We would like to thank all reviewers for their thoughtful comments that have helped improve the paper. We have implemented most of the suggestions and we answer questions below.

Our main additions are as follows. (i) Doubling the number of seeds for experiments (to 10 seeds), and including statistical tests that show the significance of the results. (ii) Providing additional ablations (pure model-free, pure model-based) (as shown in Table B) (iii) Adding a new baseline RND (Burda et al., 2018) in Table A.

Reward	Avg (μ)	StdErr	p-value
Random	3.063	0.379	0.000040
Pred. error	2.372	0.359	0.0016
RND	2.203	0.391	0.0081
BDQN	1.859	0.264	0.0064
Hash-count	1.304	0.209	0.20
Novelty	**1.000**	0.0899	-

Table A: Combined results table over both acrobot and multi-step maze environments over 10 random seeds, normalized to the mean number of steps in our Novelty approach for each environment. We provide p-values indicative of the null hypothesis $H_0 :\ \Delta\mu = \mu_1 - \mu_2 = 0$, calculated using Welch's t-test, all as per Colas et al. (2019). In this case, we do a pair-wise comparison between the central tendencies of our algorithm (Novelty) and our baselines. Full details including these statistical tests will be included in the final paper.

(All) Ablation study and contribution of the different elements. The additional ablations in Table B allows assessing the impact of some components of Eq 6. It also shows that using a combination of using a Q-value in combination with a model in the context of exploration is one key contribution as compared to ICM (Pathak et al., 2017) and RND (Burda et al., 2018). Also important, as compared to these works, we provided visualisations of the abstract representations obtained.

(R1, R2) Learning the discount factor (γ) and discussion on the different losses used. In our experiments, learning the discount factor is only used for ensuring correctness in planning for terminal states (where $\gamma = 0$). The loss associated with learning gamma is unlikely to have a significant impact as it decreases rapidly due to the simplicity of learning to map to a constant γ everywhere except for terminal transitions where it has to map to 0. We'll clarify that in the paper. We'll also mention that the loss associated with the reward function has been used previously as an auxiliary task that has the potential to *improve* learning, even in a pure model-free setting (where it is not used during planning) (Jaderberg et al., 2016). In summary, some losses in Equation 6 are easy to optimize and are likely to help learning and/or do not require particular tuning in the learning process. We will add these points to the discussion.

(R2, R4) Additional references. We will add in our related work section the relevant works of Still and Precup, Barto's work "Novelty of Surprise?" and "Never Give Up" by Badia et al. "Novelty of Surprise?" explains well the motivation of approaches based on novelty, and how they are related to "surprise-based" approaches for exploration. Still and Precup's work is related to our motivation for estimating novelty from an abstract representation that has to contain the minimal meaningful information for representing the environment. Finally, "Never Give Up" further supports our use of kNN as a novelty measure.

Ablation	Avg (μ)	StdErr	p-value
MF	944.2	136.477	0.053
MB	828.3	106.914	0.12
Full	591.2	81.86	-

Table B: A further ablation study on the multi-step maze environment. The MF (model-free) ablation does not employ any forward intrinsic reward planning ($d = 0$), while the MB (model-based) ablation only uses forward intrinsic reward planning without using or learning Q-values.

(R1, R2, R4) Choice of the abstract representation. The choice of the abstract representation and possibly the encoder architecture can be important elements. In practice, we observed that as long as the model allows sufficient capacity (e.g. at least 2 hidden neurons in the open grid world), we did not observe any consistent difference in performance. We will clarify this in the discussion and appendix of the paper.

(R1, R3) Clarifications & pseudo-code Figure 1 is obtained from environments with high dimensional observations and the visualizations are provided in Appendix J, which we will explicitly mention in the paper. We have also taken into account the other suggestions for improved clarity, such as bringing the algorithm to the main part of the paper.

(R1, R3) Deterministic environments The method as implemented is indeed currently limited to deterministic environments (as mentioned in line 43 of Section 2). That limitation could be relaxed with a generative internal model and by taking into account an expected distance for our novelty metric to handle stochastic domains.

(R1, R4) Montezuma's revenge. A full application to Montezuma's revenge is complicated due to the fact that many gradient descent steps are performed at each environment step in order to learn the model with sufficient accuracy. Preliminary results show that our approach provides a meaningful abstract representation that would allow efficient exploration in such complex games (see Appendix H), though at the cost of expensive computations.

Additional changes. We will also gladly incorporate changes about cleaning references, formalism and wording.