

A permutation-invariant surrogate for Monte Carlo k_{eff} : neutron moderation as the dominant failure mode

Brian Neely

Independent Researcher, Seattle, WA, USA

Abstract

[TODO: Write last, once the results are final. Skeleton: MCNP is accurate but expensive; surrogates exist only for fixed geometries; we generate five controlled datasets (160,000 MCNP runs) of 1–20 fissile spheres of increasing generality; a Deep Sets network reaches 432 pcm near-critical MAE on fast systems but 1,616 pcm once moderators are allowed; a control dataset that removes only the moderators isolates moderation as the cause; the measured spectrum confirms it; and the error still supports a safe screen that skips 13.6% of runs with 0 of 3,415 supercritical test configurations missed. Journal limit: 250 words, no citations.]

Keywords: criticality, neutron moderation, surrogate model, Monte Carlo, MCNP, Deep Sets, machine learning

1. Introduction

The neutron multiplication factor k_{eff} is the critical metric in criticality safety analysis. It is the ratio of neutron generation to neutron loss from one generation to the next. A system is supercritical when $k_{\text{eff}} > 1$ and subcritical when $k_{\text{eff}} < 1$. ———Change to dunderstadt and hamilton. Because neutron generation times are short, in the order of 10^{-4} s for thermal neutron and 10^{-7} s for fast neutrons —AI—[?]—/AI—, small errors in k_{eff} compound to a large errors a system’s power change or reactivity, so k_{eff} must be calculated accurately. General purpose, continuous-energy Monte Carlo transport codes such as MCNP [?] are the accepted way calculate

*Corresponding author.

an arbitrary geometry’s k_{eff} . However, these codes are also expensive computationally requiring every configuration to have a new simulation and a criticality calculation simple geometry can take minutes of CPU time. Any study that evaluates many different configurations, such as a design search or a criticality-safety screening of possible arrangements, pays that cost for every configuration. A fast surrogate model that predicts k_{eff} from input parameters that describe the geometry of the system would remove most of the CPU limitations, but requires an accurate model with well understood errors and biases. Neural Networks and other AI/ML surrogates for neutronics are now widespread in nuclear reactor physics [?], and most use fixed geometry while something else varies such as material properties, depletion, etc. With a limited sample size, Radaideh et al. [?] obtained a test error of about 166 pcm for the GODIVA bare uranium sphere by perturbing its cross-section data. —Bei et al. [?] predict k_{eff} and the three-dimensional flux of a molten salt reactor, while Molczan et al. [?] fit a Gaussian-process surrogate with calibrated uncertainty to a pressurised-water-reactor pin cell, and Dave et al. [?] compare several regressors for the flux in two fixed research facilities. In criticality safety, Grechanuk et al. [?] use machine learning to predict the bias of criticality calculations from their sensitivity profiles.—/—

The geometric configuration of nuclear systems typically varies on a fixed lattice: fuel loading patterns within a fixed core size and arrangement, predicted by a DeepONet [?], or fuel placement on a grid, where an attention-based convolutional —neural— network learn from critical experiments to predict k_{eff} [?]. Lattice geometries allow a fixed configuration to be treated as an —image of fixed size. Variable number bodies with a continuous position systems do not have a grid and this setting: varying number of fissile bodies, sizes, positions and material properties has not been tested and what we studied here.

Varying geometry creates a representation problem, because a system of multiple spheres has no natural order so any model that takes a fixed-length vector imposes an arbitrary order on the spheres while padding the absent ones. In essence, exchanging the order of sphere 1 and sphere 2 in the vector would potentially change the model prediction while have no real world change to the system. The model would need to learn the same physics for every sphere slot along with their inter-reactions, dramatically increasing the model’s dimensionality. Deep Sets [?] removes this problem by their construction by applying one shared encoder to every sphere and pooling the

results with order-independent operations. This creates a model that doesn't depend on the input spheres order or the padding of absent spheres.

This experiment and paper asks how far can such a surrogate can be pushed, and what is its failure mode. We make four contributions.

Five controlled datasets (160,000 converged MCNP criticality calculations) of 1–20 fissile spheres of differing enrichment and density, of increasing generality: equal-sized spheres in vacuum, equal-sized spheres with a surrounding universe, varying sphere sizes with a random surrounding universe, varying sphere sizes composed of two additional, randomly assigned materials from a library of 22 including moderators with a random surrounding universe, and varying sphere sizes composed of two additional, randomly assigned materials from a library of 15 excluding moderators with a gas or vacuum surrounding universe. Each dataset is designed to test one source of variation from the previous datasets, so the cost of the additional complexity can be measured separately. Every stored k_{eff} has been re-derived from the retained MCNP output. —AI— A further 10,613 calculations repeat every water-universe configuration of the fourth dataset with thermal-scattering data. —/AI—

The dominant failure mode, isolated experimentally. Accuracy falls from 432 to 1,616 pcm near-critical mean absolute error when moderating materials were introduced. A fifth, control dataset that is identical to dataset 4 except every efficient moderator is removed shows that moderation, rather than material variety or geometry, causes most of the loss.

A spectrum-resolved diagnosis. Using the fraction of fissions caused by fast neutrons, read from MCNP's own output, we show that the surrogate's error on thermally driven systems is 2.2 times its error on fast systems, without systematic bias in either.

An honest statement of what the error supports. A mean error of hundreds of pcm is inadequate to replace MCNP, but it is good enough to decide which configurations needs the enhanced accuracy of MCNP. A safe screen, with its margin chosen on the training set and applied once to a locked test set, skips 13.6% of runs on the hardest dataset while missing 0 of its 3,415 supercritical configurations; accepting one missed configuration in a thousand raises the number of skipped MCNP runs to 27.0%.

—AI— The datasets, MCNP input decks and code are released with this paper (see Data availability). This paper confines itself to what the sur-

Table 1: The five datasets. Each dataset adds one source of variation to the one above it, except series 5, which repeats series 4 with the moderating materials removed to isolate the effect of moderation.

Series	Radius (cm)	Sphere material	Universe	Runs
1	0.1–15, shared	U	near-vacuum	20,000
2	0.1–15, shared	U	random	20,000
3	0.1–15, per sphere	U	random	20,000
4	0.1–25, per sphere	U + 22 nuclides	random	50,000
5	0.1–25, per sphere	U + 15, $A \geq 40$	gas only	50,000

rogate’s error reveals about the physics; how representation, training budget and numerical precision affect it is left to future work. —/AI—

Section ?? describes the datasets and the surrogate, Section ?? the results, Section ?? their limits, and Section ?? concludes.

2. Method

2.1. Geometry and materials

All datasets are composed of a set of 1 to 20 spheres placed at random, continuous, non-overlapping positions inside a 100 cm cube filled with a surrounding medium (the “universe”). Each sphere contains uranium at an enrichment drawn uniformly between 0.01% and 99%, density drawn uniformly depending on the dataset, its radius was drawn uniformly depending on dataset and, in the later datasets, composition are drawn at random. Five datasets add one source of variation at a time (Table ??).

In series 1 the surrounding universe is helium at $10^{-4} \text{ g cm}^{-3}$, a stand-in for vacuum; every later dataset draws it at random. In series 1–3 the spheres are uranium metal with a density between 18.7 and 20 g cm^{-3} . In series 4 and 5 every sphere contains uranium and exactly two further materials, each drawn independently and uniformly from a 22 and 15 item material library, respectively (this allows for duplicate material choices). The first material has a mass fraction drawn uniformly from $[0.01, 0.70]$, the second material’s mass fraction is from $[0.01, 1 - f_2]$, where f_2 is the first fraction, and uranium takes the remainder of the mass fraction, divided between U-235 and U-238 by the uniform selection of enrichment. The uranium mass fraction can approach zero depending on the other materials mass fraction. Sphere densities in series 4 and 5 are between 8 and 20 g cm^{-3} . The series 4 material

library contains 22 nuclides: light elements (Li-7, Be-9, B-10, C, Mg-24, Si, S), structural metals (Ti, Fe, Zr, Mo, Pb), the actinides Th-232, Pu-239, Pu-240, Am-241 and Cm-242, and the neutron poisons Gd-152, Sm-149, Cd-113, Hf-178 and Eu-151. The random universe is one of water ($0.9\text{--}1.0\text{ g cm}^{-3}$), air, argon, helium, or near-vacuum helium (10^{-4} g cm^{-3}).

Series 5 is the control experiment. It repeats series 4 exactly, except that every material light enough to slow neutrons appreciably is removed: the seven nuclides with mass number $A < 40$, for which the mean logarithmic energy loss per collision $\xi \approx 2/(A + 2/3)$ exceeds about 0.05, and the water universe. What remains are the heavy metals, the actinides, the poisons, and gas fills whose densities are too low to moderate. Any difference in accuracy between series 4 and series 5 can therefore be attributed to neutron moderation rather than to geometry, size or material variety. A relabelled variant of series 4, series 6, repeats every water-universe configuration with thermal-scattering data; it is described with its results in Section ??.

2.2. Criticality calculations

Each configuration is a separate criticality (KCODE) calculation with MCNP6.1 [? ?], which reports itself as MCNP6 version 1.0. Every nuclide specified isotopically uses continuous-energy ENDF/B-VII.1 data [?] from the `endf71x` library, evaluated at 250 K (library identifier `.86c`) [?]. MCNP reports adjusting these tables from that evaluation temperature to its default cell temperature of 293.6 K through the free-gas thermal treatment, so thermal scattering is treated at room temperature even though the tables are evaluated colder. Seven materials, used only in series 4 and 5, are specified as natural elements without a library identifier. For these MCNP uses the first matching table in its data directory, and those tables come from older evaluations, all at 293.6 K [?]: Si from ENDF/B-VI.0, S and Ti from ENDF/B-VI.8, Zr from ENDF/B-VI.1, Mo from ENDF/B-VI.0, Pb from ENDF/B-V.0, and Fe from a 1984 Los Alamos evaluation. No thermal-scattering ($S(\alpha, \beta)$) data are used, so hydrogen in the water universe and the light elements of series 4 are treated as free-gas scatterers at thermal energies. The labels are internally consistent, since every configuration uses the same data, but they are not a single-library, room-temperature evaluation, and thermal systems in particular are modelled approximately; we return to this in Section ??. Every calculation runs 10^4 neutron histories per cycle, from an initial guess of $k_{\text{eff}} = 1$, and averages k_{eff} over 100 active cycles after a number of inactive cycles that depends on the dataset: 50 by default, 200 for

series 2, whose geometries converge more slowly. Fission-source convergence is checked with MCNP’s Shannon entropy diagnostic [?]. A calculation that fails this check is rerun with 200 and then 400 inactive cycles before it is rejected. The fission-source bank holds 10^5 sites.

A calculation is accepted only if (i) a final k_{eff} and standard deviation can be read from its output, (ii) the source-entropy convergence check passed, and (iii) the reported standard deviation is at most 150 pcm. The median standard deviation of the accepted calculations is 69 pcm.

2.3. Data integrity

Every stored k_{eff} and standard deviation was re-derived from the retained MCNP console output and compared with the stored label. Of the 160,000 labels, 130,117 matched exactly and 0 disagreed. The remaining 5, all in series 4, could not be checked because their console output was not retained. As a further check, one deck was resubmitted unchanged months after its original calculation and reproduced its stored k_{eff} to every recorded digit. A larger test arose in the campaign of Section ??, in which 490 decks were each calculated twice on two different machines: all 490 agreed in k_{eff} and in its standard deviation to every recorded digit. Each deck fixes its own random seed, so this is expected, but it confirms that no label depends on where it was computed.

Every aborted calculation was classified from MCNP’s own error message. Two defects in the input generator removed configurations non-randomly. First, the fission-source bank was too small. A highly supercritical system produces more fission sites per cycle than the default bank holds, and MCNP aborted 8.6% of attempted series-4 configurations and 8.0% of series-5 ones with a source overrun. Second, a constituent’s mass fraction could round to zero, which MCNP rejects. This removed 0.8% of series-4 and 0.8% of series-5 configurations, namely those containing a nearly fuel-free sphere. Both were fixed part way through generation, raising the acceptance rate per calculation from 85.3% and 87.4% to 99.9% and 99.8% for series 4 and 5; 62% and 67% of their configurations predate the fix.

Comparing the configurations accepted before and after the fix measures the resulting bias directly. The median k_{eff} rises from 1.18 to 1.21 in series 4 and from 1.11 to 1.15 in series 5, and the share of configurations above $k_{\text{eff}} = 1.5$ from 20% to 23% and from 19% to 22% (two-sample Kolmogorov–Smirnov distance 0.04, $p < 10^{-15}$ in both). Series 4 and 5 are therefore measurably but mildly depleted in highly supercritical configurations. In

series 1–3 the overrun affected at most 0.18% of configurations, and the bias is negligible. We disclose this rather than regenerate, and return to it in Section ??.

Two further exclusions do not depend on the defects. 14.8% of attempted series-1 configurations were rejected because their fission source failed to converge even with extended inactive cycles. These are excluded rather than trusted, which removes some of the most loosely coupled arrays. The remaining rejections were infrastructure failures, such as a server that could not find its data directory, and are unrelated to the physics.

2.4. Surrogate model

Each configuration is represented as a set of spheres. A sphere is described by its position, radius and density, the mass fractions of its constituents, and a one-hot encoding of each constituent’s identity over the 29-nuclide vocabulary. The universe is described the same way, by its density, its constituents’ fractions and their identities. This identity encoding was chosen over three physically derived alternatives by the same five-fold cross-validated procedure used throughout. On series 4 it gave a near-critical MAE of 1,873 pcm, against 3,519 with tabulated microscopic cross sections and atomic mass, 5,772 with cross sections but no sphere positions, and 6,158 with cross sections alone. On series 5 the four lie between 541 and 789 pcm, with identity essentially tied for best at 535 pcm. No encoding is best on every dataset. On the uranium-only series the position-free cross-section encoding is slightly better (best 323–473 pcm), but the conclusions of Section ?? do not depend on this choice: with each series’ best encoding the series-4 to series-5 ratio is still 3.5.

The network is a Deep Sets model [?]. A shared encoder of three fully connected layers of width 256 maps each sphere to a feature vector. The vectors are pooled by their sum, mean and maximum, which do not depend on the order of the spheres. The pooled vector is concatenated with the universe features and the number of spheres and passed through a head of three layers of width 512, 256 and 128 to a single output, k_{eff} . The model has 754,177 parameters.

Training minimises a squared error in which each configuration is weighted by $1+4\exp[-((k_{\text{eff}} - 1)/0.1)^2]$, concentrating accuracy near criticality, where it matters. Training is deliberately patient: up to 3000 epochs, early stopping after 80 epochs without improvement on a validation split, and the learning

rate halved after 25 stagnant epochs, down to 10^{-7} . Five independently initialised models are averaged. Training ran in TensorFlow 2.19 on an NVIDIA RTX 3090 in full 32-bit floating point. The GPU’s default TensorFloat-32 matrix mode, which keeps only a 10-bit mantissa, was explicitly disabled. On an earlier, 10,000-run version of series 3, five CPU-trained seeds of the same model gave out-of-fold MAEs of 507–534 pcm; on the GPU with TensorFloat-32 enabled it gave 579 pcm, 4.8 standard deviations outside that spread, and with it disabled 534 pcm, inside it.

2.5. Evaluation

Each dataset is divided once into a *development* set and a *locked test* set (4,000 configurations for series 1–3, 5,000 for series 4 and 5), drawn at random with a recorded seed. No model is trained on the test set and no choice in this paper — the feature set, the training schedule, the screening margin — was made with reference to it. It is scored once, and those are the numbers reported as results.

Within the development set we use five-fold cross-validation: every development configuration is predicted by models that never saw it, which gives the large, complete sample needed for the tail and subgroup analyses below. Each of the five fold models also predicts the locked test set, and a configuration’s test prediction is the mean over those five models; with five random seeds averaged as well, a test prediction is a mean of 25 models against five for a development prediction, which slightly favours the test figures. We report mean absolute error (MAE) in pcm ($1 \text{ pcm} = 10^{-5}$ in k_{eff}), over each whole set and over the near-critical band $0.95 \leq k_{\text{eff}} < 1.05$ where criticality-safety decisions are made, and alongside it MAE/σ_k , the error relative to that dataset’s spread of k_{eff} , since absolute pcm figures are not comparable between datasets of different range.

We also evaluate the surrogate as a screen. A *safe screen* skips a configuration without an MCNP calculation only if its predicted k_{eff} is below $1 - m$, for a margin m ; it never skips a configuration predicted to be supercritical. The margin is the smallest value for which no truly supercritical configuration in the *development* set would be skipped, and it is then applied once to the locked test set. That makes the reported skip rate and miss count a prospective estimate of what the screen would do on configurations it has never seen.

Table 2: Accuracy on each dataset’s locked test set, five-model ensemble, same architecture, training and feature set throughout. σ_k is the spread of k_{eff} in that dataset; MAE/σ_k puts the error on a scale that is comparable across datasets whose k_{eff} ranges differ.

Series	R^2	MAE (pcm)	Near-critical MAE (pcm)	σ_k	MAE/ σ_k	Adds
1	0.9998	413	456	0.43	0.0096	baseline
2	0.9995	503	406	0.45	0.0113	universe
3	0.9985	397	315	0.29	0.0137	per-sphere sizes
5	0.9996	520	432	0.45	0.0114	materials, fast only
4	0.9942	1,823	1,616	0.44	0.0415	+ moderation

3. Results

3.1. Accuracy across the complexity ladder

Series 1–3 share a training-set size (16,000 runs) and differ only in geometric generality: a surrounding universe, then a radius drawn per sphere rather than once per system. Near-critical error does not grow across them (456, 406 and 315 pcm), so geometric generality is close to free for this surrogate.

The raw pcm figures alone would overstate that conclusion, because the datasets do not span the same range of k_{eff} . Series 1 and 2 give every sphere in a configuration the same radius, so a configuration is either all-small and deeply subcritical or all-large and strongly supercritical, spreading k_{eff} widely ($\sigma_k = 0.43$ and 0.45). Series 3 draws each radius independently, which averages out and concentrates k_{eff} near unity ($\sigma_k = 0.29$). Measured against that spread, series 3 is in fact the hardest of the three ($\text{MAE}/\sigma_k = 0.0137$ against 0.0096 and 0.0113). Both columns are reported throughout for that reason: the pcm column is what a criticality-safety user needs, and the scale-free column is what makes datasets comparable.

Series 4 and 5 were generated at 50,000 runs against 20,000 for series 1–3, so a step in the ladder would otherwise confound complexity with training-set size. Table ?? therefore repeats series 4 and 5 trained on 16,000 development runs, the same as series 1–3, and scored on a subset of the same locked test sets.

At a matched training-set size, material variety and moderation both cost substantially: $\times 3.1$ and $\times 3.5$ respectively in pcm, and $\times 2.3$ and $\times 2.9$ scale-free. Moderation is the larger step on either measure, and it is the only one isolated by a controlled experiment (Section ??), but material variety

Table 3: The two costly steps, at a training-set size matched to series 1–3. Near-critical MAE on the locked test sets, and the same ratios measured scale-free.

Dataset (16k training)	Near-critical MAE (pcm)	Step	
series 3, uranium only	315	—	
series 5, + 14 materials	977	$\times 3.1$	($\times 2.3$ scale-free)
series 4, + moderation	3,405	$\times 3.5$	($\times 2.9$ scale-free)

is not free. Training on the full 45,000 development runs instead of 16,000 lowers the error by $\times 2.1$ (series 4) and $\times 2.3$ (series 5), so the surrogate is still data-limited at this scale.

One further difference between the datasets is the shape of the k_{eff} distribution inside the near-critical band itself, which is not the same from one dataset to the next. Re-scoring each series on a common band distribution — every 0.005-wide interval of the band weighted equally — leaves the ladder unchanged: 456, 404, 321, 443 and 1,567 pcm for series 1, 2, 3, 5 and 4 (95% bootstrap interval 1,370–1,806 pcm for series 4). The ordering is therefore a property of the datasets, not of where within the band their configurations happen to fall.

3.2. Why a set representation

A flat network that takes the spheres as a fixed-length, padded vector is the obvious alternative. Given the same training schedule, loss weighting and five-model averaging, and its own best feature set, the flat network reaches a near-critical MAE of 6,324, 8,178 and 11,497 pcm on series 3, 5 and 4, against 315, 432 and 1,616 pcm for the set model, despite having about five times as many parameters. The schedule used here was tuned for the set model, so it is fair to ask whether it handicaps a baseline with that many more parameters. On its original, shorter schedule — earlier stopping, a higher learning-rate floor and no loss weighting — the flat network reaches 6,255, 9,284 and 16,852 pcm on the same three datasets: better on some datasets and worse on others, and on every one still well behind the set model.

3.3. Moderation is the dominant failure mode

Series 5 differs from series 4 only in the removal of moderating materials. The near-critical error falls from 1,616 to 432 pcm, a factor of 3.7.



Figure 1: Near-critical mean absolute error up the complexity ladder, on each dataset’s locked test set. Solid bars use the full dataset; hatched bars repeat series 5 and 4 with the training set cut to 16,000 runs, the same as series 1–3. The dotted line is the median MCNP standard deviation, 69 pcm, the floor set by the labels themselves. Note the logarithmic scale.

3.4. The measured spectrum

The control experiment shows that removing moderators removes most of the error. To see that the error follows the neutron spectrum itself, rather than some other property of the moderating materials, we measured the spectrum directly. For a random sample of 400 series-4 configurations we recovered from MCNP’s output the fraction of fissions caused by fast neutrons, which is a property of each configuration and does not depend on the surrogate. 17% of the sample is thermal-driven, with fewer than a quarter of fissions caused by fast neutrons.

The surrogate’s error grows as the spectrum softens (Spearman correlation between absolute error and fast-fission fraction -0.18). The mean absolute error on the 68 thermal-driven configurations is 4,035 pcm, against 1,817 pcm on the 168 fast configurations (more than 90% of fissions fast): a factor of 2.2 (95% bootstrap interval 1.5–3.2). The error is not systematically signed in any band. Its median is within a few hundred pcm of zero throughout, so the surrogate is not biased on thermal systems, only less precise.

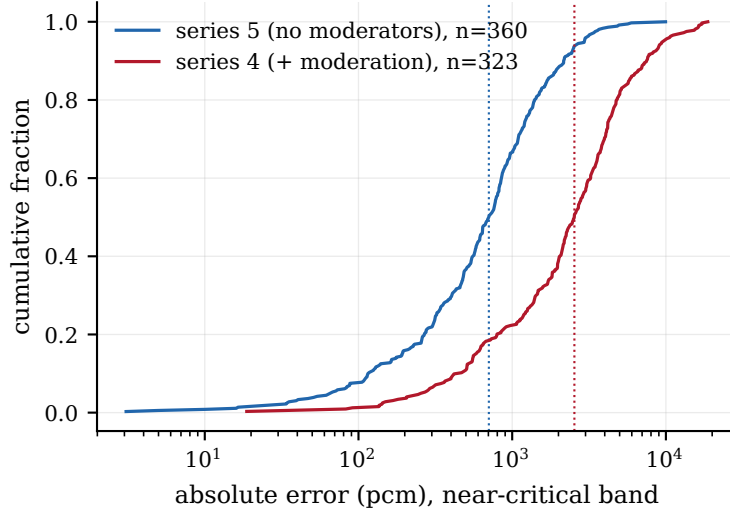


Figure 2: The moderation control at a matched training-set size: the distribution of absolute error in the near-critical band for series 4 and series 5, which differ only in whether moderating materials are present. Dotted lines mark the medians. The gap is not confined to a tail — it is the whole distribution.

3.5. What the error supports: screening

Moderation degrades the screen without destroying it. On series 5 a margin of 0.023 skips 36.6% of all configurations, which is 95% of the subcritical ones; on series 4 the margin required is 0.312 and the screen skips 13.6%. Neither missed a supercritical configuration in the locked test set.

What sets that margin is the extreme tail of the error distribution, not its typical size. The binding case in the series-4 development set is a single supercritical configuration predicted at 0.688; the next worst is predicted at 0.794. Allowing a miss budget of one supercritical configuration in a thousand — still chosen on the development set, still applied once to the locked test set — narrows the margin from 0.312 to 0.056 and raises the share of configurations skipped from 13.6% to 27.0%, at a cost of 1 missed supercritical configuration in the test set (0.3 per 1,000 supercritical ones). The same budget moves series 5 from 36.6% to 38.0% skipped, with 5 missed. Which point on that curve is acceptable is a safety judgement and not ours to make; Figure ?? gives the whole curve so that the judgement can be made explicitly.

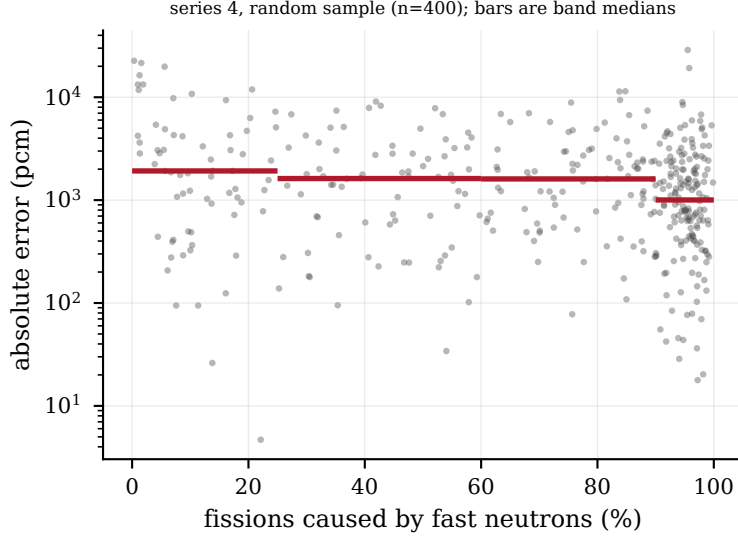


Figure 3: Absolute error against the fraction of fissions caused by fast neutrons, for the random sample of 400 series-4 configurations whose spectrum was measured from the MCNP output. Horizontal bars are the median error within each band. The spectrum is a property of the configuration, not of the surrogate.

3.6. Where the error lives

The development sets are large enough to break the error down by configuration property; the locked test sets agree but are too small to say much about any one subgroup. Two splits carry almost all of the structure. The first is the surrounding medium: in series 4 the near-critical error is 2,755 pcm for configurations in the water universe against 1,682 pcm for every other universe, a factor of 1.6. The same split appears in series 3, where the spheres are plain uranium metal and the only moderator in the problem is the water itself: 1,415 pcm against 302 pcm, a factor of 4.7 in a dataset that contains no moderating sphere material at all. The second is enrichment: the least enriched quarter of the series-4 configurations costs 1,949 pcm against 1,829 pcm for the most enriched quarter, which is the same effect seen from the other side, since a sphere of low enrichment needs a soft spectrum to approach criticality. Sphere count matters far less (2,042 pcm for 1–5 spheres against 1,914 pcm for 16–20).

Table 4: Safe screen with a zero missed-supercritical budget, on the locked test sets. The margin is chosen on the development set and applied once here, so these are prospective estimates. The last column counts missed supercritical configurations against the number of supercritical configurations in the test set.

Series	Runs skipped	Of subcritical runs	Margin m	Supercritical missed
5	36.6%	95%	0.023	0 / 3,081
4	13.6%	43%	0.312	0 / 3,415

3.7. Configurations larger than any seen in training

A randomly drawn test set asks only whether the surrogate interpolates. A criticality-safety user will also ask whether it extrapolates to systems larger than anything it was trained on. We therefore retrained the same model on the configurations of at most ten spheres (22,718 of them in series 4, 7,981 in series 3) and scored it on the 11- to 20-sphere configurations, which no model in that experiment had seen. The comparison is the ordinary model, trained across the full range of sphere counts and scored on exactly the same configurations.

On series 4 and 5 the answer is that size extrapolation is close to free. The holdout models reach 2,689 and 723 pcm against 1,836 and 554 pcm for the ordinary models, factors of 1.5 and 1.3. They also train on about half as much data, and Table ?? shows that cutting the development set from 45,000 to 16,000 runs alone costs $\times 2.1$ and $\times 2.3$; the holdout models sit between those two training-set sizes, so their deficit is accounted for by data volume with nothing left over. Their error is also flat in the number of spheres, from 2,621 pcm at 11–12 spheres to 2,592 pcm at 19–20.

Series 3 behaves differently. Its holdout model reaches 1,177 pcm against 378 pcm for the ordinary one, a factor of 3.1, and within that single model the error grows monotonically with the number of spheres: 456 pcm at 11–12 spheres against 2,352 pcm at 19–20. A gradient inside one model cannot be a training-set-size effect, so this is extrapolation failing. A plausible reason, which we state as a hypothesis, is that series 1–3 give every sphere in a configuration the same radius: the pooled sums that the set encoder produces then move steadily outside the range seen in training as spheres are added, whereas the per-sphere variety of series 4 and 5 means that ten-sphere configurations already generate pooled values overlapping those of twenty-sphere ones.

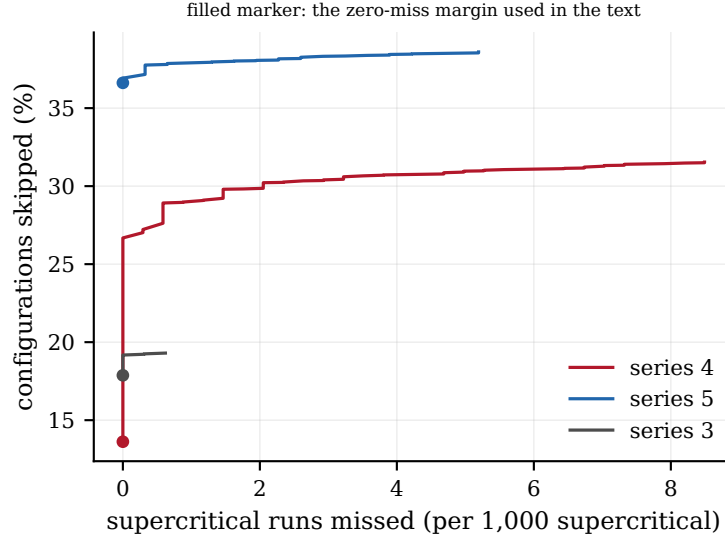


Figure 4: The screening trade-off. Each margin is chosen on the development set and its outcome plotted on the locked test set, so every point is a prospective estimate. Filled markers are the zero-miss margins of Table ???. Nearly all of the cost of the series-4 screen is paid for the last few missed configurations.

3.8. Does the surrogate know when it is wrong?

Averaging five independently initialised models gives a spread at no extra cost, and a spread is only useful if it is wider where the error is larger. It is, but weakly. On series 4 the rank correlation between the seed-to-seed spread and the absolute error is 0.27 (0.12 on series 5, 0.15 on series 3), and the median spread, 2,061 pcm, is of the same order as the mean absolute error it would have to predict, 1,873 pcm. Read as a standard deviation it is optimistic: 92% of series-4 configurations fall within two spreads of the truth, against the 95% a calibrated estimate would give. What the spread does do is mark the worst cases — the decile of widest spread has a mean error of 4,996 pcm against 1,251 pcm for the narrowest — so it is usable for deciding which configurations to send to MCNP anyway, and not as an error bar. The averaging is worth having for its own sake: it lowers the near-critical error from 2,602 to 1,873 pcm on series 4.

3.9. What the missing thermal-scattering data are worth

No deck carries an $S(\alpha, \beta)$ card, so bound-atom thermal scattering is absent from every label (Section ??). Hydrogen appears nowhere in the

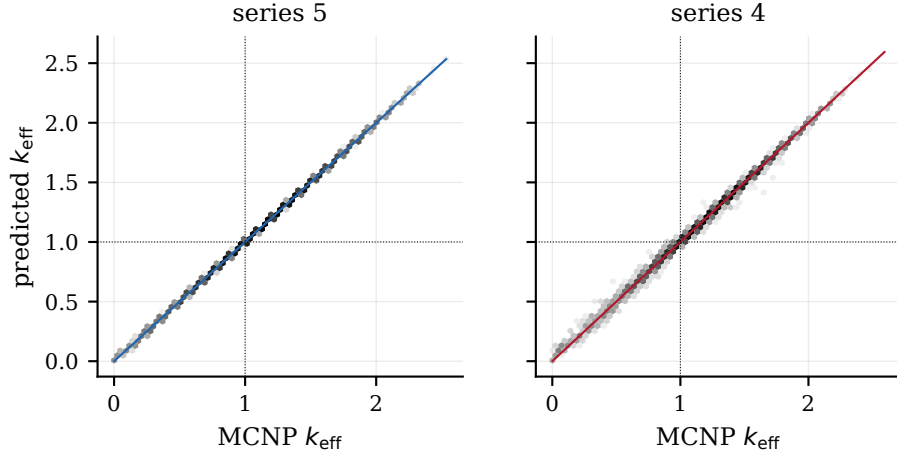


Figure 5: Predicted against calculated k_{eff} on the locked test sets (5,000 configurations each), on a logarithmic density scale. Series 5 (left) and series 4 (right) differ only in the presence of moderating materials; their near-critical errors are 432 and 1,616 pcm.

material library except in the water universe, which is 21% of series 4 and 18% of its near-critical configurations, so that is where the approximation bites. We recalculated all 10,613 of those configurations, 705 of them near critical, with MT1000 `lwtr.20t` added and nothing else changed — same geometry, same materials, same KCODE settings, same random seed.

The effect is large. Adding the thermal-scattering data lowers k_{eff} by $-2,399$ pcm on average (median $-1,956$), and by $-3,734$ pcm in the near-critical band, where the mean absolute shift is 3,749 pcm and the 95th percentile 6,780 pcm. 82% of the shifts exceed three times the combined MCNP standard deviation of the pair, so this is a difference between two physics models and not Monte Carlo noise. Rerunning one of these decks without the card reproduced its stored k_{eff} exactly, and MCNP’s free-gas temperature adjustment is identical with and without the card, so the shift is attributable to the thermal-scattering treatment alone. For scale, it is larger than the surrogate’s own near-critical error on series 4, 1,616 pcm.

Two things follow, and they should not be confused. The accuracy results of this paper are unaffected: the surrogate is trained and scored against one internally consistent set of labels, and it reproduces them as well as Section ?? reports whatever physics they encode. What does not follow is any claim about physical thermal systems. For the water-universe configurations the labels differ from a bound-scattering calculation by thousands of pcm, so

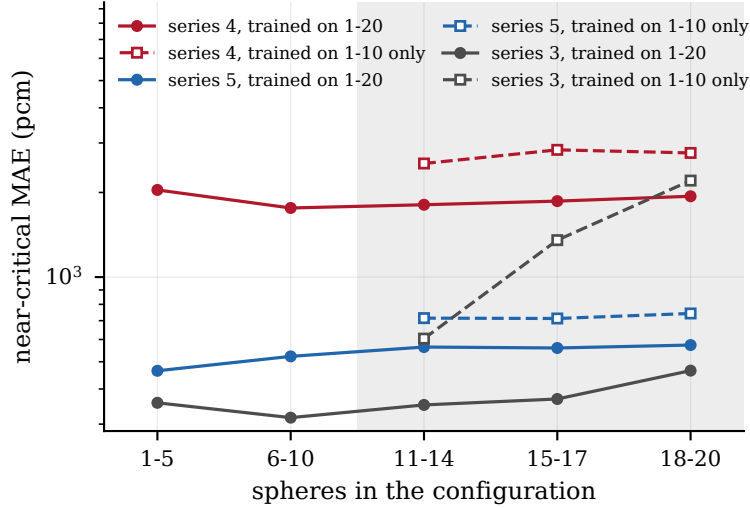


Figure 6: Near-critical error against the number of spheres. Solid lines are the ordinary models, trained on 1–20 spheres; dashed lines are the holdout models, trained only on configurations of at most ten spheres. Bins break at that limit, so the shaded region is exactly what the holdout models never saw. The series-4 and series-5 holdout curves run parallel to their ordinary counterparts; the series-3 one climbs.

those k_{eff} values should be read as the output of a defined transport model rather than as an estimate of a real system’s criticality. Whether the surrogate’s difficulty with moderation survives corrected labels is the subject of the next section.

3.10. Does the finding survive corrected labels?

If the free-gas treatment of hydrogen made moderated systems artificially hard to learn, the moderation penalty of Section ?? would be partly an artefact of the labels rather than a property of the physics. To test this we built series 6, a relabelled variant of series 4 in which every water-universe configuration is recalculated with the $S(\alpha, \beta)$ data added — 10,613 configurations, 21% of the dataset. No other label changes, because hydrogen appears nowhere else in the material library, so the configurations, the folds and the locked test sets are identical to those of series 4 and the two can be compared row by row, at the full training-set size and at the matched size of Table ?. The recalculation lowered k_{eff} by $-2,399$ pcm on average, which moved 497 configurations into or out of the near-critical band. 380 recalculations failed the acceptance rule of Section ??; they are kept and flagged rather than

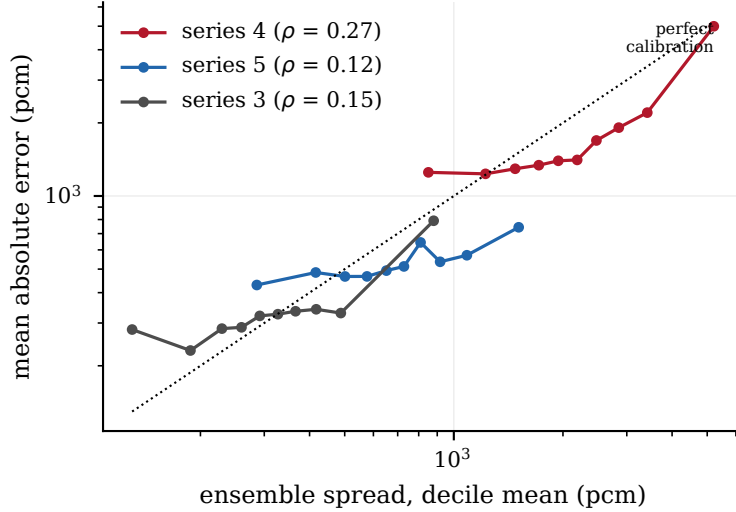


Figure 7: Ensemble spread against realised error, by decile of spread (development rows). A calibrated uncertainty would lie on the dotted line. The relationship is monotonic at the top end, which is what a triage signal needs, but the spread understates the error over most of the range.

dropped, so that the row sets stay aligned, and excluding them changes no conclusion. 26 decks that predate the generator’s fission-bank fix could only be rerun with a larger source bank, a storage setting that we verified leaves k_{eff} unchanged on a deck that runs either way.

The moderation penalty survives the correction (Figure ??). Trained identically on the corrected labels, the surrogate reaches a near-critical MAE of 1,568 pcm against 1,616 pcm on the original ones; measured against the series-5 control that is a penalty of $\times 3.6$ instead of $\times 3.7$. At the matched size the same comparison gives $\times 3.4$ against $\times 3.5$. The resolution of the comparison is set by the configurations whose labels did not change, on which the two models should agree and do to within a few per cent (ratios 0.96 and 1.01). On the near-critical configurations whose labels did change, the corrected-label model’s absolute error changes by +28 pcm on average at full size (73 configurations, 95% bootstrap interval -246 to $+305$) and by -141 pcm at matched size (57 configurations, interval -537 to $+273$); both intervals include zero. Corrected labels are therefore no easier and no harder for the surrogate to learn, and what makes moderated systems difficult for it is the moderation itself, not the approximation used to calculate it.

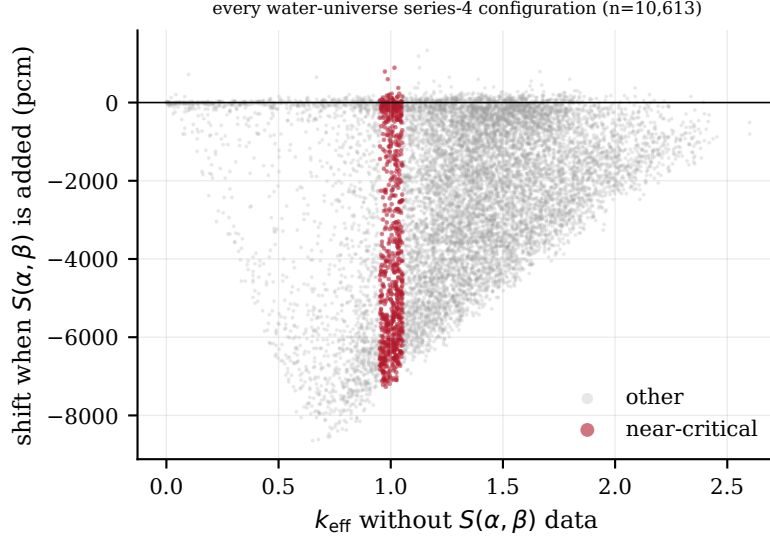


Figure 8: The change in k_{eff} when $S(\alpha, \beta)$ data for light water are added to every water-universe series-4 configuration, against the k_{eff} calculated without them. Nothing else in the deck changes.

4. Discussion and limitations

[TODO: Interpretation: why moderation is hard for this representation. A thermal system's k_{eff} depends on neutron slowing-down paths across and between spheres, which a pooled per-sphere encoding captures poorly; fast systems are closer to a sum of local contributions. Keep this as a hypothesis supported by the spectrum result, not a claim.] A natural test of that hypothesis is an architecture that models interactions between spheres explicitly, such as the attention-based Set Transformer [?], which the pooled Deep Sets encoder does not.

Limitations to state plainly:

- **Dataset bias.** The generator defects in Section ?? leave the datasets mildly depleted in highly supercritical configurations, series 1–3 entirely and series 4 and 5 in part.
- **Physically arbitrary mixtures.** Series 4 and 5 mix materials in random proportions that do not correspond to real fuels. They are a stress test of the representation, not a model of any real system.



Figure 9: The moderation penalty with and without thermal-scattering data. Near-critical MAE on the locked test sets for the series-5 control, series 4 (hydrogen treated as a free gas) and series 6 (the same configurations with bound hydrogen in water), at the matched training-set size and at full size. Error bars are 95% bootstrap intervals; the multipliers are each dataset’s penalty against series 5. Correcting the labels leaves the penalty unchanged at both sizes.

- Nuclear data.** All labels come from one code, MCNP6.1. The isotopic data are ENDF/B-VII.1 tables evaluated at 250 K, which MCNP adjusts to its default 293.6 K cell temperature through the free-gas treatment, and the seven natural elements resolve to older evaluations. No $S(\alpha, \beta)$ thermal-scattering data were used, so moderation by hydrogen and the light elements is modelled with free-gas scattering, an approximation that matters precisely in the thermal systems where the surrogate is weakest. The surrogate learns MCNP as configured, so its accuracy figures are unaffected, but the labels should not be read as a reference-quality evaluation of thermal systems, and no experimental benchmark is used beyond the code’s own validation. Section ?? measures the size of the $S(\alpha, \beta)$ omission directly: on water-universe configurations it is worth 3,749 pcm near critical, larger than the surrogate’s own error on series 4. This is the most serious limitation here. It does not affect any accuracy figure in this paper, because the surrogate is scored against the labels it was trained on, but it does mean the water-universe k_{eff} values are not estimates of a real system’s criticality.

What it does not do is create the paper’s central finding: retrained on corrected labels (Section ??), the surrogate shows the same moderation penalty, $\times 3.6$ against $\times 3.7$.

- **Closed material vocabulary.** The identity encoding cannot represent a material absent from training. A physically derived encoding could in principle, but on series 4 every one we tried was less accurate (Section ??).
- **The tail sets the margin.** The screen’s margin is set by the largest errors, not the typical one, so it is sensitive to rare configurations: on series 4 a single development configuration fixes the zero-miss margin, and a budget of one missed configuration in a thousand roughly doubles the share of runs the screen can skip (Section ??). A deployment would have to decide that budget explicitly rather than inherit it from the data.
- **Extrapolation in size.** On series 4 and 5, models trained only on configurations of at most ten spheres lose nothing beyond what their smaller training set costs anyway, and their error is flat in the number of spheres. On series 3, where every sphere in a configuration shares a radius, the same experiment degrades steadily with sphere count (Section ??). Size extrapolation is therefore not uniformly safe, and nothing here tests configurations larger than the 20 spheres the generator produces.
- **No calibrated uncertainty.** The ensemble spread ranks the hard configurations usefully but understates the error and is a weak predictor of it case by case (Section ??). A screen built on this surrogate has to rely on a margin fixed in advance rather than on a per-configuration error bar.

5. Conclusion

[TODO: Three or four sentences, written last: the surrogate handles geometric and material generality well; moderation is the dominant failure mode, shown by a control experiment and the measured spectrum; the error supports a safe screen even on moderated systems. Data and code availability statement.]

CRedit authorship contribution statement

Brian Neely: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review and editing, Visualization, Project administration.

Declaration of competing interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The sphere configurations, MCNP input decks and computed k_{eff} values for every dataset in this paper, including series 6, are available on Zenodo [TODO: dataset DOI]. The code that generated the inputs, trained and evaluated the models, and produced every number and figure in this paper is available at [TODO: GitHub URL, archived on Zenodo with its own DOI]. MCNP and the nuclear data libraries distributed with it are licensed through RSICC and are not included.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used Claude (Anthropic) in order to help design the analyses; to write and debug the software that generated the MCNP inputs, trained and evaluated the models and produced the figures; and to draft text that the author then rewrote. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- [1] J. J. Duderstadt, L. J. Hamilton, Nuclear Reactor Analysis, John Wiley & Sons, New York, 1976.
- [2] T. Goorley, M. R. James, T. E. Booth, F. B. Brown, et al., Initial MCNP6 release overview, Nuclear Technology 180 (3) (2012) 298–315. doi:10.13182/NT11-135.
- [3] R.-Z. Zhang, S.-F. Zhu, K. Wang, D. She, J.-P. Argaud, B. Bouriquet, Q. Li, H.-L. Gong, Artificial intelligence in reactor physics: current status and future prospects, Nuclear Science and Techniques 37 (6) (2026) 100. doi:10.1007/s41365-026-01928-z.
- [4] M. I. Radaideh, D. Price, T. Kozlowski, Modeling nuclear data uncertainties using deep neural networks, EPJ Web of Conferences 247 (2021) 15016, pHYSOR 2020: Transition to a Scalable Nuclear Future. doi:10.1051/epjconf/202124715016.
- [5] X. Bei, Y. Dai, K. Yu, M. Cheng, Three-dimensional surrogate model based on back-propagation neural network for key neutronics parameters prediction in molten salt reactor, Energies 16 (10) (2023) 4044. doi:10.3390/en16104044.
- [6] A. Molczan, Z. Malecha, W. Zacharczuk, Gaussian process surrogate model with uncertainty quantification for PWR pin-cell criticality prediction, Applied Sciences 16 (11) (2026) 5174. doi:10.3390/app16115174.
- [7] A. J. Dave, J. Yu, J. Wilson, B. Phillips, K. Sun, B. Forget, Empirical models for multidimensional regression of fission systems, arXiv:2105.14645 (2021). arXiv:2105.14645.
- [8] P. Grechanuk, M. E. Rising, T. S. Palmer, Using machine learning methods to predict bias in nuclear criticality safety, Journal of Computational and Theoretical Transport 47 (4–6) (2018) 552–565. doi:10.1080/23324309.2019.1585877.
- [9] Y. Lee, B. Jeon, D. Kim, Y. Yu, T. Park, DeepONet-based surrogate model for rapid predictions of neutron flux and multiplication factor in

- nuclear reactors, *Annals of Nuclear Energy* 234 (2026) 112326. doi: 10.1016/j.anucene.2026.112326.
- [10] W. Savage, L. Burnett, D. Price, Inverse critical experiment design via gradient optimization and a multigroup attention-based neural network architecture, arXiv:2606.04033 (2026). arXiv:2606.04033.
 - [11] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, A. J. Smola, Deep sets, in: *Advances in Neural Information Processing Systems* 30 (NeurIPS 2017), 2017.
 - [12] T. Goorley, M. James, T. Booth, F. Brown, et al., Initial MCNP6 release overview – MCNP6 version 1.0, Tech. Rep. LA-UR-13-22934, Los Alamos National Laboratory (2013).
 - [13] M. B. Chadwick, M. Herman, P. Obložinský, M. E. Dunn, Y. Danon, A. C. Kahler, et al., ENDF/B-VII.1 nuclear data for science and technology: Cross sections, covariances, fission product yields and decay data, *Nuclear Data Sheets* 112 (12) (2011) 2887–2996.
 - [14] J. L. Conlin, D. K. Parsons, S. J. Gardiner, M. G. Gray, A. C. Kahler, M. B. Lee, M. C. White, Continuous energy neutron cross section data tables based upon ENDF/B-VII.1, Tech. Rep. LA-UR-13-20137, Los Alamos National Laboratory (2013).
 - [15] J. L. Conlin, D. K. Parsons, S. J. Gardiner, M. G. Gray, A. C. Kahler, M. C. White, M. B. Lee, Listing of available ACE data tables, Tech. Rep. LA-UR-13-21822, Rev. 1, Los Alamos National Laboratory (2013).
 - [16] F. B. Brown, On the use of Shannon entropy of the fission distribution for assessing convergence of Monte Carlo criticality calculations, in: *PHYSOR-2006, ANS Topical Meeting on Reactor Physics*, Vancouver, BC, Canada, 2006, LA-UR-06-3737.
 - [17] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, Y. W. Teh, Set transformer: A framework for attention-based permutation-invariant neural networks, in: *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, 2019.