Space-filling experimental designs with constraints for numerical simulation in criticality safety studies

Amandine MARREL*‡

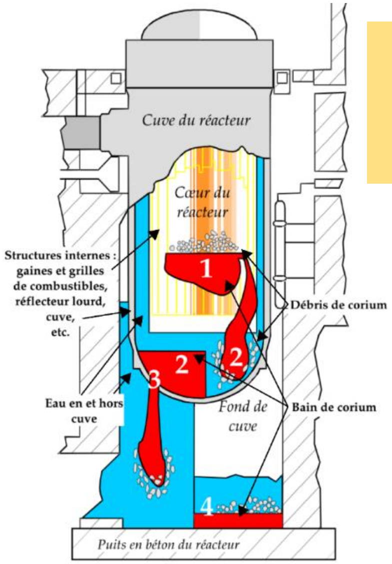
*In support of neutronics applications led by the CEA's expert Tangi NICOL**

*CEA Energy Division, IRESNE, DER, Cadarache, France

‡Avignon Université, LMA, UPR 2151, Avignon, France

Workshop - Experiments and Simulations: How to plan and optimally exploit them? Nov. 6-7th 2025, Avignon.

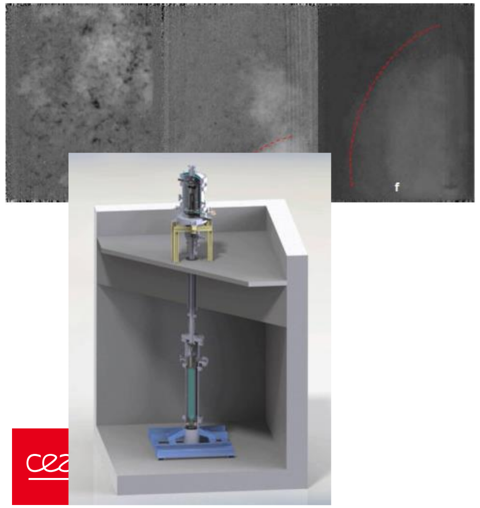
Numerical simulation for risk assessment



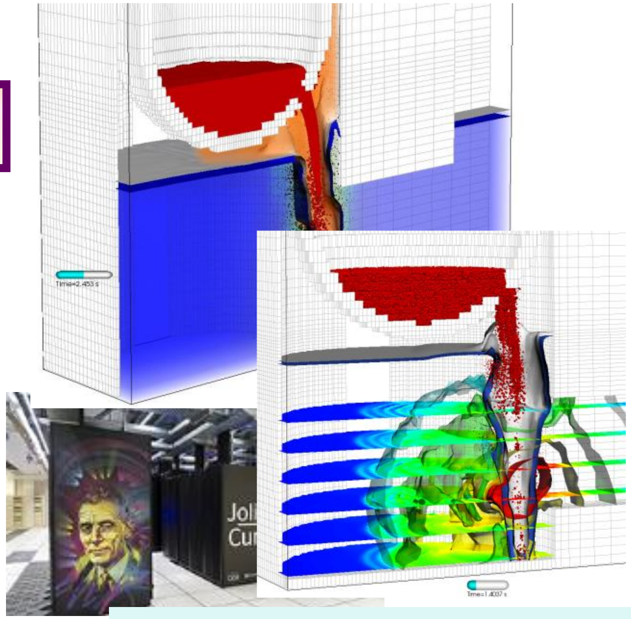
Safety studies supporting the operation and design of nuclear reactors
⇒ Simulation of accidental scenarios, material behavior, etc.

As computer simulation does not easily compare to either theory building or experimentation, several authors have suggested considering it as a "third way in science". (Heynman, 2010)

Physical Experiments



Validation



Numerical Simulation

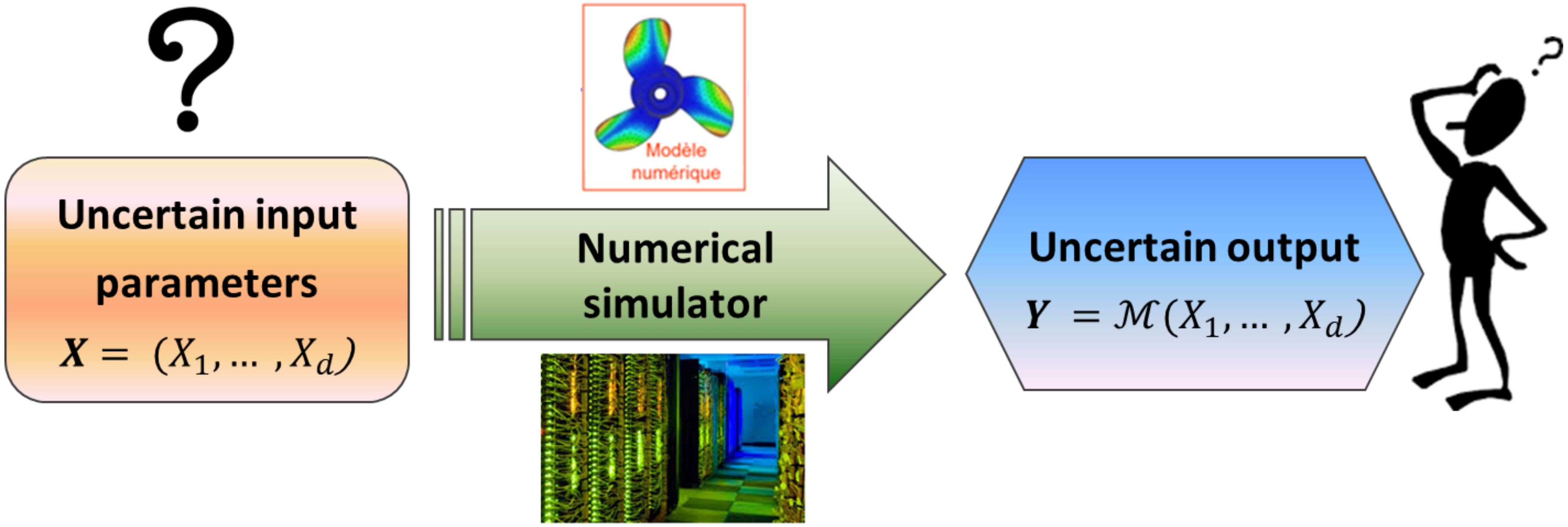
Uncertainties in numerical simulation

- **Random uncertainties:** physical parameters
- **Epistemic uncertainties:** model parameters (physical laws), design parameters, scenario parameters, etc.

Probabilistic and **statistical approaches** to take into account the different sources of **uncertainty in numerical simulation**

Numerical simulation for criticality risk assessment

- **Safety studies:** compute a failure risk (margins, rare events) with validated **computer/numerical models**
- **Numerical simulators:** fundamental tools to understand, model & predict physical phenomena
- **Uncertain input parameters,** related to physical or scenario parameters → **Explore the input domain with the simulator to ensure safety margins (even under worst configuration)**



Comparative analysis of neutronic calculation options for safety-criticality studies

- **Framewok:** two types of simulations using 2 different calculation codes are used for safety-criticality studies within the Scientific Calculation Tool dedicated to safety-criticality studies, developed by CEA and ASNR in collaboration with Orano, Framatome, and EDF.

1. **A simplified modeling** used by the industrial partner

2. **A reference modeling** without approximations, evaluated using the calculation code TRIPOLI-4®

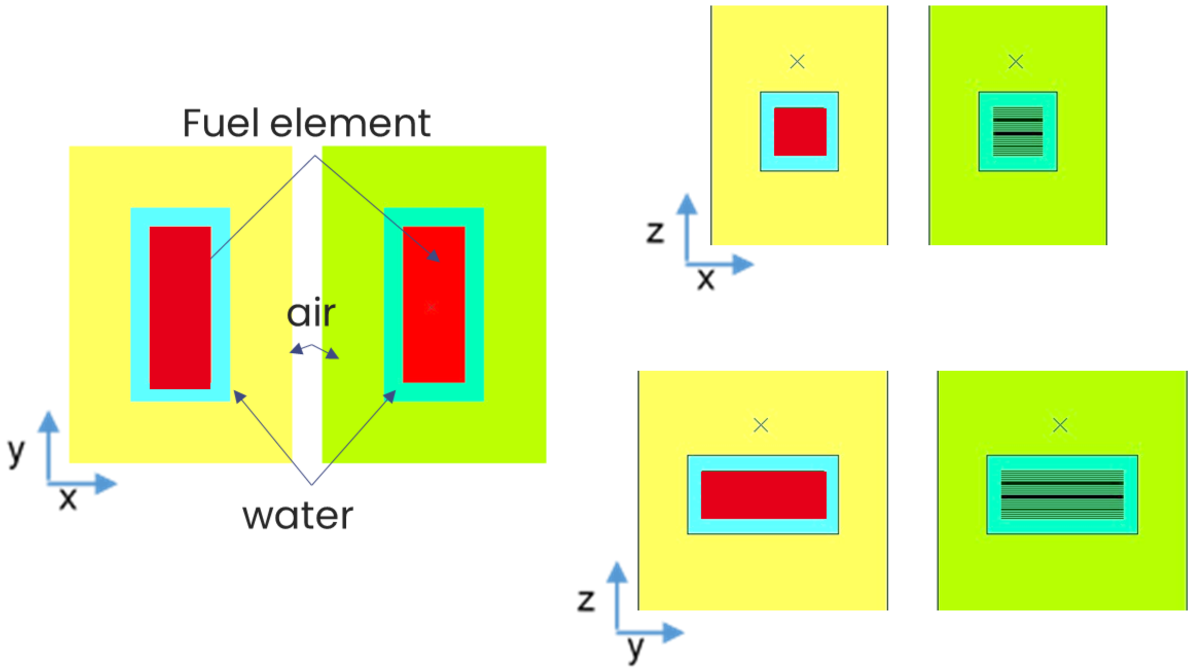
→ **(1)** allows for faster evaluations by simplifying the representation of reality

→ **(2)** provides more accurate results without approximations, but requires more CPU time, limiting the number of simulations that can be performed.

Comparative analysis of neutronic calculation options for safety-criticality studies

1. **For simplified modeling:** built upon a simplified representation where fuel is homogenized into **a single volume**

2. **For reference modeling** with TRIPOLI-4®: precise and more realistic representation of fuel
   → **Different fuel plates** surrounded by a water ring



Criticality calculation must ensures that we remain subcritical, under input parameter uncertainties.

↓

Goal here:
Compare the predictions of the two modelings to assess the calculation biases introduced by the simplified model within the domain of uncertain parameters common to both models

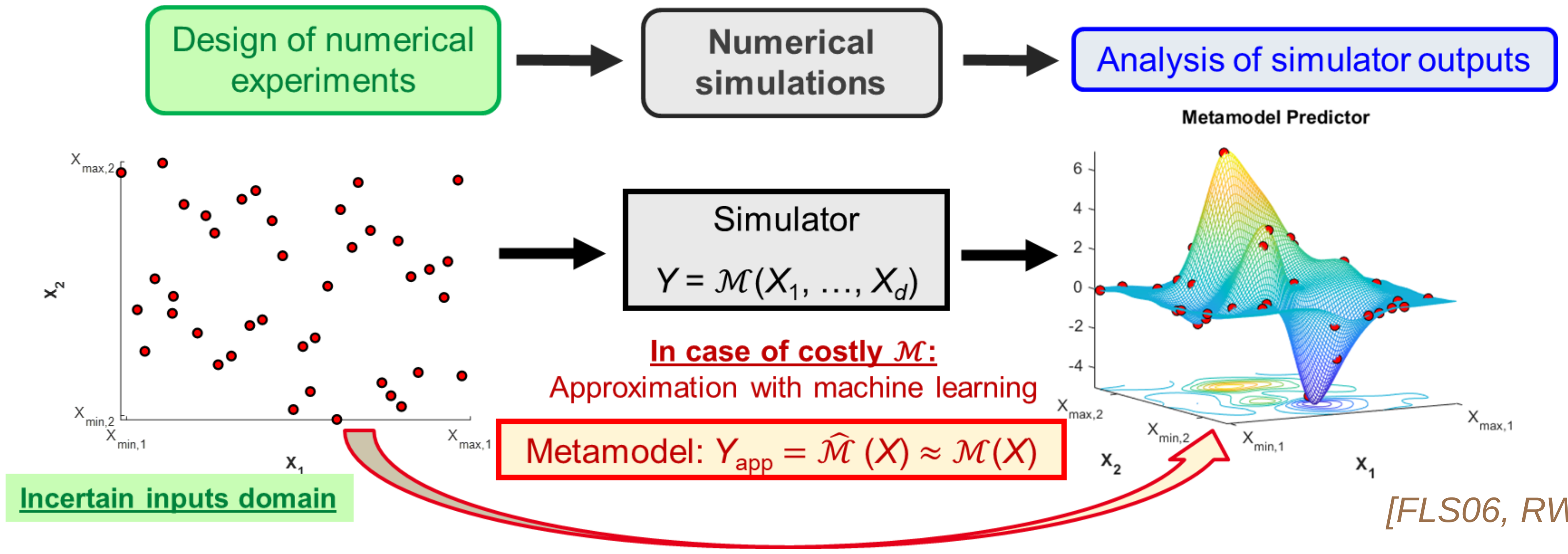
Comparative analysis of neutronic calculation options for safety-criticality studies

How to asses the calculation biases introduced by the simplified model?

→ Probabilistic framework for uncertain inputs and intensive Monte Carlo-based methods

→ **But CPU-expensive reference modeling** ⇒ number of possible simulations limited to $n \approx 300$ e.g.

→ **Solution:** build a **metamodel** to quantify the calculation discrepancies between the two modeling approaches (on the quantities of interest) and assess that the bias does not exceed a given threshold



[FLS06, RW06, MI24]

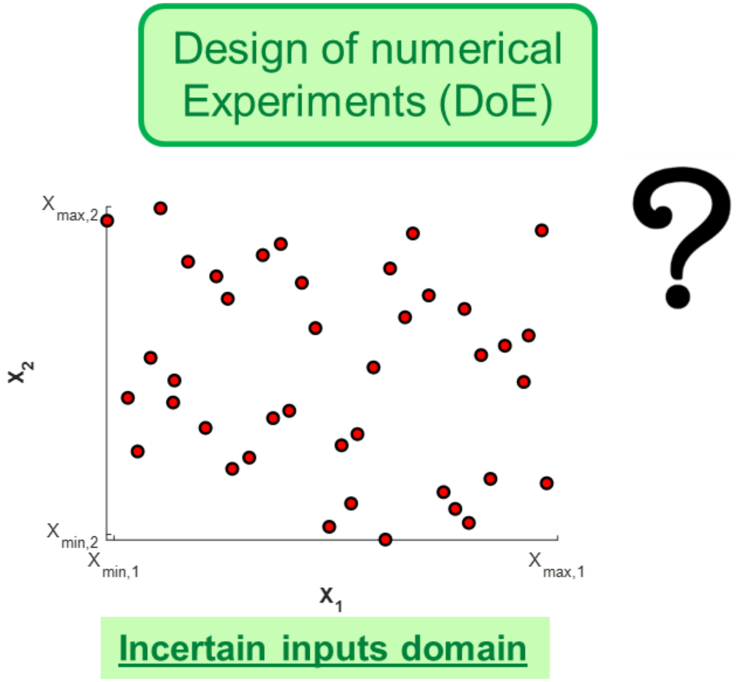
Comparative analysis of neutronic calculation options for safety-criticality studies

How to asses the calculation biases introduced by the simplified model?

→ Probabilistic framework for uncertain inputs and intensive Monte Carlo-based methods

→ **But CPU-expensive reference modeling** ⇒ number of possible simulations limited to $n \approx 300$ e.g.

→ **Solution:** build a **metamodel** to quantify the calculation discrepancies between the two modeling approaches



Comparative analysis of neutronic calculation options for safety-criticality studies

How to asses the calculation biases introduced by the simplified model?

→ Probabilistic framework for uncertain inputs and intensive Monte Carlo-based methods

→ **But CPU-expensive reference modeling** ⇒ number of possible simulations limited to $n \approx 300$ e.g.

→ **Solution:** build a **metamodel** to quantify the calculation discrepancies between the two modeling approaches

Applicative objectives and constraints for the DoE:

✓ A **single i.i.d. inputs/output sample** for both simulators, for **multi-purpose** (sensitivity analysis, uncertainty propagation...) and training a metamodel

Comparative analysis of neutronic calculation options for safety-criticality studies

How to asses the calculation biases introduced by the simplified model?

→ Probabilistic framework for uncertain inputs and intensive Monte Carlo-based methods

→ **But CPU-expensive reference modeling** ⇒ number of possible simulations limited to $n \approx 300$ e.g.

→ **Solution:** build a **metamodel** to quantify the calculation discrepancies between the two modeling approaches

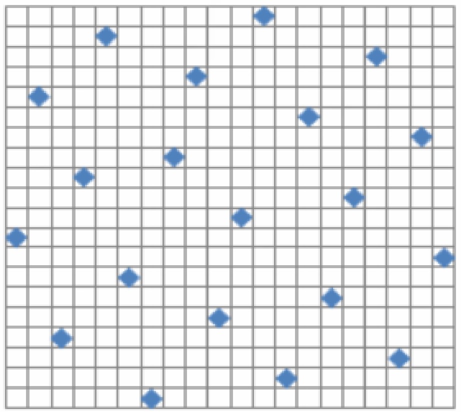
Applicative objectives and constraints for the DoE:

- ✓ A **single i.i.d. inputs/output sample** for both simulators, for **multi-purpose** (sensitivity analysis, uncertainty propagation...) and training a metamodel
- ✓ Small sample size: $n \approx 300$ simulations
- ✓ Medium dimension with d = 5 uncertain inputs

→ **Space-Filling DoE**

Optimal Space-Filling Design Sampling

(Latin Hypercube sampling with even distribution of points)



[PM12, DCI13]

cea

Comparative analysis of neutronic calculation options for safety-criticality studies

How to asses the calculation biases introduced by the simplified model?

$\longrightarrow$ Probabilistic framework for uncertain inputs and intensive Monte Carlo-based methods

$\longrightarrow$ **But CPU-expensive reference modeling** $\Rightarrow$ number of possible simulations limited to $n \approx 300$ e.g.

$\longrightarrow$ **Solution:** build a **metamodel** to quantify the calculation discrepancies between the two modeling approaches

Applicative objectives and constraints for the DoE:

✓ A **single i.i.d. inputs/output sample** for both simulators, for **multi-purpose** (sensitivity analysis, uncertainty propagation...) and training a metamodel

✓ Small sample size: $n \approx 300$ simulations

✓ Medium dimension with d = 5 uncertain inputs

✓ Respect of geometrical constraint for reference simulator: each set of inputs must **ensure that the resulting fuel geometry is feasible** →

Space-Filling DoE under equality constraint?

[SHS+03, PC17, HJR21]

How to define a S-F DoE under specific equality constraint?

1. **Mathematical formalization and insights into possible approaches**

Mathematical formalization

- **d = 5 uncertain input parameters** common to both modeling approaches:
  - X_1: Total mass of fissile component in all elements
  - X_2: Concentration of fissile component
  - X_3 and X_4: geometrical parameters of each fuel element
  - X_5: Ratio between the volume of the moderator and the fissile volume
- **Prior probability distribution** defined for simplified modeling approach
  - Uniform marginal distributions defined on given intervals $D_{X_i} = [x_{i,min} \, ; x_{i,max}] \rightarrow$ Hypercube domain $D_{\boldsymbol{X}} = \prod_{i=1}^{d} D_{X_i}$
  - Independent inputs: $\mathbb{P}_{\boldsymbol{X}} = \otimes \mathbb{P}_{X_i}$

Feasibility constraint for reference modeling → **Equality constraint**

$$N_p = f(X_1, \dots, X_5) = \sqrt{\frac{X_1}{X_2 X_3^2 X_4 (1+X_5)}} \in \mathbb{N}^*$$

⇒ **Conditional probability distribution** $\mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}} := \mathbb{P}_{\boldsymbol{X}} | \, \boldsymbol{X} \in \mathcal{C}$ with $\mathcal{C} = \{\boldsymbol{x} \in D_{\boldsymbol{X}} \subset \mathbb{R}^d \text{ such as } f(\boldsymbol{x}) \in \mathbb{N}^*\}$

How to sample from $\mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}}$?

How to define and build an « optimal » sample $\Xi_n^{\mathcal{C}} \sim \mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}}$?

→ Extension of constrained sampling methods to « stratified/quantized equality constraint »?

Mathematical formalization

- One equality constraint $f(\boldsymbol{X}) = c \Rightarrow \mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}}$: Projection of $\mathbb{P}_{\boldsymbol{X}}$ on $\mathcal{C}$
  - $\rightarrow$ Under certain regularity assumptions, $\mathcal{C}$ is a regular manifold of dimension $d-1$
  - $\rightarrow$ Conditioning by an equality constraint often yields a degenerated and analytically inaccessible distribution
  - $\rightarrow$ Possible significant modification of marginal distributions and structural dependence
- Multi-level constraint $f(\boldsymbol{X}) \in \mathbb{N}^*$: $f(\boldsymbol{X}) \in \mathcal{C}_{multi}$ with $\mathcal{C}_{multi} = \bigcup_{n \in D_{\boldsymbol{N_p}}} \mathcal{C}_n$ with $D_{N_p} = [\![N_{p,min}; N_{p,max}]\!]^*$
  - $\rightarrow$ $\bigcup_{n \in D_{\boldsymbol{N_p}}} \mathcal{C}_n$ does not constitute a global regular manifold (cf. discrete nature of the levels) $\rightarrow$ No continuous transition between two sets $\mathcal{C}_n$ and $\mathcal{C}_{n+1}$
  - $\rightarrow$ Can be interpreted as a finite stratification of the domain into geometric layers (family of independent supports without global parameterization)

$\Rightarrow$ **Conditional joint law**: despite the lack of global regularity, possible to define $\mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_{multi}} := \mathbb{P}_{\boldsymbol{X}} |\ \boldsymbol{X} \in \mathcal{C}_{multi}$

$\rightarrow$ Mixture of the conditional laws associated with each level: $\mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_{multi}} = \sum_{n \in D_{\boldsymbol{N_p}}} \mathbb{P}_{N_p}(n)\, \mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_n}$ with $\mathbb{P}_{N_p}(n)$ the induced marginal law of the integer variable $N_p := f(\boldsymbol{X})$ with $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$.

**Bounded definition domain of N_p: $D_{N_p} := \{ n \in \mathbb{N}^* \mid \exists\, \boldsymbol{x} \in D_{\boldsymbol{X}} \text{ such as } n = \lfloor f(\boldsymbol{x}) \rceil \}$*

How to sample from conditional distribution $\mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_{multi}}$?

- **Rough approach** based on a **partial recalibration** (justified by application of implicit function theorem):
  1. Generate samples of $\boldsymbol{X}$ according to prior distribution: $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$
  2. For each sample $\boldsymbol{x}$,
     a) Calculate the value of nearest integer $N_{p_{\boldsymbol{x}}} = \lfloor f(\boldsymbol{x}) \rceil$
     b) **Recalibrate a single parameter** (e.g., X_j) to obtain $\tilde{\boldsymbol{x}}$ that satisfies the constraint $f(\tilde{\boldsymbol{x}}) = N_{p_{\boldsymbol{x}}}$, while keeping **the other parameters unchanged**

- **Clear drawbacks and limitations**

  → **Alteration of local constraint geometry**: the choice of the recalibrated parameter X_j can affect the feasibility and local shape of the constraint. Imposing the constraint only on X_j partially respects the geometric structure induced by the manifold.

  → **Distortion of the conditional law:** dependencies induced by the constraint are neglected, marginal distribution of X_j is strongly deformed, while others remain unchanged.

  → **Empirical Instability:** results vary significantly depending on the chosen parameter for recalibration.

Alternative approaches

Methods for conditional simulation when dealing with **constraints domains** on real variables:

- **Implicit reparameterization (constraint absorption):** Reformulate the constraint as a change of variables where one coordinate (of the new variables) becomes the constraint value:

  $\Rightarrow$ Find $\phi: \mathbb{R}^{d-1} \to \mathbb{R}^d$ such as $X = \phi(U)$ et $\mathrm{f}(\phi(U)) = c$ $(\mathrm{i.e.}\ \phi(U) \in \mathcal{C})$

  → Enables exact sampling on the constrained manifold if ϕ is analytically invertible

  → How to find a reparametrization ϕ^{-1} where the induced law on U is tractable, well-conditioned and easy to sample?

- **Rejection-based methods:** sample from the unconstrained law, keep only points satisfying the constraint

  → Fails for real-valued equality constraints: zero probability of exact satisfaction → rejection rate is prohibitive

  → Relaxation $f(X) \in [c - \varepsilon, c + \varepsilon]$ could help but remains inefficient if constrained region $\mathcal{C}$ is narrow

Alternative approaches

- **Sequentially Constrained MC** [GC16] : sequential sampling for evolving distributions that gradually enforce the constraint
  - → Each distribution is weighted by a constraint function measuring the violation level, with increasing weight over iterations
  - → Hard indicator constraint is replaced by a probability constrained, and sampling step usually by MCMC
  - → May be ill-suited for foliated manifolds due to complex geometry

- **Constrained MCMC** [BSU12] : general constrained version of HMC algorithm
  - → Hypothesis of connected and differentiable manifold
  - → Adaptation to the union of disjoint submanifolds?

Even with adaptation of one of these methods to foliated equality constraints, how do we ensure well-distributed samples under the conditional law?

Alternative approaches

- **Selected strategy: Projection-only approach combined with subsampling techniques**
  In the same vein as [HJR21]'s approach that combines SCMC with minimum energy design

Principle: sample from the unconstrained prior and project each point onto the constraint manifold
+ find a well-spaced subsample by minimizing the Energy distance [JDT+15]

→ Constraint is implicitly enforced without explicitly sampling $N_p(\boldsymbol{x})$ or computing the conditional density

→ Approximation of joint conditional law ⇒ Introduces theoretical bias and sampling distortion

Easier to implement that MCMC strategies, but lacks theoretical convergence guarantees
Subsampling does not correct the bias, but can serve multiple goals:

→ Improve geometric coverage (e.g. space-filling design)

→ Rebalance over-represented regions

→ Approximate conditional law via importance sampling (requires conditional density evaluation)

How to define a S-F DoE under specific equality constraint?

1. Mathematical formalization and insights into possible approaches

2. Detailed proposed methodology

Proposed methodology

3-step approach to define a n-size space-filling DoE of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_{multi}}$

(under the feasibility constraint $f(\boldsymbol{X}) \in D_{N_p} \subset \mathbb{N}^*$)

1.Initial Generation: Generate a large pure Monte Carlo sample $\Xi_{n_{MC}}$ of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$, with $n_{MC} \approx 10^6$ e.g.

This initial generation is done **independently of the feasibility constraint**

Proposed methodology

3-step approach to define a n-size space-filling DoE of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_{multi}}$

(under the feasibility constraint $f(\boldsymbol{X}) \in D_{N_p} \subset \mathbb{N}^*$)

1. Initial Generation: Generate a large pure Monte Carlo sample $\Xi_{n_{MC}}$ of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$, with $n_{MC} \approx 10^6$ e.g.

2. Constraint Projection: project each point of $\Xi_{n_{MC}}$ to the nearest manifold

For each point $\boldsymbol{x}^{(i)}$ in $\Xi_{n_{MC}}^{\mathrm{MC}}$, find the closest point $\widetilde{\boldsymbol{x}}^{(i)}$ in the domain $D_{\boldsymbol{X}}$ that belongs to the nearest "leaf" manifold $\mathcal{C}_{n_i}$ where $n_i = \lfloor f(\boldsymbol{x}^{(i)}) \rceil$

For this, solve a **constrained optimization problem**:

$$\tilde{\mathbf{x}}^{(i)} = \underset{\mathbf{x} \in D_X \subset \mathbb{R}^d}{\arg\min} \|\mathbf{x} - \mathbf{x}^{(i)}\| \quad \text{such as} \quad f(\mathbf{x}) = N_{\mathbf{x}^{(i)}} \in \mathbb{N}^* \quad \text{where} \quad N_{\mathbf{x}^{(i)}} = \left[f\left(\mathbf{x}^{(i)}\right) \right]$$

$\rightarrow$ A **new DoE $\Xi_{n_{MC}}^{\mathcal{C}_{multi}}$ is obtained**, of the same size n_{MC} , where **each point satisfies the constraint** (within a numerical precision ϵ of 10^{-8} e.g.).

Proposed methodology

3-step approach to define a n-size space-filling DoE of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}^{\mathcal{C}_{multi}}$

(under the feasibility constraint $f(\boldsymbol{X}) \in D_{N_p} \subset \mathbb{N}^*$)

1. Initial Generation: Generate a large pure Monte Carlo sample $\Xi_{n_{MC}}$ of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$, with $n_{MC} \approx 10^6$ e.g.

2. Constraint Projection: **project each point of $\Xi_{n_{MC}}$ to the nearest manifold**

3. Subsampling to obtain a <u>representative</u> n-size subsample:

From $\Xi_{n_{MC}}^{\mathcal{C}_{multi}}$, select **a reduced subsample $\Xi_n^{\mathcal{C}_{multi}}$ with $n \ll n_{MC}$** (n is the target size, here n = 300).

→ Use advanced techniques such as ***Support Points*** or ***Kernel Herding*** to ensure the subsample is representative of $\Xi_{n_{MC}}^{\mathcal{C}_{multi}}$ and benefits from fast algorithms (and built-upon robust statistics and estimators)

→ To preserve statistical properties between the original sample and the subsample, ensuring good coverage in high-density regions of the target distribution

→ The **resulting subsample** defines the **final DoE for simulations with TRIPOLI-4®**.

Focus on subsampling step

Support points (SP) subsampling based on minimization of Energy Distance

Extract a representative **n-size** subset from a large **N-size** dataset by minimizing the **Energy Distance** [SR13]

Algorithm

1. **Start with a large point set** : Let $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \subset \mathbb{R}^d$

2. **Support point selection** : Find a subset of points $\mathcal{S} = \{s^{(1)}, s^{(2)}, \ldots, s^{(n)}\} \subset \mathcal{X}$ such that $m \ll N$ and that minimizes the Energy Distance $D_E(\mathcal{S}, \mathcal{X})$:

$$\mathcal{S} = \arg\min_{\mathcal{S} \subset \mathcal{X}} D_E(\mathcal{S}, \mathcal{X})$$

where D_E is defined as:

$$D_E(\mathcal{S}, \mathcal{X}) = 2\mathbb{E}[\|X - S\|] - \mathbb{E}[\|X - X'\|] - \mathbb{E}[\|S - S'\|]$$

with X, X' being independent samples from $\mathcal{X}$ and S, S' independent samples from $\mathcal{S}$.

$\Rightarrow$ Use here to obtain $\mathcal{S} = \Xi_n^{\mathcal{C}_{multi}}$ from $\mathcal{X} = \Xi_{n_{MC}}^{\mathcal{C}_{multi}}$ (and $N = n_{MC}$)

Focus on subsampling step

Support points (SP) subsampling based on minimization of Energy Distance

Extract a representative **n-size** subset from a large **N-size** dataset by minimizing the **Energy Distance** [SR13]

In practice

→ **Statistical estimator** of energy distance from two finite i.i.d. samples of $\mathcal{X}$ and $\mathcal{S}$

$$\widehat{D}_E(\mathcal{S},\mathcal{X}) = \frac{2}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\|s^{(i)} - x^{(j)}\| - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|x^{(i)} - x^{(j)}\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|s^{(i)} - s^{(j)}\|$$

→ V-statistics: asymptotically unbiased and consistent

→ **Greedy Support points algorithm** [MJ18]**:** practical heuristic to build $\mathcal{S}_n^*$ that minimizes $\widehat{D}_E(\mathcal{S}_n^*, \mathcal{X})$

1. **Initialization** : Set $\mathcal{S}_0^* = \emptyset$
2. **Iterative selection** : For $t = 1$ to n:
   - Evaluate $\widehat{D}_E(\mathcal{S}_{t-1}^* \cup \{x\}, \mathcal{X})$ for each $x \in \mathcal{X} \setminus \mathcal{S}_{t-1}^*$
   - Select $s^{(t)} = \arg\min_x \widehat{D}_E(\mathcal{S}_{t-1}^* \cup \{x\}, \mathcal{X})$
   - Update $\mathcal{S}_t^* = \mathcal{S}_{t-1}^* \cup \{s^{(t)}\}$
3. **Output** : Final support set $\mathcal{S}_n^*$

Computationally efficient, especially when $n \ll N$ (where exhaustive search becomes impractical).

Closed of **Kernel Herding** [CWS10] with **"energy-distance" kernel** [SR13]

How to define a S-F DoE under specific equality constraint?

1. Mathematical formalization and insights into possible approaches

2. Detailed proposed methodology

3. Application to test case

Application on criticality test-case

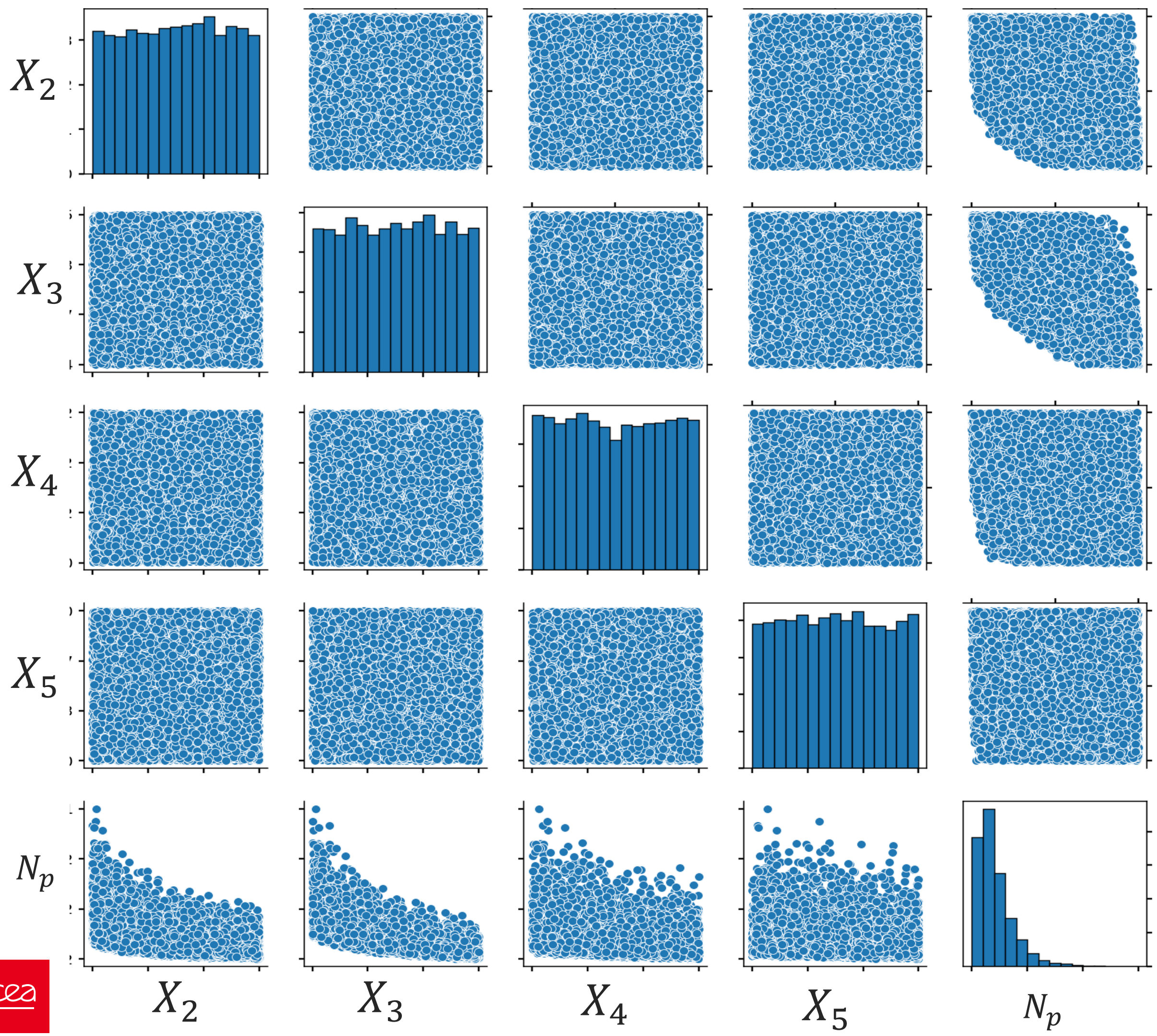
- **d = 5 uncertain input parameters** with **prior independent uniform distributions**
  - X_1: Total mass of fissile component in all elements $\rightarrow$ value of the mass is initially fixed $\Rightarrow$ 4 uncertain parameters
  - X_2: Concentration of fissile component
  - X_3 and X_4: geometrical parameters of each fuel element
  - X_5: Ratio between the volume of the moderator and the fissile volume

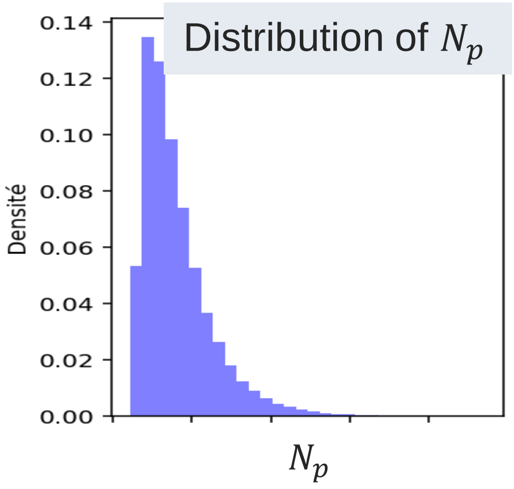
Feasibility constraint for reference modeling

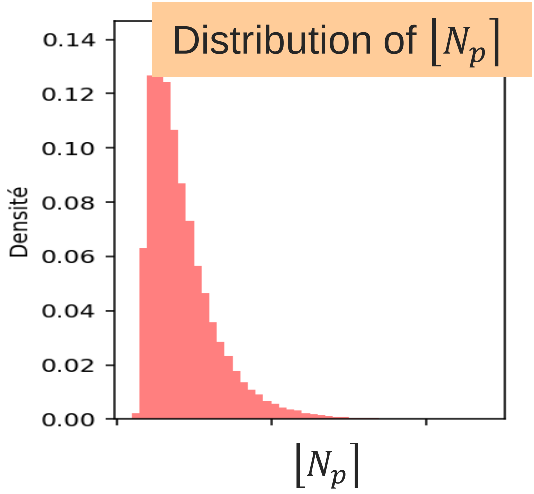
$$N_p = f(x_1, \ldots, X_5) = \sqrt{\frac{x_1}{X_2 X_3^2 X_4 (1+X_5)}} \in \mathbb{N}^*$$

Application on criticality test-case

Step 1: Large Monte Carlo sample $\Xi_{n_{MC}}$ of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$

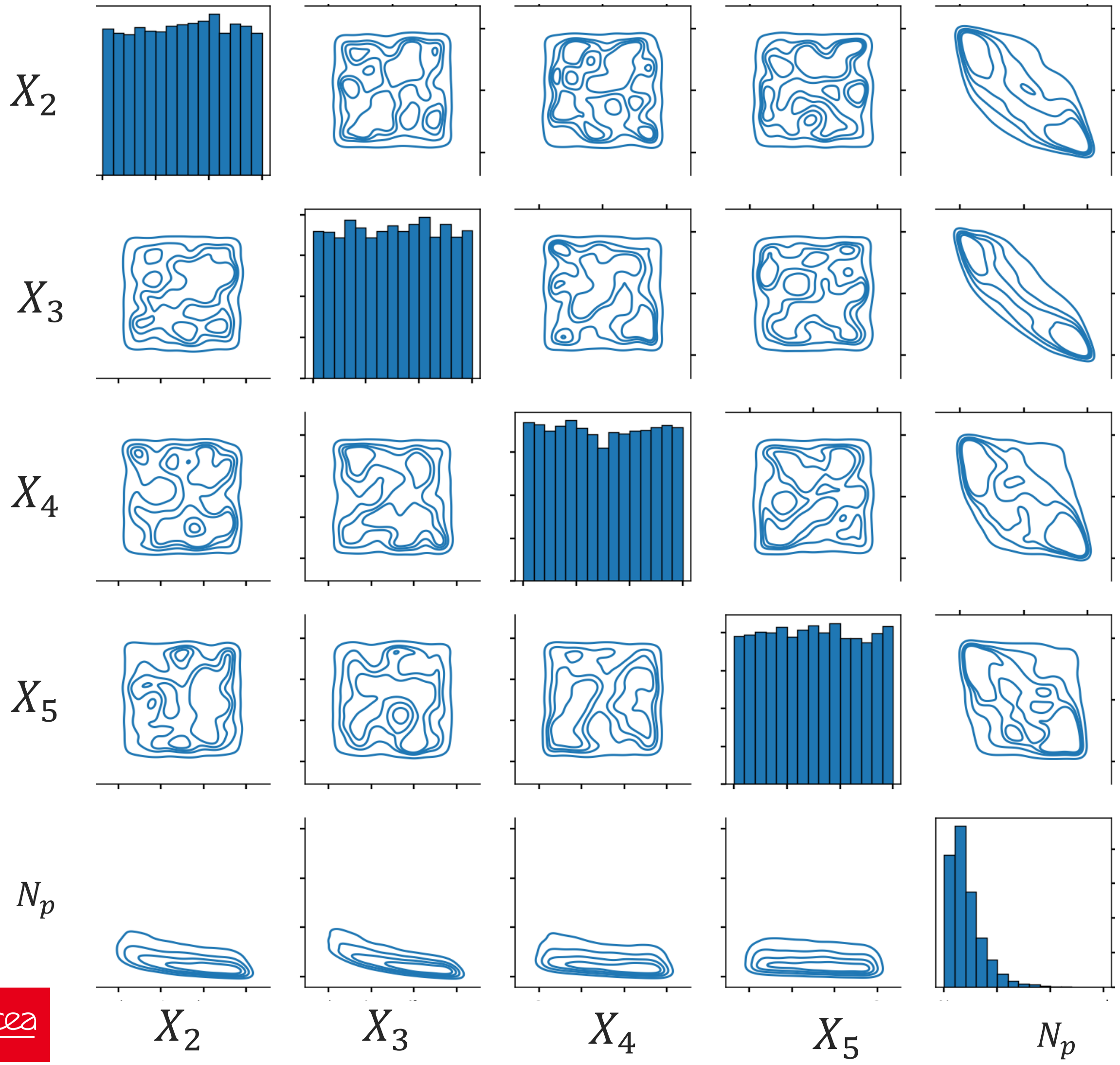




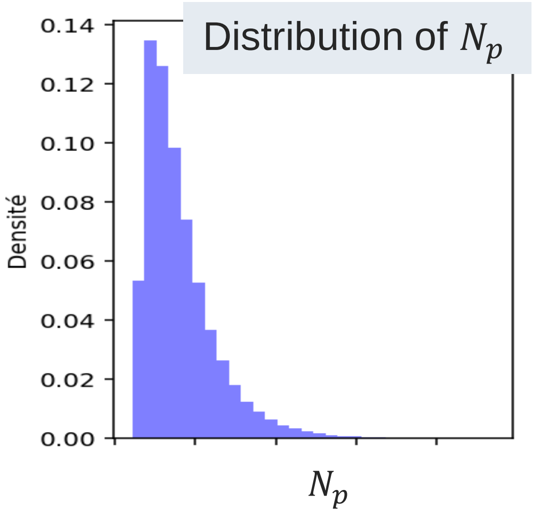


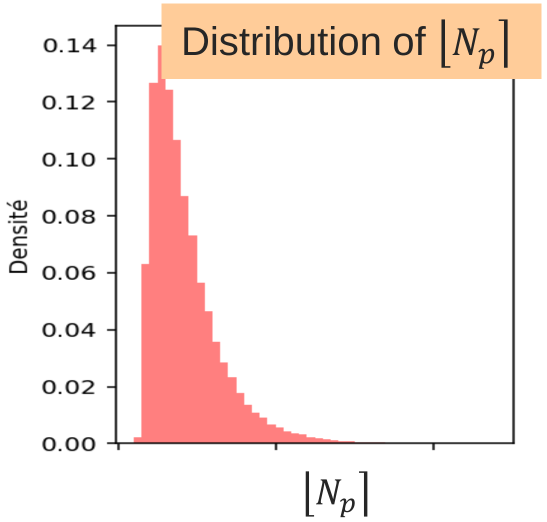
Application on criticality test-case

Step 1: Large Monte Carlo sample $\Xi_{n_{MC}}$ of $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$

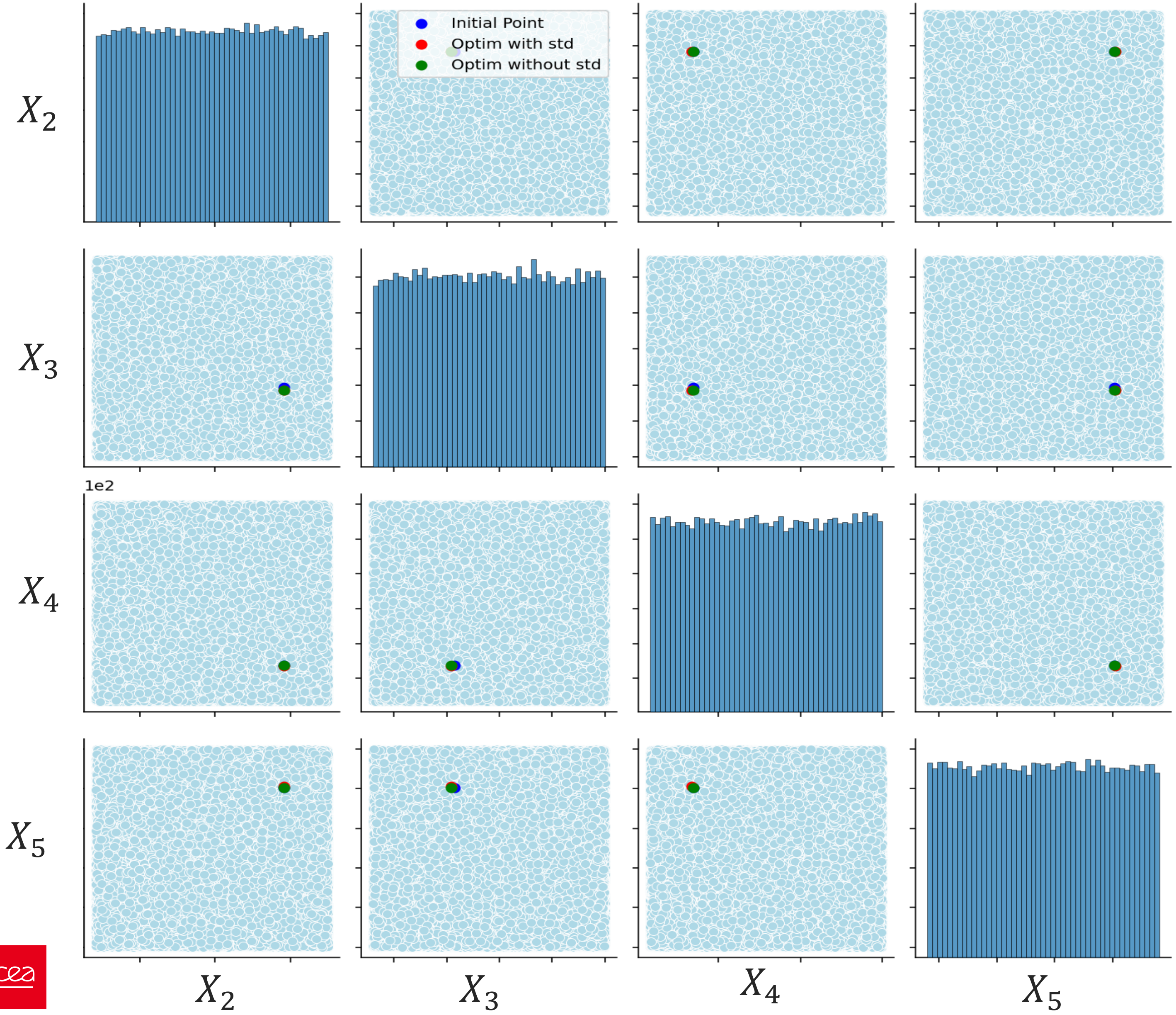


Vizualization tool with Copulogram to focus on dependence structure of marginal bivariate distribution



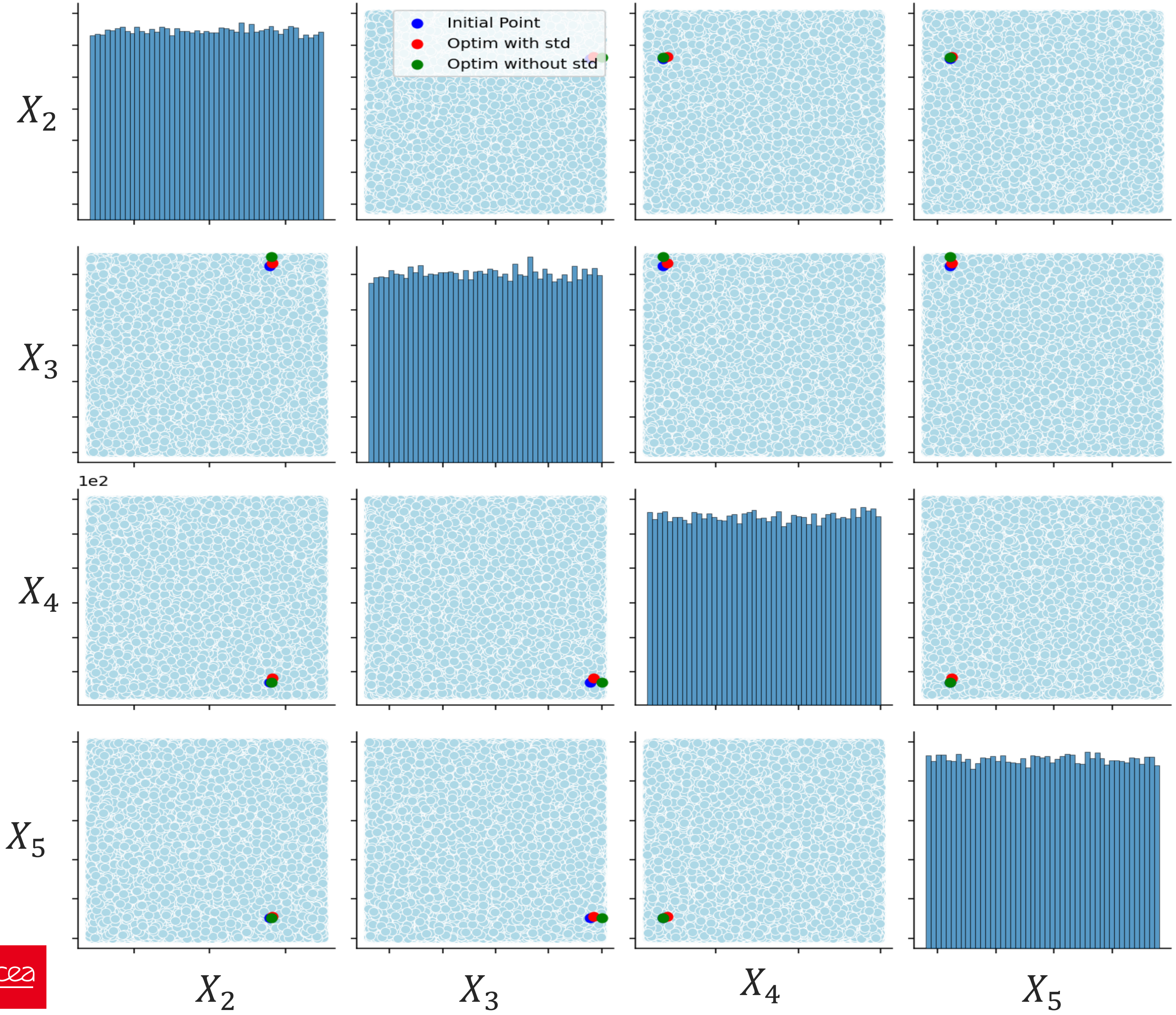


Application on criticality test-case



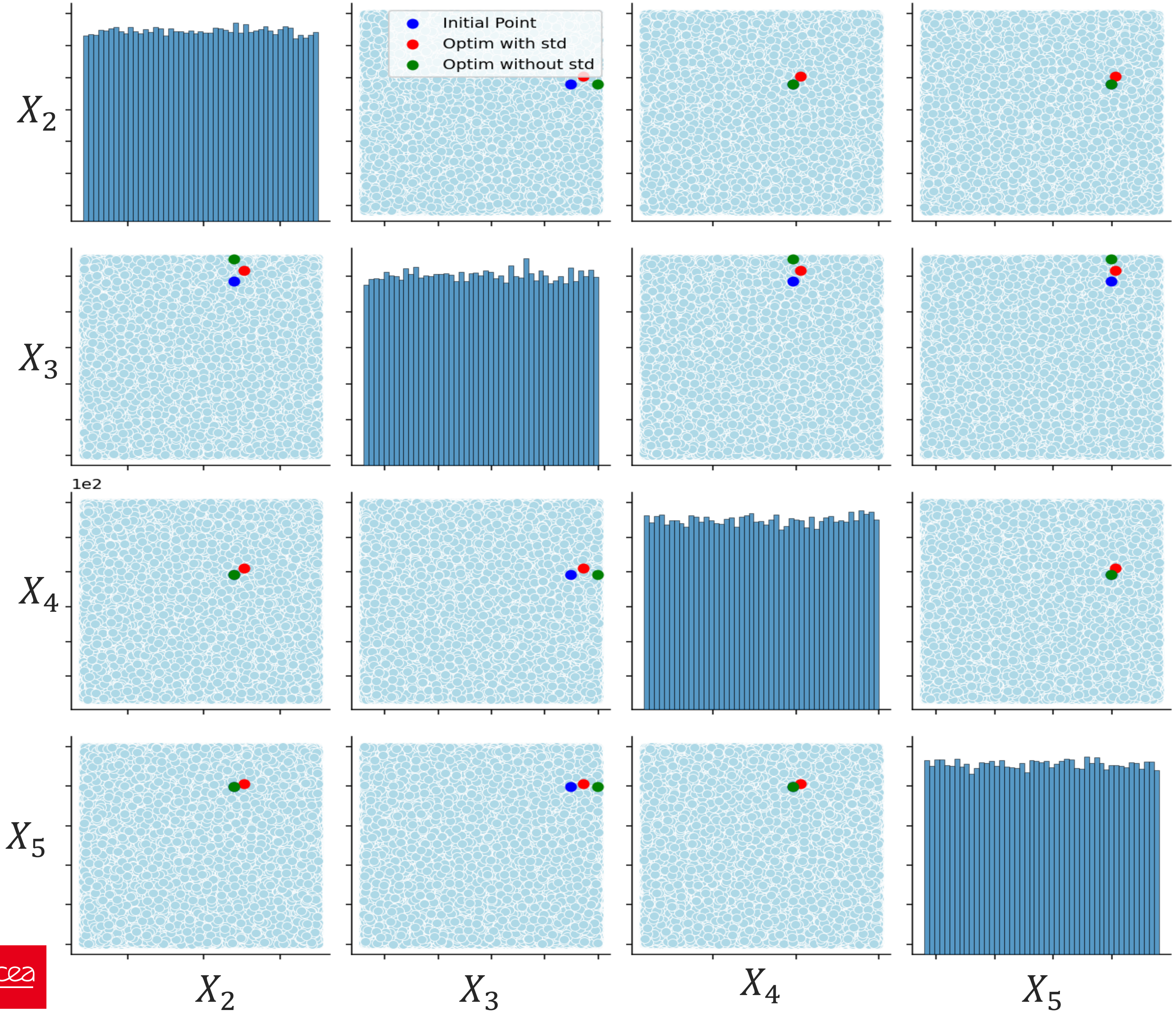
Step 2 - Constraint Projection: project each point of $\Xi_{n_{MC}}$ to the nearest manifold

Application on criticality test-case



Step 2 - Constraint Projection: project each point of $\Xi_{n_{MC}}$ to the nearest manifold

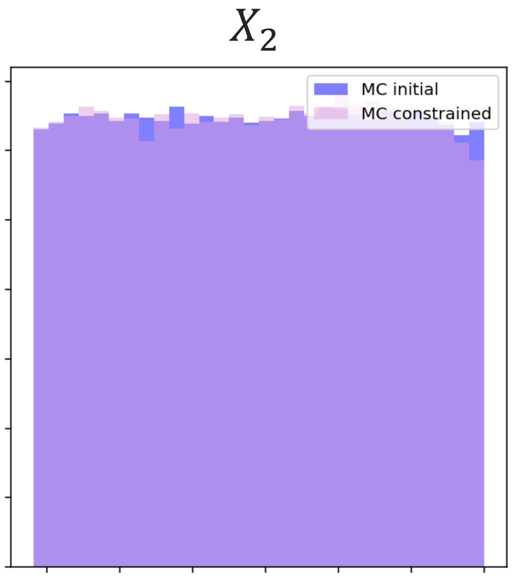
Application on criticality test-case

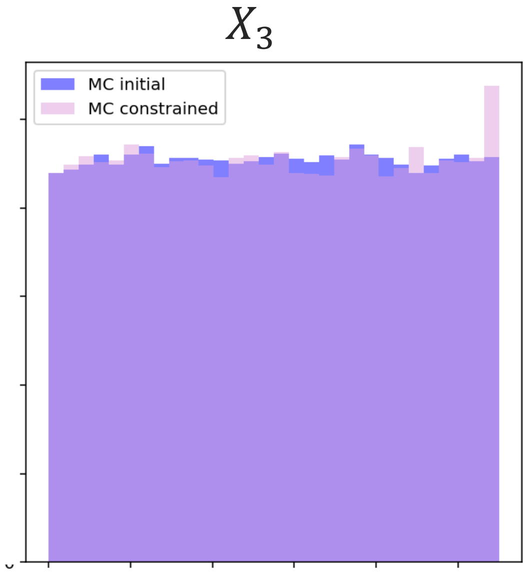


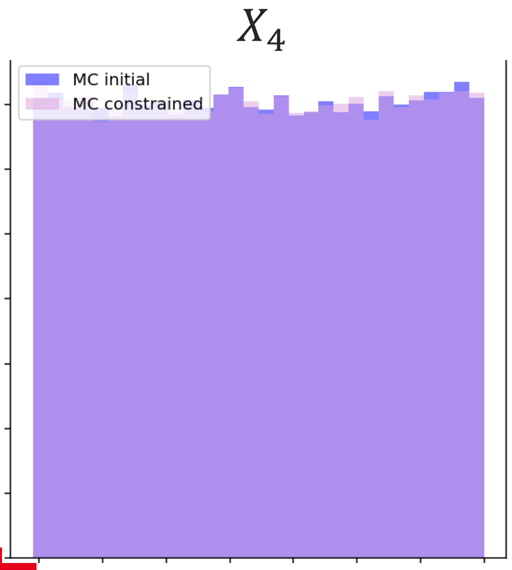
Step 2 - Constraint Projection: project each point of $\Xi_{n_{MC}}$ to the nearest manifold

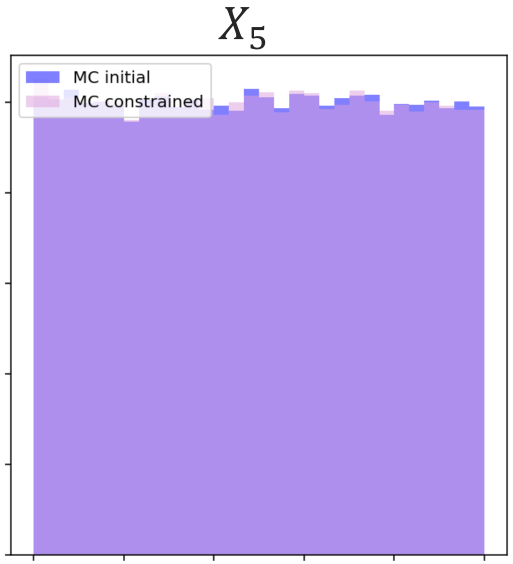
To obtain a new **DoE** $\Xi_{n_{MC}}^{\mathcal{C}_{multi}}$

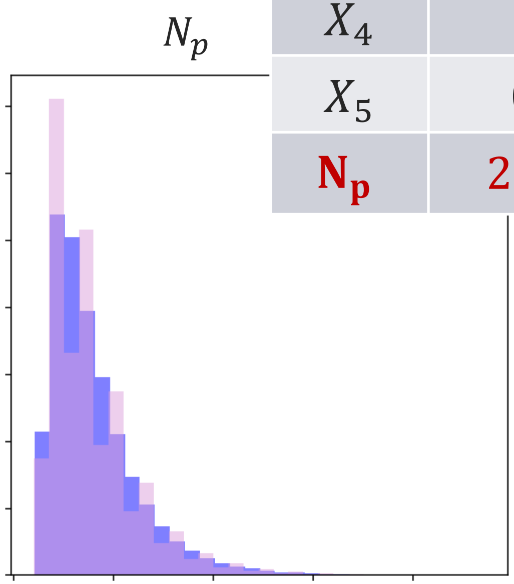
Application on criticality test-case









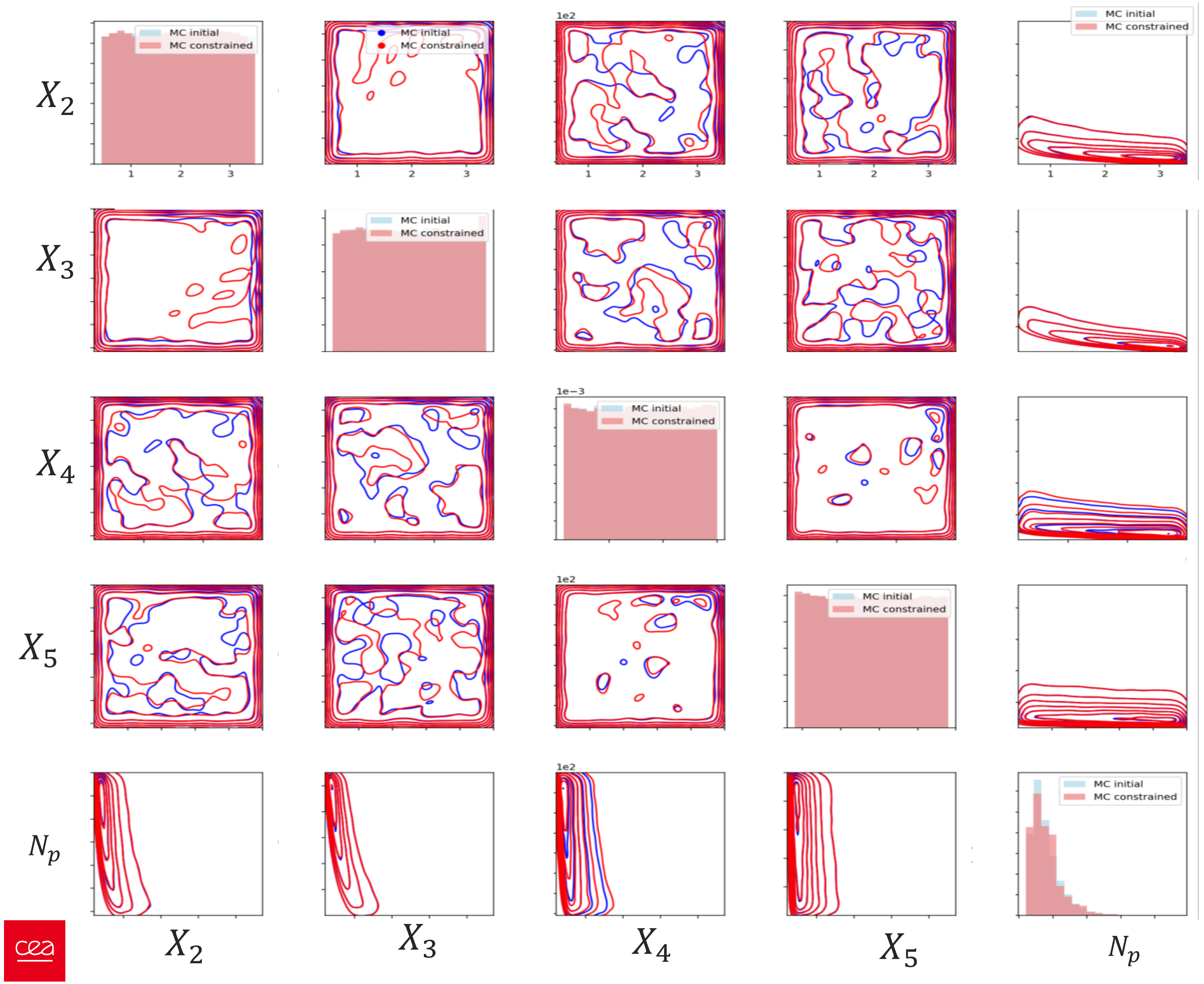


Step 2 - Constraint Projection: project each point of $\Xi_{n_{MC}}$ to the nearest manifold

Adequation test between $\Xi_{n_{MC}}$ and $\Xi_{n_{MC}}^{\mathcal{C}_{multi}}$

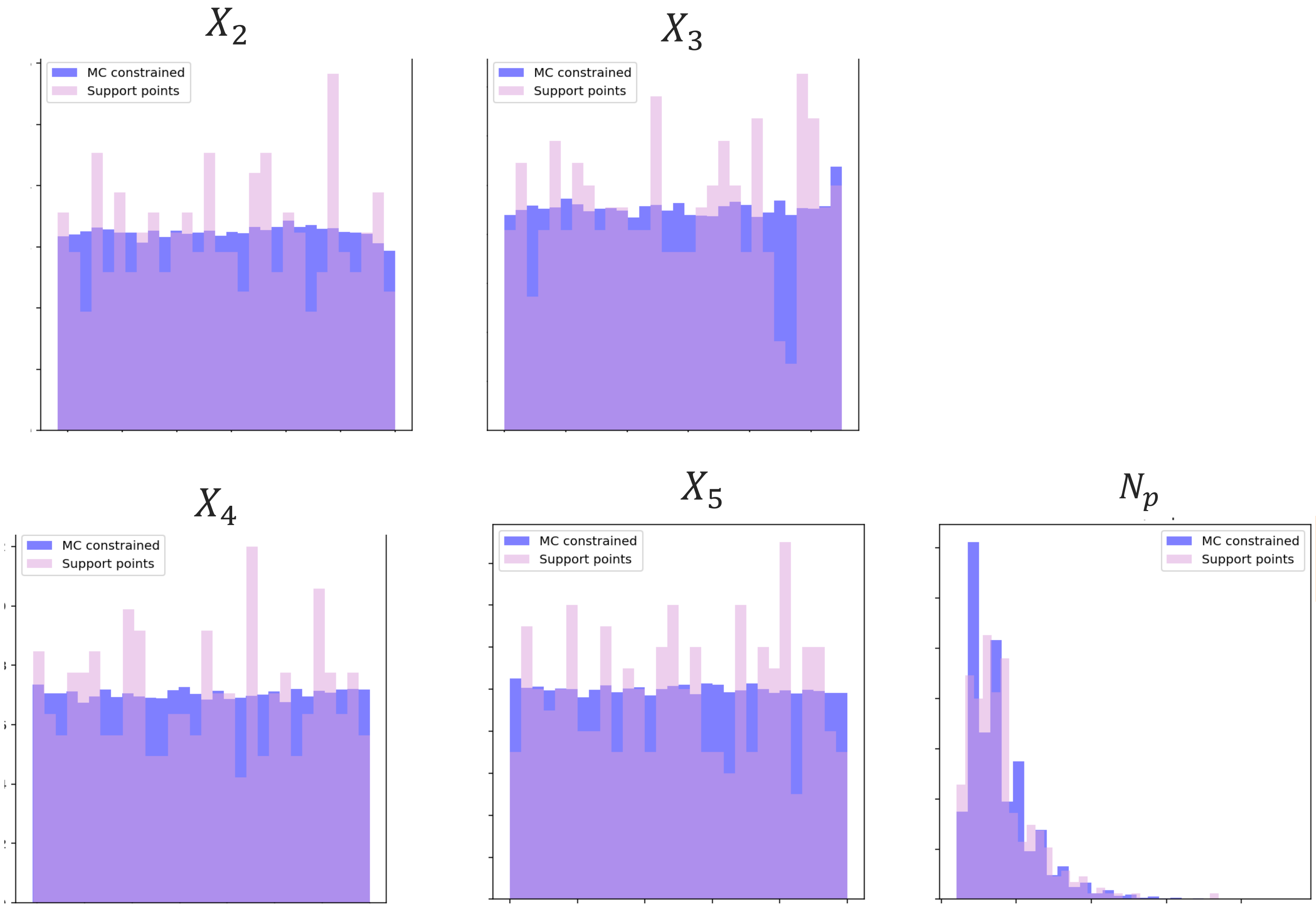
Input	KS-pval	AD-pval	CVM-pval
X_2	0.29	0.03	0.33
$\mathbf{X_3}$	10^{-4}	10^{-4}	9.10^{-3}
X_4	0.2	10^{-3}	0.9
X_5	0.36	0.05	0.62
$\mathbf{N_p}$	2.10^{-4}	2.10^{-4}	2.10^{-4}

Application on criticality test-case



Step 2 - Constraint Projection: project each point of $\Xi_{n_{MC}}$ to the nearest manifold

Application on criticality test-case



Step 3 - Subsampling: based on the greedy algorithm of support points (SP)

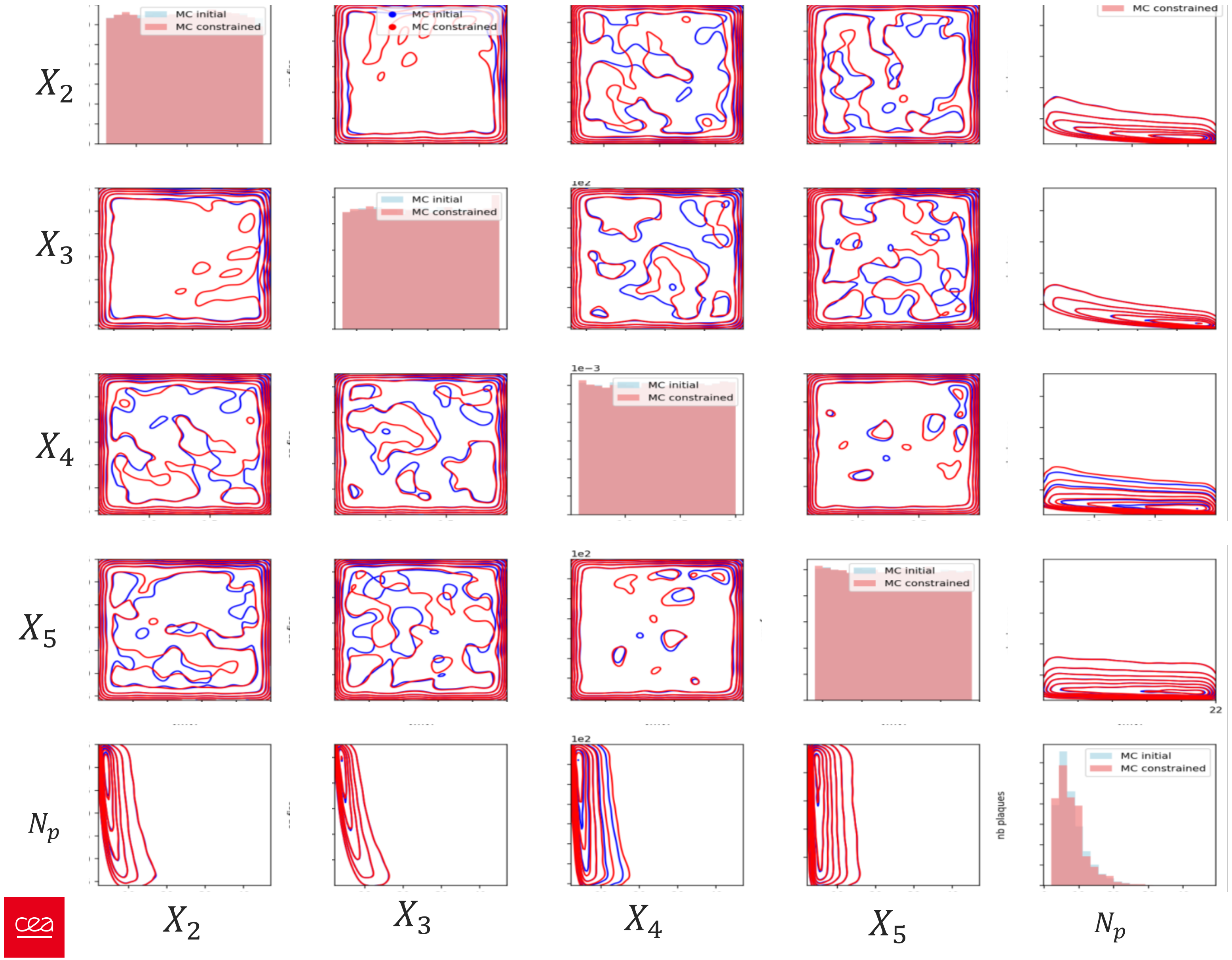
↓

Representative **reduced subsample $\Xi_n^{\mathcal{C}_{multi}}$ with $n \ll n_{MC}$**

Adequation test between $\Xi_{n_{MC}}^{\mathcal{C}_{multi}}$ and $\Xi_n^{\mathcal{C}_{multi}}$

Input	KS-pval	AD-pval	CVM-pval
X_2	0.99	0.99	0.99
$\mathbf{X_3}$	0.86	0.99	0.99
X_4	0.99	0.99	0.99
X_5	0.99	0.99	0.99
$\mathbf{N_p}$	0.90	0.96	0.88

Application on criticality test-case



Step 3 - Subsampling: based on the greedy algorithm of support points (SP)

Representative **reduced subsample $\Xi_n^{\mathcal{C}_{multi}}$** **with $n \ll n_{MC}$**

Conclusions and perspectives

- **Generic methodology:** broadly applicable to any constrained setting where one can generate large samples satisfying equality constraints (⇔ constraint is easy to evaluate)

- **Approximative but effective**: resulting sample only approximates the true conditional distribution → quality is high when the projection via Euclidean distance induces minimal distortion
  - → Application test case with foliation-based manifold ensures that nearby projections always exist

- **Flexible subsampling via Energy distance**: energy distance criterion (or more generally MMD with any characteristic kernel) offers strong flexibility in selecting representative subsets
  - → e.g.: Energy distance w.r.t. a uniform distribution if coverage of the space is prioritized (link in 1-D with discrepancy minimization and low-discrepancy designs)
  - → Here, choice motivated by the multipurpose goal (metamodeling but also direct uncertainty propagation and sensitivity analysis)

Conclusions and perspectives

- **Alternative subsampling strategies:** such as **kernel herding** (MMD with orther kernels such as Matérn kernels), though preliminary results suggest lower performance in preserving the constrained distribution
- **Goodness-of-fit testing:** 2D marginal tests between the constrained Monte Carlo sample and the subsampled set to assess how well structural dependencies induced by the constraint are preserved by subsampling techniques.
- **Further study the dependency structure** conveyed by the constraint
- **Compare (or combine) with more complex sampling methods** (e.g. SMC + CMCMC): assess what is lost with simpler [projection + subsampling] approach. Also compare with [HJR21]'s approach (SCMC with minimum energy design)

References

[BSU12] Brubaker, Salzmann and Urtasun (2012). A family of MCMC Methods on implicitly defined manifolds. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, 22:161-172.

[CWS10] Chen, Welling and Smola. (2010). Super-samples from kernel herding. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, p. 109–116.

[DCI13] Damblin, Couplet and Iooss (2013). Numerical studies of space-filling designs: Optimization of Latin Hypercube Samples and subprojection properties. Journal of Simulation, 7(4):276–289.

[Dav15] Da Veiga (2015). Global sensitivity analysis with dependence measures, Journal of Statistical Computation and Simulation, 85:1283-1305.

[FLS06] Fang, Li and Sudjianto (2006). Design and Modeling for Computer Experiments, Chapman & Hall/CRC.

[GC16] Golchi and & Campbell (2016). Sequentially Constrained Monte Carlo, Computational Statistics & Data Analysis, 97(C):98-113.

[Gra21] Gramacy (2021) Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Chapman and Hall/CRC.

[HJR21] Huang, Joseph, and Ray (2021). Constrained minimum energy designs. Statistics and Computing, 31(6):80.

[JDT+15] Joseph, Dasgupta, Tuo and Wu (2015). Sequential exploration of complex surfaces using minimum energy designs. Technometrics 57(1), 64–74.

[MI24] Marrel and Iooss (2024). Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation, Reliability Engineering and System Safety, 247:110120.

References

MJ18] Mak and Joseph (2018). Support points. The Annals of Statistics, 46:2562 – 2592.

[PC17] Perrin and Cannamela (2017). A repulsion-based method for the definition and the enrichment of optimized space filling designs in constrained input spaces. Journal de la Société Française de Statistique, 158(1):37–67.

[PM12] Pronzato and Muller (2012). Design of computer experiments: Space filling and beyond. Statistics and Computing 22(3): 681-701.

[Pro22] Pronzato (2022). Performance analysis of greedy algorithms for minimizing a maximum mean discrepancy. Statistics and Computing, 33(1):14, 1573-1375.

[PZ20] Pronzato and Zhigljavsky (2020). Bayesian Quadrature, Energy Minimization, and SpaceFilling Design. SIAM/ASA Journal on Uncertainty Quantification, 8(3):959–1011.

[RW06] C.E. Rasmussen and C.K.I. Williams (2006). Gaussian processes for machine learning. MIT Press.

[SHS+03] Stinstra, Hertog, Stehouwer and Vestjens (2003). Constrained maximin designs for computer experiments. Technometrics, 45(4):340–346.

[SR13] Székely and Rizzo (2013). Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference, 143:1249 – 1272.