20$ for APG.

It was shown in the Goody model that two other parameters are essential to model the wall pressure spectra: the wall shear stress τ_w and the Reynolds number R_T. The wall shear stress is computed from the friction velocity $u_\tau = \tau_w/\rho$ using a procedure that minimizes the difference between the experimental measurements and a boundary layer model in Eq. 9

$$u^+(y^+) = \begin{cases} u^+_{inner}(y^+) + \frac{2\Pi}{\kappa} W\left(\frac{y}{\delta}\right), & \text{if } y \leq \delta \\ U_e/u_\tau & \text{otherwise} \end{cases} \tag{9}$$

The boundary layer model is built from the theoretical law of the wall and law of the wake, shown in Fig. 12 (top). It combines u^+_{inner} with the Musker's law of the wall [31] and the Chauhan correction W [32] for the law of the wake.

The amplitude of the Chauhan correction given by the parameter Π is chosen to fit $u(\delta) = 0.99U_e$. This approach is used to scale the experimental boundary layer profile in Fig. 12 (bottom) and compute the wall shear stress in Table 4. It was observed that the values obtained for the friction velocity using the above method are identical to the ones computed experimentally in the initial paper of Salze [30] using hot film on the channel floor.

4.2. Zero pressure gradient wall pressure spectra

The microphone antenna measurements of the wall pressure spectra are used as training points for our GEP modeling approach. In total, the dataset contains four spectra of 500 frequency points each at different Reynolds numbers $R_T = [7.67, 10.53, 14.21, 16.87]$.

Table 4
Dimensionless boundary layer parameters used within the GEP dataset.

$M = \frac{U_e}{c_0}$	$\Delta = \frac{\delta^*}{\delta}$	$H = \frac{\delta^*}{\theta}$	$C_f = \frac{\tau_w}{0.5\rho u_e^2}$	$R_T = \frac{\delta^* u_{tau}^2}{\nu u_e}$	$\beta_C = \frac{\theta}{\tau_w}\frac{\partial p}{\partial s}$	Π
0.07	0.139	0.106	3.4×10^{-3}	7.67	0	0.25
0.10	0.129	0.098	2.9×10^{-3}	10.53	0	0.37
0.13	0.128	0.096	2.7×10^{-3}	14.21	0	0.45
0.17	0.124	0.094	2.5×10^{-3}	16.87	0	0.51
0.08	0.162	0.117	2.6×10^{-3}	11.48	0.42	0.69
0.11	0.107	0.080	2.6×10^{-3}	16.87	0.47	0.35
0.13	0.095	0.072	2.4×10^{-3}	20.67	0.56	0.32
0.17	0.100	0.075	2.4×10^{-3}	22.08	0.50	0.39
0.09	0.110	0.088	3.7×10^{-3}	8.03	−0.46	0.01
0.13	0.106	0.086	3.5×10^{-3}	8.61	−0.38	0.00

Table 5
Representation of four different solutions proposed by the GEP to approach Salze dataset [30].

N	Simplified	Fitness	Reynolds dependency	$\tilde{\omega} \to 0$	$\tilde{\omega} \to \infty$
Solution 1	$\frac{\tilde{\omega}(R_T^{-1.23} + \Delta + \tilde{\omega}^{-0.34})}{12.83 + 48.36\tilde{\omega} + \Delta\tilde{\omega}^{1.77} + \frac{\tilde{\omega}^{6.7}}{R_T^{5.07}}}$	3.14%	$R_T^{5.07}$	$\tilde{\omega}^{0.66}$	$\tilde{\omega}^{-6.04}$
Solution 2	$\frac{\frac{\tilde{\omega}^{0.88}}{\Delta + \tilde{\omega}^{1.42}}}{5.17\frac{\tilde{\omega}^{4.86}}{R_T^{5.17}} + \frac{R_T^{0.15}}{3.77}}$	2.84%	$R_T^{5.17}$	$\tilde{\omega}^{0.88}$	$\tilde{\omega}^{-6.28}$
Solution 3	$\frac{6.45(R_T + 1.32)\tilde{\omega}^{0.49}}{R_T + \tilde{\omega}^{1.25}(2R_T + 1) + 10\frac{\tilde{\omega}^{6.41}}{R_T^{4.2}}}$	3.24%	$R_T^{4.2}$	$\tilde{\omega}^{0.49}$	$\tilde{\omega}^{-5.96}$
Solution 4	$\frac{2.54(\tilde{\omega}^{0.97} + 0.04)}{\tilde{\omega}^{1.53} + 0.15 + \frac{0.79}{\Pi}\frac{\tilde{\omega}^{6.75}}{R_T^{3.87}}}$	2.95%	$R_T^{3.87}$	$\tilde{\omega}^{0.97}$	$\tilde{\omega}^{-5.78}$

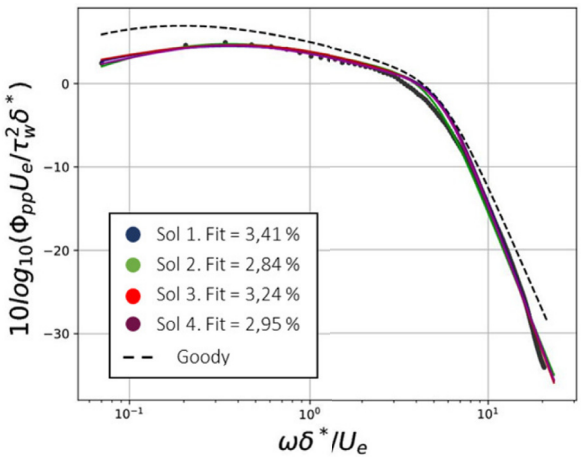
It can be observed from Fig. 13 that the experimental data follows the general trend predicted by Goody's model rescaled from the boundary layer parameters, with nevertheless some noticeable deviations: the high frequency decay does not exactly fit the -5 power law, and the low-frequency levels are slightly below those obtained by the model. It can be expected that the GEP models will differ slightly from the Goody model due to those small differences.

The GEP algorithm has been applied using the same parameters as before. The functional set, linking function and head length remain unchanged compared to the validation case on the synthetic Goody dataset. We use a local optimization loop with $p_{opt} = 1$ every 50 iterations to optimize the accuracy on the constants approximation. The results obtained from the present investigation are reported in Fig. 13 and Table 5.

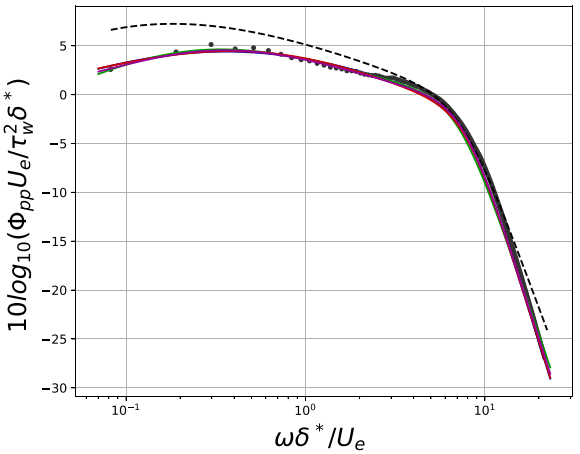
The shape of the functions found by the GEP approach remains consistent with the one found in the investigation of the synthetic dataset in the previous section. The numerator is usually a power of $\tilde{\omega} = \frac{\omega\delta^*}{u_e}$ and describes the tendency of the solution at low frequency. However, the exponent value does not fit the one predicted by the Goody model. This is likely due to the fact that the low frequencies are not as well resolved as in the synthetic dataset, making it more difficult for the algorithm to retrieve the theoretical slope of 2. It should, however, be noted that this theoretical behaviour has been rarely observed in laboratory conditions; to the authors' knowledge the only experimental confirmation has been brought by Farabee and Casarella [33] from low-speed measurements performed in a particularly quiet wind tunnel and using dedicated noise cancellation techniques [11].

In all four solutions, the denominator is composed of two terms. The first one is a combination of an exponential power of R_T and $\tilde{\omega}$, which prescribes the slope at high frequency and the Reynolds number dependency. Although the exponential values tend to vary, this term appears in all solutions. The second term of the denominator is usually a polynomial of $\tilde{\omega}$. It describes the behavior of the model at mid frequency, and its form is seen to vary significantly from one solution to another.

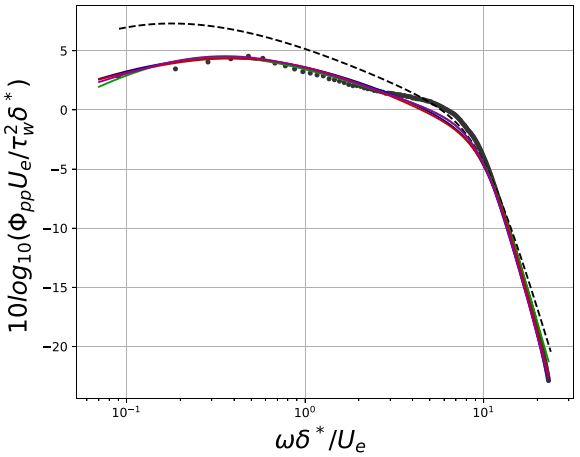
It appears in Fig. 13 that all the solutions fit equally well the data at low and high frequency but fail to capture the hump in the mid frequency region. A possible solution to provide a better fit would be to increase the head length of the individuals. However, this would increase the complexity of the mathematical expressions, making them harder to interpret.



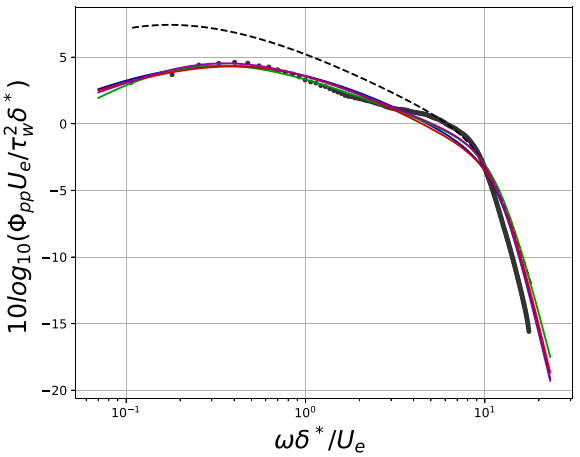
(a) $M = 0.07$ and $\beta_C = 0$



(b) $M = 0.10$ and $\beta_C = 0$



(c) $M = 0.13$ and $\beta_C = 0$

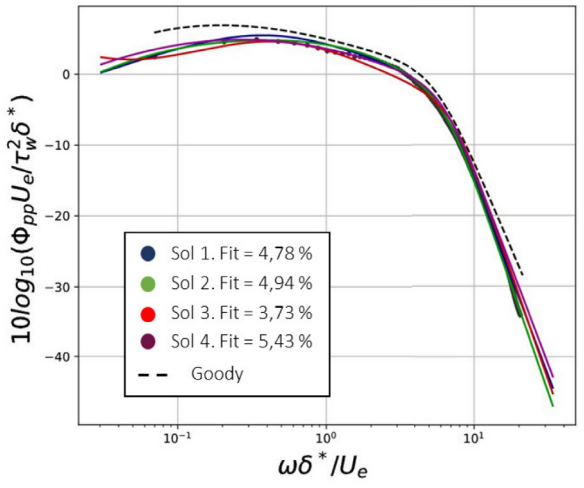


(d) $M = 0.17$ and $\beta_C = 0$

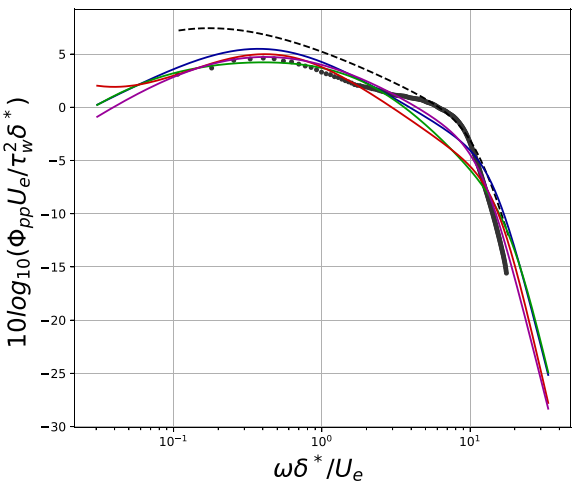
Fig. 13. Comparison between 4 GEP solutions for ZPG experimental data. The grey dot represent the training set and the black dotted line is the Goody model. The four different GEP models expressed in Table 5 are represented in colored lines.

Although all the solutions are equivalent in term of fitness, their mathematical shapes can differ as seen in Table 5. This is due to the algorithm being non-deterministic, with many combinations of operators and variables forming solutions with similar fitness. However, some features appear to be common after repeated runs. For example, the combination of an exponential power of R_T and $\tilde{\omega}$ that describes the slope at high frequency appears in all models of Table 5, as well as in the Goody model.

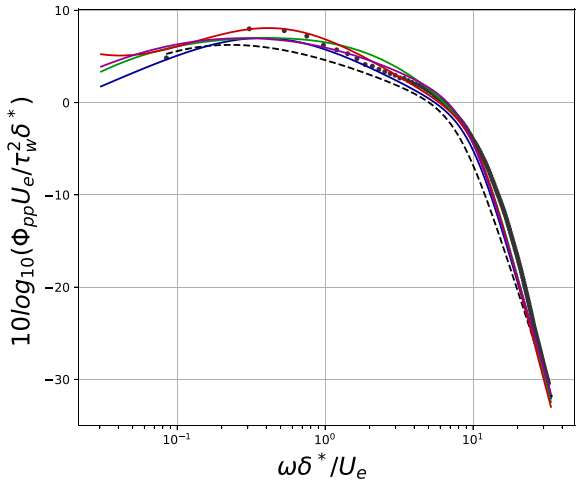
The fact that the general shape of the GEP expressions is quite similar to the Goody model constitutes an interesting result by itself. While it can be argued that the model was somehow supervised by prescribing functional and terminal sets, those sets would still allow such a broad range of solutions that the optimizer can probably be qualified as 'weakly' supervised only. It can also be observed that the small differences between the original Goody model and the experimental data are nevertheless sufficient to yield substantial variations between the low and high frequency scaling laws predicted by the GEP models. This hints at the need for high quality data at both ends of the frequency spectrum in particular, in order to obtain a reliable and robust model. Despite this sensitivity, the results indicate that the proposed approach is capable of providing physically relevant solutions while including fine features from the specific dataset used as input.



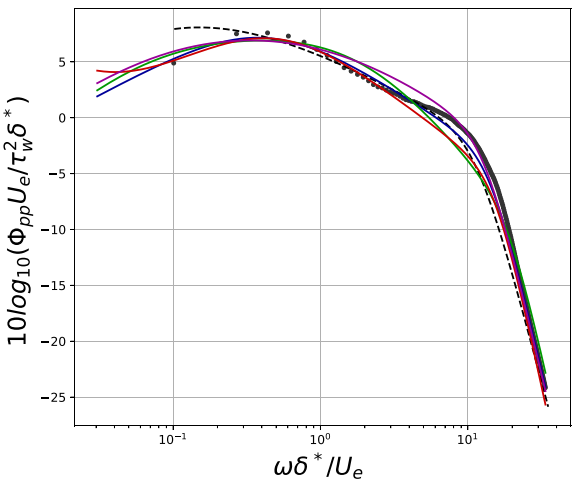
(a) $M = 0.07$ and $\beta_C = 0$



(b) $M = 0.17$ and $\beta_C = 0$



(c) $M = 0.08$ and $\beta_C = 0.42$



(d) $M = 0.11$ and $\beta_C = 0.47$

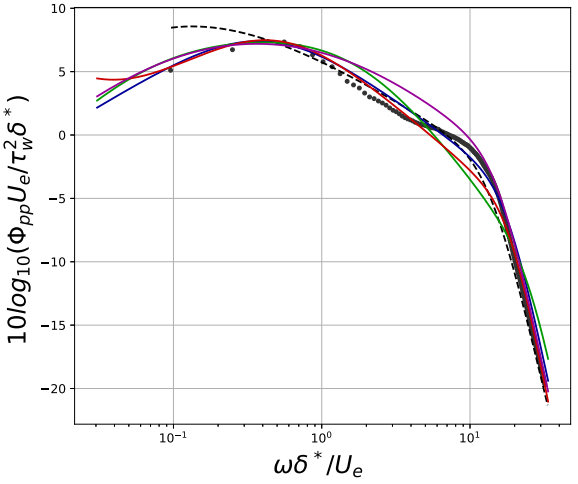
Fig. 14. Comparison between 4 GEP solutions for moderate pressure gradient. The grey dot represent the training set and the black dotted line is the Goody model. The four different GEP models expressed in Table 6 are represented in colored lines.

4.3. Moderate favorable and adverse pressure gradient wall pressure spectra

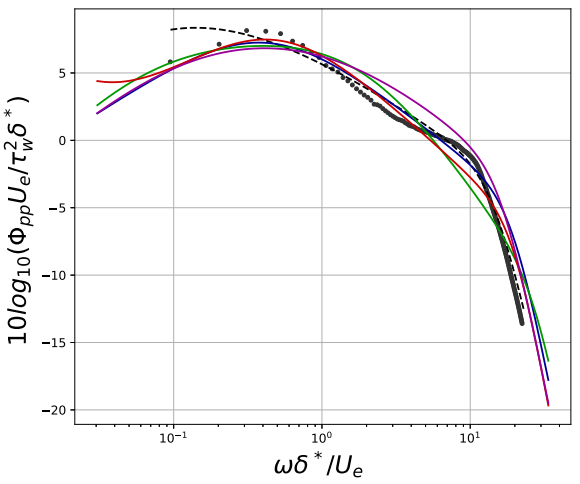
The GEP approach is now applied using the complete measurement database included in Table 4 including 10 boundary layers with adverse, favorable and zero pressure gradients. The GEP algorithm parameters, including the functional set, linking function and head length, are the same as in the previous section. The results obtained for the entire dataset are shown in Fig. 14 and Table 6.

All models fit equally well the data at low and high frequency, but are less consistent in the mid frequency region where they yield different levels and fail to properly capture the mid-frequency hump that is visible in Figs. 14(b), 14(d), 14(e), 14(f), 14(h).

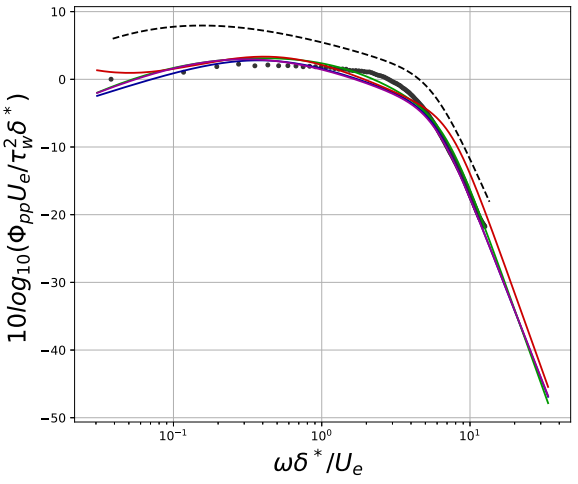
The same trends are observed for the low and high frequency ranges in Table 6 and Table 5 for both pressure gradient and zero-pressure gradient models. However the Reynolds number dependency shows differently when data with pressure gradient are included in the database. Indeed, for ZPG the Reynolds exponent was ranging from 3.87 to 5.07, while including pressure gradient data it is seen to vary between 4.06 and 6.88. This suggests that the pressure gradient affects the frequency where a transition from inner to outer variables scaling applies best. This is observed in the denominator of solution 4



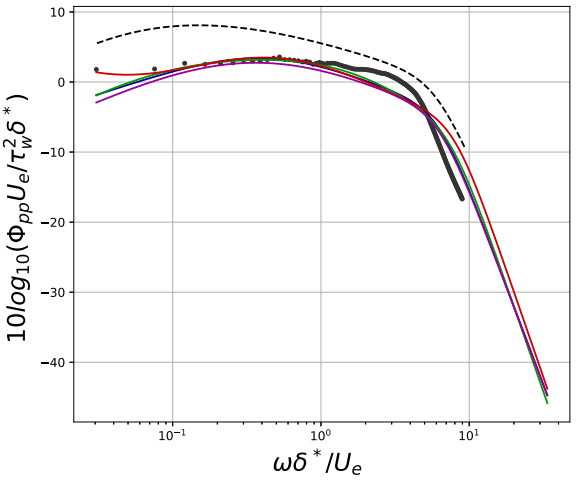
(e) $M = 0.13$ and $\beta_C = 0.56$



(f) $M = 0.17$ and $\beta_C = 0.39$



(g) $M = 0.09$ and $\beta_C = -0.46$



(h) $M = 0.13$ and $\beta_C = -0.38$

Fig. 14. Continued

where the high frequency scaling is directly impacted by the pressure gradient : $\frac{\tilde{\omega}^{6.6}}{(1+\beta_C)^{0.6}R_T^{4.06}}$. In addition, all equations show that the pressure gradient has an impact on the amplitude of the spectrum at low frequency, which increases for adverse pressure gradient and decreases for favorable pressure gradient. This is consistent with the findings of Lee [16] whose model also account for this scaling of the spectra amplitude with favorable and adverse pressure gradient. Previous authors such as Rozenberg [14] or Kamruzzaman [13] also reported an increase of pressure fluctuation level at low frequency but did not extend their analysis to favorable pressure gradients.

The influence of the pressure gradient is not accounted in the Goody model which explain the disparity between the model's predictions and the measurements points in Fig. 14. However this is captured by the different GEP models that match the training data closely in all spectra.

This approach was able to propose several models valid for the whole range of parameters considered in Table 4. The current dataset is limited to experimental measurements over a flat plate boundary layer under moderate pressure gradient. However, given new data, the GEP approach could be extended to a broader range of boundary layer parameters without additional modifications. Making this a great tool to propose models that perform well on larger dataset including different physical assumptions (e.g. pressure gradient, compressibility effects, etc.) without requiring additional effort on the methodology front.

Table 6
Representation of four different solutions proposed by the GEP to approach Salze dataset with pressure gradient [30].

N	Simplified	Fitness	Reynolds dependency	$\tilde{\omega} \to 0$	$\tilde{\omega} \to \infty$
Solution 1	$\frac{4.65\frac{1+\beta_C}{1+C_f}\tilde{\omega}^{0.72}}{1.38(\tilde{\omega}^{1.58}+0.25)+\frac{5.78\tilde{\omega}^{6.5}}{\Delta R_T^{5.78}}}$	4.78%	$R_T^{5.78}$	$\tilde{\omega}^{0.72}$	$\tilde{\omega}^{-5.78}$
Solution 2	$\frac{\left(\frac{1.26}{R_T^{1.26}}+\Delta(1+(1+\beta_C)^{2.83})\right)\tilde{\omega}^{1.22}}{C_f+H\tilde{\omega}^{1.22}+0.1C_f\tilde{\omega}^{2.78}+\frac{\tilde{\omega}^{7.25}}{R_T^{6.88}}}$	4.94%	$R_T^{6.88}$	$\tilde{\omega}^{1.22}$	$\tilde{\omega}^{-6.03}$
Solution 3	$\frac{(\Pi_C+C_f)(1+\beta_C)^{4.48}+\beta_C+1.57}{\frac{M}{R_T}+\tilde{\omega}+\frac{0.29\tilde{\omega}}{Cf+\tilde{\omega}^{1.76}}+\frac{\tilde{\omega}^{6.15}}{R_T^5}}$	3.73%	$R_T^{5.0}$	$C_f+\tilde{\omega}^{1.76}$	$\tilde{\omega}^{-6.15}$
Solution 4	$\frac{\left(\Pi_C+R_T^{0.27}+2.76(1+\beta_C)^{2.76}\right)\tilde{\omega}}{R_T+(1+\beta_C)\tilde{\omega}+\tilde{\omega}^2+\frac{\tilde{\omega}^{6.6}}{(1+\beta_C)^{0.6}R_T^{4.06}}}$	5.43%	$R_T^{4.06}$	$\tilde{\omega}^{1.0}$	$\tilde{\omega}^{-5.6}$

5. Conclusions

Various models have been proposed in the literature to describe the point spectrum of the pressure field below a turbulent boundary layer, evolved on top of each other and based on different physical assumptions to account for Reynolds number effects or the presence of pressure gradients. In this work, a different approach has been pursued, where the model is purely data-driven. Gene Expression Programming yields a mathematical expression, the complexity of which can be controlled through the length of the gene head, the functional and terminal sets. As in other machine learning techniques, a trade-off must be found to obtain a good representation of the training dataset without over-fitting.

Two well-established models have been considered in this study, in order to test the ability of GEP to yield consistent formulations and deduce the appropriate lengths of the individuals. It was found that the Chase-Howe functional form can be obtained using the original GEP algorithm reported in literature. In contrast, the more advanced model by Goody required two additions: 1) a new type of terminal called the power terminal and 2) a local optimization loop based on a least-square regression. Using those two modifications, it was possible to reproduce the functional shape proposed by Goody without any specific assumptions on the physics involved.

The algorithm was finally tested on two sets of wind tunnel measurements. For the first one, composed exclusively of zero-pressure gradient boundary layers, it provided a model that was similar to the expected Goody model, with slight changes to best match the experimental training data. On the second set, which included both favorable and adverse pressure gradient boundary layers, the GEP suggested new ways to model the impact of pressure gradient by modifying the amplitude of the spectrum at low frequency and changing the point of transition from outer to inner variable scaling. This revealed that GEP was capable of providing physically meaningful solutions while taking into account the specificities of the experiment.

In the long term, the authors conjecture that GEP can prove advantageous to devise new models, unveil key dependencies between variables in complex numerical datasets, and possibly hint at underlaying scaling laws and physical mechanisms. Simultaneously, a priori physical insight can help defining GEP meta-parameters such as linking function, mutation rates, functional and terminal sets, etc. In that sense, machine learning techniques such as the one proposed in this work could be seen as a useful complement to physics-based models, to be further investigated in more complex flow configurations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Joachim Dominique: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Julien Christophe:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration. **Christophe Schram:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Richard D. Sandberg:** Methodology, Resources, Writing - review & editing.

Acknowledgments

This research was partly supported by the French global automotive Valeo in the context of a thesis for the investigation of noise induced by HVAC systems. The authors thank Dr. Salze and Pr. Bailly from Ecole Centrale de Lyon for providing the experimental data.

References

[1] E. Ciappi, S. De Rosa, F. Franco, J. Guyader, S. Hambric, et al., Flinovia-flow induced Noise and Vibration Issues and Aspects, Springer, 2015, doi:10.1007/978-3-319-09713-8.
[2] N. Van de Wyer, A. Zapata, D. Nogueira, C. Schram, Development of a test rig for the measurement of turbulent boundary layer wall pressure statistics, 2018 AIAA/CEAS Aeroacoustics Conference (2018), doi:10.2514/6.2018-3122.
[3] M. Roger, S. Moreau, Back-scattering correction and further extensions of Amiet's trailing-edge noise model. part 1: theory, J. of Sound and Vib (286) (2005) 477,506, doi:10.1016/j.jsv.2004.10.054.
[4] H. Kraichnan R, Pressure fluctuations in turbulent flow over a flat plate, The J. of the Acoustical Society of Am. 28 (3) (1956), doi:10.1121/1.1908336.
[5] K. Bull M, Wall-pressure fluctuations beneath turbulent boundary layers: some reflections on forty years of research, J. of Sound and Vib. (190) (1996) 299,315, doi:10.1006/jsvi.1996.0066.
[6] R.L. Panton, J.H. Linebarger, Wall pressure spectra calculations for equilibrium boundary layers, J. of Fluid Mechanics (1974), doi:10.1017/S0022112074001388.
[7] W. Blake, Mechanics of flow-Induced sound and vibration - Volume 1: General concepts and elementary sources, 1, Academic Press, 1986.
[8] G. Grasso, P. Jaiswal, H. Wu, S. Moreau, M. Roger, Analytical models of the wall-pressure spectrum under a turbulent boundary layer with adverse pressure gradient, J. of Fluid Mechanics 877 (2019) 1007–1062, doi:10.1017/jfm.2019.616.
[9] K. Amiet R, Noise due to turbulent flow past a trailing edge, J. of Sound and Vib. 47 (3) (1976) 387,393, doi:10.1016/0022-460X(76)90948-2.
[10] M. Howe, Acoustics of fluid-Structure interactions, Cambridge University Press, 1998.
[11] M. Goody, Empirical spectral model of surface pressure fluctuations, AIAA J. 42 (9) (2004), doi:10.2514/1.9433.
[12] F. Hwang Y, K. William, S. Bonness, A. Hambric, Comparison of semi-empirical models for turbulent boundary layer wall pressure spectra, J. of Sound and Vib. (319) (2007) 199,217, doi:10.1016/j.jsv.2008.06.002.
[13] M. Kamruzzaman, D. Bekiropoulos, T. Lutz, W. Würz, E. Krämer, A semi-empirical surface pressure spectrum model for airfoil trailing-edge noise prediction, International J. of Aeroacoustics 14 (5–6) (2015) 833–882, doi:10.1260/1475-472X.14.5-6.833.
[14] Y. Rozenberg, G. Robert, S. Moreau, Wall-pressure spectral model including the adverse pressure gradient effects, AIAA J. 50 (10) (2012) 2168–2179, doi:10.2514/1.J051500.
[15] N. Hu, Empirical model of wall pressure spectra in adverse pressure gradients, AIAA J. 56 (9) (2018) 3491–3506, doi:10.2514/1.J056666.
[16] S. Lee, Empirical wall-pressure spectral modeling for zero and adverse pressure gradient flows, AIAA J. 56 (5) (2018) 1818–1829, doi:10.2514/1.J056528.
[17] P. Bradshaw, 'Inactive' motion and pressure fluctuations in turbulent boundary layers, J. of Fluid Mechanics 30 (2) (1958) 241–258, doi:10.1017/S0022112067001417.
[18] S. Gao, Geppy, 2019-01-22, (https://geppy.readthedocs.io/en/lates).
[19] C. Ferreira, Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence, Springer-Verlag, 1 edition. DOI: 10.1007/3-540-32849-1
[20] J. Zhong, L. Feng, Y.-S. Ong, Gene expression programming: a survey, IEEE Comput Intell Mag 12 (3) (2017) 54–72, doi:10.1109/MCI.2017.2708618.
[21] J. Weatheritt, R. Sandberg, A novel evolutionary algorithm applied to algebraic modifications of the RANS stress–strain relationship, J. of Computational Phys. 325 (2016) 22–37, doi:10.1016/j.jcp.2016.08.015.
[22] C. Lav, R.D. Sandberg, J. Philip, A framework to develop data-driven turbulence models for flows with organised unsteadiness, J. of Computational Phys. 383 (2019) 148–165, doi:10.1016/j.jcp.2019.01.022.
[23] J. Weatheritt, R.D. Sandberg, Hybrid reynolds-averaged/large-eddy simulation methodology from symbolic regression: formulation and application, AIAA J. (2017) 3734–3746, doi:10.2514/1.J055378.
[24] M. Schoepplein, J. Weatheritt, R. Sandberg, M. Talei, M. Klein, Application of an evolutionary algorithm to LES modelling of turbulent transport in premixed flames, J. of Computational Phys. 374 (2018) 1166–1179, doi:10.1016/j.jcp.2018.08.016.
[25] C. Ferreira, Automatically Defined Functions in Gene Expression Programming, in: Genetic Systems Programming, Springer, 2006, pp. 21–56, doi:10.1007/3-540-32498-4_2.
[26] T. Verstraete, Multidisciplinary turbomachinery component optimization considering performance, stress, and internal heat transfer, Karman Institute, 2008 Ph.D. thesis.
[27] J.R. Koza, Genetic programming: on the programming of computers by means of natural selection, 1, MIT press, 1992 10.1.1.44.5416.
[28] C. Ferreira, Function finding and the creation of numerical constants in Gene Expression Programming, in: Advances in soft computing, Springer, 2003, pp. 257–265, doi:10.1007/978-1-4471-3744-3_25.
[29] D. Chase, Modelling the wavevector-frequency spectrum of turbulent boundary layer wall pressure, J. of Sound and Vib. 70 (1980) 29–67, doi:10.1016/0022-460X(80)90553-2.
[30] E. Salze, C. Bailly, O. Marsden, E. Jondeau, D. Juve, An experimental characterisation of wall pressure wavevector-frequency spectra in the presence of pressure gradients, in: 20th AIAA/CEAS Aeroacoustics Conference, 2014, p. 2909, doi:10.2514/6.2014-2909.
[31] A. Musker, Explicit expression for the smooth wall velocity distribution in a turbulent boundary layer, AIAA J. 17 (6) (1979) 655–657, doi:10.2514/3.61193.
[32] K.A. Chauhan, P.A. Monkewitz, H.M. Nagib, Criteria for assessing experiments in zero pressure gradient boundary layers, Fluid Dynamics Res. 41 (2) (2009) 021404, doi:10.1088/0169-5983/41/2/021404.
[33] T. Farabee, M. Casarella, Spectral features of wall pressure fluctuations beneath turbulent boundary layers, Phys. of Fluids 3 (10) (1991) 2410–2419, doi:10.1063/1.858179.