Reviewer 2: Regarding comments on parameter $\gamma > 0$: Firstly, the condition $\gamma \in (0,\ 0.5)$ is a sufficient condition for the bound in Theorem 1 to hold, which follows from the proof (see eq. (15)). Secondly, if γ is increased, the bound tightens but on the other hand the probabilistic guarantee weakens. However, for a given data set one can readily search for the γ in $(0,\ 0.5)$ which yields the most informative (tightest) interval $\Lambda_\alpha(\boldsymbol{x})$. In the revised paper, we will elaborate on these points and amend the wording of 'tuning-free'.

Regarding comments on Fig. 1: The spatial domain in this example is in fact discretized into $R = 50$ regions whose resulting small size only make the curves *appear* continuous. The revised paper clarifies this.

Regarding comments related to the risk $\mathcal{R}(\boldsymbol{\theta})$: First, note that in eq. (5) the risk can also be writtten as $\mathcal{R}(\boldsymbol{\theta}) = -n^{-1}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{r}}[\ln p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{r})] + K$, where the constant $K = n^{-1}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{r}}[\ln p(\boldsymbol{y}|\boldsymbol{r})]$ is there only to ensure nonnegativity $\mathcal{R}(\boldsymbol{\theta}) \geq 0$ and the natural property $\mathcal{R}(\boldsymbol{\theta}) = 0 \Leftrightarrow p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{r}) \equiv p(\boldsymbol{y}|\boldsymbol{r})$. Thus including K in the definition of risk is natural and it requires no knowledge of $p(\boldsymbol{y}|\boldsymbol{r})$ in order to define the unknown $\boldsymbol{\theta}^\star$ in eq. (6) as it does not affect the optimization problem. Secondly, since $\ln p(\boldsymbol{y}|\boldsymbol{r}) = \sum_{i=1}^n \ln p(y_i|r_i)$ and $\ln p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{r}) = \sum_{i=1}^n \ln p_{\boldsymbol{\theta}}(y_i|r_i)$ are a sum of n terms and division by n in eq. (5) is merely a natural normalization to make derivation and final expressions neater. The risk $\mathcal{R}(\boldsymbol{\theta})$ is therefore the Kullback-Leibler divergence *per sample*. We will clarify this is the revised manuscript.

Regarding comments related to empirical coverage: Let $\Lambda_\alpha(\boldsymbol{x})$ and $\Lambda'_\alpha(\boldsymbol{x})$ denote the prediction intervals for the proposed regularized approach (in eq. (7)) and the standard likelihood approach, respectively. In Section 4.1, we conduct a small but illustrative simulation study in which both intervals are computed using the conformal prediction framework (Algorithm 1). Both intervals have empirical coverages that are valid, that is, exceed $1 - \alpha = 80\%$. The difference between the intervals lies in their sizes as shown in Fig. 2c, where $\Lambda_\alpha(\boldsymbol{x})$ is found to be much smaller than $\Lambda'_\alpha(\boldsymbol{x})$ for the same coverage. In the interest of more details, we will include additional simulations in the supplementary material that evaluate the empirical coverages of the two intervals under different scenarios.

Regarding comments related to notational issues: r has now been added in union in line 44. The intensity function $\lambda(\boldsymbol{x})$ is defined as the number of events *per unit area*. This has now been clarified on line 44. Capital Y is the maximum number of events in a region and only plays the role of a practical upper limit on the conformal prediction interval so as to terminate the for-loop in Algorithm 1. Given that it is set reasonably high for the problem at hand, it does not affect the results and in practice we simply set it to an integer multiple of the maximum number of counts observed in any region. Details regarding $\boldsymbol{\phi}(r)$ have now been added on lines $89 - 90$. In Algorithm 2, **'repeat'** has now been replaced by **'while'** and the convergence criteria specified. Other notational issues have also been addressed and will appear in the revised paper.

Reviewer 3: Regarding comments on utility of the method in extrapolation scenarios: Unless a method is based on some plausible physical model of the point process, one cannot expect it to yield valid *and* informative intervals far from the observed data. Indeed the proposed method yields valid intervals but they become increasingly uninformative for regions further from the observed data. By contrast, the standard kernel density-based methods (e.g. [6]) may wrongly infer the absence of events outside data as zero intensity in this case, whereas the credible intervals obtained from Bayesian method become less informative but do not exhibit any statistical validity (see Fig. 2b). Hence proposed method yields intervals that reflect uncertainity due to missing data in a statistically appropriate way (see Fig. 2a and 2b).

Regarding comments on the balance of the terms in the fitting criterion in eq. (7): Note that $-n^{-1}\ln p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{r}) = -n^{-1}\sum_{i=1}^n \ln p_{\boldsymbol{\theta}}(y_i|r_i)$ is a sum of n terms therefore it will not decay as the number of datapoints n increase. On the other hand, the second regularization term will decay with n, as desired. Thus the fitting criterion in eq. (7) is appropriately balanced.

Regarding comments on comparison with other methods: In the literature, the Cox process is the default model and state-of-the-art methods are based using log-Gaussian link function (LGCP in Section 4.1). Given the intractibility of this model, most work is centered on computational approximations ([9], [13], [20], [2]). In this work, we have compared our method with the popular integrated Laplace approximation (INLA [17], [11]). Less tractable Monte Carlo approximations, e.g. [9], [2] were not implemented in this study.

Regarding comments on the computational complexity of the proposed method: Algorithm 2 solves a series of weighted lasso problems and can therefore be solved in a runtime that scales as $\mathcal{O}(nR^2)$ where n is the number of datapoints and R is the dimension of $\boldsymbol{\theta}$. This has been clarified in the revised paper.

Reviewer 4: Regarding comments related to effect of discretization: the manner in which space is discretized will affect the size of intervals but not their statistical validity. That is, irrespective of how space is discretized, the out-of-sample guarantees eqs. (3) and (8) will still hold, whether the resulting model is misspecified or not. We will elaborate on the effect of spatial discretization in the revised paper.