Bayesian averaging of computer models with domain discrepancies: a nuclear physics perspective

Vojtech Kejzlar,¹ Léo Neufcourt,¹,² Taps Maiti,¹ and Frederi Viens¹

¹Department of Statistics and Probability, Michigan State University
²Facility for Rare Isotope Beams, Michigan State University

Abstract

This article studies Bayesian model averaging (BMA) in the context of several competing computer models in nuclear physics. We quantify model uncertainty in terms of posterior prediction errors, including an explicit formula for their posterior variance. We extend BMA when the competing models are defined on non-identical study regions. Any model’s local forecasting difficulty is offset by predictions obtained from the average model, extending individual models to the full domain. We illustrate our methodology via simulations and applications to forecasting nuclear masses. We find significant improvement in both the BMA prediction error and the quality of the corresponding uncertainty quantification.

Keywords—Bayesian model calibration; Computer models; Gaussian process; Model mixing; Model uncertainty; Nuclear masses; Uncertainty quantification.

1 Introduction

Model averaging arises in situations with several competing models available to solve the same or similar problems, and no single model can be selected at a desired level of certainty. This is a common scenario in many scientific fields concerned with modeling complex systems. One of the historically dominant solutions is Bayesian Model Averaging (BMA) (Kass and Raftery, 1995; Hoeting et al., 1999), which is a natural Bayesian framework when facing a finite or countable number of alternative models. The seminal review work by Geweke (1999) introduced BMA in the field of econometrics and later in other fields such as in the political and social sciences; BMA has also been applied to the medical sciences (Balasubramanian et al., 2014; Schorning et al., 2016), ecology and evolution (Silvestro et al., 2014; Hooten and Hobbs, 2015), genetics (Wei et al., 2011; Wen, 2015), and machine learning (Clyde et al., 2011; Hernández et al., 2018).

Contemporary developments in computing capabilities have meanwhile brought new perspectives to model averaging. The rapid surge of models implemented on a computer, which we shall refer to as computer models, opened several challenges to BMA. Nuclear structure provides stimulating examples (Erler et al., 2012; Piekarewicz et al., 2012; Afanasjev et al., 2015; Neufcourt et al., 2019; Olsen and Nazarewicz, 2019) which illustrate the canonical difficulties one faces with BMA of computer models that might require hundreds to thousands of core hours for a single evaluation.
Consider a general situation where experimental measurements \((x_i, y_i)_{i=1}^n\) of a physical process \(x \mapsto y(x)\) are used to predict its values \(y^* = y(x^*)\) at a new input value \(x^*\). In the nuclear physics context, \(x = (Z, N)\) marks the isotopes, defined by the proton number \(Z\) and neutron number \(N\). The output \(y(x)\) is the corresponding value of an observable such as a nuclear mass or radius, and \(y(x^*)\) are typically extrapolations of these observables. The Bayesian inference approach is to consider a quantity of interest \(\Delta\) and derive its posterior probability given the data \(y = (y_1, \ldots, y_n)\) and model \(\mathcal{M}\). This is simply obtained from Bayes' formula as long as one can express the likelihood \(p(y|\Delta, \mathcal{M})\) and provide a prior distribution \(\pi(\Delta|\mathcal{M})\). Establishing the likelihood \(p(y|\Delta, \mathcal{M})\) is done through a model \(\mathcal{M}\). The quantity \(\Delta := y^* = y(x^*)\) is typically of universal interest, but \(\Delta\) can represent any latent quantity at any level of the model. Various well-known methods such as the Metropolis-Hastings algorithm (Chib and Greenberg 1995) or more advanced Monte Carlo Methods such as Hamiltonian Monte-Carlo, or No U-Turn Sampler (Homan and Gelman 2014) can be applied to obtain samples from the posterior distribution \(p(\Delta|y, \mathcal{M})\).

Now, let us consider a scenario with \(K\) competing models \(\mathcal{M}_1, \ldots, \mathcal{M}_K\), and let us assume that there exists a true model \(\mathcal{M}\). The Bayes formula can be applied to express the posterior probability of a model \(\mathcal{M}_k\), given observations \(y\), as

\[
p(\mathcal{M}_k|y) = \frac{p(y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{\ell=1}^K p(y|\mathcal{M}_\ell)\pi(\mathcal{M}_\ell)} \propto_k p(y|\mathcal{M}_k)\pi(\mathcal{M}_k),
\]

where the symbol \(\propto_k\) indicates proportionality with respect to \(y\) for fixed \(k\). Here, the difficulty lies in evaluation of the evidence integral \(p(y|\mathcal{M}_k)\) for each model under the prior probability \(\pi(\mathcal{M}_k)\) that \(\mathcal{M}_k\) is the true model \(\mathcal{M}\), after which the BMA posterior distribution for any quantity of interest \(\Delta\) can be derived as

\[
p(\Delta|y) = \sum_{k=1}^K p(\Delta|y, \mathcal{M}_k)p(\mathcal{M}_k|y).
\]

This formula shows that the actual posterior probability of an observable \(\Delta\) is the average of \(\Delta\)'s the posterior distributions given each model, weighted by the model posterior probabilities. In other words \((\frac{3}{2})\) is simply a mixture of \(K\) distributions which makes sampling from the BMA posterior density immediate once we obtain posterior samples under each model.

In particular, the posterior mean and variance of \(\Delta\) are given by

\[
\mathbb{E}[\Delta|y] = \sum_{k=1}^K \mathbb{E}[\Delta|y, \mathcal{M}_k]p(\mathcal{M}_k|y),
\]

and the standard result for the conditional variance of a mixture yields the posterior variance of \(\Delta\) given \(y\)

\[
\text{Var}[\Delta|y] = \sum_{k=1}^K p(\mathcal{M}_k|y)\text{Var}[\Delta|y, \mathcal{M}_k] + \text{Var}[\mathbb{E}[\Delta|y, \mathcal{M}]|y].
\]

Note that the term \(\text{Var}[\mathbb{E}(\Delta|y, \mathcal{M})|y]\) is the variance of a function of the discrete random variable \(M\), which accounts for the model uncertainty. This model uncertainty is not accounted for by individual models. Its inclusion thus allows for a more honest uncertainty quantification.

In this work, we first describe the methodology of Bayesian analysis of computer models with generalization to account for experimental data coming from several sources (several observables). Then, we perform a systematic analysis of the prediction errors, establishing that BMA is the
optimal linear combination (projection) in the $L^2$ sense under the posterior probability distribution, among all possible mixtures of models. Motivated by recurrent scenarios in nuclear structure, we subsequently extend BMA to situations when different models constrain different subsets of the data. Lastly, we present an application of the BMA methodology to two scenarios, a pedagogical "toy" example based on synthetic data and an illustration of the methodology in a setting with real nuclear-physics data. We frequently use nuclear physics terminology, but we hope the paper will be transparent to scientists from other disciplines.

2 Review of Bayesian Analysis of Computer Models

2.1 Calibration

Over the past two decades, there has been a considerable amount of research dedicated to Bayesian calibration of computer models starting with the seminal work of Kennedy and O’Hagan (2001) with extensions and applications provided by Higdon et al. (2005, 2008, 2015). We briefly discuss the standard Bayesian framework and refer to Kennedy and O’Hagan (2001) for full treatment.

Let us consider a single computer model $f(x, \theta)$ relying on a parameter vector $\theta$ of dimension $d$ and an input point $x$ in a finite dimensional space (for simplicity). Model calibration corresponds to determining the unknown and hypothetical *true* value $\theta^*$ of the parameter $\theta$, at which the physical process would satisfy $y(x) = f(x, \theta^*) + \delta(x)$; $\delta(x)$ is the systematic discrepancy of the model whose form is generally unknown. The term "calibration" is broader than the term "estimation", which can imply the use of some well-established statistical methodology. Herein, our notion of calibration is a specific Bayesian estimation methodology, which include a full evaluation of uncertainty for every parameter. Thus we can write the complete statistical model as

$$y_i = f(x_i, \theta) + \delta(x_i) + \sigma(x_i)\epsilon_i, \quad (5)$$

at each of the $n$ observations $y_1, \ldots, y_n$. For brevity, we use $y_i$ for $y(x_i)$. The term $\sigma(x) > 0$ is a scale function to be parametrized and inferred, and $\epsilon_i$ are independent standard random variables representing measurement errors, which we assume to be Gaussian.

One may expect the function $f(x, \theta)$ representing the output of the computer model to be deterministic, given a model and its parameters. This holds in the case of inexpensive computer models, where the evaluation of $f(x, \theta)$ takes a reasonably "constant" time. For computationally expensive models the evaluations of $f(x, \theta)$ cannot be reasonably performed at calibration runtime, and needs to be done beforehand, typically on a grid. Hence it has become a common practice to emulate the computer model by a Gaussian process $\mathcal{GP}(m(x, \theta), k((x, \theta), (x', \theta'))) \| m$ and covariance function $k$. The data in this situation also include a number $N$ of runs of the computer model at pre-determined points $\{(\bar{x}_1, \bar{\theta}_1), \ldots, (\bar{x}_N, \bar{\theta}_N)\}$. Refined approaches typically decompose $m$ on a dense family of basis functions (wavelets, Fourier, polynomials) across the domain of $x$ and $\theta$. As for the covariance function, the most popular choices are the quadratic exponential kernel, the Matern kernels, and fractional Gaussian noise (Rasmussen and Williams 2006).

The discrepancy function $\delta(x)$ represents the systematic error between the computer model and the physical process. While it is intrinsically deterministic, a Gaussian process model is typically imposed for inference. However, the inference becomes tricky when there is only a single observation $y$ for every $x$. This is often the reality for nuclear physics datasets, and one cannot hope to distinguish $\delta(x)$ from the noise $\sigma(x)\epsilon_i$. Indeed one could take $\delta(x_i) := y_i - f(x_i, \theta)$ for any value of $\theta$ as well as $\sigma(x) = 0$. It is easy to see that such confounding between the discrepancy term...
and calibration parameters makes $\theta^*$ non-identifiable in general. Several authors have pointed this out and proposed various methods to mitigate the problem including Brynjarsdóttir and O’Hagan (2014); Plumlee (2017); Tuo and Wu (2015, 2016); Bayarri et al. (2007). Our main goal here, nonetheless, is not correct the identification of $\theta^*$, but a prediction of new observations with honest uncertainties quantification. For prediction, the identifiability of calibration parameter is not particularly a concern.

Rapid calibration via linearization Here, we propose a variation to the calibration method in Kennedy and O’Hagan (2001) based on a linearization of the computer model. We consider situation where $f(x, \theta) = x^T \theta$, i.e. a linear model. This simple framework is valid in practice, as the approximation of a function $f(x, \theta)$, up to a reparametrization, as long as one can a priori localize the parameters, typically around a special value $\theta_0$, such as the maximum likelihood estimator or the solution of a least-squares optimization of the computer model. Indeed, one can replace the computer model by a linear function of $\theta$ in the range of a first order Taylor expansion of $f(x, \theta)$ around $\theta_0$

$$f(x, \theta) \approx f(x, \theta_0) + \nabla_{\theta} f(x, \theta_0)^T \cdot (\theta - \theta_0),$$

as long as one has access to the gradients $\nabla_{\theta} f$. The advantage of this method is that it is intuitive, computationally inexpensive, and can be used either as a mean function of a GP emulator for computationally expensive model, or as a surrogate in the procedure outlined herein for calibration of inexpensive computer models.

Several observables for the same model. The simple notational for of model (5) obscures the situation where $y$ mixes data of different nature, which can have critical impact on both calibration and model averaging. This can represent different types of observables in nuclear physics.

Suppose that the same model is fitted to $q$ types of observables $y_1^{o_1}, \ldots, y_q^{o_q}$. To recognize the dependency on observable types, we can write

$$y_i^{o_l} = f^{o_l}(x_i, \theta) + \delta_i(x_i) + \sigma_l(x_i) \epsilon_i, \quad l = 1, \ldots, q.$$  

and the corresponding likelihood can be computed naturally for each data $y_i$ corresponding to an observable $y^{o_l}$. Allowing different types of observables doesn’t raise significant differences with the standard case. Particular attention is to be brought to the noise scaling parameter $\sigma_l(x)$ which can now vary across observables. In this case the weighting between different observables is done through the relative importance of the scaling functions $\sigma_k(x)$ which can either be fixed or left as parameters. In the context of nuclear physics we refer to the discussion in Afanasjev et al. (2014, Section 3).

2.2 Computing the Evidence Integral

Suppose the $k$-th model is parametrized by $\phi_k \in \mathbb{R}^{d_k}$ for $k = 1, \ldots, K$ that consists of both calibration parameters and hyperparameters (i.e. parameters that go into defining prior distributions for other parameters) with likelihoods $p(y|\mathcal{M}_k, \phi_k)$. The evidence integral [Kass and Raftery, 1995; Fragoso et al., 2018; Aitken and Akman, 2013] of the model $\mathcal{M}_k$ is defined as

$$p(y|\mathcal{M}_k) = \int_{\phi_k} p(y|\phi_k, \mathcal{M}_k) \pi(\phi_k|\mathcal{M}_k) d\phi_k.$$  

(8)
The numerical evaluation of evidence integrals is challenging in practice, and requires approximation. A natural idea is to apply the Monte Carlo method, as most commonly used in the literature, and which we have adopted in our applications:

\[
p(y|\mathcal{M}_k) = \frac{1}{n_{MC}} \sum_i p(y|\phi^{(i)}_k, \mathcal{M}_k).
\]

Here \(\phi^{(i)}_k\) are i.i.d. samples from the prior \(\pi(\phi_k|\mathcal{M}_k)\) for \(i = 1, \ldots, n_{MC}\). While the Monte Carlo approximation yields reasonable results, it requires new evaluations of the likelihood on new samples from the prior \(\pi(\phi_k|\mathcal{M}_k)\) which can be very costly in computing time.

Another frequently used method is the Laplace approximation, which relies on the fact that the integration in (8) can be performed formally in case of a linear regression with Gaussian noise. It corresponds to a second order Taylor approximation of the log-likelihood around its maximum, in which case the likelihood becomes Gaussian. The Laplace method typically gives very good results for likelihoods which are very peaked. We refer the reader to Kass and Raftery (1995) for a survey of different methods used to compute evidence integral.

### 3 BMA and Prediction Error

BMA is only one of various natural ways to deal with several alternative models and to account for model uncertainty, but it does have the property of reducing the Posterior Mean Square Error (PMSE) of prediction of a new observation \(y^*\). In this section, we illustrate this property in a clear and concise way.

Let us, for simplicity of notation, consider two competing models \(\mathcal{M}_1\) and \(\mathcal{M}_2\) - the treatment of multiple models follows from a similar argument, and our verbal descriptions below in this section occasionally refer to the general case without further comment. Denote \(\hat{y}^*_1 := E[y^*|y, \mathcal{M}_1]\) and \(\hat{y}^*_2 := E[y^*|y, \mathcal{M}_2]\) as the posterior means of \(y^*\) under each model, and let \(y^* := E[y^*|y]\). We also define \(p_k := p(\mathcal{M}_k|y)\) for \(k = 1, 2\) for the posterior probability of each model. Thus the BMA posterior mean estimator (3) can be written as \(\hat{y}^* = p_1\hat{y}^*_1 + p_2\hat{y}^*_2\). The PMSE of \(y^*\) is then defined naturally as \(E[(y^* - y^*)^2|y]\) with the following decomposition.

**Lemma.** For every \(\lambda_1, \lambda_2 \geq 0\) satisfying \(\lambda_1 + \lambda_2 = 1\), we have

\[
E[(y^* - \hat{y}^*)^2|y] = E[(y^* - \lambda_1 y^*_1 - \lambda_2 y^*_2)^2|y] - [(\lambda_1 - p_1)y^*_1 - (\lambda_2 - p_2)y^*_2]^2
\]

(9)

This lemma, proved in Appendix A, shows explicitly that the BMA PMSE is smaller than the PMSE associated with any convex combination \(\lambda_1\hat{y}^*_1 + \lambda_2\hat{y}^*_2\) of the each of the two models’ posterior means. It also measures how much smaller it is, and shows that equality holds as soon as the convex coefficients \(\lambda_k\) are equal to the posterior probabilities \(p_k\) of each model, \(k = 1, 2\). Now, applying the Lemma twice, with \((\lambda_1, \lambda_2) = (1, 0)\) and with \((\lambda_1, \lambda_2) = (0, 1)\) we obtain the following dual expressions for the BMA PMSE, involving each individual model’s PMSE, showing how much smaller the former is compared to the two latter:

\[
E[(y^* - \hat{y}^*_1)^2|y] - p_2^2(\hat{y}^*_1 - \hat{y}^*_2)^2 = E[(y^* - \hat{y}^*_2)^2|y] - p_1^2(\hat{y}^*_1 - \hat{y}^*_2)^2.
\]

(10)

To be more descriptive about these properties, first we record the optimality of BMA’s PMSE as the following inequality:

\[
E[(y^* - \hat{y}^*_1)^2|y] \leq E[(y^* - \hat{y}^*_i)^2|y], \quad i = 1, 2.
\]

(11)
The previous inequality is the clearest way to state that the BMA estimator \( \hat{y} \) produces the smallest prediction error, in the PMSE sense, among all the individual models’ posterior mean estimators, as long as all those models are used in creating the BMA estimator. We interpret this as a translation of the fact that each model that goes into creating the BMA estimator necessarily ignores model uncertainty. Note that this says nothing about how the BMA estimator would compare to a model not used in its definition.

But more can be said. As we mentioned implicitly, the previous inequality is a weaker statement than the statement in the Lemma, since the latter covers optimality over all convex combinations of the original models, not just the individual models themselves: the Lemma shows that BMA achieves the following minimum

\[
(p(M_k|y))_{k=1,2} := \arg\min_{\lambda \in [0,1]^2; \lambda_1 + \lambda_2 = 1} \mathbb{E}[(y^* - (\lambda_1 \hat{y}_1 + \lambda_2 \hat{y}_2))^2[y]].
\] (12)

Hence, the BMA estimator is actually optimal over all convex combinations of the individual estimators \( \hat{y}_1 \) and \( \hat{y}_2 \).

We can also express the reduction of the PMSE for the BMA estimator, compared to the best (lowest) PMSE among all of the individual models’, as

\[
r_{BMA}^2 := 1 - \frac{\mathbb{E}[(\hat{y}^* - y^*)^2[y]]}{\min_i \mathbb{E}[(\hat{y}_i^* - y^*)^2[y]]}.
\] (13)

In the specific case of two competing models, if we assume for instance that the 'best' model is \( M_2 \), we can obtain an even more explicit expression for \( r_{BMA}^2 \) which provides the relative gain attained by BMA, namely

\[
r_{BMA}^2 = p(M_1|y)^2 \frac{(\hat{y}_1 - \hat{y}_2)^2}{\mathbb{E}[(y_2^* - y^*)^2[y]]}.
\] (14)

Below in the Application section, we denote the sample version of the expression in (13) as \( \hat{r}_{BMA}^2 \), which we will use to evaluate the performance of BMA quantitatively.

To finish this section, we decompose the quantity \( \mathbb{E}[(\hat{y}^* - y^*)^2[y]] \) against the residuals \( (\hat{y}_i^* - y^*) \), \( i = 1, 2 \) from each individual model. This is easily done by symmetrizing formula (10) via reintroducing \( y^* \) to identify these residuals, and then taking another conditional expectation with respect to \( y \) to avoid an expression which depends on unobserved data. We obtain

\[
\mathbb{E}[(y^* - \hat{y}^*)^2[y]] = (p_1 - p_2^2)\mathbb{E}[(y^* - \hat{y}_1^*)^2[y]] + (p_2 - p_2^2)\mathbb{E}[(y^* - \hat{y}_2^*)^2[y]]
\]

\[\quad - (p_1^2 + p_2^2)\mathbb{E}[(\hat{y}_1^* - y^*)(y^* - \hat{y}_2^*)[y]].
\] (15)

Formula (15) shows that the PMSE of the BMA estimator is an explicit linear combination of the prediction errors of estimators for each constituent model, those that do not consider model uncertainty, but that one must subtract a coupling correction term on the right hand side of (15).

It is interesting to note that the weights in the aforementioned linear combination can be interpreted as the variances of Bernoulli random variables with the posterior model probabilities \( p_1 \) and \( p_2 \) as their success probabilities. Also note that, since these variances \( p_k - p_k^2 < p_k \), the linear combination is not convex, but is smaller. The correction term is not necessarily a subtraction of a positive term, but it is likely to be so in some generic cases, for example, when both individual models have significant biases in opposite directions for prediction of \( y^* \). This is particularly interesting when both models have similar posterior performances. Indeed, then, both values of \( p_k \) will be close to \( 1/2 \), which minimizes the values of \( p_k - p_k^2 \) for both \( k = 1, 2 \). This is a scenario where using the BMA model will significantly improve prediction errors even when each model is
competitive compared to the other, regardless of how large the individual models’ biases are, and without knowing in what direction they go, as long as the two models are known or assumed to have significant defects that work in opposite directions.

The PMSE corresponds to an MSE under the posterior probability. It is of different nature that the MSE one would get if the predicted values could be compared to experimental measurements. Nevertheless it is a good proxy for an actual MSE, when posterior predictions are close to actual measurements.

4 Domain-Corrected Model Averaging

Motivation. It is easy to imagine scenarios where alternative models are defined on different subsets of the same input space: this can typically arise with local models or with numerical models with different constraints. It is the case for instance for nuclear mass models: e.g. Energy Density Functionals (EDF) primarily cover heavier nuclei, while computations based on two- and three-body interactions (typically known as A-body models) range over lighter nuclei; this even occurs for different EDFs (see Kortelainen et al. (2010); Klupfel et al. (2009)). This can also easily happen for models constrained by observables of different type where one type of observable constrains one parameter, and each model has a wider parameter space for a specific variable. For instance, some nuclear models are mostly fitted on masses, and others on masses and radii. Surprisingly, BMA has not been studied or even proposed in this context; in particular, in nuclear physics, to our knowledge, one does not find studies which provide principled, fully satisfactory answers on how to compared models with similar, overlapping, but significantly non-identical domains. Other applications of our framework could include time series with missing data, or different time scales, e.g. in a financial setting where different classes of assets can also be described as observables.

To address this “domain discrepancy”, we present a method which relaxes the requirement that all models cover the same domain, and we name our procedure domain-corrected BMA.

We start by considering two models \( M_A \) and \( M_B \), which we will also denote by \( (A) \) and \( (B) \), or even merely \( A \) and \( B \), for simplicity depending on the context, and we assume that they are respectively defined only on subsets \( y^{(A)} \) and \( y^{(B)} \) of the data \( y \), denoting \( y^{(-A)} \) and \( y^{(-B)} \) for their respective complements in \( y \). The actual Bayesian evidence for each of these models are the probabilities \( p(y|A) \) and \( p(y|B) \), but a ”classical” BMA construction would have to be based on \( p(y^{(A)}|A) \) and \( p(y^{(B)}|B) \) instead, i.e. when each model refers only to its original range of validity. These quantities are related by conditioning as follows:

\[
p(y|A) = p(y^{(A)}, y^{(-A)}|A) = p(y^{(A)}|A)p(y^{(-A)}|y^{(A)}, A).
\]

This expression emphasizes that, to obtain Model \( (A) \)’s actual Bayesian evidence, \( p(y^{(A)}|A) \) must be multiplied by a correction factor \( p(y^{(-A)}|y^{(A)}, A) \) which represents the information one has on \( y^{(-A)} \) assuming that Model \( (A) \) holds and that it does not provide any prediction at the data points in \( y^{(-A)} \). Note that the distribution \( p(y|A) \) is meaningful only to the extent that \( y \) – and thus \( y^{(-A)} \) – is measurable in the underlying probability space, which implies the existence of underlying distributions \( p(y^{(-A)}) \) and subsequently of \( p(y^{(-A)}|y^{(A)}) \) and \( p(y^{(-A)}|y^{(A)}, A) \). To that extent, the problem of averaging models with different domains can be ill posed, if these distributions cannot be defined convincingly.

Depending on the precise meaning and understanding of the domain of definition of a model, there are two easy situations covering two extreme scenarios.
• If Model (A) does not predict the values $y^{(-A)}$ to exist, then this factor should be $p(y^{(-A)}|y^{(A)}, A) = 0$ and the model is ruled impossible.

• If neither Model (A) nor the data $y^{(A)}$ constrain the values $y^{(-A)}$ at all, then the factor should be $p(y^{(-A)}|y^{(A)}, A) = 1$.

Inadequacies of the classical BMA. The second scenario above is helpful to understand the modeling implications of confounding the narrow and broad meanings of Bayesian evidence for models with restricted domains: in the extreme case where no information can be gleaned about $y^{(-A)}$ from $y^{(A)}$ or from (A), it is legitimate to ignore the aforementioned correction factor. This has been proposed as a methodological matter of convenience: in the absence of any convincing information to the contrary, one may set the correction factor to $p(y^{(-A)}|y^{(A)}, A) = 1$, and this choice is easily adopted. Trotta (2008) contains one of the rare discussions on averaging models with different domains. The following passage from this reference is consistent with our comments above:

On the other hand, it is important to notice that the Bayesian evidence does not penalize models with parameters that are unconstrained by the data. It is easy to see that unmeasured parameters (i.e. parameters whose posterior is equal to the prior) do not contribute to the evidence integral, and hence model comparison does not act against them, awaiting better data (Trotta, 2008).

This corresponds to positing that, given Model (A), $y^{(-A)}$ should be deterministically equal to its sample value. Let us illustrate how this approach can fail to provide a satisfactory ranking of models in two examples where a model takes a shortcut by ‘refusing’ to predict difficult/challenging points.

Scenario 1. Consider the situation where one model $M_0$ is empty, in the sense that $p(y|M_0) = p(\emptyset|y) = p(\emptyset) = 1$ and $p(M_0|y) = \pi(M_0)$. On the other hand, any other model which constrains any part of the data will have an evidence lower than 1 which implies that the model will end up with lower posterior weights when starting from equal prior weights. Thus any predictive model will be deemed inferior to a non-predictive one.

Scenario 2. Take two models A and B with input space (domain of $x$) $X := \{a, b\}$. Assume model A does well at location $a$, but poorly at location $b$. Assume model B does passably well at location $a$, but does not predict anything at location $b$. One can easily choose the numbers to get an extreme situation (e.g. make A’s prediction at location $b$ to be extremely poor) where model B ends up with a much higher Bayes factor than model A, while the common sense idea by which no prediction is a form of extremely poor prediction, would always imply that model A is better than model B.

These examples show how important it is to acknowledge that a model’s inability to make predictions in some locations is not a neutral property.

Domain correction with two models. The above discussion motivates a modification of the procedure to amend the model weights to account for their (in-)ability to provide predictions at locations of interest. To account reasonably for the effect of a model’s domain into the computation
of the model weights, we propose the weaker assumption that \( p(y^{(-A)}|y^{(A)}, A) \) is independent from the model, i.e. we assume

\[
p(y^{(-A)}|y^{(A)}, A) = p(y^{(-A)}|y^{(A)}).
\]

This is quite natural considering that it implies a distribution \( p(y^{(A)}|A) \) but model \( (A) \) provides no information on \( y^{(-A)} \), leaving \( y^{(-A)} \) unconstrained by \( (A) \). This natural, intermediate solution can be shown to correspond to the following choice (details omitted). One replaces the factors of the likelihood corresponding to the missing model predictions by a log-average of the likelihoods over the models which do produce predictions, where the average is based on the predictive models’ posterior weights.

The evidence \( p(y|A) \) is now given by

\[
p(y|A) \propto_A p(y^{(A)}|A)p(y^{(-A)}|y^{(A)})\pi(A).
\]  \hspace{1cm} (16)

Our assumption \( y^{(A)} \cup y^{(B)} = y \) implies that \( y^{(-A)} \) can only be informed by \((B)\). Hence

\[
p(y^{(-A)}|y^{(A)}) = p(y^{(-A)}|y^{(A)}, B) = p(y^{(-A)}|y^{(A)} \cap (B), B)
\]  \hspace{1cm} (17)

which can be written as an explicit integral with respect to model \((B)\)'s parameter \( \phi_B \),

\[
\int p(y^{(-A)}|B, \phi_B)p(\phi_B|y^{(A)} \cap (B), B)d\phi_B,
\]  \hspace{1cm} (18)

and computed similarly to a classical evidence integral (see Section 2.2).

**Domain correction in the general case.** In the general case, each model \( \mathcal{M}_k \) constrains a subset \( y^{(k)} \) of the data \( y \) (for \( k = 1, \ldots, p \)); as in the case of two models, \( y^{(-k)} \) denotes the complement subset of \( y^{(k)} \) in \( y \). We also introduce \( y^{(c)} := \bigcap_k y^{(k)} \) as the set of data which is common to all individual models. Moreover we assume that \( y = \bigcup_k y^{(k)} \), i.e. every datapoint is covered by at least one model. We also assume, up to taking equivalence classes on models (see Appendix B for details), that for each pair of models there exists a chain of models joining them where each model shares a data point in its domain with each of its neighbours. Relying on the same principles, we set

\[
p(y^{(-k)}|y^{(k)}, \mathcal{M}_k) = p(y^{(-k)}|y^{(k)}),
\]

which leads to the model weights \( w_k \) of the form

\[
p(\mathcal{M}_k|y) \propto_k p(y^{(-k)}|y^{(k)})p(\mathcal{M}_k|y^{(k)}) = w_k.
\]  \hspace{1cm} (19)

In the computation of the corrective factors, there is the additional difficulty that, when there is more than one model constraining \( y^{(-k)} \), the factor \( p(y^{(-k)}|y^{(k)}) \) is no longer equal to a single \( p(y^{(-k)}|y^{(k)}, \mathcal{M}_k) \) but rather to the Bayesian average of all models constraining \( y^{(-k)} \). The notation for a given corrective factor can become quite cumbersome, or could be ambiguous, when model domains have very general intersections, but these corrective factors can still be computed recursively rather than directly. We relegate the calculations of the general case to the appendix, for the sake of readability.
A call for localization. Let us take now model weights $w_k$ as given, by our domain corrected BMA method or another. The posterior distribution for a general quantity of interest $\Delta$ can be calculated from

$$p(\Delta|y) \propto \sum_{k=1}^{K} w_k p(\Delta|y^{(k)}, M_k).$$  \hspace{1cm} (20)$$

Suppose that $\Delta$ is $y^*: = y(x^*)$. Then

$$p(y^*|y) \propto \sum_k w_k p(y^*|y^{(k)}, M_k) = \sum_{x^* \in D_k} w_k p(y^*|y^{(k)}, M_k) + \sum_{x^* \notin D_k} w_k p(y^*|y^{(k)}, M_k).$$

In the traditional BMA framework, $x^*$ would naturally belong to the domain $D_k$ of every model $M_k$, but this does not hold in our framework. Let us consider the problem of model selection. Picking one best model and using it for predictions at all locations would amount to a likelihood of 0 at all locations outside of the domain of the selected model, which is rather extreme. This calls for a local variation of model selection. For instance, instead of selecting a single best model, one can rank the models such that $M_0$ is the best model and $M_K$ is the worst model and use the following prediction procedure: at each location $x$, the true model is the model $M$ with the smallest $k$ which contains $x$ in its domain. One could easily think of more complex procedures. This also suggest local variations of model mixing.

Discussion. In the naive procedure a model which does not include an existing data point in its domain would have weight 0. With the domain-corrected procedure a model which does not include a given data point in its domain would have the same weight as the same model which would predict the average value over the other models. In that sense, the procedure is neutral towards (does not penalize or favor) a model not producing any prediction and is still conservative.

We can also note that our procedure questions that the models are mutually exclusive (i.e. that there exist a true model $M$ such that each $M_k$ is of the form $1_{\{M = k\}}$). Specifically, our procedure is incompatible with global exclusivity of models in the classical BMA framework, in the sense that the classical BMA assumes that one of the model is true on the whole domain. In our new procedure, we replaced the model prediction with a “default” value at locations where it would otherwise predicts nothing, given by the Bayesian average prediction. Consequently our model averages predict something in any given location, even when the set of predictive models depends on the location. Applying the same principle to the germane problem of model selection would lead to producing an ”optimal” model replaced by the BMA outside its own domain.

To be absolutely clear, it is of course always possible to bypass the domain issues by restricting estimations to the data which is common to the domains of all models. This circumvents having to deal with models withholding their predictions, but it can leave significant amounts of data unused. That disadvantage can be particularly stark in nuclear physics, when comparing EDF models with $A$-body models, as we mentioned at the start of this section, since these two model classes have such narrow overlap.

5 Examples and Applications

To illustrate the methodology described in the previous sections, we present two generic instances under which BMA of computer models leads to a reduction in prediction error. Our first illustration offers a pedagogical examples with an error reduction $\hat{r}^2_{BMA}$ exceeding 0.5. Our second set of examples focuses on application of the methodology in a setting with nuclear mass models.
Each of the examples in this section look at a situation with several competing models without any prior knowledge of which is better; thus we set the prior model weights to be uniform over the model space. All the posterior samples were computed using a Hamiltonian Markov-Chain Monte-Carlo algorithm. The evidence integrals were approximated using basic Monte-Carlo integration.

### 5.1 A Pedagogical Example

We assume here that the models under consideration are computationally expensive, and that we have access to their parameter gradients, which is often the case in typical applications. Therefore, we may define two competing computer models as Gaussian processes with means determined by their first order Taylor expansions. In this fashion, the calibration procedure falls under the rapid calibration via linearization. In the spirit of making the examples transparent, we choose to ignore the systematic discrepancy between the computer model and the physical model.

Let us consider a synthetic dataset $y$ with 10 identically 0 observations, whose noise is generated independently from the normal distribution $N(0, 10^{-3})$, at input points $x = (\pm 1, \pm 2, \pm 3, \pm 4, \pm 5)$. We will see clearly how the BMA intuitively leads to improvement in predictive power. We define the mean values $f_1$ and $f_2$ for the two computer models via their first order Taylor expansions, as follows:

$$f_i(x, \theta_i) \approx \alpha_i x^2 + \theta_i, \quad i \in \{1, 2\} \quad (21)$$

where $\alpha_1 = 0.5$ and $\alpha_2 = -0.5$ and $\theta_1$ and $\theta_2$ represent unknown calibration parameters. One easily conceptualizes this as a scenario with two computationally expensive models (only 10 observations) with quantitatively comparable $x$-dynamics acting in the opposite directions. It is natural to consider mixing both models since neither expansion in (21) provides significant a priori knowledge on which model is better. In addition, this is supported by the calculated posterior model probabilities, found below in Table 1 to be essentially equal. To begin with, we also assume that both models are defined on a common domain.

| Model  | RMSE  | $P(\mathcal{M}_k|y)$ | $\hat{r}_{BMA}^2$ |
|--------|-------|----------------------|-------------------|
| $\mathcal{M}_1$ | 0.027 | 0.5 | 0.691 |
| $\mathcal{M}_2$ | 0.023 | 0.5 | 0.546 |
| $\mathcal{M}_{BMA}$ | 0.015 | - | - |

Table 1: BMA weights, RMSE and RMSE reduction versus individual models for the pedagogical example.

After calibrating the two models on the full dataset, we sampled from posterior predictive densities of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_{BMA}$ at all available input points.

Figure 1 shows how the structural differences in $\mathcal{M}_1$ and $\mathcal{M}_2$ are reflected in the predictions they yield. Typically, where $\mathcal{M}_1$ overestimates the predicted value, $\mathcal{M}_2$ tends to underestimate. On the other hand, the predictions from $\mathcal{M}_{BMA}$ give a tighter fit to the original dataset and provide a better predictive ability on average. The estimated root MSE (RMSE) reduction for the BMA posterior mean estimator is 52.5%, a considerable improvement compared to the better of the two models in (21).
To further demonstrate the benefits of BMA, let us now extend this scenario to a situation where models in (21) are defined on different domains, and the corrective terms \( p(y^{(-k)}|y^{(k)}) \) need to be computed (e.g. the models are defined on different subsets of the nuclear landscape). We again consider a synthetic dataset with observations equal to 0 with measurement error generated independently from \( N(0, 10^{-3}) \), here at input points \( x = (\pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, \pm 8, \pm 9) \). Each of the models \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) was assigned a different training dataset \( y^{(k)} \). We constructed four different scenarios, for various proportions of shared observations \( (D_{\text{shared}}) \), from 20\% to 80\%, according to the scheme in Table 2.

| \(D_{\text{shared}}\) | \(\mathcal{M}_1\) | \(\mathcal{M}_2\) |
|------------------------|-----------------|-----------------|
| 0.2                    | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
| 0.4                    | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
| 0.6                    | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
| 0.8                    | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |
|                        | x               | x               |

Table 2: The employed scheme depicting the observations contained in a training dataset \( y^{(k)} \) for given model \( \mathcal{M}_k \). The symbol “x” denotes the inclusion.

For each four values of \( D_{\text{shared}} \), we carried out the BMA procedure the same way we did in the case where the models share the same domain, with the additional step involving the corrective terms \( p(y^{(-k)}|y^{(k)}) \). Note that the approximate computation of these terms is typically more demanding than the computation of evidence integrals \( (8) \), because it requires integration over a posterior distribution of parameters.

The RMSE was calculated at the same input points as in the case of the pedagogical example above. The BMA posterior mean estimator consistently outperforms estimators based on each
model individually, regardless of the proportion $D_{\text{shared}}$ of shared training data. Table 3 gives quantitative results which are comparable with the results in Table 1. In this table, $Q$ denotes the posterior odds ratio $p(y^{(1)}|y^{(1)})p(M_1|y^{(1)})/p(y^{(2)}|y^{(2)})p(M_2|y^{(2)})$ for sampling from mixture distribution (20).

| $D_{\text{shared}}$ | Model | RMSE | $p(y^{(k)}|M_k)$ | $p(y^{(k)}|y^{(k)})$ | $Q$ | $\hat{\tau}^2_{\text{BMA}}$ |
|---------------------|-------|------|----------------|----------------|----|----------------|
| 0.2                 | $M_1$ | 4.69 | 2.78 · 10^{-21} | 1.98 · 10^{-19} | 0.96 | 0.507 |
|                     | $M_2$ | 4.62 | 2.73 · 10^{-21} | 2.11 · 10^{-19} | -   | 0.492 |
|                     | $M_{\text{BMA}}$ | 3.29 | -               | -               | -   | -               |
| 0.4                 | $M_1$ | 4.66 | 7.99 · 10^{-20} | 4.33 · 10^{-16} | 1.10 | 0.513 |
|                     | $M_2$ | 4.52 | 7.95 · 10^{-20} | 3.96 · 10^{-16} | -   | 0.482 |
|                     | $M_{\text{BMA}}$ | 3.25 | -               | -               | -   | -               |
| 0.6                 | $M_1$ | 4.36 | 3.32 · 10^{-18} | 8.59 · 10^{-12} | 1.11 | 0.504 |
|                     | $M_2$ | 4.32 | 3.29 · 10^{-18} | 7.84 · 10^{-12} | -   | 0.494 |
|                     | $M_{\text{BMA}}$ | 3.07 | -               | -               | -   | -               |
| 0.8                 | $M_1$ | 3.60 | 1.45 · 10^{-16} | 2.99 · 10^{-6}  | 1.03 | 0.508 |
|                     | $M_2$ | 3.56 | 1.42 · 10^{-16} | 2.98 · 10^{-6}  | -   | 0.496 |
|                     | $M_{\text{BMA}}$ | 2.53 | -               | -               | -   | -               |

Table 3: BMA results for pedagogical example with domain correction.

The odds ratios stay expectedly close to 1, since the deviations from out-of-domain data are comparable across the models under consideration. Notice that the values of the corrective terms increase (they are closer to 1) as $D_{\text{shared}}$ increases. This can be expected, since the extreme case of $D_{\text{shared}} = 1$, where the domain is shared, corresponds to the classical BMA framework where no correction is needed. Also note that, the larger the overlap in models’ domains, the smaller the posterior mean squared error is, while keeping the estimated RMSE reduction close to 50%.

5.2 Averaging Nuclear (Mass) Models

An important task in the field of theoretical nuclear structure is the prediction of various physical observables with quantified uncertainties (Afanasjev et al., 2014). Nuclear mass models approximate these observables and other properties of atomic nuclei, for each pair $(Z, N)$ of proton number $Z$ and neutron number $N$. Of particular interest are the binding energy and the root mean square (rms) proton charge radius, which measures the size of an atomic nucleus. Mass models are typically a combination of a non-linear function $m_\theta(Z, N)$ with a parameter vector $\theta$ to be determined and a discrepancy term $\delta(Z, N)$, of the form

$$f(Z, N; \theta) = m_\theta(Z, N) + \delta(Z, N),$$

where $\delta(Z, N)$ is a correction due to physics that is not accounted for by the function $m_\theta(Z, N)$.

For the purpose of a naive exercise we propose the following family of functions for $m_\theta(Z, N)$:

$$m_\theta^l(Z, N) = \frac{\alpha(Z + N)}{1 + \gamma(Z + N)} + \beta_\theta^l, \quad l \in \{1, 2\},$$

where $\alpha$ is a linear or non-linear function representing basic physics based on the atomic number $A := Z + N$, and $\gamma$ is a constant which allows a modeler some additional flexibility to choose the
convexity of the mass model depending on their range of interest for \( A \). To fix ideas, let us assume that \( \alpha \) is a constant, thus becoming akin to a scale parameter, to adjust the model to realistic ranges of atomic masses. The mass model (23) describes a scenario where the model behaves linearly for nuclei with small atomic number \( A \), i.e. light nuclei, but the denominator on the right-hand side of (23) starts to dominate with increasing atomic number. The function becomes markedly convex for large \( A \) when \( \gamma < 0 \), and similarly concave for positive \( \gamma \), for heavier nuclei. The model is also allowed to depend on two types of observables through the constant additive parameters \( \beta_{\alpha 1}, \beta_{\alpha 2} \).

It is not uncommon for the data used to fit nuclear mass models to be scarce (expensive), and practitioners estimate subsets of the model parameters based on external sources of prior knowledge, before fitting the model to the data. We present a BMA analysis of two models according to (23), which differ in the value of the shape parameter \( \gamma \), as chosen by two hypothetical practitioners who want specific features in the regime of large \( A \). Both models are meant to agree largely for small \( A \), and disagree significantly for large \( A \). Specifically we choose \( \gamma_1 = 1/50 \), and \( \gamma_2 = -1/600 \). In the range of realistic \( A \)’s, i.e. integers from 1 to 300, both models are increasing, the one with \( \gamma_2 = -1/600 \) is roughly linear with a slight convexity for large \( A \), the other is strongly concave for large \( A \), which forces it to deviate strongly from linearity in that range. The choice of these two values of \( \gamma \) was designed to produce two models which are not comparable intuitively. The one with \( \gamma = -1/600 \) represents a simplistic linear approach, which can be interpreted as an effort to cover a wider range of atomic numbers in the absence of any specific physics. One may presume that it operates poorly for very large \( A \). The other, with \( \gamma_1 = 1/50 \), represents a model which attempts to cover and patch together two different regimes, where a specific effort is made to handle large \( A \). The absolute numerical values of gamma used here are designed to differentiate these two scenarios. The case \( \gamma = -1/600 \) could be replaced by \( \gamma = +1/600 \) without major changes to our analysis. The negative sign, which changes the functions convexity, allows for models which disagree in several respects. The \( \theta_i = (\alpha_i, \beta_{i, \alpha 1}, \beta_{i, \alpha 2}) \) are the calibration parameters for models \( M_i, i = 1, 2 \), while we assume here that the parameter \( \gamma \) has been determined ahead of time without the use of Bayesian statistics.

In this example, we used a data set of measurements of 20 binding energies and of 15 charge radii which are publicly available on the site of the National Nuclear Data Center based at the Brookhaven National Laboratory corresponding to nuclei of metals whose predictions have been challenging to nuclear physicists [Wang et al., 2017; Angeli and Marinova, 2013]. Gaussian process emulators with means according to (23) were fitted on a training set of 10 measurements of binding energies (\( ^{198,200,202}\text{Pb} \), \( ^{174}\text{Hf} \), \( ^{170,172}\text{Yb} \), \( ^{172}\text{Er} \), \( ^{156,168}\text{Dy} \), and \( ^{160}\text{Gd} \)) and 10 measurements of charge radii (\( ^{198,200,202,206,210,214}\text{Pb} \) and \( ^{198,112,122,124}\text{Sn} \)). To aid with the transparency of the results, we again choose to ignore the systematic discrepancy \( \delta \).

| Dataset | Model | \( P(M_k|y) \) | RMSE | \( \tilde{r}_{BMA}^2 \) | RMSE | \( \tilde{r}_{BMA}^2 \) |
|---------|-------|---------------|-------|-----------------|-------|-----------------|
| Train   | \( M_1 \) | 0.18 | 0.034 | 0.455 | 0.013 | 0.255 |
|         | \( M_2 \) | 0.82 | 0.032 | 0.394 | 0.013 | 0.223 |
|         | \( M_{BMA} \) | - | 0.025 | - | 0.011 | - |
| Test    | \( M_1 \) | - | 0.434 | 0.147 | 0.024 | 0.879 |
|         | \( M_2 \) | - | 0.563 | 0.563 | 0.009 | 0.118 |
|         | \( M_{BMA} \) | - | 0.400 | - | 0.008 | - |

Table 4: BMA weights, RMSE and RMSE reduction versus individual models for binding energies and proton rms charge radii measurements.
From the point of view of the RMSE, the BMA predictor again provides on average a considerably better fit than \( M_1 \) or \( M_2 \) for both observables. Table 4 shows that in general, both models produce relatively small RMSE for the training data, however, the BMA methodology still produces a significant improvement. Especially in the case of binding energies, the improvement is close to 40% relative to the better of the two models. It is also good to know that the posterior probabilities on \( M_1 \) and \( M_2 \) do not discard either of the models, despite \( M_2 \) being presumably superior. This supports our decision to take both models into account, rather than to pick one model over the other.

We further considered a testing data set consisting of 10 nuclei for prediction of binding energies \(^{204,206}\text{Pb, }^{158,162,164,166}\text{Dy, and }^{154,156,158,162}\text{Gd}\) and 5 nuclei for prediction of proton rms charge radii \(^{204,208,212}\text{Pb and }^{114,116}\text{Sn}\). This set of nuclei was disjoint from those in the training data. Again, the BMA posterior mean predictor performed better than either of the two models according to the estimated \( \hat{r}^2_{BMA} \) as in the case of the training data. In addition to the enhanced predictive ability of the BMA posterior mean predictor, BMA also increases the honesty of uncertainty quantification for the prediction, by accounting for modeling uncertainty as demonstrated by the decomposition of the total variance in (4). The data-driven averaging of uncertainties on observables according to the posterior probabilities \( P(M_k|y) \) yields a compromise between possibly over-confident credible intervals provided by the model \( M_2 \) and unnecessarily wide intervals of \( M_1 \). See Appendix C for results on selected isotopes that illustrate this phenomenon.

5.3 Averaging Nuclear Mass Emulators in the Ca Region

In their recent work, Neufcourt et al. (2019) use Gaussian Processes to model the discrepancies between experimental data and theoretical calculations for several nuclear models from the density functional theory. Their objective is to obtain quantified extrapolations for nuclear masses in the regions where no experimental data are available while providing robust uncertainty bounds. Such predictions are of direct interest to guide future nuclear experiments or to feed astrophysical models for the abundance of elements in the universe. Neufcourt et al. (2019) compute a simplified BMA of 9 models (Cao et al., 2019; Möller et al., 2016; Goriely et al., 2013), where the weights are based on the ability of the model to predict positive separation energies for all nuclei which have been observed. Here, we perform a full BMA analysis of the same models and compare our results to those obtained by Neufcourt et al. (2019).

We consider the same training dataset of one-neutron \((S_{1n})\) and two-neutron \((S_{2n})\) separation energies AME2003 (Audi et al., 2003) restricted to the calcium (Ca) region on the nuclear landscape with \( Z \geq 14 \) and \( N \leq 22 \). The predictive performance of each model (augmented with a GP model for systematic discrepancies) and the BMA posterior mean predictor is evaluated on both the training dataset and a testing dataset of new measurements in AME2016 (Wang et al., 2017).

Similarly to Neufcourt et al. (2019) we calculate model posterior probabilities independently on four non-overlapping nuclear domains according to the parity of numbers \( Z \) and \( N \). We assess the performance of the BMA with two metrics, the RMSE improvement and the fidelity of the empirical coverage probability. These quantities are calculated independently on the four parity domains and aggregated. See the results in Table 5.

The predictions based on the full BMA \((M_{BMA})\) outperform the simplified method of Neufcourt et al. (2019) \((M_{BMA(simple)})\) by 11% on the training dataset and 13% on the testing one, as measured by \( \hat{r}^2_{BMA} \). The lowest RMSE on the training dataset was attained by SLy4 and UNEDF1 respectively for AME2016 \( \setminus \) AME2003. This result should not discourage practitioners from using BMA posterior mean predictors, because the BMA methodology outlined in this paper allows for existence of a ”best” model for a particular data domain. However, such a model does not account
Table 5: Model posterior weights for the 9 nuclear mass models under consideration with RMSE and $\hat{r}_{BMA}^2$ values for training (AME2003) and testing (AME2016 \ AME2003) datasets. The last three rows correspond respectively to averaging with prior weights, simplified BMA [Neufcourt et al., 2019], and full BMA.

for modeling uncertainty whereas BMA does, and therefore the BMA posterior mean estimator performs consistently well irrespective of the dataset. In fact it attains the second lowest RMSE on both AME2003 and AME2016 \ AME2003. Moreover, if we consider only a subset of the whole model space, the BMA attains the lowest root RMSE. See Table 6 in Appendix C for the results with restricted model space.

Fig. 7 in Appendix C shows the empirical coverage probabilities (ECP), i.e. the proportion of independent testing points falling into the respective credibility intervals with nominal value given in the horizontal axis - ideally a straight line. While it is not clear that the BMA has an improved ECP compared to each individual models, its ECP is certainly significantly better than the worst models and similar to the models with highest fidelity.

6 Conclusion

Motivated by nuclear physics research problems, we analyzed the Bayesian Model Averaging setup - the natural Bayesian framework to infer any unknown quantity of interest when several models are competing. We focused on the specific challenges arising from BMA of computer models such as the calibration of both computationally inexpensive and expensive models as well as the computation of the evidence integral in this context.

The nuclear physics perspective led us to study the special case of averaging models with different domains, which has not been explored in the literature to our knowledge. We also gave a theoretical justification for the use of BMA posterior mean predictor in terms of PMSE reduction. While this predictor does not guarantee a universal improvement in predictive ability, on average, it performs at least as well as the best models under consideration.

Finally, we applied the methodology outlined in this paper under two general scenarios that lead to considerable RMSE reduction; one set of pedagogical examples with a synthetic dataset...
and one real data analysis of a pair of simple models fitted to a set of binding energies and proton rms charge radii. These illustrative examples provide insight into the benefits of BMA and serve as a "proof of concept." We also provided a full-scale BMA analysis of 9 state-of-the-art nuclear mass models. Fully documented Python code that reproduces all the examples in this paper is available at https://github.com/kejzlarv/BA_of_computer_models and can be easily modified to serve practitioners.

There are several opportunities to further explore BMA of computer models, within and beyond nuclear physics. As any other Bayesian method, it hinges on efficient sampling from posterior distributions. Direct sampling from the BMA posterior distribution could significantly improve the methodology ease of implementation. While our theoretical basis for BMA comes from potential reduction of PMSE, a more universal argument could seek consistency of an estimator of $\Delta$ based on its BMA posteriors. Finally our domain correction to BMA corresponds to an elementary and constrained localization. Developing a more elaborated local BMA procedure could answer a wider range of practical challenges in model mixing.

Acknowledgement. The authors thank Witold Nazarewicz for providing us with motivating examples and questions, and for insightful discussions and comments on an early version of our work.
Appendix A  BMA and Prediction Error Lemma

Lemma. For every $\lambda_1, \lambda_2 \geq 0$ satisfying $\lambda_1 + \lambda_2 = 1$, we have

$$\mathbb{E}[(y^* - \hat{y})^2|y] = \mathbb{E}[(y^* - \lambda_1 \hat{y}_1 - \lambda_2 \hat{y}_2)^2|y] - [(\lambda_1 - p_1)\hat{y}_1 - (\lambda_2 - p_2)\hat{y}_2]^2. \quad (24)$$

Proof. This follows from taking conditional expectation in the following expression, derived from standard factorization identities, and notice that the right hand side is $y$-measurable:

$$(y^* - \hat{y})^2 - (y^* - \lambda_1 \hat{y}_1 - \lambda_2 \hat{y}_2)^2 = [2y^* - (\lambda_1 + p_1)\hat{y}_1 - (\lambda_2 + p_2)\hat{y}_2][(\lambda_1 - p_1)\hat{y}_1 + (\lambda_1 - p_1)\hat{y}_1].$$

\qed

Appendix B  Averaging models with different domains: the general case

Suppose given data $(x_i, y_i)_{i=1}^n$ and $p$ models $(\mathcal{M}_k)_{k=1}^p$, and assume that each model $\mathcal{M}_k$ is defined on a subset $x^{(k)}$ of $x$. Denote also $y^{(k)}$ the subset of $y$ corresponding to $x^{(k)}$, $y^{(-k)}$ the complementary subset as well as $y^{(\ominus)} := \bigcap_k y^{(k)}$. Suppose naturally that all data locations are in the domain of at least one model so that $y = \bigcup_k y^{(k)}$. This situation arises naturally in nuclear physics, where $y$ is the value of a given observable (e.g. energy) corresponding to a nuclear configuration $x_i := (Z_i, N_i)$ and where commonly used models are defined on a restricted nuclear domain where specific physical interactions prevail.

Note that if the datasets are disjoint, there is simply no basis to compare the models. Given a set of models, the reader can easily get convinced that one can define a unique minimal equivalence relationship $\star$ on the models (i.e. with a number of equivalence classes maximal) satisfying $\mathcal{M} \star \mathcal{M}'$ if $\mathcal{M}$ and $\mathcal{M}'$ share at least one data point, i.e. $\mathcal{M} \star \mathcal{M}'$ if and only if there exists $r \geq 0$ and a sequence of models $\mathcal{M} =: \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_r := \mathcal{M}'$ such that $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ have a common data point for each $0 \leq i < r$. The computation of the posterior weights of the models can then be done within each class of equivalence, and we will therefore assume that there is only one such equivalence class.

In ‘classical’ Bayesian model averaging where all models share the same domain $y$ one can express the posterior probabilities on the models $p(\mathcal{M}_k|y)$ using Bayes formula

$$p(\mathcal{M}_k|y) \propto_k p(y|\mathcal{M}_k) \pi(\mathcal{M}_k) \quad (25)$$

and estimates the evidence integral (Bayes factor) $p(y|\mathcal{M}_k)$ as detailed in Section [2.2]. In our situation, however, the model $\mathcal{M}_k$ provides an expression $p(y^{(k)}|\mathcal{M}_k)$ instead of $p(y|\mathcal{M}_k, \phi_k)$, so that the standard procedure cannot be applied without a further argument.

Starting from (25), we expand $p(y|\mathcal{M}_k)$ similarly to the two-model case as

$$p(y|\mathcal{M}_k) = p(y^{(k)}, y^{(-k)}|\mathcal{M}_k) = p(y^{(-k)}|y^{(k)}, \mathcal{M}_k)p(y^{(k)}|\mathcal{M}_k).$$

Instead of setting $p(y^{(-k)}|y^{(k)}, \mathcal{M}_k) = 1$ which unsatisfactorily advantages models which withhold their predictions at difficult locations (see the example scenarios and discussion in Section [3]) our domain-corrected model averaging estimates

$$p(y^{(-k)}|y^{(k)}, \mathcal{M}_k) = p(y^{(-k)}|y^{(k)}).$$
This yields evidence and posterior weights given respectively by

\[
p(y|M_k) = p(y^{(-k)}|y^{(k)}) p(y^{(k)}|M_k)
\]

\[
p(M_k|y) \propto_k p(y^{(-k)}|y^{(k)}) p(y^{(k)}|M_k) \pi(M_k),
\]

similarly to the two-model case. All that is left now is to evaluate the corrective likelihood \( p(y^{(-k)}|y^{(k)}) \).

Letting \( S \) be the subset of models constraining \( y^{(-k)} \) we can compute the corrective likelihood \( p(y^{(-k)}|y^{(k)}) \) by conditioning with respect to the model:

\[
p(y^{(-k)}|y^{(k)}) = \sum_{l=1}^{p} p(y^{(-k)}|y^{(k)}, M_l) \pi(M_l)
\]

\[
= \sum_{l \in S} p(y^{(-k)}|y^{(k)}, M_l) \pi(M_l) + \sum_{l \not\in S} p(y^{(-k)}|y^{(k)}) \pi(M_l)
\]

\[
= \frac{1}{\sum_{l \in S} \pi(M_l)} \sum_{l \in S} p(y^{(-k)}|y^{(k)}, M_l) \pi(M_l)
\]

(a) The simple case is when \( y^{(-k)} \) is non-divisible, in the sense that for every \( l \) we have \( y^{(-k)} \subset y^{(l)} \) or \( y^{(-k)} \cap y^{(l)} = \emptyset \): in that case, \( p(y^{(-k)}|y^{(k)}, M_l) \) and the sum above have explicit expressions.

(b) In the general case some models may be defined only on a strict subset of \( y^{(-k)} \). In that case we have

\[
p(y^{(-k)}|y^{(k)}, M_l) = p(y^{(-k) \cap(l)}, y^{(-k) \cap(-l)}|y^{(k)}, M_l)
\]

\[
= p(y^{(-k) \cap(l)}|y^{(k)}, M_l) p(y^{(-k) \cap(-l)}|y^{(k)}, y^{(-k) \cap(l)}, M_l)
\]

\[
= p(y^{(-k) \cap(l)}|y^{(k) \cap(l)}, M_l) p(y^{(-k) \cap(-l)}|y^{(l)}\cap(k), y^{(k) \cap(-l)})
\]

The first term is explicit given a model. Hence we can compute inductively \( p(y^{(-k_1) \cap(-k_2) \cap \ldots \cap(-k_q)}|y^s) \) for all \( q \)-tuples \((k_1, \ldots, k_q)\) and subset of data \( y^s \subset y \) with a decreasing recursion on \( q \), where the first iteration corresponds to the simple case (a).

For practicality purposes it is important to notice that the complexity of the underlying algorithm is at most exponential in the number of models, where each iteration requires the computation of a predictive posterior of decreasing subsets of data given decreasing subsets of the data, plus \( N \) computations of corrective likelihoods as in (a), where \( N \) is the number of non-divisible subsets.
Appendix C  Examples and Applications Charts

C.1 Domain-corrected BMA

Here we present the fit results of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_{BMA}$ for the pedagogical example under the scenario where the domain correction is needed. The BMA provides on average better fit than each of the models individual.

![Graphical representation of predictions](image)

Figure 2: Posterior mean predictions for the 10 observations $y$ for the two models in (21) as well as their BMA. The dashed line segments represent the values of the original observations.
C.2 Averaging nuclear (mass) models

To illustrate how BMA increases the honesty in uncertainty quantification of the prediction by accounting for modeling uncertainty (see (4)), Figures 3, 4, and 5 show predicted values with 95% credible intervals for binding energies of selected lead and gadolinium isotopes and proton rms charge radii for some tin and lead isotopes.

Figure 3: Fit for binding energies of lead isotopes $^{198,200,202}$Pb (in MeV) and proton rms charge radii for tin isotopes $^{108,112,122,124}$Sn (in fm). The dashed line segments represent the values of the original observations.
Figure 4: Predictions for binding energies of lead isotope $^{204}\text{Pb}$, and gadolinium isotopes $^{156,162}\text{Gd}$ (all in MeV). The dashed line segments represent the values of the original observations.

Figure 5: Predictions for proton rms charge radii for tin isotopes $^{114,116}\text{Sn}$ and lead isotope $^{212}\text{Pb}$ (all in fm). The dashed line segments represent the values of the original observations.
C.3 Averaging nuclear mass emulators in the Ca region

Figure 6: Empirical coverage probability. (AME2016 + AME2018) \ AME2003

| Model posterior weights | Errors |
|-------------------------|--------|
|                         | Training | Testing |
|                         | RMSE $\hat{\mu}_{BMA}$ | RMSE $\hat{\mu}_{BMA}$ |
| $S_{1n}$ (odd N) | $S_{2n}$ (even N) |
| even Z  | odd Z  | even Z  | odd Z  | even Z  | odd Z  | even Z  | odd Z  |
| SkM*    | 0.000  | 0.001  | 0.000  | 0.000  | 0.142  | 0.375  | 0.925  | 0.413  |
| FRDM-2012 | 1.000  | 0.997  | 0.900  | 0.399  | 0.114  | 0.031  | 0.808  | 0.231  |
| HFB-24  | 0.000  | 0.002  | 0.100  | 0.601  | 0.146  | 0.405  | 0.806  | 0.227  |
| $\mathcal{M}_{BMA}$ | 0.112  | -      | 0.709  | -      |

Table 6: Model posterior weights for a subset of nuclear mass models under consideration with RMSE and $\hat{\mu}_{BMA}$ values for training (AME2003) and testing (AME2016 \ AME2003) datasets.
References

Afanasjev, A. V., Agbemava, S. E., Ray, D., and Ring, P. (2014). Error estimates of theoretical models: a guide. *Journal of Physics G: Nuclear and Particle Physics*, 41:074001.

Afanasjev, A. V., Agbemava, S. E., Ray, D., and Ring, P. (2015). Neutron drip line: Single-particle degrees of freedom and pairing properties as sources of theoretical uncertainties. *Physical Review C*, 91:014324.

Aitken, S. and Akman, O. E. (2013). Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. *BMC Systems Biology*, 7:72–83.

Angeli, I. and Marinova, K. (2013). Table of experimental nuclear ground state charge radii: An update. *Atomic Data and Nuclear Data Tables*, 99:69 – 95.

Audi, G., Wapstra, A., and Thibault, C. (2003). The ame2003 atomic mass evaluation: (ii). tables, graphs and references. *Nuclear Physics A*, 729:337–676.

Balasubramanian, J. B., Visweswaran, S., Cooper, G. F., and Gopalakrishnan, V. (2014). Selective model averaging with bayesian rule learning for predictive biomedicine. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2014:17–22.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49:138–154.

Brynjarsdóttir, J. and O’Hagan, A. (2014). Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30:114007.

Cao, Y., Nazarewicz, W., Olsen, E., Savanur, A., and Schunck, N. (2019). Mass explorer.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20:80–101.

Erler, J., Birge, N., Kortelainen, M., Nazarewicz, W., Olsen, E., Perhac, A., and Stoitsov, M. (2012). The limits of the nuclear landscape. *Nature*, 486:509–512.

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86:1–28.

Geweke, J. (1999). Using simulation methods for bayesian econometric models: inference, development, and communication. *Econometric Reviews*, 18:1–73.

Goriely, S., Chamel, N., and Pearson, J. M. (2013). Further explorations of skyrme-hartree-fock-bogoliubov mass formulas. xiii. the 2012 atomic mass evaluation and the symmetry coefficient. *Physical Review C*, 88.

Hernández, B., Raftery, A. E., Pennington, S. R., and Parnell, A. C. (2018). Bayesian additive regression trees using bayesian model averaging. *Statistics and Computing*, 28:869–890.
Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103:570–583.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2005). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26:448–466.

Higdon, D., McDonnell, J. D. M., Schunck, N., Sarich, J., and Wild, S. M. (2015). A bayesian approach for parameter estimation and prediction using a computationally intensive model. *Journal of Physics G: Nuclear and Particle Physics*, 42:034009.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–401.

Homan, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, 15:1351–1381.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to bayesian model selection for ecologists. *Ecological Monographs*, 85:3–28.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kennedy, M. C. and O’Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:425–464.

Klupfel, P., Reinhard, P. G., Burvenich, T. J., and Maruhn, J. A. (2009). Variations on a theme by skyrme: A systematic study of adjustments of model parameters. *Physical Review C*, 79:034310.

Kortelainen, M., Lesinski, T., Mor, J., Nazarewicz, W., Sarich, J., Schunck, N., Stoitsov, M. V., and Wild, S. (2010). Nuclear energy density optimization. *Physical Review C*, 82:024313.

Möller, P., Sierk, A., Ichikawa, T., and Sagawa, H. (2016). Nuclear ground-state masses and deformations: Frdm(2012). *Atomic Data and Nuclear Data Table*, 109-110:1–204.

Neufcourt, L., Cao, Y., Nazarewicz, W., Olsen, E., and Vien, F. (2019). Neutron drip line in the ca region from bayesian model averaging. *Physical Review Letters*, 122:062502.

Olsen, E. and Nazarewicz, W. (2019). α-decay energies of superheavy nuclei: Systematic trends. *Physical Review C*, 99:014317.

Piekarewicz, J., Agrawal, B. K., Colò, G., Nazarewicz, W., Paar, N., Reinhard, P.-G., Roca-Maza, X., and Vretenar, D. (2012). Electric dipole polarizability and the neutron skin. *Physical Review C*, 85:041302.

Plumlee, M. (2017). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112:1274–1285.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Schomning, K., Bornkamp, B., Bretz, F., and Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, 35:4021–4040.
Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., and Salamin, N. (2014). Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Systematic Biology*, 63:349–367.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49:71–104.

Tuo, R. and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43:2331–2352.

Tuo, R. and Wu, C. F. J. (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4:767–795.

Wang, M., Audi, G., Kondev, F. G., Huang, W. J., Naimi, S., and Xu, X. (2017). The ame2016 atomic mass evaluation (II). tables, graphs and references. *Chinese Physics C*, 41:030003.

Wei, W., Visweswaran, S., and Cooper, G. F. (2011). The application of naive bayes model averaging to predict alzheimer’s disease from genome-wide data. *Journal of the American Medical Informatics Association*, 18:370–375.

Wen, X. (2015). Bayesian model comparison in genetic association analysis: linear mixed modeling and snp set testing. *Biostatistics*, 16:701–712.