We thank all the reviewers for excellent questions and many relevant remarks.

[[Reviewer 1]] ■ **On extensions to other domains (e.g. CNN):** Thank you for this remark. While our method can in principle be used for various datasets and black-box architectures, we are focusing on tabular datasets as it is the case in [1]. Media-specific tasks, such as image classification and natural language processing, are beyond the scope of this paper. One of the reason for this is that our method produces interpretations directly in terms of the input features. In the case of CNN, for instance, we believe that meaningful interpretations should involve hidden units explicitly, see [2] for instance.

[[Reviewer 2]] ■ **Usage of the word faithful:** Thank you for pointing this out, we agree that faithful is not best. As you point out, our models are indeed symbolic models; for this reason, we will rename our algorithm Symbolic Pursuit in the final manuscript. ■ **Hyperparameters in Table 2:** We did not include these details in Table 2 for the sake of conciseness. We simply wanted to demonstrate that the interpretable models we obtained are accurate (have high R^2 scores) and concise (have a small number of terms). We shall report the details in the Supplementary Material of the final manuscript. ■ **Interpretation of Figure 3:** The explanation provided for Figure 3 from the Supplementary Material in the review is correct; we will make this more explicit in the final manuscript. ■ **Meaning of global interpretations:** In the context of this work, we mean that our models are good global approximations of the black-boxes (see the high R^2 scores in Table 2). This is not the case for local models such as LIME.

[[Reviewer 3]] ■ **Quantitative comparison with other methods:** Many thanks for this suggestion; we will include such experiments in the final manuscript. In the mean time, we ran a toy experiment: We drew 100 inputs $(x_1, x_2) \sim U\left([0,1]^2\right)$ and set, as a pseudo black box function, $f(x_1, x_2) = 0.3 \cdot x_1 + 0.6 \cdot x_2$. In this simple setup, the feature x_2 is everywhere twice as important than the feature x_1; this corresponds to having a normalized feature importance vector everywhere equal to $(0.44, 0.88)$. We used our method and two others to produce feature importance vectors at four test points; the results are reported in Table 1. We see that our method produces feature importance vectors that are very similar and close to the actual vectors; the other methods produce feature importance vectors that are very different from the actual vectors and very different at the various test points. We will add more sophisticated experiments in the final manuscript. ■ **Advantages over Symbolic Metamodels [1]:** There are two main advantages. The first one, as discussed in lines 221-226, is that our method produces expressions with many fewer terms than those produced by Symbolic Metamodels. For instance, on the experiments that are synthesized in Table 2, a Symbolic Metamodel would need 78 terms for Wine, 28 for Yacht, 105 for Boston, 36 for Energy, 45 for Concrete; we think that interpreting an expression with so many terms would be very difficult. By contrast (see Table 2), for these datasets our method almost always produces models with only 1, 2 or 3 terms. The second advantage is that we address hyperparameter selection (see Section 3). This allows more flexibility than Symbolic Metamodels in the function(s) that we are able to identify (see especially lines 226-232). The comparison for the breast cancer dataset is unfortunately impossible since the dataset is not public. ■ **Precision on the notion of interpretability:** We believe that our method helps interpretability by outputting a model whose mathematical expression can realistically be written and manipulated by a human subject (which is obviously not the case for the black-box). An illustration of this, as mentioned in the review, is the possibility of building a Taylor approximation to extract linear effects (such as feature importance) and non-linear effects (such as feature interactions). We also suggest that various linear combinations of the features appearing in the model at Equ. (10) can be interpreted as new variables in which the model takes a simple form. This offers additional conciseness (on top of the parsimonious size of the model), which facilitates the analysis by a human subject. ■ **On the interpretability of polynomial splines:** Thank you for this remark, we shall clarify in the final manuscript. We would like to avoid restricting to polynomials for the reasons explained in line 28-46, but it is more a matter of representation power rather than interpretation. ■ **On the definition of R^2:** By R^2 we mean the usual coefficient of determination; we will give a precise definition in the revision. Thank you for pointing this out.

Table 1: Normalized feature importance vector.

Test point	Our Method	LIME	SHAP
1	$(0.44, 0.92)$	$(0.84, 0.52)$	$(0.89, 0.45)$
2	$(0.42, 0.90)$	$(0.20, 0.97)$	$(0.28, 0.95)$
3	$(0.44, 0.92)$	$(0.85, 0.51)$	$(0.49, 0.86)$
4	$(0.44, 0.92)$	$(0.55, 0.83)$	$(0.70, 0.70)$

[[Reviewer 4]] ■ **Notational inconsistencies:** Many thanks for pointing out the notational inconsistencies; we will correct them in the final manuscript.

[1] Ahmed M. Alaa and Mihaela van der Schaar. Demystifying black-box models with symbolic metamodels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11304–11314. Curran Associates, Inc., 2019.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition*