

Chapter 1

BASELINE GENETIC PROGRAMMING SYMBOLIC REGRESSION ON BENCHMARKS FOR SENSORY EVALUATION MODELING

Pierre-Luc Noel¹, Kalyan Veeramachaneni² and Una-May O'Reilly²

¹*Swiss Federal Institute of Technology, Switz.* ²*Massachusetts Institute of Technology, USA*

Abstract We introduce hedonic modeling benchmarks for the field of sensory science evaluation. Our benchmark framework provides a general means of defining a response surface which we call a “sensory map”. A sensory map is described by a mathematical expression which rationalizes domain specific knowledge of the explanatory variables and their individual or higher order contribution to hedonic response. The benchmark framework supports the sensory map’s so-called *ground truth* to be controllably distorted to mimic the human and protocol factors that obscure it. To provide a baseline for future algorithm comparison, we evaluate a public research release of genetic programming symbolic regression algorithm on a sampling of the framework’s benchmarks.

Keywords: symbolic regression, benchmarks, sensory evaluation, hedonic modeling

1. Introduction

Benchmark problems (or simply “benchmarks”) allow the evaluation of algorithms. The GP research community has a variety of useful, realistic general GP symbolic regression benchmarks including non-linear non-polynomials (Keijzer, 2003; Vladislavleva et al., 2009), publicly obtainable financial market data (Nikolaev and Iba, 2001; Becker et al., 2006; Becker et al., 2007; Becker and O’Reilly, 2009), the Mathematica-Wolfram data set (Kotanchek et al., 2009) and the accuracy problems of (Korns, 2011).

To systematically design a general benchmark for GP symbolic regression is straightforward. One creates a response surface which is a function of explanatory variables. The function is executed to obtain observational samples of the response surface. Samples are usually collected all at once and split into training and cross validation sets before the GP symbolic regression algorithm executes. When the algorithm completes and produces a predictive model of the response surface, this model can be queried with a set of unseen samples (i.e. the test set) and its predictive accuracy on the benchmark response surface can be ascertained.

We have developed a suite of GP symbolic regression and complementary algorithms (Veeramachaneni et al., 2010; Vladislavleva et al., 2010a; Vladislavleva et al., 2010b) to knowledge-mine hedonic preferences data collected when multiple assessors (also called panelists) are each asked how much he or she likes a set of food or highly aromatic stimuli. Figure 1-1 depicts the general sensory evaluation process. In a study (or experiment) each *assessor* is presented, in succession, with a limited quantity of randomly ordered, pre-selected, different products from a design space. On each presentation, the assessor must sense the product (by taste or smell) and respond to a query designed to elicit information as to how much s/he likes it. In the benchmarks we present here, the query to the assessor is “how much do you like X?” and the response structure (or format) is that the assessor must respond with one of nine discrete choices from the range bounded by *extremely dislike* through *neutral* to *extremely like*. An assessor’s responses are collected for one experiment. These responses are observational samples of the assessor’s (unknown) hedonic response surface. GP symbolic regression can be used to model an assessor’s hedonic response to the product space by training with some (or all) of these numerically converted observations and knowledge of the design inputs (i.e. explanatory variables, ingredients of the food or the constituents of the aroma) in the product space (Vladislavleva et al., 2010a).

Validating this hedonic modeling of real data is virtually impossible because the number of queries is extremely low - around 10 for taste and around 40 for smell due to sensory fatigue. Using precious data samples for testing and cross validation is of little value because frequently the queries are de-

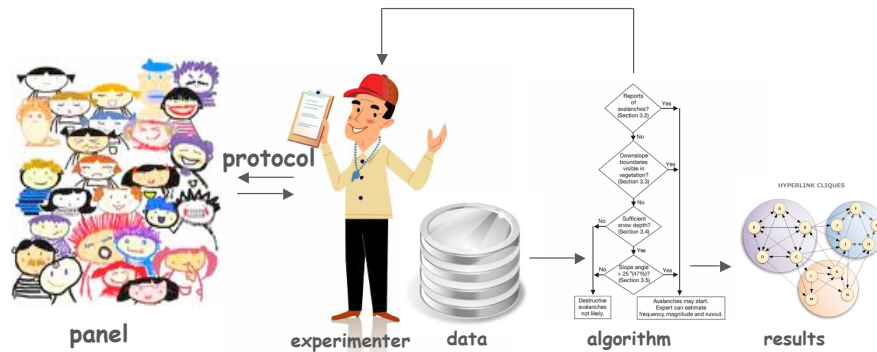


Figure 1-1. The sensory protocol and experimental analysis process which situates the context of the presented benchmark design and framework.

terminated by experimental design and are very distant from one another in the design space (i.e. they are at design corners). They are better exploited for training.

Other means of validation might be to use domain experts' experience, ask assessors to test optimized product designs derived using the predictive models or go back to the original assessors to confirm unseen sample predictions. These options turn out to be infeasible because assessors individually tend to not be consistent across days with their hedonic responses to a set of products (though as an aggregate there is stability), experts do not match with naive assessors, and, often neither are available. Thus, in the domain of sensory evaluation, there is no means of evaluating modeling methods.

Using a general symbolic regression benchmark is also insufficient. First, there is no rational basis for using any existing benchmark. They do not rationally express domain specific knowledge of how the ingredients, individually or in higher order combinations, contribute to hedonic response. Second, a benchmark should express one defining aspect of the sensory evaluation domain: a human is in the loop and introduces "noise" into the sampling for a number of different reasons. These range from human factors like fatigue, moodiness, inconsistency, perceptual confusion and memory loss to how humans deal with the protocol's response format, e.g. the 9 value hedonic range.

As a solution, we present a means of systematically designing, in a parameterized, controllable manner, domain-informed sensory evaluation benchmarks. To accomplish this, a benchmark expresses a sensory map **plus** what an assessor will report given this ground truth sensory map and the distortions arising from human judgment and protocol-driven response decision. The benchmark framework simulates the end to end process of sensory evaluation

(see Figure 1-2): Queries, each accompanied by a sample, are formulated according to the protocol. Each sample is presented to the assessor, the latter sniffs or tastes it (which is a physical stimulus) and responds to the query. The benchmark framework assumes the physical stimulus generates a raw sensory interpretation in the brain which it represents via the sensory map. It assumes the way the assessor reports each response, which also depends on the protocol, is the combination of this raw information, and the distortions involved by the human judgment and reporting.

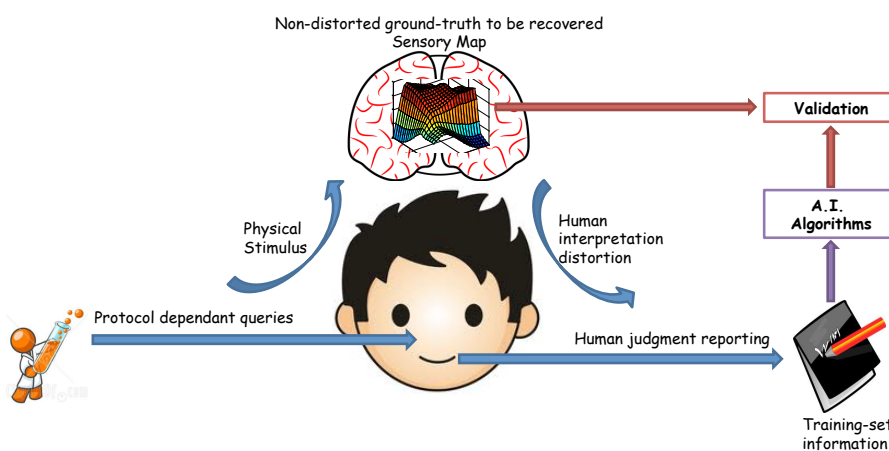


Figure 1-2. Simulation scope of the benchmark framework.

We proceed thus: Section 2 describes sensory map construction. Section 3 describes how we tunably model distortion. Section 4 uses the benchmark framework to evaluate the predictive accuracy of simple GP symbolic regression models trained on successively less training samples and on successively more distortion. Section 5 concludes.

2. Generation of a sensory map (Ground truth)

The benchmark framework assumes there exists a “ground-truth” definition of the non-distorted hedonic function, a so-called sensory map. It is convenient to think of the map as being in the assessor’s brain. The sensory map describes the hedonic response of the assessor for each possible product in a design space. For example, in (Vladislavleva et al., 2010b), the design (or product) space is all possible combinations of seven flavoring ingredients, referred to as keys. Here, the sensory map is a function of seven variables (the volume of each key) and one output which is the hedonic score. Each key level is normalized to range continuously from 0 to 1, meaning respectively *ingre-*

dient not present and clearly too much of this ingredient. For comparison, the range of a map is always $[-4, 4]$ where -4 means *extremely dislike*, 4 means *extremely like*, and a 0 -value means neutrally like.

Naive sensory maps

A naive sensory map is a completely analytic function of explanatory variables (e.g. the key levels) with one output ranging from -4 to 4 . It is not intended to represent a plausible human sensory response to flavors. Because of this drawback, we present only one example, Equation 1.1, (of 7 dimensions):

$$h_{naive}(\vec{k}) = \text{sigmoid}(\sin(10k_1) + \cos(10k_3) + \sin(7k_2 + 3k_4) + \sin(3k_7 + k_5k_6), [-4, 4], 0.75) \quad (1.1)$$

where \vec{k} is the sample i.e. the vector of the key levels k_i and sigmoid is:

$$\text{sigmoid}(x, [a, b], \beta) = (b - a) \frac{1}{1 + e^{-2\beta x}} + a \quad (1.2)$$

where a and b respectively are the lower and upper asymptotes and β controls the steepness of the curve. This easily keeps the hedonic score within the range of $[-4, 4]$.

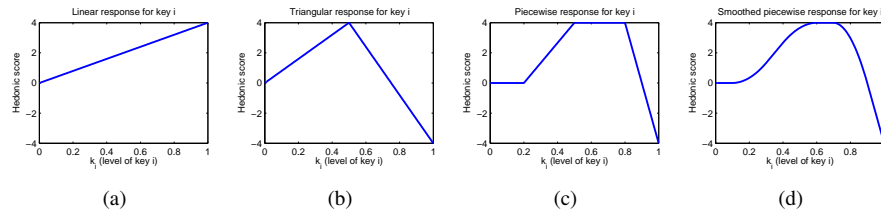


Figure 1-3. Examples of sensory response to individual keys (a) Linear response (b) Roof response (c) Piecewise response and (c) smoothed piecewise response

Rational Sensory Maps

Rational sensory maps exploit domain specific knowledge expressing an assessor's hedonic response to a single ingredient (a.k.a. key). Founding our approach on this knowledge provides plausibility to the extent this knowledge is acquired or estimated. Our approach is to rationally combine the response function of each ingredient (domain knowledge) into a comprehensive one for multiple ingredients. We mathematically construct an n -dimensional sensory map which uses lower dimensional information which is known, or easily accessible. Two domain specific properties are integrated as design invariants:

Invariant #1 If all key levels are 0, meaning that only the base is present, the liking score has to be neutral, i.e. 0.

Invariant #2 For each key i , the single response for this specific key has to be continuously recovered when all other keys are at their zero levels.

Design Steps. Designing a rational sensory map involves 3 steps: **(1)** design 1-dimensional sensory responses **(2)** combine these single responses to form a multi-dimensional map, **(3)** design a merging function for the combination to preserve the two invariant properties. Use a post-merging coefficient γ to scale the merged map relative to the combined responses, then, if necessary, apply a sigmoid function to ensure responses are within the range $[-4, 4]$.

Step 1. 1-Dimensional Sensory Response Design. A 1D sensory response expresses how an assessor responds to an increasing level of one flavor ingredient. We describe a 1D response to key i as h_i , with one of four parameterized functions:

- 1 **Linear response:** A linear sensory response either increases or decreases linearly with the increase of an ingredient. The response is parameterized by the slope and intercept of the linear function given in the following equation.

$$h_i = a_i \times k_i + b_i \quad (1.3)$$

- 2 **Piecewise linear (2-piece):** This map models the following hedonic response behavior: as the key level k_i increases, the assessor likes it more until a point of maximal preference. Then, when the key level further increases, the flavor is too strong and the assessor actually starts to dislike it. This is characterized by a *piecewise* linear function with 2-pieces. For any key k_i this function is parameterized by $\{l_i, a_i^{(1)}, b_i^{(1)}, a_i^{(2)}, b_i^{(2)}\}$.

$$h_i = \begin{cases} a_i^{(1)} \times k_i + b_i^{(1)} & \text{if } k_i \leq l_i \\ a_i^{(2)} \times k_i + b_i^{(2)} & \text{if } l_i < k_i < 1 \end{cases} \quad (1.4)$$

- 3 **Piecewise linear (n-piece):** This map models hedonic response behavior in which an assessor changes from positive to negative or vice versa, or has a constant hedonic response at multiple volume intervals between the minimum and maximum of the key. In any volume interval the response is a linear function with a specific slope. This is thought to be more realistic for many ingredients and assessors. For any key k_i this function is parameterized by a set of interval extrema $\{l_i^{(1)}, \dots, l_i^{(n)}\}$ and the coefficients for the linear function that describes each segment.

These are a_1, \dots, a_n and, b_1, \dots, b_n .

$$h_i = \sum_{p=1}^{n-1} a_p \times k_i + b_p \text{ if } l_i^{(p)} \leq k_i \leq l_i^{(p+1)} \quad (1.5)$$

- 4 **Piecewise linear smooth:** Piecewise linear single responses are defined by their intervals and slopes within each interval. However, smoothed single responses are presumably even more plausible. To model this, we apply a smoothing mean filter, Equation 1.6), over a range controlled by a so-called smoothing coefficient s_c . The smoothed single response of Figure 1-3 (c) is Figure 1-3(d) when a smoothing coefficient of 0.1 is used.

$$h_{i_{smoothed}}(k) = \int_{k-s_c}^{k+s_c} h_i dk, \quad (1.6)$$

Examples of these single response models are shown in Figure 1-3-bottom. We can design a *homogeneous* set of 1D responses for the keys in which all the keys have similar sensory response functions but the parameters are varied. A *heterogeneous* alternative mixes multiple kinds of 1D response functions.

Step 2. Multi-dimensional Sensory Map Design. These sensory maps are a combination of the 1D response functions. We call the functional relationship linking the lower dimensional functions to the higher dimensional sensory map a *combining function*. Combining function outputs are scaled with λ , a scaling coefficient before merging. We defer explaining λ until the end of Step 3 where we also explain a second scaling coefficient, γ .

Three kinds of *combining functions* are:

- 1 **Additive:**

$$h_{1,\dots,d} = \lambda \sum_{i=1}^d h_i, \quad (1.7)$$

- 2 **Multiplicative**

$$h_{1,\dots,d} = \lambda * \prod_{i=1}^d h_i \quad (1.8)$$

- 3 **Second order combining function** Pairs of keys may also generate second-order responses, on top of first order ones. This 2^{nd} order combining function is:

$$h_{1,\dots,d} = \lambda \Sigma ((H^T H) * K), \quad (1.9)$$

where K is the coefficient matrix for the second order terms and $*$ the term-by-term multiplication of it with the single response functions in

H. An example of a 7-dimensional function is given by

$$H = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \end{pmatrix}, \text{ and } K = \begin{pmatrix} 0 & 2 & 0 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

Step 3. Merging to Preserve Map Invariants. In all but additive combining functions, *merging* is required to preserve design invariant #2. Recall that this invariant maintains that the 1D response to a key i is recoverable when all other key levels are 0. To understand a merging function, first consider a simpler case where there is only two ingredients. Then, the merging procedure can be formulated as follows:

$$\begin{aligned} h'_{1,2} &= m_{down}^{r_2}(k_2)h_1 \\ &+ m_{down}^{r_1}(k_1)h_2 \\ &+ (1 - m_{down}^{r_1}(k_1))(1 - m_{down}^{r_2}(k_2))h_{1,2} \end{aligned} \quad (1.10)$$

with $h_{1,2}$ as the combining function from Step 2 and $m_{down}^{r_i}$ the "merging-down" function defined by:

$$m_{down}^{r_i}(k_i) = \begin{cases} 1 - \left(\frac{k_i}{r_i} - \frac{1}{2\pi} \sin(2\pi \frac{k_i}{r_i})\right), & \text{if } 0 \leq k_i \leq r_i \\ 0, & \text{if } r_i < k_i \leq 1 \end{cases} \quad (1.11)$$

with r_i the *range* of the merging in the direction of k_i . Figure 1-4 depicts this merging down function with a range of 0.5.

Correspondingly, note that $1 - m_{down}^{r_i}(k_i)$ can be seen as a "merging-up" coefficient.

The idea of this merging procedure is that the 2d sensory map $h_{1,2}$ (which is the result of a combination of both single responses h_1 and h_2) is valid when both key levels k_1 and k_2 are sufficiently large. However, when a key level, say k_1 , approaches zero, one wants to smoothly recover the 1d response h_2 . This is implemented using the merging coefficient of equation (1.10). In the example, as k_1 approaches 0, the coefficient $m_{down}^{r_1}(k_1)$ approaches 1, emphasizing the influence of the single response h_2 , while at the same time, the coefficient $(1 - m_{down}^{r_1}(k_1))$ approaches 0 diminishing the influence of $h_{1,2}$. The same logic applies when key level k_2 approaches zero. This approach can be generalized to the n-dimensional case as follows:

$$\begin{aligned} h'_{1,\dots,d} &= \sum_{i=1}^d M_i(k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_d)h_i \\ &+ \prod_{i=1}^d (1 - M_i) h_{1,\dots,d} \end{aligned} \quad (1.12)$$

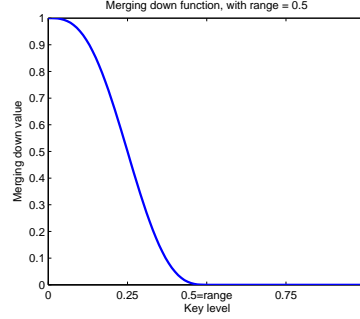


Figure 1-4. Merging down function with $r_i = 0.5$.

with $M_i(k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_d) = \prod_{j \neq i} m_{down}^{r_j}(k_j)$. Maintaining consistency of the 2-d hedonic response 'building blocks' (i.e. when other key levels approach 0), can be achieved with the following equation:

$$\begin{aligned}
 h'_{1, \dots, d} &= \sum_{i=1}^d M_i h_i \\
 &+ \prod_{i=1}^d (1 - M_i) \\
 &\quad \left[\sum_{(p,q) \in S} h_{pq} M_{pq} \right. \\
 &\quad \left. + \prod_{(p,q) \in S} (1 - M_{pq}) h_{1, \dots, d, \dots, p_1 q_1, \dots, p_s q_s} \right]
 \end{aligned} \tag{1.13}$$

with $M_{pq} = \prod_{j \neq p, j \neq q} m_{down}^{r_j}(k_j)$, and S the set of the s pairs (p_i, q_i) whose 2-d responses are known and used to build the d dimensional sensory space.

A post-merging coefficient γ is used to scale the merging function. If responses are outside the range $[-4 \ 4]$, a sigmoid function with parameter β is further applied:

$$h''_{1, \dots, d} = \text{sigmoid}(\gamma \times h'_{1, \dots, d}, [-4 \ 4], \beta) \tag{1.14}$$

Scaling Coefficients, λ, γ : We "tune" a map by choosing the λ and γ coefficients in a coupled manner. We aim for a factor of 2 between the entire range of the map, $h_{1, \dots, d}$, and the sub-range when only a single key level is non-zero. The range of a combining function can be calculated in a straight-forward analytical manner for any heterogenous or homogeneous set of 1D responses. This allows us to pick one of λ or γ and set the other accordingly. For example, in map 1 of Table 1-2, the range of h before scaling with λ is $[-8 \ 8]$ so, when we pick $\gamma = 1/2$, we set $\lambda = 1$. In map 2 of Table 1-2, the maximum of h happens to be 176 (occurring at key levels equaling $[0.5, 0.25, 0.75, 1.0, 0.5, 0.5, 0]$) so, picking $\gamma = 1/2$, implies we set $\lambda = 4/88 = 8/176$.

When using a second order combining function, the choice of coefficients is arbitrary. We proceed by identifying “interesting” 2D cuts of the multi-dimensional surface and use particle swarm optimization to find the range of h' . We loosely aim for a factor of 2 between this range and the range of the responses when only a single key level is non-zero when picking λ , γ and β .

3. Sample Distortion

When reporting their judgment, humans are sources of error (Leibowitz and Post, 1982). These errors or distortions arise from: (1) assessors’ intrinsic characteristics/abilities, and (2) protocol induced factors. The literature on psychological analysis of hedonics and hedonics protocols describes such errors and human characteristics. As well, experts who conduct many hedonic sensory evaluation protocols record many examples. Certain error sources are well understood and can be avoided very easily. For instance, expectation of the observer can be avoided by labeling samples in a neutral manner (Leibowitz and Post, 1982). This eliminates expectations and *a priori* biases in judgment. Table 1-1 mathematically describes our framework’s distortions. We use h_j^t where h denotes the hedonic response (undistorted at first, later successively changed by each distortion), the superscript t indexes the position of the sample in the query sequence and subscript j indexes each successive distortion.

Human Induced Distortion

The first distortion factor is the *inconsistency* in the ratings from an assessor. For instance in (Costello et al., 2007), only roughly 1/3 of assessors gave the exact same rating to actually identical flavors. The second factor is *mood*, which increases or decreases the hedonic response. The third factor is the *sensitivity* of an assessor to distinguishing among samples.

Protocol Induced Distortion

How a judgment is reported can lead to different information (Moskowitz, 1982). While this is not strictly “error”, it implies that ground truth will be transformed in a protocol dependent way before it is reported. For instance, when using fixed 9-point category scaling (a.k.a. a hedonic scale), assessors must respond within this range. This causes *clipping* and *discretization* distortion. In addition, an assessor may use an extremum of the scale early on one sample then later like (or dislike) another sample more. Assessors “solve” this dilemma with truncation – i.e. re-using an extremum to express a more extreme response. For the forced response hedonic protocol we initially simulate, we model these distortions as *first query ignorance*.

Sensory judgment will be biased by previously tested samples. For instance, a given sample will tend to be judged saltier when presented in a set that in-

cludes many low salt concentration samples, while it will tend to be judged less salty when presented in a distribution including many high salt concentration samples (Riskey, 1982). We call this type of distortion *contextual exaggeration*.

Table 1-1. Models for a variety of distortion sources

Distortion	Source	Model	Parameter
Inconsistency	Human	$h_1^t = h_0^t + \eta(0, \sigma_1(t))$	$\sigma_1(t)$
Sensitivity	Human	$h_2^t = s(t) \times h_1^t$	$0 \leq s(t) \leq 1$
Mood	Human	$h_3^t = h_2^t + m(t)$	$m(t)$
Ist Query Ignorance	Protocol	$h_4^1 = h_3^1 + \eta(0, \sigma_4)$	σ_4
Contextual Exaggeration	Protocol	$h_4^t = h_3^{t-1} + \alpha(\delta, t) \times \delta$ with $\delta = h_3^t - h_3^{(t-1)}$ and $\forall t > 1$	$\alpha(\delta, t)$

Time Varying Fatigue Distortion

Sensory fatigue that increases over causes time varying error. We assume time to be discrete and corresponding to successive presentation of samples. The quality and accuracy of human input greatly degrades with repeated prompts for input (Schmidt and Lipson, 2006). Even though protocols are often designed to keep sample quantity low, fatigue will eventually alter assessors' capabilities. For instance, the average person can sample only up to seventy aromas before they become biased (Costello et al., 2007). In our framework, this is modelled by giving a temporal dimension to every source of distortion. This means that the parameters described for different distortions in Table 1-1 change over time. $\sigma_1(t)$, $s(t)$ and $m(t)$ change using a temporal evolution function defined below. α is a function of t and δ . Its values with changing t for different δ 's is shown in the figure 1-5(b).

Starting with an initial value of a distortion parameter r_i and a set final value r_f of this same parameter, the temporal evolution of the parameter corresponds to equation 1.15 and is depicted in figure 1-5. In this equation, τ is a time constant that can be interpreted in term of number of sample tested.

$$r(t) = r_i + (r_f - r_i)(1 - e^{-t/\tau}) \quad (1.15)$$

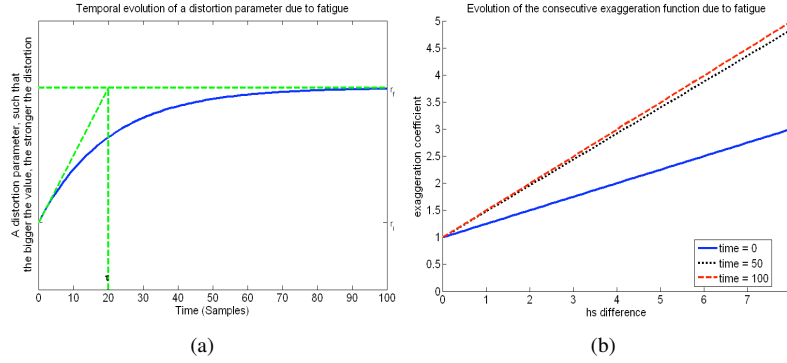


Figure 1-5. Temporal evolution of a distortion parameter due to (a) panelist fatigue, (b) contextual exaggeration.

4. Baseline GP symbolic regression experiments

We now apply “standard” symbolic regression to sensory evaluation benchmarks of varied difficulty. We use the GPLAB toolbox (Silva and Almeida, 2003; Silva, 2011). We control modeling difficulty via sensory map choice, presence or absence of distortion, and, the number of samples recorded. We set the dimensionality of all maps to 7 and use the four sensory maps described in Table 1-2. We then set the initial and final levels for the distortion as detailed in Table 1-3 or choose not to have distortion. We use sample sizes of [40, 100, 1000]. Thus we have experiments pertaining to 36 datasets. A benchmark experiment proceeds per Figure 1-6.

Table 1-2. The 4 sensory maps used.

Sensory Map	Single Responses	Combining Function	Merging Range	Scaling Parameters
1	Linear with different coefficients	Additive	No merging needed	$\lambda = 1$ $\gamma = \frac{1}{2}$
2	2-piece-linear with different coefficients	2nd order, K per example in 2.0	$r_i = 0.5$	$\lambda = \frac{4}{88}$ $\gamma = \frac{1}{2}$
3	Diverse and more complex	Involving $+$, \times , \cos , \sin , $\sqrt{\quad}$, and $ $	$r_i = 0.3$	$\lambda = \frac{4}{5}$ $\gamma = 3.5$ $\beta = 0.1$
4	Naive Sensory Map, Equation 1.1			

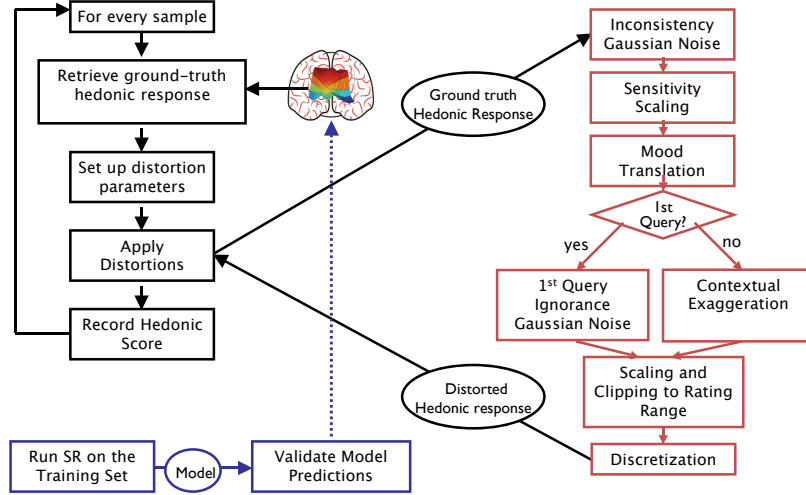


Figure 1-6. Flow of the benchmark framework.

Table 1-3. Parameter settings for distortion

	No Distortion	Strong Distortion With Fatigue ($\tau = 20$)	Strong Distortion Without Fatigue
Parameter	r	$\{r_i, r_f\}$	r
σ_1	0	$\{0.75, 2.25\}$	0.75
s	1	$\{0.5, 0.2\}$	0.5
m	0	$\{1, -1\}$	1
σ_4	0	2	2
$\alpha(\delta)$	1	$\{1 + \frac{1}{4}\delta, 1 + \frac{1}{2}\delta\}$	$1 + \frac{1}{4}\delta$
Rating range	$[-4, 4]$ Cont.	$[1, 9]$ Discrete	$[1, 9]$ Discrete

Results

We perform 30 runs for each experiment and collect statistics on the GP symbolic regression performance. We evaluate model predictive accuracy. Predictive accuracy compares the predictions from the best evolved model to the hedonic response values (ground truth) in the sensory map (without distortion). Our definition of accuracy uses a tolerance of 1, meaning that if the predicted liking score of a testing sample is in the range of the actual ground-truth liking score plus or minus 1, the prediction is considered as being correct. Accuracy is defined as a ratio of quantity of correctly predicted samples to the total quantity of testing samples. We use 3000 samples that are generated using a uniform random distribution for testing. We run GP symbolic regression with a population of 500, for 39 generations after random population initialization. During each run we compute the predictive accuracy (using the testing dataset) of the best-so-far evolved model every second generation. At the experiment's end we compute the mean and variance of accuracy.

Figure 1-7 shows how well the 4 sensory maps can be predicted without any distortion. GP symbolic regression is able to very accurately predict map 1 which is an additive map of 7 1D linear response functions after approximately 20 generations with just 40 samples. For the map 2 which is a 2nd order map composed of 2-piece linear 1D responses, the more samples, the higher the prediction accuracy. That is, for 40 and 100 samples, prediction accuracy is above 0.9 and with 1000 samples accuracy hits 1.0. The 1000 sample experiment's accuracy is achieved by the end of 39 generations. For the two smaller sample sizes, it is possible that the accuracy could improve more if GP symbolic regression had been run for additional generations.

Maps 3 and 4, '*diverse and more complex*' and *naive* respectively, cannot be accurately modeled even with 1000 training samples. Their accuracy varies over sample size from 40 to 1000 between (0.4...0.55) and (0.3...0.4) respectively. However, more samples definitely improve their prediction performance. The reason for lower performance on these maps is the complexity of the underlying sensory map. E.g., the third map, whose features are presented in Table 1-2 is more complex and diverse due to the presence of *cos*, *sin*, *square* root and *absolute* functions. It doesn't appear that more evolutionary generations would substantially improve accuracy way for either map.

Consider the same 4 maps, but now with the strong distortion and fatigue we choose to model as arising from the protocol and human behavior. Modeling results are shown in Figure 1-8. There is a large drop in prediction accuracy compared to the distortion-free, fatigue-free experiment set. For map 1 and map 2, prediction accuracy, with distortion and fatigue present, drops from perfect (samples=1000) or near perfect (samples=100 or 40) to 0.2, 0.3, and 0.3 respectively. Map 3's modeled accuracy is similar to those of maps 1 and

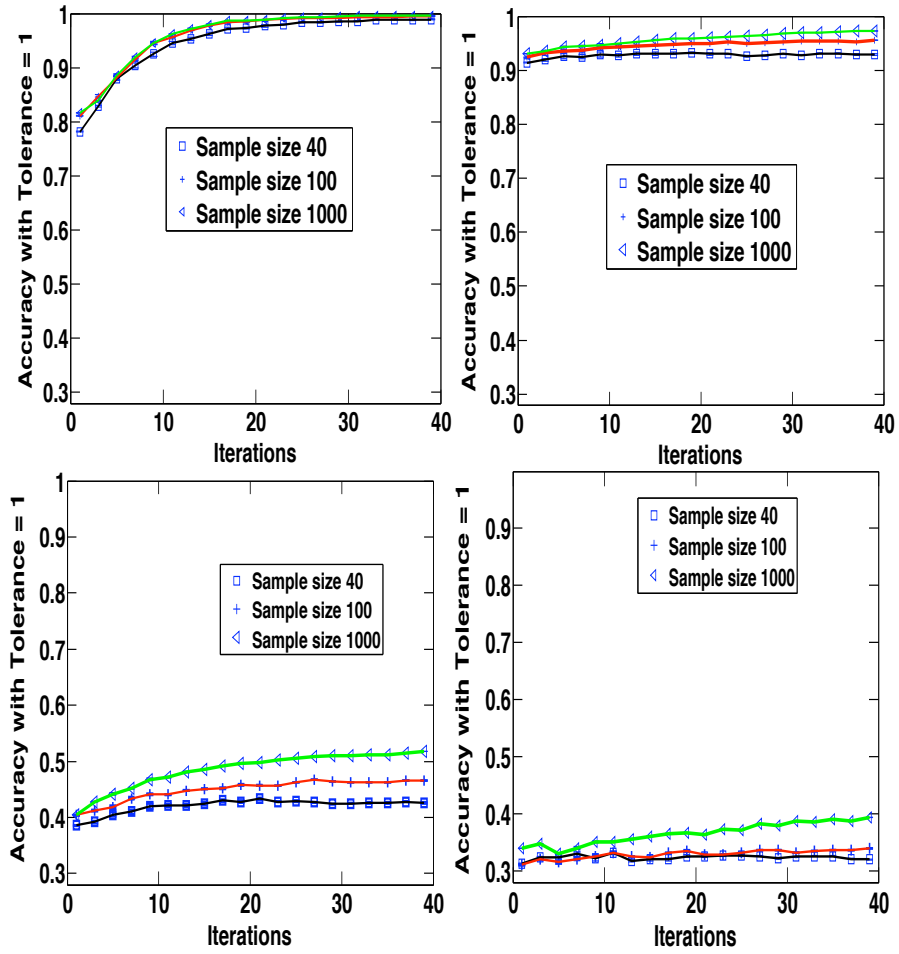


Figure 1-7. GP symbolic regression modeling performance without distortion or fatigue and three training sample sizes. Clockwise from top left: map 1, map 2, map 3, map 4.

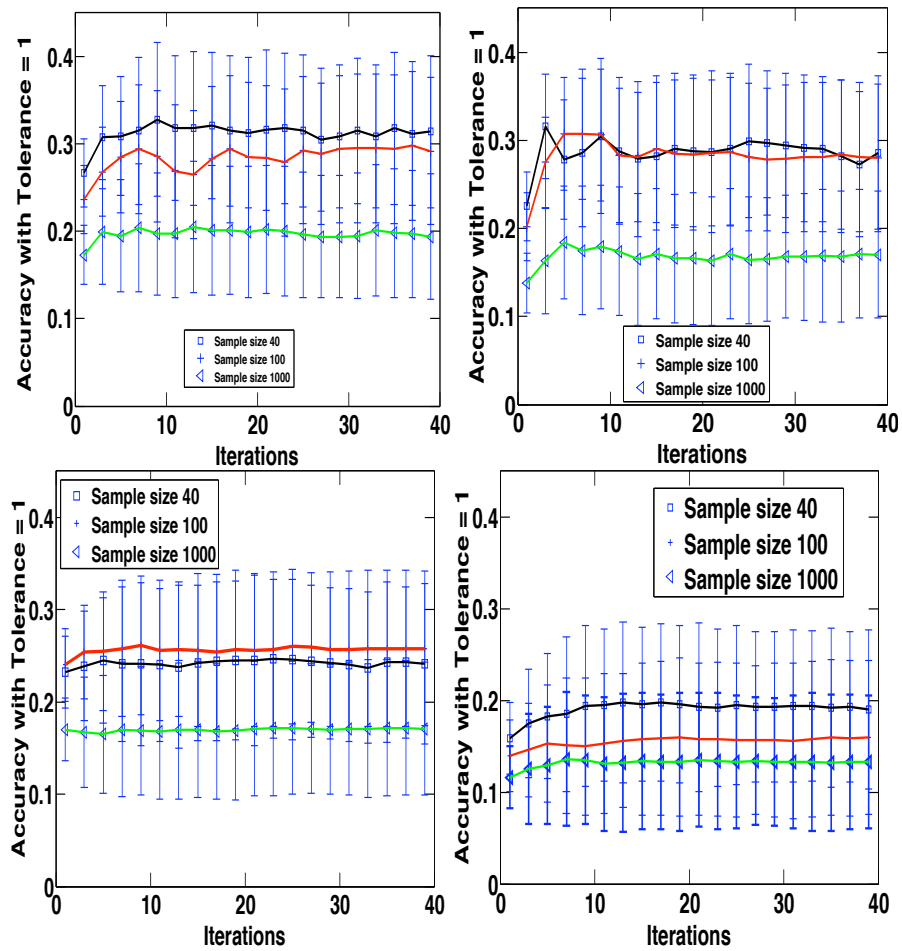


Figure 1-8. GP symbolic regression modeling performance with distortion and fatigue and three training sample sizes. Clockwise from top left: map 1, map 2, map 3, map 4.

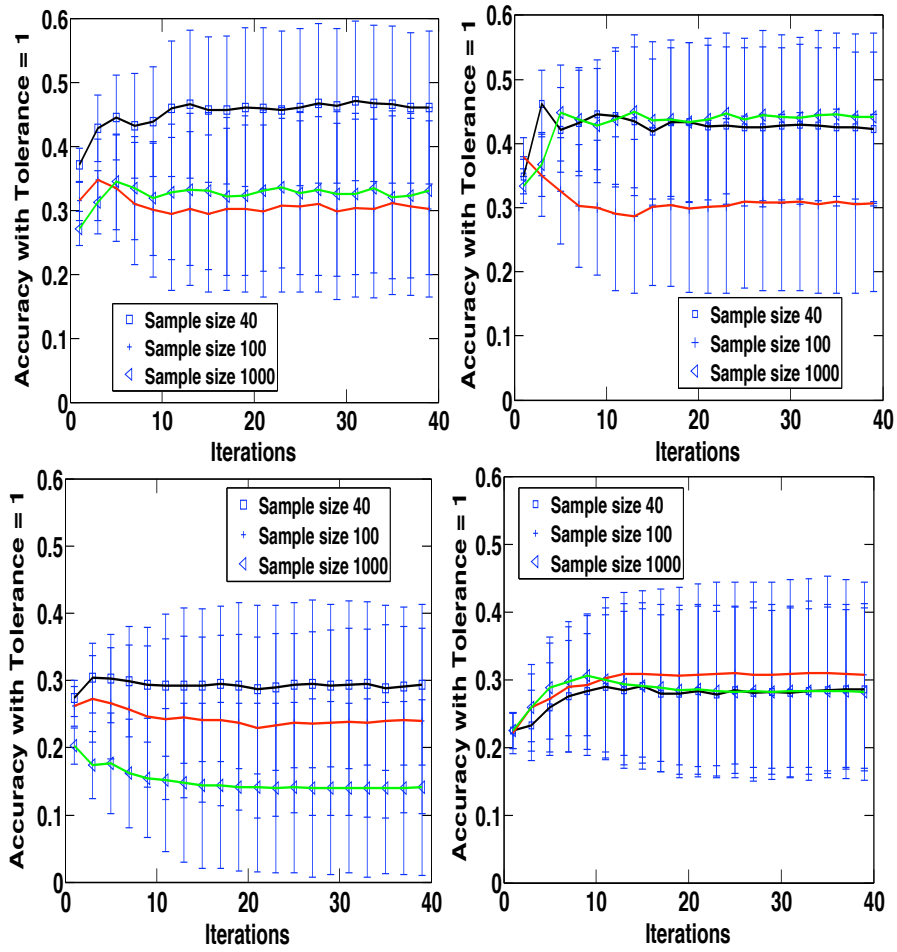


Figure 1-9. GP symbolic regression modeling performance with distortion but not fatigue and three training sample sizes. Clockwise from top left: map 1, map 2, map 3, map 4.

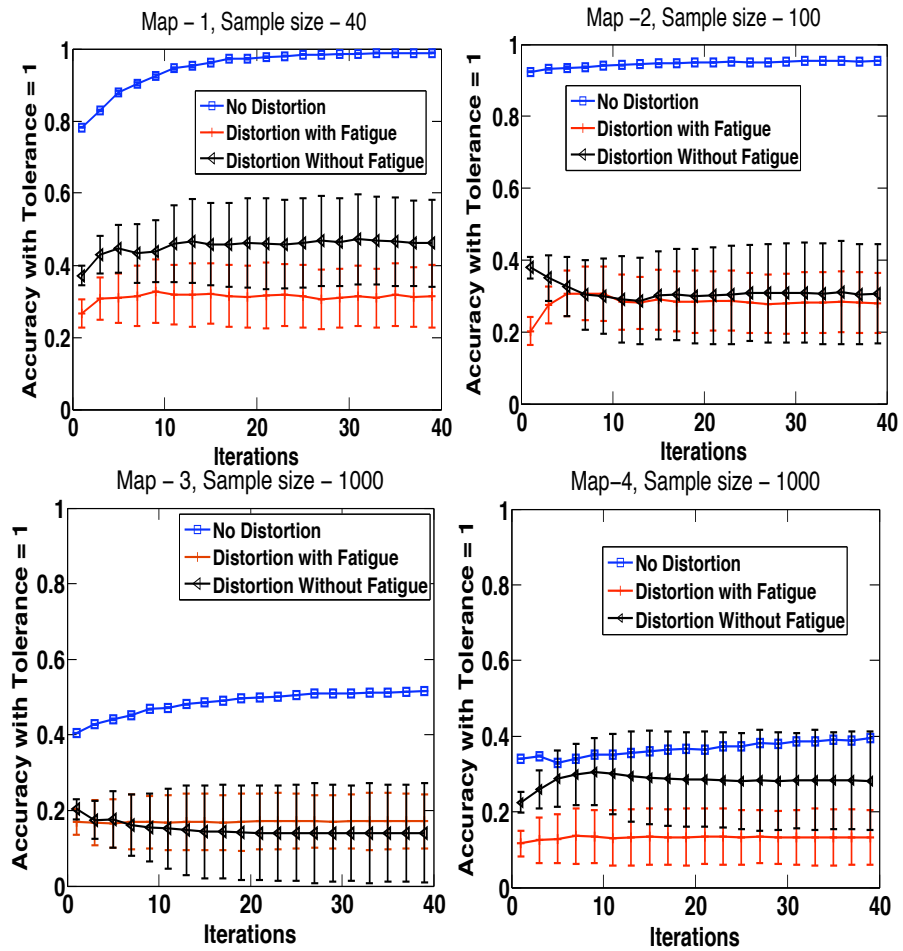


Figure 1-10. GP symbolic regression modeling performance comparing cases of distortion-noise-free, distortion only, distortion+noise.

2, but, compared to its distortion-free, fatigue free equivalent, this represents a drop from 0.4, 0.4 and 0.5 for sample sizes 40,100,1000 respectively to 2.5, 2.4 and 1.9.

This implies that protocol and human “noise” are the key contributors to inability to recover ground truth. Together, distortion and fatigue sufficiently reduce the information content of later samples in the query chain to make them not worth soliciting (under the distortion and fatigue levels modeled with the parameters of Table 1-3). When the samples increase from 40 to 100, the information content of the extra 60 samples is relatively neutral or slightly detrimental toward recovering the ground truth of the map. When the samples increase from 100 to 1000, there is always a decrease in model accuracy. The results call for further work to understand the sensitivity of sample size and map properties to fatigue and distortion levels.

Figure 1-9 shows the results of experiments that model only distortion. Let ss denote sample size. In the case of map 1, for $ss = 100$, predictive accuracy improves by about 0.15 (to 0.48 from 0.32) without fatigue. For $ss = 100$, the predictive accuracy of 0.3 does not change. For $ss = 1000$, predictive accuracy improves to 0.1 (from 0.2 to 0.3). Map 2 results are essentially similar to those of map 1. With map 3 there is a clearer distinction between $ss = 40$ and $ss = 100$, and the order based on predictive accuracy changes from 100 being better with fatigue to 40 being better without fatigue. The predictive accuracy ranking according to sample size of map 4 is $\{40, 100, 1000\}$ with distortion and fatigue. When fatigue was not modeled, the ordering changes to $ss = 100$ with highest predictive accuracy and no difference between $ss = 40$ or $ss = 1000$. Figure 1-10 cross-references data from each of Figures 1-7 to 1-9 to show the impact of fatigue and/or distortion on predictive accuracy for one sample size for each map.

5. Summary, Discussion and Future Work

Our aims in designing this specialized benchmark framework for sensory evaluation are:

- 1 to rationally support the description of sensory maps which are hedonic response surfaces as a function of product design inputs
- 2 to rationally express, with mathematics, the distortions which arise when using a specific hedonic sensory evaluation protocol
- 3 to support the assessment of the performance of GP symbolic regression algorithms with respect to how predictive accuracy scales with:
 - sampling quantity for training
 - sensory maps with different degrees of ruggedness

- different sources and levels of distortion which arise when using a specific hedonic sensory evaluation protocol

Our goal is a level of generality which:

- supports the modeling of different sources of human sensory distortion manifested when assessors are queried about hedonic response
- supports modeling how a protocol's query-response component contributes to distorting the ground truth of the sensory map, in addition to human sourced distortions

We have demonstrated the benchmark framework in a baseline evaluation of GP symbolic regression on sensory maps while controlling distortion and the simulation of human sensory fatigue. We have gained insights into predictive accuracy and determined in what cases more samples help in increasing predictive capability. When there is no distortion, additional samples always increase predictive capability. However, when distortion is present, this is not necessarily the case. With an additional factor of fatigue, the performance declines with additional number of samples, as the data becomes more and more noisy with each additional sample.

Experts with commercial interest in sensory evaluation have noted that the benchmark framework can also assist with the specialized co-design of algorithms and protocols for a given product design space. When these domain experts have evidence of assessor characteristics and feedback on how protocols are interpreted by assessors, or when they have deeper knowledge of the sensory map for the product design space, this knowledge can be expressed via the benchmark framework. This allows them to explore the limits of what they can learn if they use a protocol, with some limited number of samples, on a specific quality of assessor, given the product space and an algorithm for analyzing the data after the experiment. Interactively, they can gain insight into the impact of how assessor errors (which partially relate to the number of samples) makes some surveys impractical. They can also vary sensory maps under the same set of protocol and assessor conditions to see how a product design space may be suitable for a real experiment. Alternatively, for a design space they can describe, they can determine how many samples are needed by the algorithm to derive models that predict accurately.

We plan to extend the algorithm suite we can run on the benchmarks. As well, we will extend the benchmark framework with additional protocols.

Acknowledgments

We would like to thank Dr. Hansruedi Gygax and Dr. Guillaume Blancher of Givaudan Flavors Corporation plus our reviewers.

References

- Becker, Ying, Fei, Peng, and Lester, Anna M. (2006). Stock selection : An innovative application of genetic programming methodology. In Riolo, Rick L., Soule, Terence, and Worzel, Bill, editors, *Genetic Programming Theory and Practice IV*, volume 5 of *Genetic and Evolutionary Computation*, chapter 12, pages -. Springer, Ann Arbor.
- Becker, Ying L., Fox, Harold, and Fei, Peng (2007). An empirical study of multi-objective algorithms for stock ranking. In Riolo, Rick L., Soule, Terence, and Worzel, Bill, editors, *Genetic Programming Theory and Practice V*, Genetic and Evolutionary Computation, chapter 14, pages 241–262. Springer, Ann Arbor.
- Becker, Y.L. and O'Reilly, U.M. (2009). Genetic programming for quantitative stock selection. In Xu, L., Goodman, E.D., Chen, G., Whitley, D., and Ding, Y., editors, *GEC '09: Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, pages 9–16, Shanghai, China. ACM.
- Costello, E., McGinty, L., Burland, M., and Smyth, B. (2007). The role of recommendation for flavor innovation and discovery. In *IC-AI*, pages 463–469.
- Keijzer, Maarten (2003). Improving symbolic regression with interval arithmetic and linear scaling. In Ryan, Conor, Soule, Terence, Keijzer, Maarten, Tsang, Edward, Poli, Riccardo, and Costa, Ernesto, editors, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 70–82, Essex. Springer-Verlag.
- Korns, Michael (2011). Accuracy in symbolic regression. Genetic and Evolutionary Computation, chapter 8. Springer, Ann Arbor.
- Kotanchek, M.E., Vladislavleva, E.Y., and Smits, G. (2009). Symbolic regression via GP as a discovery engine: Insights on outliers and prototypes. In Riolo, R.L., O'Reilly, U.M., and McConaghy, T., editors, *Genetic Programming Theory and Practice VII*, Genetic and Evolutionary Computation, chapter 4, pages 55–72. Springer, Ann Arbor.
- Leibowitz, HW and Post, RB (1982). Capabilities and limitations of the human being as a sensor. *Selected sensory methods: problems and approaches to measuring hedonics*, page 2.
- Moskowitz, HR (1982). Utilitarian benefits of magnitude estimation scaling for testing product acceptability. In *Selected sensory methods: problems and approaches to measuring hedonics: a symposium*, page 11. ASTM International.
- Nikolaev, Nikolay and Iba, Hitoshi (2001). Genetic programming using chebyshev polynomials. In Spector, Lee, Goodman, Erik D., Wu, Annie, Langdon, W. B., Voigt, Hans-Michael, Gen, Mitsuo, Sen, Sandip, Dorigo, Marco,

- Pezeshk, Shahram, Garzon, Max H., and Burke, Edmund, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 89–96, San Francisco, California, USA. Morgan Kaufmann.
- Riskey, DR (1982). Effects of context and interstimulus procedures in judgments of saltiness and pleasantness. In *Selected sensory methods: problems and approaches to measuring hedonics: a symposium*, page 71. ASTM International.
- Schmidt, Michael D. and Lipson, Hod (2006). Actively probing and modeling users in interactive coevolution. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, GECCO '06, pages 385–386, New York, NY, USA. ACM.
- Silva, S. (2011). <http://gplab.sourceforge.net/index.html>. GPLab v.3 April 2007.
- Silva, S. and Almeida, J. (2003). GPLAB—a genetic programming toolbox for MATLAB. In *Proceedings of the Nordic MATLAB Conference*, pages 273–278.
- Veeramachaneni, Kalyan, Vladislavleva, Katya, Burland, Matt, Parcon, Jason, and O'Reilly, Una-May (2010). Evolutionary optimization of flavors. In et al, Juergen Branke, editor, *GECCO '10: Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 1291–1298, Portland, Oregon, USA. ACM.
- Vladislavleva, Ekaterina J., Smits, Guido F., and den Hertog, Dick (2009). Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *IEEE Transactions on Evolutionary Computation*, 13(2):333–349.
- Vladislavleva, Katya, Veeramachaneni, Kalyan, Burland, Matt, Parcon, Jason, and O'Reilly, Una-May (2010a). Knowledge mining with genetic programming methods for variable selection in flavor design. In et al, Juergen Branke, editor, *GECCO '10: Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 941–948, Portland, Oregon, USA. ACM.
- Vladislavleva, Katya, Veeramachaneni, Kalyan, and O'Reilly, Una-May (2010b). Learning a lot from only a little: Genetic programming for panel segmentation on sparse sensory evaluation data. In Esparcia-Alcazar, Anna Isabel, Ekart, Aniko, Silva, Sara, Dignum, Stephen, and Uyar, A. Sima, editors, *Proceedings of the 13th European Conference on Genetic Programming, EuroGP 2010*, volume 6021 of LNCS, pages 244–255, Istanbul. Springer.