Local Bayesian Dirichlet mixing of imperfect models

Vojtech Kejzlar [1,†], **Léo Neufcourt** [2], **and Witold Nazarewicz** [3,*]

[1]Mathematics and Statistics Department, Skidmore College, Saratoga Springs, New York 12866, USA
[2]FRIB Laboratory, Michigan State University, East Lansing, Michigan 48824, USA
[3]Department of Physics and Astronomy and FRIB Laboratory, Michigan State University, East Lansing, Michigan 48824, USA
[†]vkejzlar@skidmore.edu
[*]witek@frib.msu.edu

ABSTRACT

To improve the predictability of complex computational models in the experimentally-unknown domains, we propose a Bayesian statistical machine learning framework utilizing the Dirichlet distribution that combines results of several imperfect models. This framework can be viewed as an extension of Bayesian stacking. To illustrate the method, we study the ability of Bayesian model averaging and mixing techniques to mine nuclear masses. We show that the global and local mixtures of models reach excellent performance on both prediction accuracy and uncertainty quantification and are preferable to classical Bayesian model averaging. Additionally, our statistical analysis indicates that improving model predictions through mixing rather than mixing of corrected models leads to more robust extrapolations.

Introduction

When considering predictions or extrapolations of physical quantities into unknown domains, a reliance on a single imperfect theoretical model can be misleading. To improve the quality of extrapolated prediction, it is in fact advisable to use several different models and mix their results[1–4]. In this way, the "collective wisdom" of several models could be maximized by providing the best prediction rooted in the most current experimental information. To carry out the mixing, statistical machine learning (ML) methods, with their ability to capture the local features, are tools of choice. Specifically, Bayesian Model Averaging (BMA) can be used to combine multiple models to produce more reliable predictions since it is the natural Bayesian framework to account for the uncertainty on the model itself[1–3]. In absence of another established methodology, the application of BMA to scenarios where several models compete to describe the same phenomenon has been relatively widespread in such diverse fields as weather forecasting[5], political science[6], transportation[7], and nuclear physics[8,9].

It is important to remember, however, that BMA relies on the assumption that one of the proposed models is the true model (i.e., a model that perfectly describes the physical reality), which is clearly inappropriate when dealing with complex systems and approximate modeling. In practice, it often happens that none of the competing state-of-the-art models can be dominated by the others, in the sense that each model does something better than the others. In such a setup, models should not be viewed as exclusive but as complementary, and BMA seems theoretically ill-grounded. In addition, in the standard implementations of the BMA, the weights are global, i.e., they are constant over the input domain, and thus unable to catch local model preferences.

Besides BMA, there exist other methods to combine results of different models. In fact, combining models has been the subject of much research in ML that has led to the development of the topical "ensemble learning methods" (bagging and boosting). These methods remain in their spirit and purpose very close to BMA and typically do not fix the inadequacies mentioned above: their goal is to identify the best performing model given a set of models. See Ref.[10] for review of additional approaches.

In this work, we develop and apply the Local Bayesian Model Mixing (LBMM), an extension of Bayesian stacking[11–15], for managing competing models. Under the Bayesian stacking framework, one assumes that the true model is a linear combination of the models instead of being one of the models. The extrapolations are thus obtained via a direct mixture of the models, as compared to the mixture of posterior distributions under the standard BMA. Unlike the BMA weights which reflect the fit of a statistical model to data, *independently of the set of available models* except for normalization, the weights based on model mixing or stacking reflect the model's contribution to the final predictions[13]. The LBMM used in this study makes the use of Dirichlet distribution to infer stacking weights which hierarchically depend on the model input space and thus highlight the local fidelity of theoretical models. Additionally, the LBMM framework well captures uncertainties of individual models and their mixing weights through the proposed hierarchical structure. Below, we first present the general LBMM framework

followed by a pedagogical case of global mixture of models that corresponds to classical Bayesian stacking. Subsequently, we let the model weights vary across the model input space and consider several hierarchical Bayesian models based on the Dirichlet distribution.

As an example, we apply the new method to predicting nuclear mass, or binding energy, which is the basic property of the atomic nucleus. Since we consider BMA to be a point of comparison for our LBMM methodology, we briefly review the general predictive framework of BMA in Methods section. The binding energy determines nuclear stability as well as nuclear reactions and decays. Quantifying the nuclear binding is important for many nuclear structure and reaction questions, and for understanding the origin of the elements in the universe. The astrophysical processes responsible for the nucleosynthesis in stars often take place far from the valley of beta stability, where experimental masses are not known. In such cases, missing nuclear information must be provided by extrapolations. Accurate values for nuclear masses and their uncertainties beyond the range of available experimental data are also used in other scientific fields, such as atomic and high-energy physics, as well as in many practical applications. In order to improve the quality of model-based predictions of masses of rare isotopes far from stability, ML approaches can be used that utilize experimental and theoretical information. A broad range of ML tools have been used to mine unknown nuclear masses, including Gaussian processes (GPs), neural networks, frequency-domain bootstrap and kernel ridge regression[16–26] (see the recent review[27] for more references). In a series of papers[8,28–31], the BMA methodology has been applied to nuclear mass predictions. In this work, we propose the LBMM approach to produce model-informed extrapolations of nuclear masses that overcome the limits of BMA mentioned above.

Bayesian model mixing

Let us consider experimental observations $y(x_i)$ of a physical process at locations $x_i \in \mathscr{X} \subset \mathbb{R}^q$, $i = 1, \ldots, n$, governed by a "true model" $y^*(x)$, and let us assume that the true model is not fully captured by one of the proposed models $f_k, k = 1, \ldots, p$, but rather a combination of these models. It is then natural to consider a statistical mixture model of the general form:

$$y(x_i) = \sum_{k=1}^{p} \omega_k(x_i) f_k(x_i) + \sigma \varepsilon_i, \tag{1}$$

where σ represents the scale of the error of the mixture model, $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, and $f_1(x_i), \ldots, f_p(x_i)$ are theoretical values for the datum $y(x_i)$ provided by the p theoretical models considered.

In practice, the weights $\boldsymbol{\omega}(x_i) = (\omega_1(x_i), \ldots, \omega_p(x_i))$ must be taken in a space where inference is possible. This can be done in many ways. In this work, we will highlight a few alternative models for $\boldsymbol{\omega}(x_i)$ which are tractable, suggestive, and fully Bayesian.

Additionally, one can improve the models by accounting for systematic errors. This can be done by adding to each model the systematic error correction $\delta_{f,k}$:

$$y(x_i) = \sum_{k=1}^{p} \omega_k(x_i) \left(f_k(x_i) + \delta_{f,k}(x_i) \right) + \sigma \varepsilon_i. \tag{2}$$

Global mixtures
First, we present the simplest application of Bayesian model mixing (BMM) where one assumes global weights (GBMM), i.e., weights that are *constant over the input domain*.

Linear model (L)
Let us first model the underlying physical process by a global (linear) mixture of the individual models:

$$y(x_i) = \sum_{k=1}^{p} \omega_k \left(f_k(x_i) + \delta_{f,k}(x_i) \right) + \sigma \varepsilon_i. \tag{3}$$

Using Eq. (3) the log-likelihood of the model can be written as:

$$\log p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\delta}_f, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y(x_i) - \sum_{k=1}^{p} \omega_k \left(f_k(x_i) + \delta_{f,k}(x_i) \right) \right]^2,$$

where $\boldsymbol{y} = (y(x_1), \ldots, y(x_n))$. To ensure that the weights $\boldsymbol{\omega}$ have the same support as the model weights in BMA, it may also be justified to assume the simplex constraints $\omega_k \geq 0$ and $\sum_k \omega_k = 1$. In that case, the posterior distributions should also satisfy the

simplex constraints. While the first condition is easily met using non-negative priors, the second can be more challenging to enforce with priors. Nevertheless the naive idea of projecting the unconstrained posteriors appears to be relatively efficient in the case of simple linear models[32]. Here projecting the simplex constraints corresponds to substituting

$$\omega_k \longleftarrow \frac{max(\omega_k,0)}{\sum_{\ell} \max(\omega_{\ell},0)}. \tag{4}$$

Dirichlet model (D)

As a refinement of the global linear mixture and a step towards local mixtures with simplex constraints, one can suppose that the weights $\boldsymbol{\omega}$ are given *hierarchically* by a Dirichlet distribution:

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) \propto \prod_k \omega_k^{\alpha_k-1} \tag{5}$$

with the hyperprior $\pi(\boldsymbol{\alpha})$ on the hyperparameter $\boldsymbol{\alpha}$. The reason for this additional modeling layer is twofold. First, it allows us to express uncertainty about the prior model weighing imposed by $\boldsymbol{\alpha}$. Note that a Dirichlet distribution with size p and the parameter vector $\boldsymbol{\alpha} > 0$ is a multivariate continuous distribution on the simplex $\{\omega_1,...,\omega_p \geq 0 : \sum_{k=1}^{p} \omega_k = 1\}$ where the average value of ω_j is $\alpha_j/\sum_{k=1}^{p} \alpha_k$. Looking at the shape of the distribution in Fig. 1, it is clear that $\boldsymbol{\alpha} < 1$ is close to model selection while $\boldsymbol{\alpha} > 1$ encourages true mixing. The hyperprior $\pi(\boldsymbol{\alpha})$ allows us to quantify our uncertainty about these two regimes. Secondly, the hierarchical model for $\boldsymbol{\omega}$ permits, with slight modification, a heterogeneity of weights based on the value of x, which we shall exploit shortly.
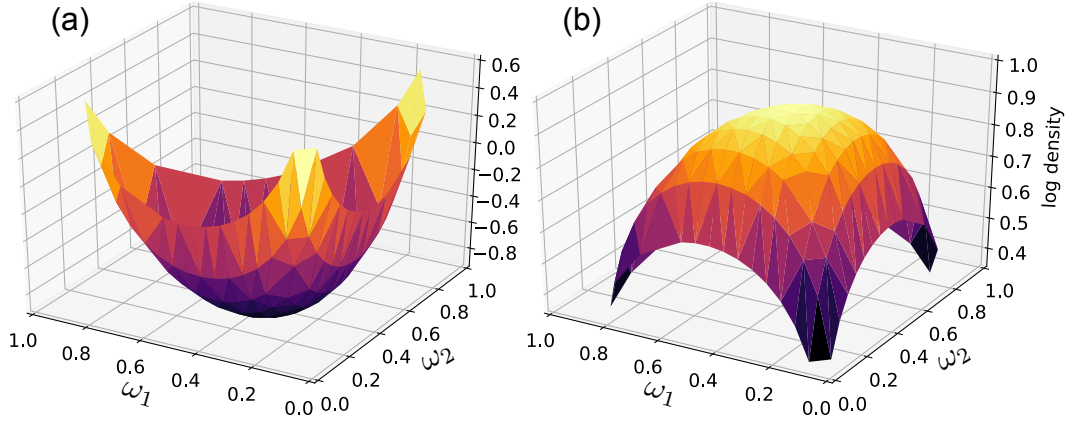

Figure 1. Log density of the Dirichlet distribution when $p = 3$ as a function of ω_1 and ω_2 ($\omega_3 = 1 - \omega_1 - \omega_2$). The parametrization of Dirichlet distribution is such that $\boldsymbol{\alpha} < 1$ is close to model selection (one dominant weight with a high probability) and $\boldsymbol{\alpha} > 1$ leads to true mixing (several large-probability weights). Left: $\boldsymbol{\alpha} = (0.3, 0.3, 0.3)$; Right: $\boldsymbol{\alpha} = (1.3, 1.3, 1.3)$.

Consequently, up to a choice of prior $\pi(\boldsymbol{\alpha}, \boldsymbol{\delta}_f, \sigma)$, the joint posterior distribution of $(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\delta}_f, \sigma)$ is given by

$$p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\delta}_f, \sigma|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\delta}_f, \sigma)p(\boldsymbol{\omega}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}, \boldsymbol{\delta}_f, \sigma), \tag{6}$$

which does not have a closed form in general and needs to be approximated using MCMC. Predictions of observations from a physical process at new locations are then obtained by propagating the posterior samples of $(\boldsymbol{\alpha}, \boldsymbol{\delta}_f, \sigma)$ through the hierarchy described above.

The Dirichlet weights encapsulate the contribution of each model to the mixture, which makes the interpretation of the weights probabilistic. In that sense, they carry a different meaning than the BMA weights, which measure the fidelity of individual models. Still, both weights can be compared as they play the same role in the the final predictions – keeping in mind that the information contained in the posterior distribution of the Dirichlet weights is richer than the point values produced by BMA.

Local mixtures

Let us now consider that the observations $y(x_i)$ follow the general statistical model given by Eq. (2). The key feature here is that now the weights depend on the location x. Without additional information, the functions $\boldsymbol{\omega}(x_i)$ shall be estimated with a non-parametric estimator satisfying the simplex constraints.

In order to account for the local dependency of the model weights while satisfying the simplex constraints, we propose a hierarchical framework based on the Dirichlet distribution. Specifically, we take for every x weights $\boldsymbol{\omega}(x)$ as Dirichlet random variables defined by parameters $\boldsymbol{\alpha}(x) = (\alpha_1(x), \ldots, \alpha_p(x))$. The correlations between $\omega_k(x)$'s for different values of x shall be contained in the corresponding $\alpha_k(x)$-correlations. We investigate two models for $\alpha_k(x)$: a Generalized Linear Dirichlet model (GLD) and a Gaussian Process Dirichlet model (GPD).

In particular, we assume that at every location x the model weights $\boldsymbol{\omega}(x)$ follow a Dirichlet distribution with parameters $\alpha_k(x)$, for $k = 1, \ldots, p$, now depending on x, that will encode the spatial relationships between the model weights over the input space. Since the Dirichlet distribution is defined for parameter values $\alpha_k > 0$, we additionally apply an exponential *link* function, i.e., we consider $\gamma_k(x) := \log(\alpha_k(x))$ which can be regressed symmetrically. Thus, the purpose of the link function is to allow for unconstrained modeling of $\gamma_k(x)$.

The GLD version of our local mixing framework represents $\gamma_k(x)$ parametrically as

$$\gamma_k(x) = \boldsymbol{\beta}_k^T x, \tag{7}$$

where $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) = (\beta_{1,1}, \ldots, \beta_{1,q}, \ldots, \beta_{p,1}, \ldots, \beta_{p,q})$ is a parameter vector. The linear nature of Eq. (7) corresponds to the assumptions that the correlations between local weights have a relatively large spatial range.

As a finer version of LBMM, we propose a non-parametric GPD model for $\gamma_k(x)$ defined by a Gaussian process prior parametrized with a constant mean γ_k^∞ and covariances $c_k(x, x')$ given by quadratic exponential kernels[33]. Additionally, we assume that $\gamma_k(x)$ and $\gamma_j(x)$ are statistically independent for $k \neq j$. Note that the proposed hierarchical structure takes into account not only the relationship between $\omega_k(x)$'s for different values of x (via $\alpha_k(x)$) but also the correlations between the weights of models at given spatial location x (via Dirichlet distribution). This would not be possible if one choose to model $\omega_k(x)$'s directly, let's say with a GP over $\omega_k(x)$.

Application: nuclear mass extrapolation

As a case study for the Bayesian model mixing framework described above, we consider the separation energies of atomic nuclei, which were the subject of our previous investigations[8, 18, 28, 29, 31]. Our particular goal is to compare the following alternatives:

  (i) Raw models without statistical correction δ_f (/w δ_f) vs. models corrected with δ_j (w/ δ_f);

 (ii) BMA vs. global BMM;

(iii) Global BMM vs. local BMM.

The two-neutron separation energy (S_{2n}) is a fundamental property of the atomic nucleus defined as the energy required to remove two neutrons from the nucleus. It can be expressed as a difference of nuclear masses. Here, the input space $\mathscr{X}$ is represented by the numbers of protons Z and neutrons N. Consequently, in this study $q = 2$, $x_i := (Z_i, N_i)$ and y_i is the observed two-neutron separation energy at x_i. We are particularly interested in even-even nuclei, for which both N and Z are even numbers. We use the most recent measured values of two-neutron separation energies for nuclei from the AME2003 dataset[34] as training data ($n = 521$) for BMM and the GP systematic corrections; for BMA calculations we retain as evidence dataset a subset of this training data consisting of 8 nuclei: 3 proton-rich nuclei ^{148}Er, ^{188}Po, ^{242}Cf, and 5 neutron-rich nuclei ^{64}Cr, ^{116}Ru, ^{160}Nd, ^{168}Hf, ^{232}Ra. We keep additional data tabulated in AME2020[35] for an out of sample extrapolative testing dataset ($n = 59$). These three domains are depicted in Fig. 2.

As for prediction, we will use the largest domain on which two-neutron separation energies are positive, i.e., the corresponding nuclei are predicted to exist. In line with our previous studies, we consider seven theoretical models based on the nuclear density functional theory (DFT) which is capable of describing the whole nuclear chart: SkM*[36], SkP[37], SLy4[38], SV-min[39], UNEDF0[40], UNEDF1[41], and UNEDF2[42]. The DFT data were taken from the theoretical database[43]. The above set of DFT models was augmented by two well-fitted mass models FRDM-2012[44] and HFB-24[45] that have significantly more parameters than the (less phenomenological) DFT models, resulting in a better fit to measured masses.

In the subsequent Bayesian analyses, we use independent priors for the different statistical parameters and keep consistent notation throughout the model mixing variants. In general, we use normal priors for unconstrained parameters, Gamma priors for positive parameters, and uniform priors for bounded parameters. Recall that a Gamma distribution is parametrized by a shape parameter a and a rate parameter b, and has its mean given by a/b and variance given by a/b^2. For the error scale parameters σ, we use Gamma priors with scale parameter 5 and rate parameter 10, with mean 0.5 MeV and standard deviation 0.22 MeV.
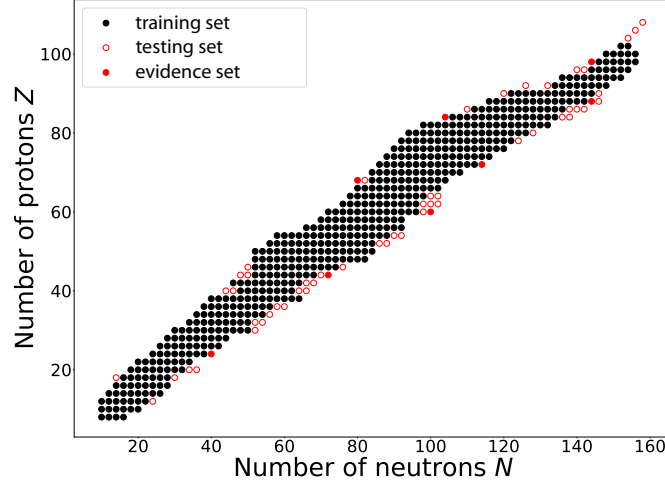
Figure 2. Training (black dots), testing (red circles), and evidence (red dots) datasets of two-neutron separation energies of even-even nuclei used in this study. The eight evidence nuclei are also included in the testing dataset. Each nucleus is represented by the number of protons Z and neutrons N. See text for details.

In the case of LBMM variant with GLD defined by Eq. (7), we take independent normal prior distributions with mean 0 and standard deviation 1 for the elements of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$. For LBMM with GPD, we used independent squared exponential kernels for the GP: namely,

$$c_k(x, x') = \eta_k e^{-\frac{(Z-Z')^2}{2\rho_Z^2} - \frac{(N-N')^2}{2\rho_N^2}}$$

characterized by three hyperparameters η_k, ρ_Z, and ρ_N. We have chosen to take the length-scale parameters ρ common to all nuclear models, but leave the GP intensity parameters η be different for each model, in order to ensure stability and convergence. As a result the GPD weights follow the frequency of the residuals for each model. In the case of Dirichlet distribution with GPD local mixture, we take independent normal priors with mean 0 and variance 1 for the GP mean parameter γ_k^∞, and independent Gamma priors for the three scale parameters η_k, ρ_Z and ρ_N. These priors are taken with respective parameters $(10, 2)$, $(5, 2)$, $(5, 2)$; this corresponds to slightly informative priors which helped to ensure convergence towards weights localized on an appropriate scale. The parameter γ_k^∞ determines the long range weight of each model, i.e., far from the training data. Note that taking a zero-mean GP, i.e., setting $\gamma_k^\infty = 0$, would amount to uniform weights far from the data.

When it comes to the global GBMM+L mixture, Eq. (1), we take for $\boldsymbol{\omega}$ independent uniform priors on $[0, 1]$. In practice, the simplex constraint is satisfied implicitly, without the need to apply Eq. (4). This confirms that all the individual models are well conceived. For the GBMM+D variant, we take for α a half-normal prior with standard deviation 2.

For $\delta_{f,k}$, we use the systematic correction for two-neutron separation energy residuals (i.e., differences between theoretical and measured two-neutron separation energies) computed in[29] using Bayesian Gaussian processes; these are fixed with no priors. The training dataset in[29] agrees with Fig. 2 up to a set of 5 additional nuclei. In what follows, we do not use these nuclei during training whenever uncorrected models are considered since they would be extrapolative from the GP's perspective. The consequences of this omission is negligible due to the overall size of the training set.

All model weights were trained with the S_{2n} values from the full training dataset with the exception of BMA, where we have used only the evidence set shown in Fig. 2 that consists of 8 nuclei. Similar to[8,28,29], we compute the BMA weights only on a set of representative nuclei because computing evidences on a large dataset inevitably leads to model selection. This happens due to the exponential and multiplicative nature of Gaussian likelihood which punishes large deviations more than it favors good fits (see[4,30] for details). These weights are in turn applied to obtain predictions for S_{2n}. The proton-rich limit of $\mathcal{X}$, determined by two-proton separation energies, was identified in the previous study[28].

Tables 1 and 2 summarize the results of our model variants and are discussed in the following paragraphs.

Uncorrected models versus corrected models

We first discuss the fidelity of individual models. To this end, we study root-mean-square (rms) deviations for all the modeling variants discussed in this paper. We consider raw model predictions and the predictions including the systematic corrections $\delta_{f,k}$. It is seen that for the individual models, the corrected variants generally outperform the raw predictions, see Ref.[18] for

Table 1. Rms deviations (in MeV) for all individual models, global (BMA, GBMM) and local (LBMM) mixtures. Values are provided with and without systematic corrections. For corrected models, we show only the test rms as the train rms and σ values are negligible after the GP fit. For abbreviations of BMA variants, see text.

| Model | Uncorrected models | | | Corrected models |
	Train	Test	σ	Test
SkM*	1.19	1.14	1.19(4)	0.66
SkP	0.84	0.74	0.83(3)	0.64
SLy4	0.99	0.81	0.99(3)	0.68
SV-min	0.77	0.63	0.77(2)	0.55
UNEDF0	0.77	0.63	0.77(2)	0.61
UNEDF1	0.75	0.50	0.75(2)	0.48
UNEDF2	0.85	0.67	0.84(3)	0.54
FRDM-2012	0.48	0.45	0.47(1)	0.38
HFB-24	0.42	0.40	0.42(1)	0.40
BMA(ex)	0.38	0.32	0.55(16)	0.35
BMA(MC)	0.39	0.32	0.56(17)	0.35
BMA(Lap)	0.40	0.32	0.57(17)	0.35
GBMM+L	0.33	0.31	0.33(1)	0.41
GBMM+D	0.33	0.31	0.33(1)	0.46
LBMM+GLD	0.29	0.35	0.30(1)	
LBMM+GPD	0.25	0.33	0.26(1)	

discussion. The exception is HFB-24, which has been carefully calibrated to experimental masses; in this case the correction term δ_f does not lead to a lower rms deviation on the testing dataset. We want to point out that mixing of corrected models should be done with caution. It is our empirical experience that *mixing (GBMM or LBMM) of previously corrected models can lead to overfitting* as one tends to fit the statistical models to the small leftover noise since the residuals of all corrected models on the training dataset are practically zero. This can be clearly observed in Table 1: the rms deviations for both local and global mixtures slightly outperform the combinations of corrected models on the the testing dataset. For instance, GBMM+D of uncorrected models gives 0.31 MeV rms deviation on the testing dataset as compared to 0.46 MeV on the corrected models and also as compared to 0.35 MeV rms deviations of BMA on corrected models. Since providing accurate extrapolations is the main focus of this work, in the following, we focus the discussions primarily on the uncorrected models.

BMA versus Global Mixtures

The BMA evidence integrals were calculated on the evidence dataset by means of Monte Carlo (MC), Laplace approximations, and in a closed form under conjugate priors. We denote the corresponding BMA variants as follows: BMA(MC), BMA(Lap) and BMA(ex), respectively (see Methods section for the calculations of BMA weights). In Table 2, we see that the model weights produced by BMA are consistent across all three evidence computation approaches, irrespective of whether the systematic correction has been applied. Averaging corrected models is more democratic as compared to the raw models: this is expected since the GP-based $\delta_{f,k}$ corrections fit the training data closely irrespective of the theoretical model. Still, the BMA testing rms with uncorrected and corrected models are very similar, with a slight preference for the uncorrected models.

The global mixtures of uncorrected models are generally slightly outperforming BMA on both training and testing datasets (see Table 1). This is in fact expected, given that these weights are designed to maximize the predictive power of the model mixture. Indeed, the GBMM+L model is the Bayesian counterpart to a frequentist linear regression against the different nuclear model predictions that minimizes the rms on the training set. This principle still holds despite the uniform prior used for the GBMM+L model which is very informative and plays a regularizing role that reduces overfitting and favors mixing. We can see that the Dirichlet mixture model yields very similar weights, with the benefits of having its weights natively located on the simplex. This comparison of global weights already speaks in favour of ruling out BMA for the purpose of combining imperfect models, in the favor of a Bayesian Dirichlet model. Table 1 also shows the posterior mean of the noise scale parameter σ for comparison. As a rule of thumb, a statistical model with a conservative (liberal) uncertainty quantification (UQ) would have σ above (below) the test rms, and a statistical model with high-fidelity UQ has σ close to the test rms. A more comprehensive view of UQ that reflects the fully propagated prediction uncertainty can be gleaned from Fig. 3 that shows the empirical
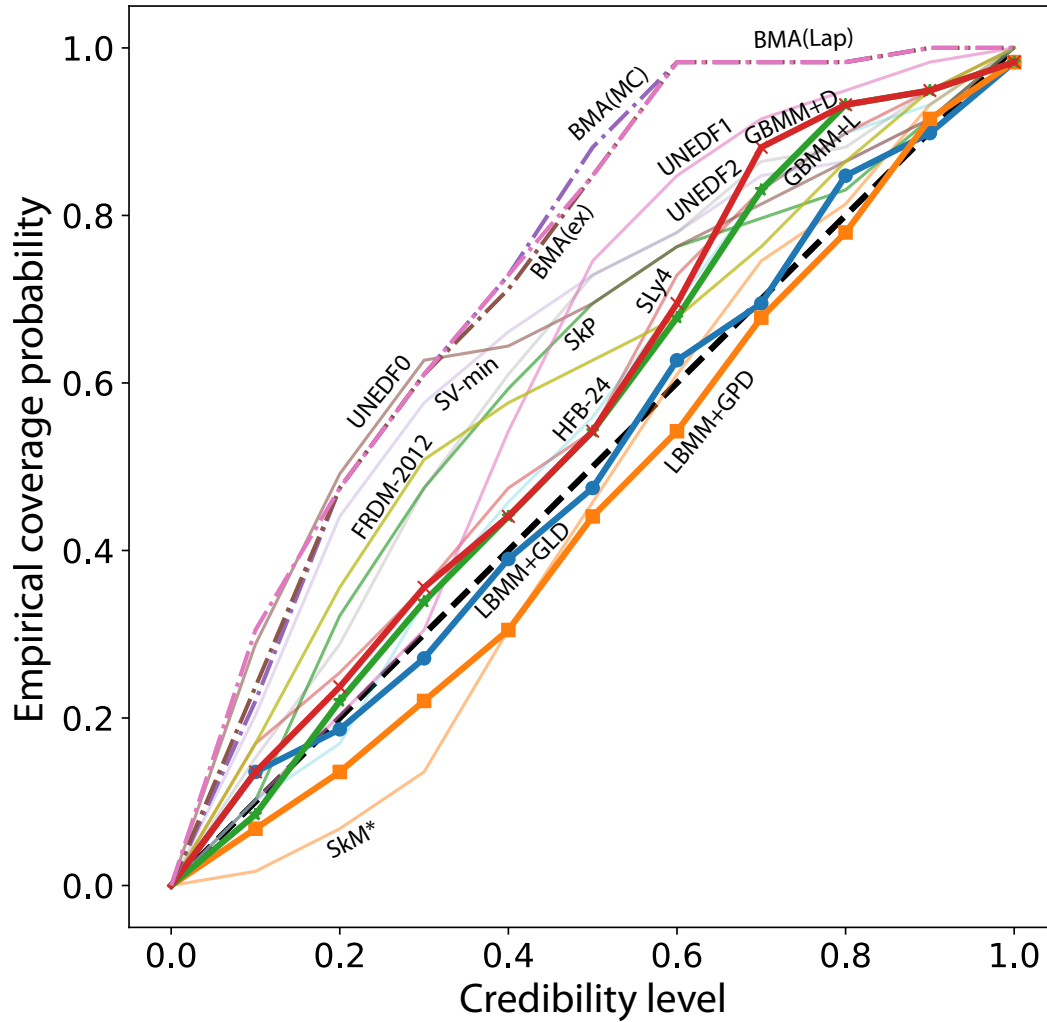
Figure 3. Empirical coverage probability for raw models without statistical correction together with BMA and BMM variants. The empirical coverage probability was calculated with equal-tailed credibility intervals. The reference line (diagonal) is marked by a dashed line.

coverage probability[46,47] (ECP). Each curve in Fig. 3 corresponds to the proportion of predictions in the testing dataset falling into the respective credible intervals (equal-tailed credible intervals). If the ECP curve closely follows the diagonal, then the actual fidelity of the credible interval corresponds to the nominal value. Thus we see that the *GBMM has both a superior prediction performance and a better UQ then BMA and individual models.*

Global versus Local Mixtures

The posterior mean of the LBMM+GPD weights are shown in Fig 4. The same plots but for LBMM+GLD are given in the Supplementary Information. As discussed earlier, mixing models locally corrected for systematic errors is highly susceptible to overfitting and we therefore focus on uncorrected models, i.e. without δ_f. Both LBMM variants show the dominance of the well-fitted HFB-24 mass model throughout the nuclear landscape. As expected, the simplistic linear dependence of weights $\boldsymbol{\omega}$ on (Z, N) in the GLD variant is insufficient to fully capture the complex local behaviour of the mass models learned by a more flexible GPD variant. While the HFB-24 mass model dominates, the final LBMM+GPD results involve other models, primarily FRDM2012, UNEDF0, and SkM*. The weight distribution naturally depends on the choice of models involved in the analysis. This suggests, that a preselection of diversified models to be used in LBMM could also be considered beforehand.

In terms of the rms deviations, the GPD variant does better than the GLD local mixture, which reflects the ability of the GPD to capture the local performance of mass models. Local mixtures perform better than global mixtures on the training

Table 2. Global weights calculated on the training dataset with different methods: BMA on the evidence subset (see Fig 2) obtained from a closed form computation as well as Monte Carlo and Laplace approximations, and the two global mixtures GBMM+L and GBMM+D obtained over the whole training set. For compactness, the following abbreviations are used: SV=SV-min, UNEn=UNEDFn (n=0,1,2), FRDM=FRDM-2012, and HFB=HFB-24.

Model	SkM*	SkP	SLy4	SV	UNE0	UNE1	UNE2	FRDM	HFB
				Uncorrected models					
BMA(ex)	0.00	0.00	0.00	0.03	0.07	0.02	0.13	0.48	0.27
BMA(MC)	0.00	0.00	0.00	0.04	0.09	0.04	0.15	0.42	0.26
BMA(Lap)	0.00	0.00	0.00	0.03	0.09	0.04	0.16	0.46	0.22
GBMM+L	0.01	0.02	0.01	0.01	0.02	0.02	0.06	0.28	0.57
GBMM+D	0.01	0.02	0.00	0.00	0.02	0.01	0.07	0.29	0.57
				Corrected models					
BMA(ex)	0.00	0.00	0.00	0.01	0.00	0.44	0.06	0.47	0.02
BMA(MC)	0.00	0.00	0.00	0.02	0.01	0.41	0.09	0.43	0.03
BMA(Lap)	0.00	0.00	0.00	0.02	0.01	0.42	0.09	0.44	0.02
GBMM+L	0.11	0.12	0.11	0.13	0.08	0.11	0.13	0.10	0.11
GBMM+D	0.24	0.15	0.20	0.13	0.03	0.02	0.10	0.00	0.12

set, and than BMA on both training and testing sets; however they fall slightly behind global mixtures on the testing set. We attribute this to the difficult tuning of the statistical model which is very sensitive to the variations of the parameters and a limited testing dataset in terms of its distribution across the nuclear landscape. In fact, Markov chain Monte Carlo (MCMC) sampling from the Bayesian posterior distributions can be numerically unsatisfactory with conventional Metropolis samplers. This is due to the relative large number of parameters of LBMM. In order to achieve satisfactory convergence, we recommend using more sophisticated No-U-Turn sampler[48] that tends to perform well in scenarios with moderately large parameter spaces (see Methods section for more details). In terms of UQ fidelity, the LBMM variants clearly dominate over the global mixtures, BMA, and individual models since the ECP of their respective predictive credible intervals closely matches the nominal values (see Fig. 3).

Discussion

In this work, we propose and implement a Bayesian Dirichlet model mixing framework. The proposed method is illustrated by applying it to nuclear mass models to assess their local fidelity and improve predictability. Raw theoretical models and their statistically-corrected versions were considered to better understand the interplay between GP modeling, BMA, and the BMM frameworks.

Bayesian model mixing of raw models results in testing rms that are at least as good or better than Bayesian model averaging (irrespective of models being corrected) with clearly superior UQ. Thus, *improving model predictions through mixing rather than mixing of corrected models* leads to the best performance in terms of both prediction accuracy and UQ. Since BMM is trained on a sizable training set, it is also more robust to the choice of priors than BMA which can be very prior sensitive if the evidence in data is weak[1].

BMM of corrected models should be performed with caution as it may lead to overfitting. In this case, one likely achieves a better improvement with standard BMA based on a well chosen evidence set.

The LBMM+GPD variant achieved the smallest training error on the training dataset (0.25 MeV) which demonstrates that the LBMM well captures the local presences of individual models. Furthermore, the local mixtures clearly surpass all the other modeling strategies explored in this work in terms of UQ fidelity. This shows that the proposed hierarchical Dirichlet model for LBMM effectively represents and propagates uncertainties which is essential for mass modeling into unexplored domains[29].

The results of BMA depend on the choice of the evidence dataset. That is, by increasing the density of the evidence data in the region of interest, e.g., for applications or extrapolations, one can improve the predictive power of averaging procedure. Improvement in the performance of BMM can also be achieved by restricting the training dataset to the region of interest as opposed to training on the whole domain; this motivates our introduction of local BMM models.

The distributions of BMA and BMM weights also depend on the choice of theoretical models. Table 2 and Fig. 4 show that mixing a large set of models results in some having minimal contributions and point out to the existence of a class of models with
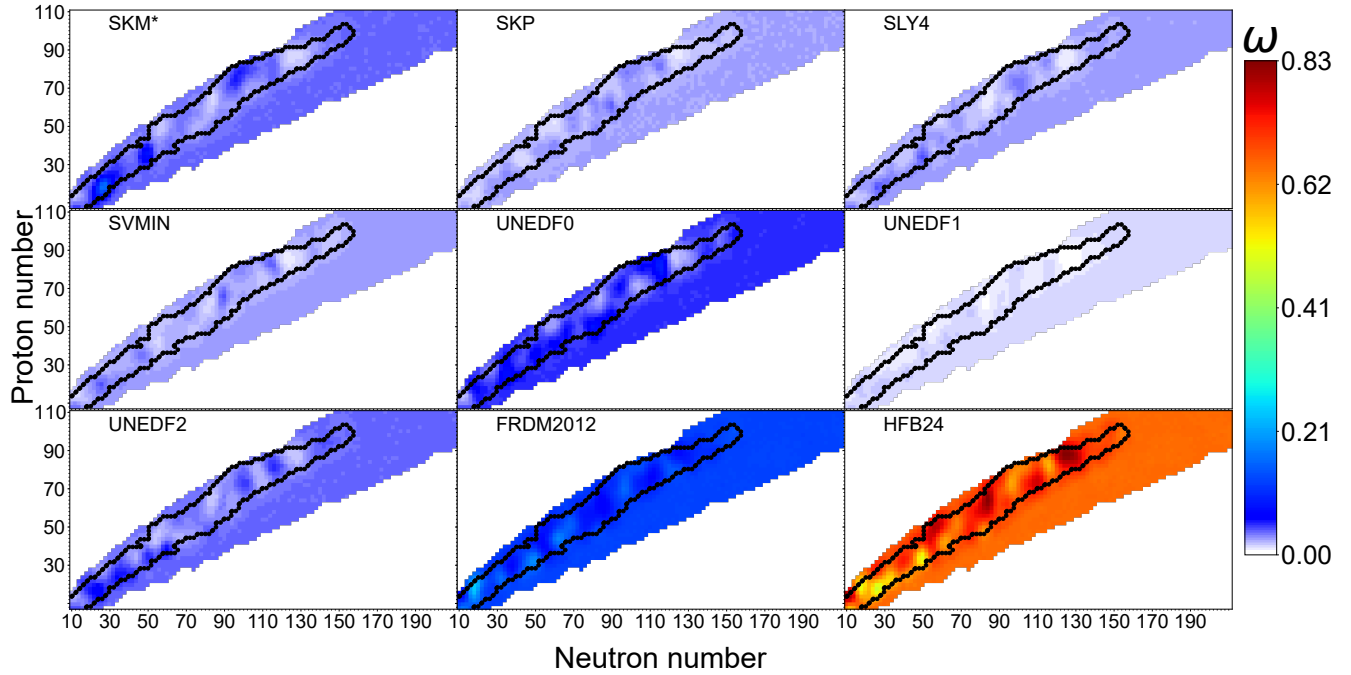
Figure 4. Posterior means of the local model weights in the LBMM+GPD variant across the nuclear landscape.

similar local preferences (e.g., UNEDFn class). This indicates that adding model preselection and model orhogonalization[49] to the BMM pipeline could lead to a further improvement in predictive performance. In fact, in the context of our GBMM+L model, it is well known that collinearity between the proposed theoretical models is a source of major instabilities.

References

1. Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Stat. Sci.* **14**, 382–417, DOI: 10.1214/ss/1009212519 (1999).

2. Wasserman, L. Bayesian model selection and model averaging. *J. Math. Psych.* **44**, 92–107, DOI: 10.1006/jmps.1999.1278 (2000).

3. Fragoso, T., Bertoli, W. & Louzada, F. Bayesian model averaging: A systematic review and conceptual classification. *Int. Stat. Rev.* **86**, 1–28, DOI: 10.1111/insr.12243 (2018).

4. Phillips, D. R. *et al.* Get on the BAND wagon: A Bayesian framework for quantifying model uncertainties in nuclear dynamics. *J. Phys. G* **48**, 072001, DOI: 10.1088/1361-6471/abf1df (2021).

5. Chmielecki, R. M. & Raftery, A. E. Probabilistic visibility forecasting using Bayesian model averaging. *Mon. Wea. Rev.* **139**, 1626–1636, DOI: 10.1175/2010MWR3516.1 (2011).

6. Montgomery, J. M. & Nyhan, B. Bayesian model averaging: Theoretical developments and practical applications. *Polit. Anal.* **18**, 245–270, DOI: 10.1093/pan/mpq001 (2010).

7. Zou, Y., Lord, D., Zhang, Y. & Peng, Y. Application of the Bayesian model averaging in predicting motor vehicle crashes. *US Dep. Transp.* (2012).

8. Neufcourt, L., Cao, Y., Nazarewicz, W., Olsen, E. & Viens, F. Neutron drip line in the Ca region from Bayesian Model Averaging. *Phys. Rev. Lett.* **122**, 062502, DOI: 10.1103/PhysRevLett.122.062502 (2019).

9. Everett, D. *et al.* Phenomenological constraints on the transport properties of qcd matter with data-driven model averaging. *Phys. Rev. Lett.* **126**, 242301, DOI: 10.1103/PhysRevLett.126.242301 (2021).

10. Clemen, R. T. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* **5**, 559–583, DOI: 10.1016/0169-2070(89)90012-5 (1989).

11. Le, T. & Clarke, B. A Bayes Interpretation of Stacking for $\mathscr{M}$-Complete and $\mathscr{M}$-Open Settings. *Bayesian Anal* **12**, 807 – 829, DOI: 10.1214/16-BA1023 (2017).

12. Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Anal.* **13**, 917 – 1007, DOI: 10.1214/17-BA1091 (2018).

13. Yao, Y., Pirš, G., Vehtari, A. & Gelman, A. Bayesian Hierarchical Stacking: Some Models Are (Somewhere) Useful. *Bayesian Anal.* **17**, 1043 – 1071, DOI: 10.1214/21-BA1287 (2022).

14. Semposki, A. C., Furnstahl, R. J. & Phillips, D. R. Interpolating between small- and large-g expansions using Bayesian model mixing. *Phys. Rev. C* **106**, 044002, DOI: 10.1103/PhysRevC.106.044002 (2022).

15. Yannotty, J. C., Santner, T. J., Furnstahl, R. J. & Pratola, M. T. Model mixing using Bayesian additive regression trees (2023). 2301.02296.

16. Utama, R. & Piekarewicz, J. Validating neural-network refinements of nuclear mass models. *Phys. Rev. C* **97**, 014306, DOI: 10.1103/PhysRevC.97.014306 (2018).

17. Niu, Z. M., Fang, J. Y. & Niu, Y. F. Comparative study of radial basis function and Bayesian neural network approaches in nuclear mass predictions. *Phys. Rev. C* **100**, 054311, DOI: 10.1103/PhysRevC.100.054311 (2019).

18. Neufcourt, L., Cao, Y., Nazarewicz, W. & Viens, F. Bayesian approach to model-based extrapolation of nuclear observables. *Phys. Rev. C* **98**, 034318, DOI: 10.1103/PhysRevC.98.034318 (2018).

19. Wu, X. H. & Zhao, P. W. Predicting nuclear masses with the kernel ridge regression. *Phys. Rev. C* **101**, 051301, DOI: 10.1103/PhysRevC.101.051301 (2020).

20. Yüksel, E., Soydaner, D. & Bahtiyar, H. Nuclear binding energy predictions using neural networks: Application of the multilayer perceptron. *Int. J. Mod. Phys. E* **30**, 2150017, DOI: 10.1142/S0218301321500178 (2021).

21. Gao, Z.-P. *et al.* Machine learning the nuclear mass. *Nucl. Sci. Tech.* **32**, 109, DOI: 10.1007/s41365-021-00956-1 (2021).

22. Shelley, M. & Pastore, A. A new mass model for nuclear astrophysics: Crossing 200 keV accuracy. *Universe* **7**, DOI: 10.3390/universe7050131 (2021).

23. Sharma, A., Gandhi, A. & Kumar, A. Learning correlations in nuclear masses using neural networks. *Phys. Rev. C* **105**, L031306, DOI: 10.1103/PhysRevC.105.L031306 (2022).

24. Pérez, R. N. & Schunck, N. Controlling extrapolations of nuclear properties with feature selection. *Phys. Lett. B* **833**, 137336, DOI: 10.1016/j.physletb.2022.137336 (2022).

25. Lovell, A. E., Mohan, A. T., Sprouse, T. M. & Mumpower, M. R. Nuclear masses learned from a probabilistic neural network. *Phys. Rev. C* **106**, 014305, DOI: 10.1103/PhysRevC.106.014305 (2022).

26. Mumpower, M. *et al.* Bayesian averaging for ground state masses of atomic nuclei in a machine learning approach. *Front.Phys.* **11**, DOI: 10.3389/fphy.2023.1198572 (2023).

27. Boehnlein, A. *et al.* Colloquium: Machine learning in nuclear physics. *Rev. Mod. Phys.* **94**, 031003, DOI: 10.1103/RevModPhys.94.031003 (2022).

28. Neufcourt, L. *et al.* Beyond the proton drip line: Bayesian analysis of proton-emitting nuclei. *Phys. Rev. C* **101**, 014319, DOI: 10.1103/PhysRevC.101.014319 (2020).

29. Neufcourt, L. *et al.* Quantified limits of the nuclear landscape. *Phys. Rev. C* **101**, 044307, DOI: 10.1103/PhysRevC.101.044307 (2020).

30. Kejzlar, V., Neufcourt, L., Nazarewicz, W. & Reinhard, P.-G. Statistical aspects of nuclear mass models. *J. Phys. G* **47**, 094001, DOI: 10.1088/1361-6471/ab907c (2020).

31. Hamaker, A. *et al.* Precision mass measurement of lightweight self-conjugate nucleus ^{80}Zr. *Nat. Phys.* **17**, 1408–1412, DOI: 10.1038/s41567-021-01395-w (2021).

32. Patra, S. *Constrained Bayesian Inference through Posterior Projection with Applications*. Ph.D. thesis (2019).

33. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).

34. Audi, G., Wapstra, A. & Thibault, C. The AME2003 atomic mass evaluation: (II). Tables, graphs and references. *Nucl. Phys. A* **729**, 337 – 676, DOI: 10.1016/j.nuclphysa.2003.11.003 (2003).

35. Wang, M., Huang, W., Kondev, F., Audi, G. & Naimi, S. The AME 2020 atomic mass evaluation (II). Tables, graphs and references. *Chin. Phys. C* **45**, 030003, DOI: 10.1088/1674-1137/abddaf (2021).

36. Bartel, J., Quentin, P., Brack, M., Guet, C. & Håkansson, H.-B. Towards a better parametrisation of Skyrme-like effective forces: A critical study of the SkM force. *Nucl. Phys. A* **386**, 79 – 100, DOI: 10.1016/0375-9474(82)90403-1 (1982).

37. Dobaczewski, J., Flocard, H. & Treiner, J. Hartree-Fock-Bogolyubov description of nuclei near the neutron-drip line. *Nucl. Phys. A* **422**, 103 – 139, DOI: 10.1016/0375-9474(84)90433-0 (1984).

38. Chabanat, E., Bonche, P., Haensel, P., Meyer, J. & Schaeffer, R. New Skyrme effective forces for supernovae and neutron rich nuclei. *Phys. Scr.* **1995**, 231 (1995).

39. Klüpfel, P., Reinhard, P.-G., Bürvenich, T. J. & Maruhn, J. A. Variations on a theme by Skyrme: A systematic study of adjustments of model parameters. *Phys. Rev. C* **79**, 034310, DOI: 10.1103/PhysRevC.79.034310 (2009).

40. Kortelainen, M. *et al.* Nuclear energy density optimization. *Phys. Rev. C* **82**, 024313, DOI: 10.1103/PhysRevC.82.024313 (2010).

41. Kortelainen, M. *et al.* Nuclear energy density optimization: Large deformations. *Phys. Rev. C* **85**, 024304, DOI: 10.1103/PhysRevC.85.024304 (2012).

42. Kortelainen, M. *et al.* Nuclear energy density optimization: Shell structure. *Phys. Rev. C* **89**, 054314, DOI: 10.1103/PhysRevC.89.054314 (2014).

43. Mass Explorer (2020). http://massexplorer.frib.msu.edu.

44. Möller, P., Sierk, A., Ichikawa, T. & Sagawa, H. Nuclear ground-state masses and deformations: FRDM(2012). *At. Data Nucl. Data Tables* **109-110**, 1 – 204, DOI: 10.1016/j.adt.2015.10.002 (2016).

45. Goriely, S., Chamel, N. & Pearson, J. M. Further explorations of Skyrme-Hartree-Fock-Bogoliubov mass formulas. XIII. the 2012 atomic mass evaluation and the symmetry coefficient. *Phys. Rev. C* **88**, 024308, DOI: 10.1103/PhysRevC.88.024308 (2013).

46. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.* **102**, 359–378, DOI: 10.1198/016214506000001437 (2007).

47. Gneiting, T., Balabdaoui, F. & Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **69**, 243–268, DOI: https://doi.org/10.1111/j.1467-9868.2007.00587.x (2007).

48. Homan, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1351–1381 (2014).

49. Clyde, M., Desimone, H. & Parmigiani, G. Prediction via orthogonalized model mixing. *J. Am. Stat. Assoc.* **91**, 1197–1208, DOI: 10.1080/01621459.1996.10476989 (1996).

50. Gelman, A. *et al. Bayesian Data Analysis* (CRC Pres, 2013), third edn.

51. Goriely, S., Chamel, N. & Pearson, J. M. HFB-24 mass formula (2020). http://www.astro.ulb.ac.be/bruslib/nucdata/hfb24-dat.

52. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in python using PyMC3. *PeerJ Comp. Sci.* **2:e55**, 1351–1381, DOI: 10.7717/peerj-cs.55 (2016).

Methods

BMA highlights

Let us consider the task of predicting observations from a physical process at new locations x^* using the observations $\boldsymbol{y} = (y(x_1), \ldots, y(x_n))$. The BMA posterior predictive distribution is

$$p(y(x^*)|\boldsymbol{y}) = \sum_{k=1}^{p} p(y(x^*)|\boldsymbol{y}, \mathcal{M}_k) p(\mathcal{M}_k|\boldsymbol{y}).$$

This is simply a linear combination of individual models' posterior predictive distributions. The global model weights are taken as the posterior probabilities $p(\mathcal{M}_k|\boldsymbol{y})$ that the model $\mathcal{M}_k$ is the true model as given by the Bayes' theorem:

$$p(\mathcal{M}_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{\ell=1}^{p} p(\boldsymbol{y}|\mathcal{M}_\ell)\pi(\mathcal{M}_\ell)},$$

where

$$p(\boldsymbol{y}|\mathcal{M}_k) = \int p(\boldsymbol{y}|\sigma_k, \boldsymbol{\delta}_{f,k}, \mathcal{M}_k) \pi(\sigma_k, \boldsymbol{\delta}_{f,k}|\mathcal{M}_k) d\sigma_k d\boldsymbol{\delta}_{f,k}$$

is the evidence (integral) of model $\mathcal{M}_k$ and $\pi(\sigma_k, \boldsymbol{\delta}_{f,k}|\mathcal{M}_k)$ is the prior density of model's parameters (noise scale σ_k and systematic discrepancy $\boldsymbol{\delta}_{f,k}$), $p(\boldsymbol{y}|\sigma_k, \boldsymbol{\delta}_{f,k}, \mathcal{M}_k)$ is the data likelihood, and $\pi(\mathcal{M}_k)$ is the prior probability that $\mathcal{M}_k$ is the true model – assuming that one of the models is true.

There is only a handful of statistical distributions under which the evidence integral can be expressed in a closed form. One such scenarios is linear regression models with conjugate priors; the statistical model $\mathcal{M}_k$ with a constant discrepancy term $\delta_{f,k}$ belongs to this case. For each model, let us consider the prior

$$\pi(\delta_{f,k}, \lambda_k|\mathcal{M}_k) = \pi(\delta_{f,k}|\lambda_k, \mathcal{M}_k)\pi(\lambda_k|\mathcal{M}_k),$$

where $\delta_{f,k}$, conditionally on a theoretical model choice $\mathcal{M}_k$ and a precision parameter (the inverse of the variance) λ_k, follows a normal distribution with mean μ and variance $1/\lambda_k$. Let us further assign to the precision λ_k a gamma prior with shape parameter a and rate parameter b. Then, the evidence integral has the closed form solution:

$$p(\boldsymbol{y}|\mathcal{M}_k) = \frac{\Gamma(a_n)b^a}{\Gamma(a)b_n^{a_n}\kappa_n^{\frac{1}{2}}}(2\pi)^{-\frac{n}{2}},$$

where $a_n = a + \frac{n}{2}$, $b_n = b + \frac{1}{2}\sum_{i=1}^{n}(d_i - \bar{d})^2 + \frac{n(\bar{d}-\mu)^2}{2(1+n)}$, $\kappa_n = 1 + n$, while denoting $d_i := y_i - y_k(x_i)$ and $\bar{d} := (\sum_i d_i)/n$. This solution can be obtained by simple but tedious algebraic manipulations, see Ref.[50] for details. As stated in the main manuscript, we use for the parameter σ a gamma prior with scale and rate parameters 5 and 10. In order to match the mean and standard deviation of $1/\sigma^2$ when σ is distributed according to the common Gamma prior with shape and scale parameters 5 and 10, the results for the closed form BMA were obtained under a gamma prior for the precision (inverse variance) parameter with shape and scale parameters 0.252 and 0.030.

When evidences cannot be obtained explicitly, a MC estimate can be computed as

$$\widehat{p_{MC}(\boldsymbol{y}|\mathcal{M}_k)} = \frac{1}{n_{MC}}\sum_{i=1}^{n_{MC}} p(\boldsymbol{y}|\boldsymbol{\delta}_{f,k}^{(i)}, \sigma_k^{(i)}, \mathcal{M}_k),$$

where $\{(\boldsymbol{\delta}_{f,k}^{(i)}, \sigma_k^{(i)}) : i = 1, \ldots, n_{MC}\}$ are samples from the prior distribution of model parameters $\pi(\boldsymbol{\delta}_{f,k}, \sigma_k|\mathcal{M}_k)$.

Alternatively, when the discrepancy term is considered constant, the evidence integral can be approximated by a closed form expression. A technique frequently used is Laplace's method of integral quadrature[1]:

$$\widehat{p_L(\boldsymbol{y}|\mathcal{M}_k)} = 2\pi|\widetilde{\Sigma}_k|^{\frac{1}{2}} p(\boldsymbol{y}|\tilde{\delta}_{f,k}, \tilde{\sigma}_k, \mathcal{M}_k)\pi(\tilde{\delta}_{f,k}, \tilde{\sigma}_k|\mathcal{M}_k),$$

where $\tilde{\sigma}_k$ and $\tilde{\delta}_{f,k}$ represent the posterior modes and $\widetilde{\Sigma}_k = (-\boldsymbol{D}^2 l(\tilde{\delta}_{f,k}, \tilde{\sigma}_k))^{-1}$ is the inverse of the Hessian matrix of second

derivatives of $l(\delta_{f,k}, \sigma_k) = \log p(\mathbf{y}|\delta_{f,k}, \sigma_k, \mathcal{M}_k) + \log \pi(\delta_{f,k}, \sigma_k|\mathcal{M}_k)$. For $\sigma_k \sim \text{Gamma}(a, b)$ and $\delta_{f,k} \sim N(\mu, s^2)$, we have

$$\frac{\partial^2 l(\delta_{f,k}, \sigma_k)}{\partial \sigma_k^2} = \frac{n - a + 1}{\sigma_k^2} - \frac{3\sum_i(y(x_i) - y_k(x_i) - \delta_{f,k})^2}{\sigma_k^4},$$

$$\frac{\partial^2 l(\delta_{f,k}, \sigma_k)}{\partial \sigma_k \partial \delta_{f,k}} = -\frac{2\sum_i(y(x_i) - y_k(x_i) - \delta_{f,k})}{\sigma_k^3} - \frac{\delta_{f,k} - \mu}{s^2},$$

$$\frac{\partial^2 l(\delta_{f,k}, \sigma_k)}{\partial \delta_{f,k}^2} = -\frac{n}{\sigma_k^2} - \frac{1}{s^2}.$$

The computation of evidence integral is simplified in the scenarios without the constant discrepancy term $\delta_{f,k}$ as the statistical model contains only a single parameter σ_k. We leave the details of this simple exercise in probability to the reader.

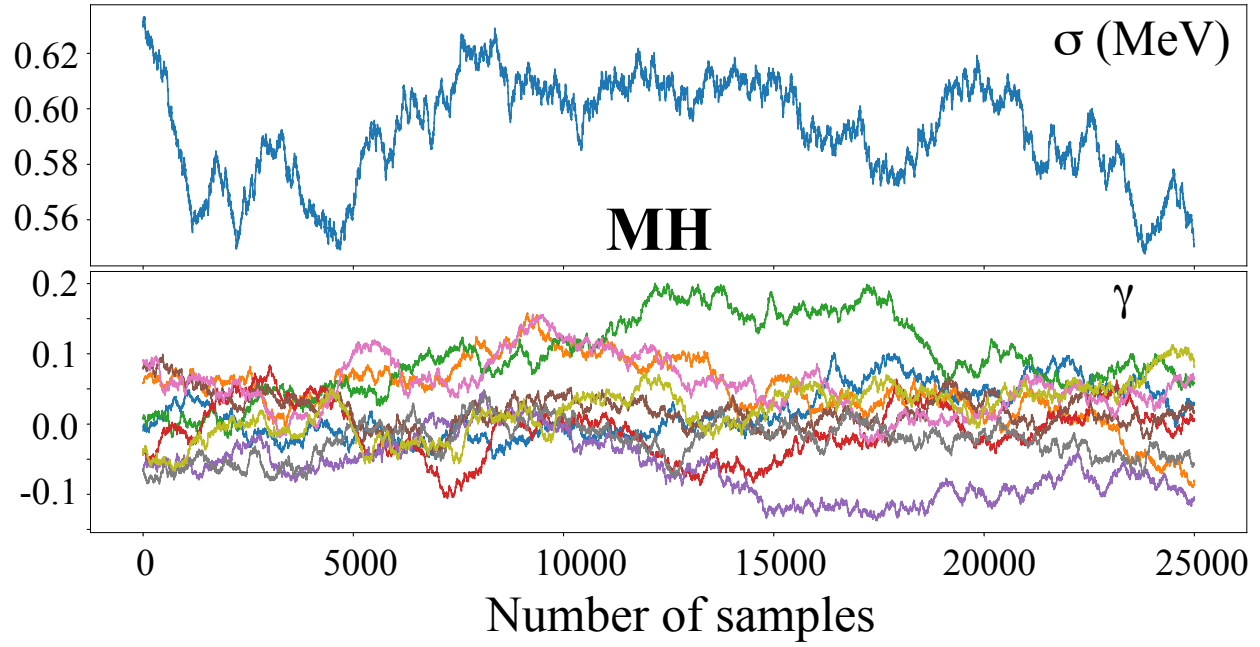### Application: summary of modeling choices

As a matter of clarity and to guarantee reproducibility of the results presented in section *Application: nuclear mass exploration*, Table I lists parameter choices and priors for each of the modeling variants discussed. Note that when we consider theoretical model without statistical correction, $\delta_{f,k}(x_i) := 0$.

METHODS Table I. Summary of statistical models, their parameters, and priors used in section *Application: nuclear mass exploration*.
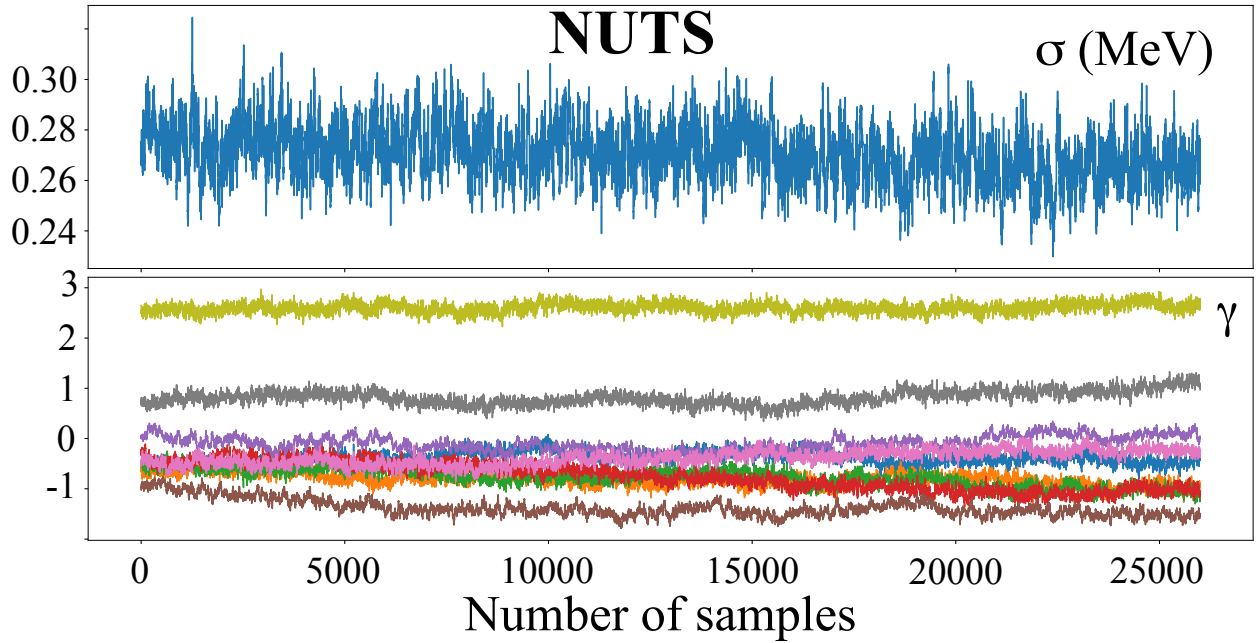
	Statistical Model	Prior distributions
BMA	$\mathcal{M}_k : y(x_i) = f_k(x_i) + \delta_{f,k}(x_i) + \sigma\varepsilon_i$	$1/\sigma \sim \text{Gamma}(0.252, 0.030),$ $\pi(\mathcal{M}_k) = 1/9$
GBMM+L	$y(x_i) = \sum_{k=1}^p \omega_k \left(f_k(x_i) + \delta_{f,k}(x_i)\right) + \sigma\varepsilon_i$	$\sigma \sim \text{Gamma}(5, 10),$ $\boldsymbol{\omega} \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$
GBMM+D	$y(x_i) = \sum_{k=1}^p \omega_k \left(f_k(x_i) + \delta_{f,k}(x_i)\right) + \sigma\varepsilon_i$	$\sigma \sim \text{Gamma}(5, 10),$ $\boldsymbol{\omega}|\boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}),$ $\boldsymbol{\alpha} \overset{\text{iid}}{\sim} \text{Half-Normal}(2^2)$
LBMM+GLM	$y(x_i) = \sum_{k=1}^p \omega_k(x_i) \left(f_k(x_i) + \delta_{f,k}(x_i)\right) + \sigma\varepsilon_i$	$\sigma \sim \text{Gamma}(5, 10),$ $\boldsymbol{\omega}(x_i)|\boldsymbol{\alpha}(x_i) \sim \text{Dirichlet}(\boldsymbol{\alpha}(x_i)),$ $\log(\alpha_k(x_i)) = \boldsymbol{\beta}_k^T x_i,$ $\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} N(0, 1)$
LBMM+GPM	$y(x_i) = \sum_{k=1}^p \omega_k(x_i) \left(f_k(x_i) + \delta_{f,k}(x_i)\right) + \sigma\varepsilon_i$	$\sigma \sim \text{Gamma}(5, 10),$ $\boldsymbol{\omega}(x_i)|\boldsymbol{\alpha}(x_i) \sim \text{Dirichlet}(\boldsymbol{\alpha}(x_i)),$ $\log(\alpha_k(x_i)) \sim \text{GP}(\gamma_k^\infty, c_k(x_i, x_i'))$, where $c_k(x_i, x_i') = \eta_k e^{-\frac{(Z_i - Z_i')^2}{2\rho_Z^2} - \frac{(N_i - N_i')^2}{2\rho_N^2}}$ and $\gamma_k^\infty \sim N(0, 1),$ $\eta_k \sim \text{Gamma}(10, 2),$ $\rho_N \sim \text{Gamma}(5, 2),$ $\rho_Z \sim \text{Gamma}(5, 2)$

MCMC computations

The MCMC approximate posterior distributions for all the modeling variants discussed in this work were obtained using the Hamiltonian Monte Carlo based No-U-Turn sampler (NUTS)[48]. In general, we obtained at least 50×10^3 samples from the posterior distributions after which we discarded half as a burn-in. While more conventional samplers such as Metropolis-Hastings (MH) algorithm[50] would be sufficient for both BMA and global mixtures, using NUTS is essential to achieve satisfactory convergence when it comes to LBMM. To illustrate this, we provide selected MCMC traceplots for LBMM+GPD variant with MH and NUTS approximations in METHODS Figs. 1 and 2, respectively. The NUTS whose performance tends to be superior to MH in scenarios with moderately large parameter spaces clearly shows the convergence of Markov Chain while the MH displays poor mixing.

METHODS Figure 1. Traceplots of the scale parameter σ and GP mean parameters γ_k^∞ obtained via the Metropolis-Hastings algorithm in the LBMM+GPD variant.

METHODS Figure 2. Similar as in METHODS Fig. 1 but for the No-U-Turn sampler.

Data availability

The experimental data used in this study comes from the publicly available measurements collected in AME2003[34] and AME2020[35]. The results of the nuclear DFT mass models are publicly available from the theoretical database MassExplorer[43]. The FRDM-2012 results were taken from the supplementary data of Ref.[44]. The HFB-24 mass predictions were taken from the Brusslib website[51].

Code availability

MCMC sampling from the Bayesian posterior distributions was performed with the Python package PyMC3 (version 3.3)[52]. The Bayesian calculations of the theoretical binding energies that support the findings of this study will be made publicly available in https://github.com/kejzlarv upon publication of this article.

Supplementary Information

Supplementary information is linked to the online version of the paper at https://www.nature.com/srep/.

Acknowledgements

Author contributions

V.K. and L.N. performed the Bayesian analysis. All authors discussed the results and prepared the manuscript.

Author information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.N. (witek@frib.msu.edu)

Competing interests statement

The authors declare that they have no competing financial interests.