RADAI: A Large-Scale Realistic Dataset for Radiation Detection Algorithm Development

James M. Ghawaly Jr.,[*,a,b] Daniel E. Archer,[b] Andrew D. Nicholson,[b] Douglas E. Peplow,[c] Nicholas J. Prins,[c] Tenzing H. Y. Joshi,[d] Mark S. Bandstra,[d] Andrew C. Jones,[d] Brian J. Quiter,[d] and Abigael C. Nachtsheim[e]

[a] *Louisiana State University*
Division of Computer Science and Engineering
Patrick F. Taylor Hall, Baton Rouge, LA 70726

[b] *Oak Ridge National Laboratory*
Physics Division
P.O. Box 2008, Oak Ridge, TN 37830

[c] *Oak Ridge National Laboratory*
Nuclear Energy and Fuel Cycle Division
P.O. Box 2008, Oak Ridge, TN 37830

[d] *Lawrence Berkeley National Laboratory*
Nuclear Science Division
1 Cyclotron Road, Berkeley, CA 94720

[e] *Los Alamos National Laboratory*
Computer, Computational, and Statistical Sciences Division
Bikini Atoll Rd, Los Alamos, NM 87545

[*]Email: jghawaly@lsu.edu

Number of pages:	45
Number of tables:	9
Number of figures:	17

Abstract

Locating radioactive material that may be used for illicit purposes is an important national security capability. Doing so in complicated and dynamic environments, such as urban centers with freely-moving radiation detection systems represents a specific challenge whose general solutions may be applied to related problems. The so-called urban radiological search problem is therefore one of ongoing study. In recent decades, new detector systems and analysis techniques have been developed to address urban search. A major complicating factor is the spatial and temporal variability of natural background radiation, which strongly affects the ability of any approach to achieve its statistical limits in sensitivity and false alarm rate. Deployed instruments and algorithms are typically best developed, characterized, and tested by using large amounts of data, but data collection is time-consuming and expensive, idiosyncratic instrumental effects such as gain drift will inevitably creep in, and a large portion of the problem's phase space is unlikely to ever be explored. Large datasets with high coverage of the phase space are especially important for developing algorithms based on machine learning/artificial intelligence methods. To this end, an approach has been developed to generate synthetic data for urban search, where arbitrarily long data collections that feature realistic ambient background rates and variability can be generated. The approach incorporates large-scale Monte Carlo simulations of the gamma-ray background, including the effects of building structures and clutter; and of threat and nuisance sources, which are embedded within the three-dimensional scene for maximal realism. The data are being shared publicly to spur development of new techniques as well as to allow for fair comparisons between developed algorithms.

Keywords — radiation detection dataset, Monte Carlo simulation, artificial intelligence, machine learning, radiation detection algorithm, design of experiments

I. INTRODUCTION

Radioactive and nuclear materials are used across a wide variety of benign applications, including medical diagnostics, industrial radiography, and power generation. However, if such sources are lost, stolen, or diverted, it is critical to national, homeland, and global security that they be rapidly detected, identified, and localized. To accomplish this, algorithms must be developed to analyze data generated by radiation detectors, most commonly spectroscopic detectors with medium-energy resolution such as thallium-loaded sodium iodide scintillators, or NaI(Tl), that are typically flown, driven, and carried through the search environment. The algorithms are typically designed to notify the operator when sources inconsistent with expected background are present through the issuance of an alarm. These algorithms often also aim to classify or identify the detected source. If such a scenario occurs in an urban environment, however, the problem becomes significantly more complex. Naturally occurring radioactive material (NORM) in buildings and surrounding infrastructure, as well as topography, human activity, and environmental conditions such as rainfall and cosmic-ray variations, contribute to a dynamic and spatially heterogeneous background that can easily obscure or mimic source signatures. As a result, algorithms must not only be sensitive and specific but also robust to a wide range of confounding conditions.

Developing and evaluating such algorithms ideally leverages large, realistic datasets that represent the full variability of operational conditions. However, acquiring real-world data is expensive, time-consuming, and fundamentally limited by instrumentation, environmental unpredictability, and safety constraints. Additionally, true ground truth is often unavailable in real data, further complicating algorithm training, assessment, and comparison. Algorithms based on artificial intelligence (AI)/machine learning (ML) are especially dependent on large datasets that adequately cover the phase space of the variables encountered in the real world.

To address these limitations, we have developed the Radiological Anomaly Detection and Identification (RADAI) dataset, a large-scale synthetic dataset designed to support the development, training, and rigorous evaluation of radiation detection algorithms, particularly those leveraging machine learning and artificial intelligence. RADAI builds upon and significantly expands the capabilities of earlier datasets used in competitions such as Chameleon Street [1, 2], introducing longer runs, more varied urban geometries, richer environmental modeling (e.g., cosmic rays and rain), and more realistic threat and nuisance sources.

By employing a modular, physics-based Monte Carlo simulation framework, RADAI enables controllable yet realistic generation of gamma-ray detection data in complex urban environments. The dataset includes list-mode gamma-ray detector data featuring vehicle motion, detector response, dynamic background effects, and various threat scenarios with detailed ground truth. This makes RADAI not only a valuable resource for algorithm development but also a reproducible benchmark for fair performance comparisons.

The rest of this paper is organized as follows:

- Section II discusses prior work in urban source search dataset development.
- Section III describes the methods used to simulate urban environments, detector responses, and radioactive sources.
- Section IV details the structure and content of the dataset as well as how it can be accessed.
- Section V discusses potential applications and concludes the paper.

II. PRIOR WORK

Ghawaly et al. [1] introduced a large synthetic urban search dataset, which featured Monte Carlo–simulated list-mode data from a 2"×4"×16" NaI(Tl) detector moving at constant speed through a simplified virtual street environment. This environment, known as Chameleon Street, was designed to be reconfigurable, allowing the simulated city blocks to be arranged into different street configurations. The simulations included variable NORM background, six types of threat sources (both bare and shielded), and full spectral and temporal labeling of the detector data. The dataset was used in a public Topcoder competition [3], and its validity was supported by comparison with experimental measurements collected at Fort Indiantown Gap [2]. The rigorously controlled ground truth made the dataset a valuable resource for developing and evaluating radiation detection and identification algorithms. The majority of top-performing approaches in the Topcoder competition employed AI/ML methods [3], primarily deep learning. This benchmark dataset laid the groundwork for RADAI's more expansive and realistic data generation framework, described in this paper.

The Chameleon Street dataset was produced in two primary stages. First, a large suite of high-fidelity Monte Carlo simulations using the SCALE/MAVRIC code system [4] computed

the flux spectra arriving at a series of positions along four traffic lanes through various multi-block streets. The pulse-height spectra within a $2'' \times 4'' \times 16''$ NaI(Tl) detector were computed for each position using a detector response function. These simulations independently described contributions to background and the point-like nuisance and threat sources. Second, time-ordered gamma-ray energy depositions were sampled from the pulse-height data to emulate a vehicle driving along a specified lane. This process generated three datasets: (1) a labeled training set indicating whether a source was present and, if so, the time of closest approach; (2) an unlabeled public test set that provided real-time scoring feedback on the competition leaderboard; and (3) a private test set whose results were withheld from participants but used by organizers for final scoring.

Several features were incorporated into the simulations to discourage overfitting to the background of any single geometry. The virtual street was based on idealized versions of seven blocks of Gay Street in Knoxville, TN, with streets, sidewalks, parking lots, and buildings atop a thick slab of soil. These blocks could be arranged into eight different sequences, referred to as *instances*, and periodic boundary conditions were implemented to allow runs to wrap around the street geometry. For background modeling, each NORM was simulated independently at a concentration of $1\,\mathrm{Bq\,kg^{-1}}$ in buildings, sidewalks, streets, and soil, enabling flexible scaling during sampling. Threat sources were positioned at fifteen different locations within the scene, including building fronts, alleyways, and parking lots. In total, the simulations consumed approximately 94,400 CPU hours (more than 10 CPU years), generating 7,424 pulse-height mesh tallies (each with 768 voxels and 2,000 energy bins), amounting to roughly 130 GB of raw simulation output.

III. METHODOLOGY

III.A. Mobile Search Monte Carlo Simulations Geometries

To support the development, testing, and comparison of radiation detection algorithms, especially those based on machine learning, a more comprehensive and realistic synthetic dataset was needed than what was provided by the earlier Chameleon Street effort. Several key enhancements were identified to increase the utility of the new dataset:

1. A broader set of potential nuisance and threat sources, including common industrial isotopes, medical sources, and special nuclear material (SNM);

2. A larger and more diverse selection of urban block types to reflect a wider range of environments;

3. Extended list-mode data durations to better support false alarm rate (FAR) analysis;

4. Longer and more varied street geometries with less out artificial repetition;

5. Realistic clutter such as vehicles, pedestrians, and urban infrastructure, to better reflect the scattering and shielding seen in real environments;

6. Background contributions from atmospheric cosmic-ray interactions; and

7. Time-varying backgrounds induced by precipitation, specifically from ^{214}Pb and ^{214}Bi.

III.A.1. Block Adjacency

Modeling large, contiguous city environments with high fidelity in Monte Carlo simulations presents a major computational challenge. Fully simulating an extended street with multiple radiation sources—both background and threat—quickly becomes impractical due to the memory and runtime requirements. While it is possible to simulate individual city blocks, the challenge is that radiation in one block can influence the detection signature in adjacent blocks. As a result, to maintain physical accuracy, neighboring blocks must be present in any simulation where interactions extend beyond a single block.

This observation was first noted during the Chameleon Street simulations, where most city blocks were approximately 100 m in length. It was found that radiation sources affected only their immediate neighboring blocks, and had negligible influence beyond that.

Building on this insight, the RADAI simulations adopt a modular approach based on three-block segments. In this scheme, each Monte Carlo simulation models a central block with its immediate neighbors, but only includes active radiation sources in the center block. Radiation flux quantities are recorded across all three blocks, enabling accurate representation of both local and adjacent contributions to the radiation field.

To simulate an arbitrarily long street, random sequences of blocks can be assembled, with each segment drawn from the precomputed library of three-block combinations. This approach significantly reduces computational overhead while preserving realism and spatial context. We refer to this method as *The Never-Ending Street*.

To illustrate the concept of *The Never-Ending Street*, consider a vehicle traveling through a series of blocks: $\ldots\ b_{i-2}\ b_{i-1}\ b_i\ b_{i+1}\ b_{i+2}\ \ldots$ While in block b_i the total count rate spectra seen at a particular position would be the sum of the count rate spectra from three 3-block combinations:

from b_{i-1}	sources in the combination	b_{i-2}	b_{i-1}	b_i			
from b_i	sources in the combination		b_{i-1}	b_i	b_{i+1}		and
from b_{i+1}	sources in the combination			b_i	b_{i+1}	b_{i+2}	

As the vehicle moves from block b_i to block b_{i+1}, the next block in the sequence is picked at random from the group of all 3-block combinations that start with blocks b_i and b_{i+1}. This process continues for as long as needed.

To support *The Never-Ending Street*, three new urban blocks (bridge, park, and tunnel, described in the following section) were added to the seven city blocks originally used in the *Chameleon Street* simulations. This expanded the total to ten distinct blocks ($N = 10$). In theory, this results in $N^3 = 1000$ possible 3-block combinations. If block repetition is not allowed (i.e., a block cannot immediately follow itself), the number reduces to $N(N-1)^2 = 900$.

To further limit the number of required Monte Carlo simulations, an adjacency matrix was defined to restrict which blocks are allowed to follow each other. This matrix, shown in Table I, defines the valid transitions between blocks. The constraints were selected to both preserve realistic sequences and also reduce the combinatorics.

With each block allowed to be followed by half of the others, the number of 3-block combinations requiring simulation drops significantly, from 900 to $N(N/2)(N/2) = 250$. To further reduce the computational burden, threat sources were simulated in only 90 of these 250 combinations. These 90 combinations were selected to ensure that each block appears as the center block in at least one 3-block simulation, with a representative mix of surrounding blocks.

III.A.2. New Urban Blocks

Three new blocks were added to the Chameleon Street city block models: a bridge over seawater, a park, and a tunnel. Each was designed to introduce features commonly found in larger coastal cities and to expand the environmental diversity of the dataset. The bridge block provides reduced gamma-ray background due to the absence of nearby NORM-containing structures. The park lacks large buildings facing the road, resulting in lower structural shielding and a more open

TABLE I
The adjacency matrix for the Never-Ending Street model

Block		Can be followed by block								
0	Bridge		1		3		5	6		9
1	Park			2		4	5	6	8	
2	Tunnel	0	1				5	6		9
3	3000 Gay St.		1			4	5	6	7	
4	4000 Gay St.	0		2			5		7	9
5	5000 Gay St.		1			4			7	8	9
6	6000 Gay St.			2	3	4			7	8	
7	7000 Gay St.	0		2	3					8	9
8	8000 Gay St.	0		2	3	4			7		
9	9000 Gay St.	0	1		3			6		8	

geometry. The tunnel block produces a more spatially uniform background due to its enclosed and repetitive structure. The tunnel and bridge geometries also represent an abrupt geometric change, which can be challenging for algorithms. These new blocks are schematically shown in Figure 1.

The original city blocks were mostly 320–360 ft (97.5–109.7 m) long, with the exception of the 4000 block, which was 500 ft (152.4 m). In contrast, the newly added bridge and tunnel blocks are each 1,000 ft (304.8 m) long, while the park block is 660 ft (201.2 m) long. All blocks, new and old, share a standard width of 660 ft (201.2 m) and include a cross street at one end.

The vertical reference plane ($z = 0$) corresponds to the top of the sidewalks along the main roads. Road surfaces are positioned 6 in (0.1524 m) below the sidewalks. As with the original blocks, the new blocks are constructed using asphalt, concrete, soil, and granite (gravel beneath the asphalt), but also feature water and trees. The original set also included brick and granite as common building materials.

The bridge block has a simple 2-foot thick, 4-lane concrete roadway 40 ft (12.192 m) wide over a large expanse of sea water that is 10 ft (3.048 m) deep. The sea water surface is 30 ft (9.144 m) below the top of the roadway. The only NORM source included in the sea water is ^{40}K.

The park block is primarily composed of soil and features a central roadway identical in structure to those in the original city blocks. The road is 40 ft (12.192 m) wide and constructed from 1 ft (0.3048 m) of asphalt over 1 ft (0.3048 m) of granite gravel. It is flanked by 6 ft (1.8288 m) wide sidewalks, each 0.5 ft (0.1524 m) thick. A freshwater lake, 18 ft (5.4864 m) deep, is located 50 ft (15.24 m) from the centerline of the roadway. This lake contained no radioactive sources. The park also includes gently sloped hills that are spherical caps 20 ft (6.096 m) high with 50 ft

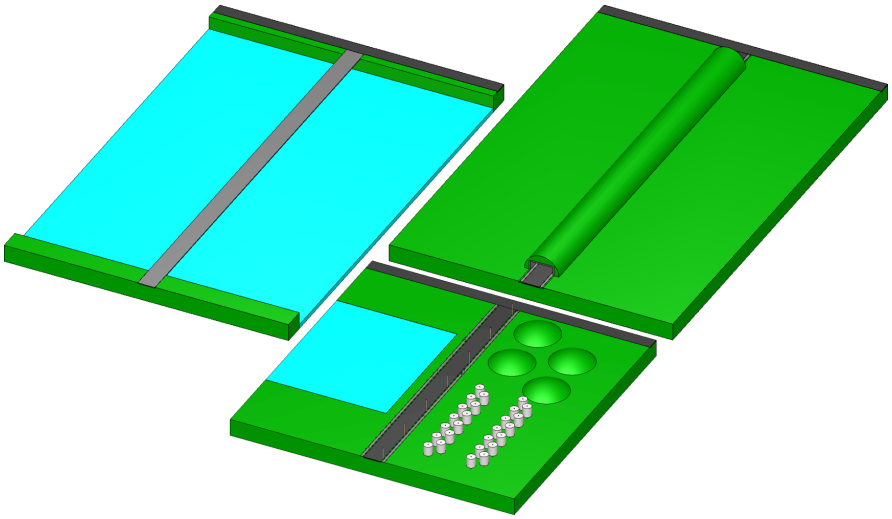


Fig. 1. The new blocks are the bridge block (upper left), the park block (lower center), and the tunnel block (upper right). Materials are soil as green, concrete as medium gray, asphalt as dark gray, and water as light blue and tree canopy as white.

(15.24 m) radii at their bases. Trees in the park are modeled as 30 ft (9.144 m) tall cylinders with a 2 ft (0.6096 m) diameters, surrounded by low-density canopies that extends down to 8 ft (2.438 m) above the ground.

The tunnel block features roadways and sidewalks consistent with those in the other blocks. The tunnel structure includes 3 ft (0.9144 m) thick interior concrete side walls and a roof, all encased within a half-cylinder of soil with a radius of 40 ft (12.192 m).

All three new blocks include the same type of cross-street as in the original models: a 30 ft (9.144 m) wide roadway composed of 1 ft (0.3048 m) of asphalt over 1 ft (0.3048 m) of granite gravel. As with the original city blocks, material compositions and densities are sourced from the Compendium of Material Composition Data for Radiation Transport Modeling [5].

III.A.3. Clutter

To better reflect real-world urban environments, several types of clutter were added to the seven original city blocks and to the park block. These additions include pedestrians and telephone poles along the sidewalks, as well as parked cars in the streets and in selected parking lots.

The pedestrian models, shown in Figure 2, were based on stylized representations of the adult male and female phantoms commonly used in health physics [6]. Each figure was composed

of simplified homogeneous soft tissue volumes: cone-shaped legs, an elliptical cylindrical torso, and an elliptical neck and head with a half-ellipsoid crown. The female phantom is 5 ft 4.5 in (1.6383 m) tall and weighs 121.9 lbs (55.31 kg), while the male phantom is 5 ft 8.5 in (1.74 m) tall and weighs 160.3 lbs (72.71 kg). Along the sidewalks, two rows of evenly spaced positions were defined every 3 ft (0.9144 m). Each position was randomly assigned to be occupied by a female phantom (2% probability), a male phantom (2%), or left empty (96%). Once selected, the placement of each phantom remained fixed for all simulations.

The vehicle model was based on a Dodge minivan [7], and includes a steel frame, engine, side panels, roof, and glass windows. Each component was modeled with simplified geometry and assigned scaled-down material densities to approximate realistic weight distributions. The initial model had a total mass of approximately 7,000 lbs (3,175 kg), which was heavier than a typical minivan. To better match a realistic curb weight, the densities of the steel components were reduced, resulting in a final model mass of 4,300 lbs (1,950 kg). Wheels were added for visual completeness. The final vehicle geometry is shown in Figure 3. When parked on the street, vehicles were placed in 21 ft (6.4008 m) long spaces. In parking lots, spaces were 9 ft (2.7432 m) wide. Not all parking spaces were filled.

Telephone poles were modeled as tapered wooden cylinders, 1 ft (0.3048 m) in diameter at the base, narrowing to 8 in (0.2032 m) at the top, with a total height of 30 ft (9.144 m). Each pole included two wooden crossbars near the top, as shown in Figure 4. Poles were placed on one side of the street at intervals of 100 ft (30.48 m).

A detailed view of the 7000 block, including pedestrians, telephone poles, and parked vehicles, is provided in Figure 5. Figure 6 shows all Chameleon Street blocks that were enhanced with clutter.

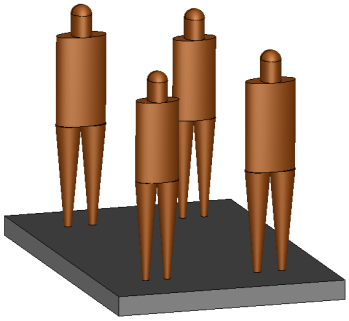


Fig. 2. Models of people along the sidewalks (two sizes: female and male).

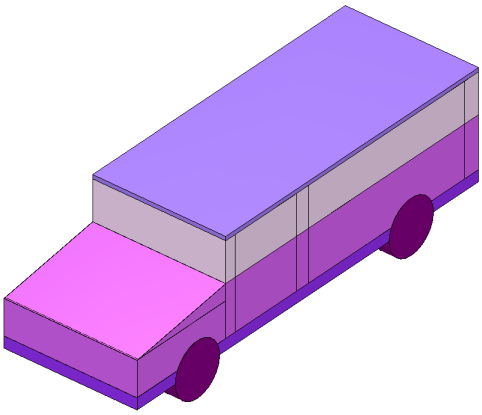


Fig. 3. The minivan model.

Fig. 4. The telephone pole model.

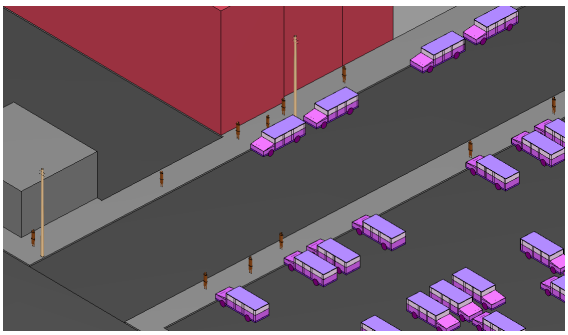


Fig. 5. Clutter in block 7 showing people, telephone poles and parked cars.

Terrestrial Background Radiation

Terrestrial background sources were simulated separately for each NORM component in each material within each block at a concentration of $1\,\mathrm{Bq\,kg^{-1}}$. The actual strengths of each are later set through multiplicative scaling in the synthetic data generation codes as described in

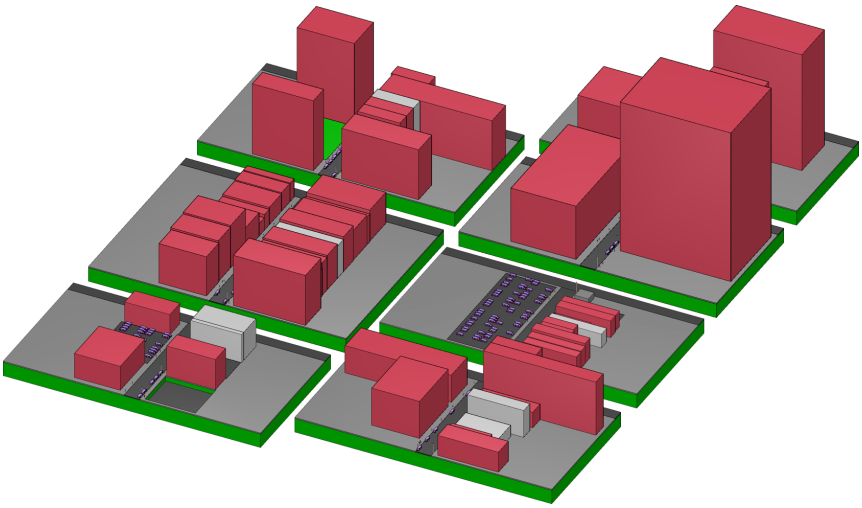


Fig. 6. The seven Chameleon Street blocks with clutter. Materials are soil as green, granite as light gray, concrete as medium gray, asphalt as dark gray, and brick as red. The left side from front to back shows blocks 3, 4, and 5. The right side from front to back shows blocks 6, 7, 8 and 9.

Section III.D), allowing for user-defined variation in the background spectra. The materials used in the different blocks are listed in Table II with identification numbers (IDs) from Ref. [5]. The NORM components contained in each material are listed in Table III.

III.B. Cosmic Background Modeling

To account for the contribution of cosmic-ray secondaries to the background radiation field, we conducted a two-stage simulation campaign using MEGAlib [8], built on Geant4 [9]. In the

TABLE II
Materials in each block

Block		Length (cm)	Material IDs present					
0	Bridge	30480.0	19		97	105	267	279
1	Park	20116.8	19		97	105	267	
2	Tunnel	30480.0	19		97	105	267	
3	3000 Gay St.	10972.8	19	46	97		267	
4	4000 Gay St.	15240.0	19	46	97		267	
5	5000 Gay St.	10820.4	19	46	97	105	267	
6	6000 Gay St.	10820.4	19	46	97		267	
7	7000 Gay St.	10515.6	19	46	97		267	
8	8000 Gay St.	10515.6	19	46	97			
9	9000 Gay St.	9753.6	19	46	97			

TABLE III
NORM components in each material

Material[a]	ID	density (g cm^{-3})	Volumetric background NORM components[b]			
Asphalt Pavement	19	2.5784	^{40}K	^{232}Th	^{238}U/^{235}U	
Brick, Common Silica	46	1.8	^{40}K	^{232}Th	^{238}U/^{235}U	
Concrete, Ordinary (NIST)	97	2.3	^{40}K	^{232}Th	^{238}U/^{235}U	
Earth, US. Average	105	1.51	^{40}K	^{232}Th	^{238}U/^{235}U	^{137}Cs/^{137m}Ba
Rock, Granite	267	2.69	^{40}K	^{232}Th	^{238}U/^{235}U	
Sea Water, Standard	279	1.023343	^{40}K			

[a]Material composition and density taken from Ref. [5] .
[b]Thorium and uranium sources included all daughters in secular equilibrium .

first stage, primary protons with energies above 1 GeV were simulated incident on a 42-layer planar atmospheric model, producing photon fluences at sea level as a function of energy and zenith angle. These tallies were then processed to obtain spectral irradiance distributions across 5° zenith angle bins.

In the second stage, each angle-energy bin was used to construct a far-field photon source incident on a virtual spherical shell surrounding a $2'' \times 4'' \times 16''$ NaI(Tl) detector. The resulting deposited-energy spectra were analyzed to estimate the detector response to the cosmic photon field. These simulations were validated against real measurements from a detector deployed for 3.8 days on the roof of Building 50B at Lawrence Berkeley National Laboratory (LBNL). The measured and simulated spectra show strong agreement in spectral shape above 4 MeV, though a scale factor of 2.197 was required to match the measured intensity (see Figure 7). This scaling discrepancy may be attributable to local environmental effects not captured in the simulation (e.g., ground plane reflection, altitude variation, or contributions from non-photon secondaries).

To adapt the cosmic background for site-specific urban geometries, each zenith-angle bin was weighted according to the fraction of visible sky computed for that angle at each source position in the RADAI simulation grid. The resulting cosmic pulse-height contribution was then added to the total background spectrum at each position. A full description of the cosmic simulation and validation methodology will be presented in a separate manuscript.

III.C. Rain-Induced Background Modeling

Rain-induced background sources in the dataset consist of ^{214}Pb and ^{214}Bi, modeled independently without their respective decay chains. Simulated activity was confined to flat ground-level

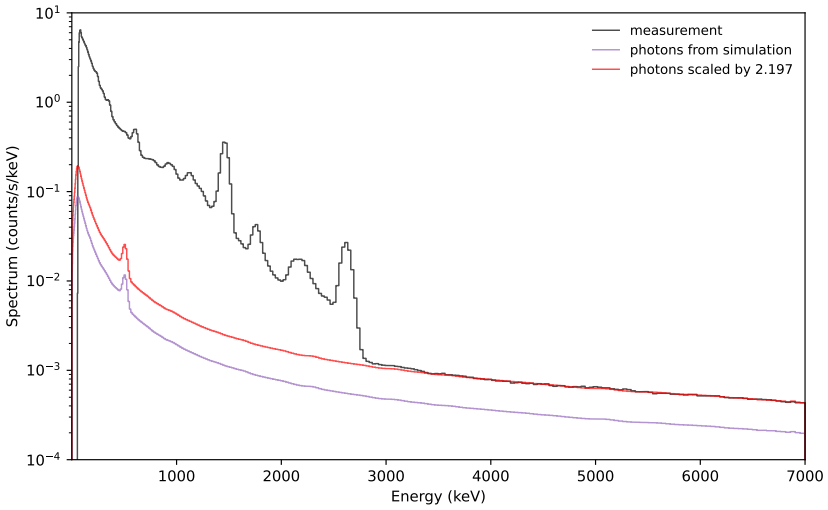


Fig. 7. Measured $2'' \times 4'' \times 16''$ NaI(Tl) data from the roof of LBNL Building 50B, shown with the results of the cosmic-ray secondary photon simulations (magenta) and the scaled photon spectrum (red). Detector energy resolution has not been applied to the simulation.

surfaces—such as streets and sidewalks—where rainwater and deposited radionuclides would realistically accumulate. Rooftop sources were excluded under the assumption that gamma emissions from elevated surfaces would have minimal impact on street-level detector response. Each isotope was simulated with an activity of $1\,\mathrm{Bq\,m^{-2}}$, allowing for subsequent time-dependent scaling based on modeled rain deposition.

To derive the temporal source terms (i.e., activity in $\mathrm{Bq\,m^{-2}}$ as a function of time), we developed a dynamic model of radon progeny washout and surface accumulation. This model was calibrated against field measurements collected over three years using a $2'' \times 4'' \times 16''$ NaI(Tl) detector deployed at Oak Ridge National Laboratory (ORNL). Monte Carlo simulations of the detector were performed to estimate its energy-dependent response to uniform surface sheets of ^{214}Pb and ^{214}Bi separately. These detector responses were combined with an estimate of the static background spectrum to construct time-evolving full-spectrum models driven by the decay kinetics of the radon daughters.

The model also incorporated a rainfall-driven source term, allowing for time-varying input of activity during precipitation events. In most cases, the inferred source terms closely followed the temporal structure of corresponding rain gauge data. A total of 30 distinct rain events of varying durations and intensities were modeled from the ORNL dataset and used as representative rain

profiles in the RADAI simulation framework. The rain events ranged from few-minute downpours to day-long extended events. Figure **??** shows the measured and fitted spectra for a representative real rain event, alongside a its time-series count-rate curve. The full simulation and validation methodology is described in detail in a separate manuscript [10].

III.C.1. Nuisance and Threat Sources

The list of threat sources is described in Table IV and is mostly derived from the standards developed by the National Committee on Radiation Instrumentation, N42 [11, 12, 13, 14, 15]. This includes nine 'basic' isotopes, including NORM and those that are commonly used for industrial and research applications. These isotopes were simulated either as nearly bare or with a thick coating to simulate a shielded container. Four medical sources, including their minor contaminants, were also developed. These were modeled as emitted from within a 8 cm-thick poly(methyl methacrylate) (PMMA) shield. The basic sources and medical sources were all simulated at a strength of 1 Bq, so they could be scaled up to any strength by synthetic data generation codes.

Each of the seven SNM threat source types were simulated for three different masses and with/without 1 cm of steel shielding. The isotopic compositions of the uranium materials were taken from Ref. [16] for depleted uranium (DU), natural uranium (with all daughters included), refined uranium (no daughters), low-enriched uranium (3% in ^{235}U), and highly enriched uranium (HEU) (93.3% in ^{235}U). The compositions for fuel grade (12% ^{240}Pu) and weapons grade plutonium (WGPU) (6% ^{240}Pu) are taken from the DOE standard [17].

Several sources were suggested by the Detecting Radiation Algorithms Group (DRAG) [18, 19]: ^{67}Cu, a ^{90}Sr/^{90}Y bremsstrahlung source, and additional shielding configurations of ^{192}Ir. A total of 72 threat sources were used. More detail is given on the basic, medical and SNM sources in a previous report [20]. The DRAG sources were developed using the same methods as the basic isotopes.

The potential source locations for each block are listed in Table V. These are the same locations used in making the Chameleon Street synthetic data, with new locations chosen for the new blocks. All 72 threat sources were simulated in each source location of a given block.

Simulated detector responses for point sources of the bare isotopic sources and the medical sources (in 8 cm of PMMA) are shown in Figures 8–10. Figure 11 shows two examples of plutonium

TABLE IV
Source types

Basic (9×2): 1 Bq
Isotopes: ^{40}K, ^{57}Co, ^{60}Co, ^{137}Cs, ^{133}Ba, ^{192}Ir, ^{226}Ra, ^{232}Th, ^{241}Am
Shielding: bare (0.025 cm steel) and shielded (1 cm steel)
Medical (6): 1 Bq
Isotopes: ^{18}F, ^{99m}Tc, ^{131}I, ^{133}Xe, ^{177}Lu, ^{201}Tl
Shielding: always in 8 cm PMMA
Nuclear Material (7×3×2):
materials: DU, U ore, Refined U, LEU, HEU, FGPu, WGPu
Uranium sizes: 2.5, 10, and 25 kg; Plutonium sizes: 0.5, 2, and 8 kg
Shielding: bare and shielded (1 cm steel)
DRAG Sources (3×2): 1 Bq
^{67}Cu: bare (0.025 cm steel) and shielded (1 cm steel)
^{90}Sr/^{90}Y in Al: 1 cm and 2 cm thicknesses
^{192}Ir in steel: 2 cm and 5 cm thicknesses

TABLE V
Threat source locations by block

Block	ID	Description of location	Location (cm) x	y	z	Distance from road centerline (feet)	(m)
0	01	middle of the bridge, just inside of the road	594.36	15239	100	19.5	5.9436
	02	halfway between cliff edge and side street	1524	29260	100	50	15.24
1	11	centered in the trees	4572	5486	100	150	45.72
	12	centered in the four hills	4724.4	14630	100	155	47.244
2	21	10 feet inside tunnel, 2 feet from wall	731.52	1830	100	24	7.3152
	22	middle of the tunnel, 2 feet from wall	731.52	15239	100	24	7.3152
3	31	side of building	-1524	4267.2	100	50	15.24
	32	parking lot (10' 4" below sidewalk)	1524	1813.56	-341.96	50	15.24
4	41	front of building	1127.76	10059.4	100	37	11.2776
	42	between two brick buildings*	1219.2	2849.88	100	40	12.192
	43	interesting corner	-1066.8	2377.44	100	35	10.668
5	51	interesting corner	1219.2	5821.68	100	40	12.192
	52	between two brick buildings	1828.8	2484.12	100	60	18.288
6	61	between two granite buildings	1524	4770.12	100	50	15.24
	62	side of building	-1828.8	1524	100	60	18.288
7	71	parking lot	-3048	4800.6	100	100	30.48
	72	interesting corner	1066.8	2590.8	100	35	10.668
8	81	side of building	2133.6	2834.64	100	70	21.336
	82	side of building	-2438.4	487.68	100	80	24.384
9	91	side of building	3048	8229.6	100	100	30.48
	92	front of building	-2438.4	4099.56	100	80	24.384

*Source location 42 is between two brick buildings and behind a brick column.

spectra and Figure 12 shows examples of uranium spectra. Note the similarity in the shape of the depleted, refined and low-enriched uranium spectra, which overlap except for below 200 keV. Detail below 200 keV for the lower enrichment uranium spectra is shown in Figure 13. For the other masses of SNM sources, please see the previous report [20].

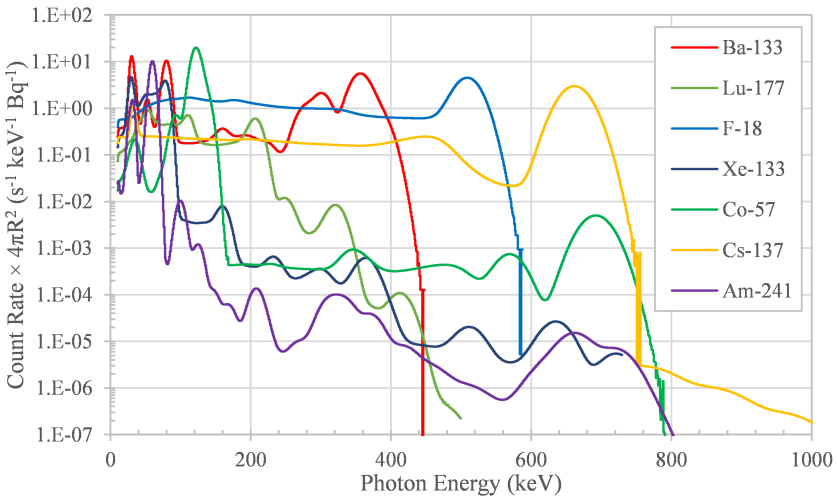


Fig. 8. Simulated $2'' \times 4'' \times 16''$ NaI(Tl) responses for point sources of ^{133}Ba, ^{18}F, ^{57}Co, ^{137}Cs and ^{241}Am.

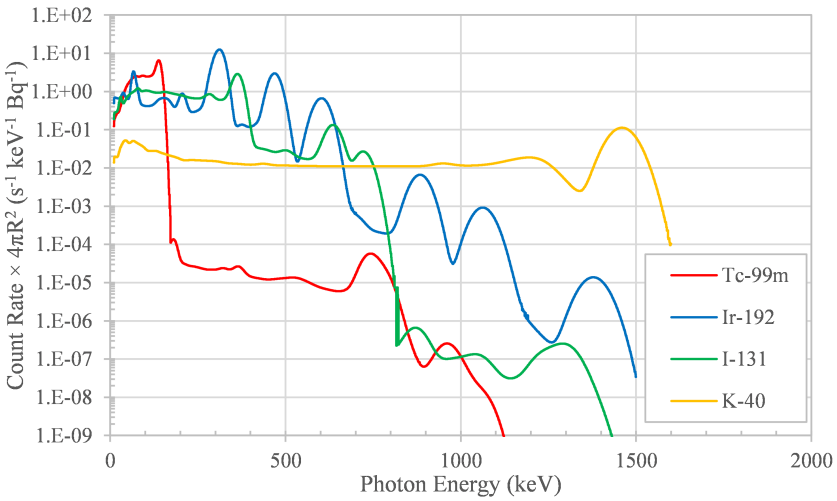


Fig. 9. Simulated $2'' \times 4'' \times 16''$ NaI(Tl) responses for point sources of ^{99m}Tc, ^{192}Ir, ^{131}I and ^{40}K.

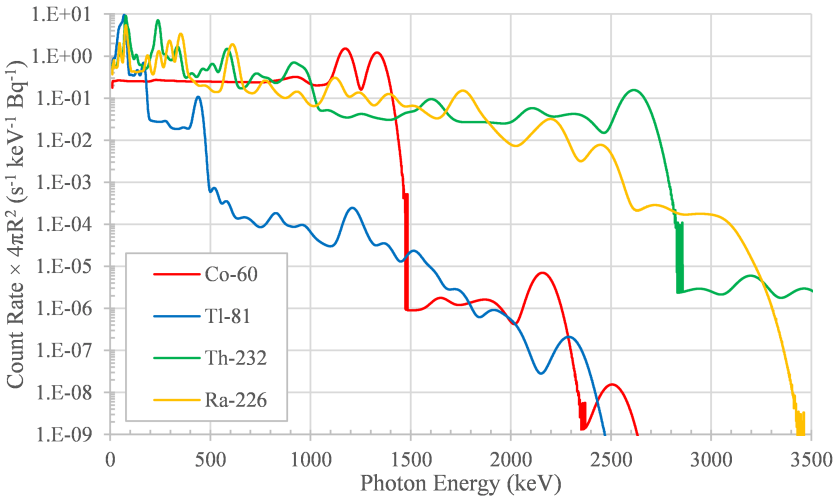


Fig. 10. Simulated $2'' \times 4'' \times 16''$ NaI(Tl) responses for point sources of ^{60}Co, ^{201}Tl, ^{232}Th, and ^{226}Ra.

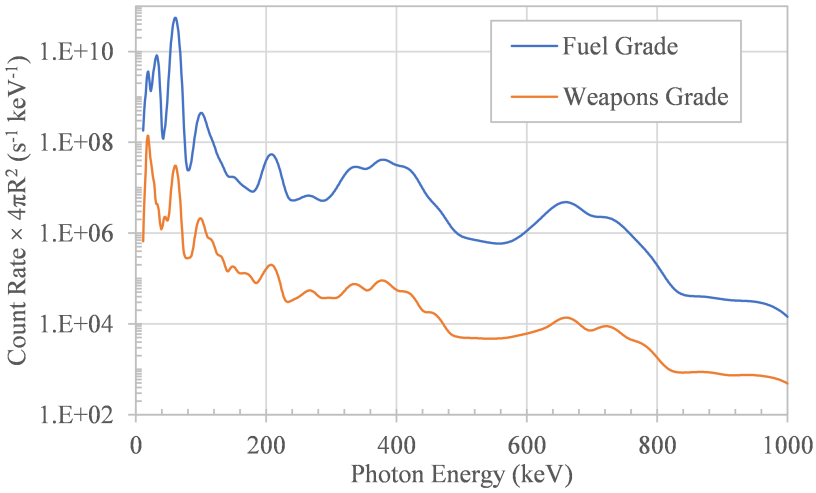


Fig. 11. Simulated $2'' \times 4'' \times 16''$ NaI(Tl) responses 2 kg of fuel grade and 2 kg of weapons grade plutonium.

III.C.2. Simulation Outputs

The results of an individual Monte Carlo simulation for a given 3-block combination are two mesh tallies of energy-dependent flux. The mesh tallies extend the length of the 3-block combination in the center of each lane of travel in the roadway and comprise contiguous cubic voxels of 1 m side-length. Each voxel of each mesh tally is post-processed to compute the energy-dependent mean energy deposition spectrum, using a convolution of the energy-dependent flux with

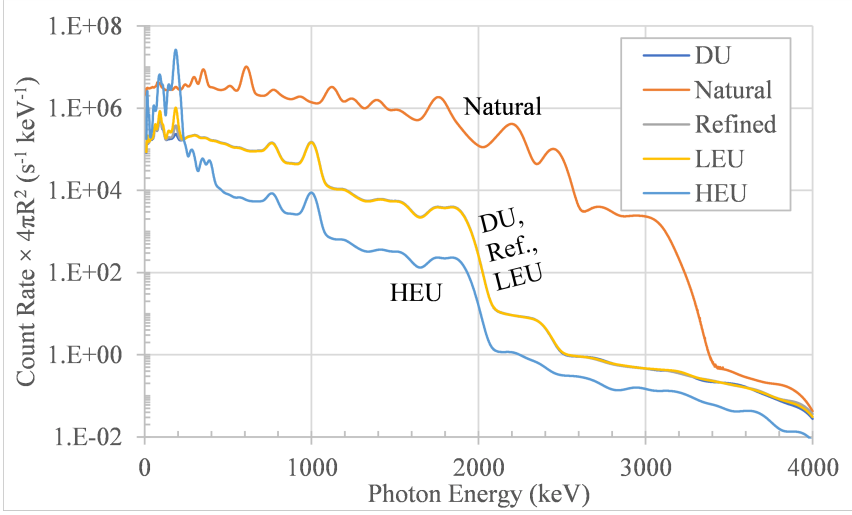


Fig. 12. Simulated $2'' \times 4'' \times 16''$ NaI(Tl) responses 5 kg of different types of uranium. Note, depleted-, refined-, and low-enriched-uranium spectra are nearly overlapping.

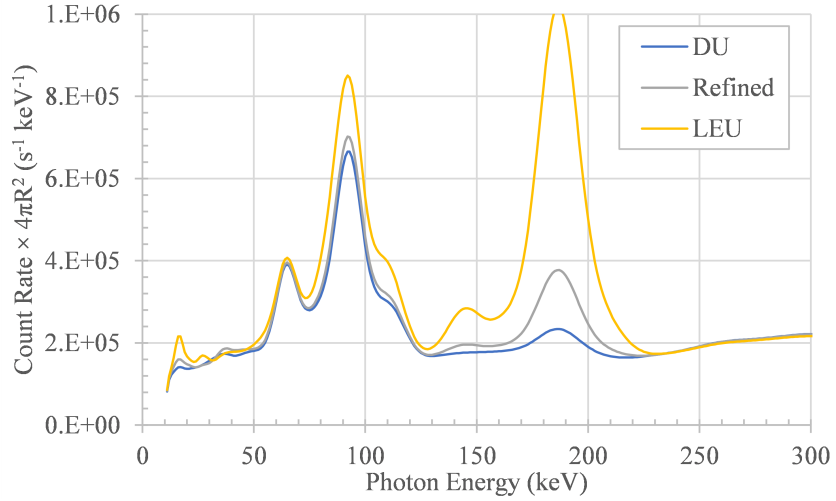


Fig. 13. Simulated $2'' \times 4'' \times 16''$ NaI(Tl) responses 5 kg of low-enriched uranium zoomed-in on lower energies.

a precomputed $2'' \times 4'' \times 16''$ NaI(Tl) response function. At this point, the only effect necessary to create a predicted detector spectrum is the detector's energy resolution, which is subsequently applied by the synthetic data generation codes (section III.D).

The real goal of the Monte Carlo simulations was to create realistic count rate spectra to sample from, at each location along the path of the vehicle. Like real measurements, Monte Carlo is subject to stochastic error which is reduced with longer run times. The goal then is to allow

the Monte Carlo enough time that the stochastic noise does not manifest in the computed count rate spectra in a manner that could confuse spectroscopic algorithms. In the Chameleon Street simulations, background calculations were run for 100 hours and threat source calculations were run for 24 hours. The background calculations need more time because they must converge well over the length of the street, while the threat sources only need to be converged near the source, where the flux is not smaller than some fraction of the background flux. Test runs for this project showed that these times were reasonable for simulations using the 3-block combinations.

For the 250 3-block combinations, 90 of which have threat sources, the number of background simulations required is (250 combos) $\times$ (12.2 volumetric material/isotope sources + 2 rain sources) = 3,550 and the number of threat source simulations is [81 $\times$ (2 locations) + 9 $\times$ (3 locations)] $\times$ (72 sources) = 13,608. With times of 100 hours for background simulations and 24 hours for threat source simulations, as well as some time for 250 $\times$ 6 = 1,500 hours for adjoint calculations for variance reduction, this adds up to 683,092 cpu·hours (or about 77.9 cpu·years). A total of 34,441 mesh tally files of energy-dependent flux were created. These were convolved with a $2'' \times 4'' \times 16''$ NaI(Tl) response function to give the same number of pulse height spectra mesh tally files. The total file size of all the mesh tallies was 1,980 GB.

III.D. Generation of Synthetic List-mode Detector Data

The process of generating a simulated dynamic detector measurement has five main steps listed below.

1. Street geometry and travel kinematics sampling;
2. Material background composition sampling;
3. Environmental sampling;
4. Spectrum processing; and
5. List-mode sampling and aggregation

Each of these five steps is carried out in order, with each step having a set of user-defined parameters for designing the contiguous set of data, each of which is referred to as a run. The following sections describe the methodology, user-defined inputs, and final outputs for each of these steps.

III.D.1. Street Geometry and Travel Kinematics

For a given run, the sampling framework must first sample a chain of blocks through which the simulated detector will traverse. This chain of blocks is herein referred to as the street geometry, with a specific street geometry for run r, G_r, given as a list of block IDs with n blocks, $G_r = \{b_0, b_1, b_2, \ldots, b_n\}$. Unlike in the previous dataset (Ref. [1]), the detector speed profile is stochastic, nonlinear and variable throughout the run. As such, n is unknown *a priori*, being dependent on both the sampled speed profile and the desired run time, t_r. Runs are generated by sampling the street geometry and detector travel kinematics (velocity, acceleration, stopping, etc.) parameterized as listed in Table VI. The mechanisms by which these variables are used and their effects on the generated run are outlined in the following sections.

TABLE VI
User-defined parameters for street geometry and detector kinematics sampling

Symbol	**Description**	**Units**
C	Set of all 3-block combinations available in a run	N/A
t_r	Target duration of the simulation run	s
v_{min}	Minimum target velocity in a speed zone	m/s
v_{max}	Maximum target velocity in a speed zone	m/s
$l_{\text{zone,min}}$	Minimum length of a speed zone	blocks
$l_{\text{zone,max}}$	Maximum length of a speed zone	blocks
p_{stop}	Probability of stopping at a stoplight at the end of a block	%
$t_{\text{stop,min}}$	Minimum stop time at a stoplight	s
$t_{\text{stop,max}}$	Maximum stop time at a stoplight	s
$d_{\text{stop,max}}$	Max distance from end of block where the detector may stop	m
P_B	Constraints on min/max entry probabilities for each block b_i used to construct the transition matrix	%

Let B be the set of all possible unique blocks (of length m) for which radiation data were simulated (in this case, the 11 blocks shown in Figs. 1 and 6). At the beginning of run generation, the first block in G_r is randomly sampled from B, with each block having equal sampling probability. After the first block is randomly sampled, a block transition matrix, unique to the run, $T_{B,r}$ is generated from P_B, which is a set of conditions describing the minimum and maximum probabilities of entering a given block b_i for all possible block transitions. Note that some block transitions are not possible. For example, one such condition might be $0.1 \leq P_{b_4} \leq 0.5$, which states that the probability of entering block b_4 must be between 10 and 50%. After this matrix is created, it is normalized so that block transition probabilities for each block sum to one.

The process of iteratively sampling G_r from $T_{B,r}$ is modeled as a discrete-time Markov chain, with $T_{B,r}$ being a right stochastic matrix. The stationary distribution of a given $T_{B,r}$, π_r, is a row vector where each element represents the total probability of entering each block. π_r is calculated by performing eigendecomposition of $T_{B,r}$ to yield the eigenvectors, v. π_r is then determined by $\pi_r = \frac{v}{\sum_i v}$.

For a given run, the algorithm for generating a $T_{B,r}$ that meets the criteria in P_B is given in Algorithm 1. For some specific combinations of P_B, there may be no solution found by the algorithm but the values of P_B used for the datasets generated in this project were solvable.

Algorithm 1 Algorithm for generating a transition matrix based on block probability constraints.

procedure GENERATETRANSITIONMATRIX(P_B,m) ▷ m is the number of unique blocks available to this run
  $valid \leftarrow$ false
  while not $valid$ **do** ▷ loop until transition matrix yields valid stationary distribution
    $T_{B,r} \leftarrow \text{random}(m, m)$ ▷ generate a $m \times m$ matrix of random values between 0 and 1
    $v \leftarrow \text{eigenvec}(T_{B,r})$ ▷ calculate eigenvectors of $T_{B,r}$
    $\pi_r \leftarrow v/\text{sum}(v)$ ▷ calculate stationary matrix of $T_{B,r}$
    $valid \leftarrow$ true
    for $i \leftarrow 1, m$ **do**
      if $\pi_{r,i} \leq \min(P_{B,i})\pi_{r,i}$ **or** $\pi_{r,i} \geq \max(P_{B,i})$ **then**
        $valid \leftarrow$ false
        break
  return $T_{B,r}$

III.D.2. Kinematic Modeling of Detector Motion

As previously noted, the speed and acceleration profile of the detector—referred to here as the kinematics profile—as it traverses the street network is stochastic and variable. To ensure realistic representation of urban driving behavior, we adopted the empirical polynomial model of acceleration developed by Akçelik and Biggs [21], specifically their Equations 4 and 5, which are based on high-fidelity driving data collected under urban, suburban, and rural conditions by the University of Sydney.

The simulation divides the street network into speed zones, each consisting of a contiguous sequence of blocks with uniform target velocity. The target velocity for a given zone, $v_{t,\text{zone}}$, is sampled from a uniform distribution $U(v_{\min}, v_{\max})$, and the number of blocks in the zone, l_{zone}, is sampled from $U(l_{\text{zone,min}}, l_{\text{zone,max}})$. The total physical length of the speed zone is denoted d_{zone}. For each speed zone, the algorithm determines whether the detector can reach $v_{t,\text{zone}}$ within d_{zone},

and decelerate (if needed) to the target velocity of the next zone, $v_{t,\text{zone}+1}$. If not, a lower reachable velocity is calculated and used as the effective target.

To simulate interruptions such as stoplights, a stopping event may be inserted within any block with probability p_{stop}. When a stop occurs, the stop location is sampled from $U(0, d_{\text{stop,max}})$, and the dwell time is sampled from $U(t_{\text{stop,min}}, t_{\text{stop,max}})$. The algorithm checks whether deceleration to a full stop is physically feasible given the current velocity and distance remaining; if not, upstream adjustments to the velocity profile are made. To anticipate such events, the software samples the upcoming street geometry and stoplight configuration at least two blocks ahead. This is required to safely decelerate from the maximum allowed speed of 13.4 m s^{-1} to a full stop over a worst-case distance of 30 m. Acceleration and deceleration distances are computed numerically using the velocity profile $v(t)$, integrated via the trapezoidal rule with a timestep of $\Delta t = 10^{-5}$ s.

The full block sampling and kinematic simulation proceeds iteratively until the desired run duration t_r is reached. At the end of each run, the velocity v_t, acceleration a_t, distance d_t, and sampled street geometry G_r are saved for subsequent use in spectrum synthesis. Additionally, the kinematics simulation provides the duration of the detector in each mesh tally voxel, which is used to scale the count rate spectra for Poisson sampling as described in Section III.D.6. The full kinematics simulation algorithm is provided in pseudocode in Algorithm 2.

Algorithm 2 Urban Detector Kinematics Simulation

Initialize $t \leftarrow 0$, $v_t \leftarrow 0$, $d_t \leftarrow 0$, $a_t \leftarrow 0$
Sample initial speed zone parameters: $v_{t,\text{zone}} \sim U(v_{\min}, v_{\max})$, $l_{\text{zone}} \sim U(l_{\text{zone,min}}, l_{\text{zone,max}})$
Compute total zone length: $d_{\text{zone}} \leftarrow \text{block\_length} \times l_{\text{zone}}$
Pre-sample $v_{t,\text{zone}+1}$ and stoplight state for next two blocks
while $t < t_r$ **do**
    if detector must stop in current block with probability p_{stop} **then**
        Sample stop location $d_{\text{stop}} \sim U(0, d_{\text{stop,max}})$
        Sample stop duration $t_{\text{stop}} \sim U(t_{\text{stop,min}}, t_{\text{stop,max}})$
        if braking feasible over d_{stop} **then**
            Adjust velocity profile to decelerate to 0 before d_{stop}
        else
            Adjust upstream velocity target
    Use Akçelik–Biggs model to compute acceleration a_t at current v_t
    Update v_t, d_t via trapezoidal integration: $\Delta t = 10^{-5}$ s
    if current zone end reached or v_t cannot reach $v_{t,\text{zone}}$ in d_{zone} **then**
        Sample new $v_{t,\text{zone}+1}$, l_{zone}, and update d_{zone}
    Advance time: $t \leftarrow t + \Delta t$
Store final time series: velocity v_t, acceleration a_t, distance d_t, and geometry G_r

III.D.3. Material NORM Composition

After the street geometry and detector kinematics profile are sampled, the concentrations of NORM in each material in the model (see Table III) are selected. These concentrations are fixed for a sampled number of blocks, then abruptly change to a new set of concentrations. The road-length in units of blocks of each NORM concentration zone is determined by uniform sampling between user-specified minimum and maximum lengths.

In the *Chameleon Street* dataset, the material NORM concentrations were varied uniformly within 80% of the mean concentrations measured at a specific real-world location [22]. For this dataset, however, a literature review was performed and 156 individual measurements of NORM concentrations from materials around the globe were compiled [23, 22, 24, 25, 26, 27]. The NORM

components are abbreviated as K, U, T, and C for ^{40}K, ^{238}U/^{235}U + progeny, ^{232}Th + progeny, and ^{137}Cs, respectively. The NORM concentrations used in this work are shown in Figure 14 and most often show fairly strong correlations between K, U, and T for each material type. In order to preserve these correlations and to produce probability distributions from which to sample, two-component Gaussian mixture models were fit to data from each material type. Correlation matrices from the literature survey data and the Gaussian mixture model probability distributions are shown in Table VII. The only material for which a distribution was not empirically fit to the discrete histogram data was seawater. Seawater is unique in that uranium and thorium isotopes (and their progeny) contribute less than 20% of the net activity from seawater and the concentration of ^{40}K is fairly constant globally at a mean concentration of 12.056 Bq kg^{-1} with a standard deviation of 1.2056 Bq kg^{-1} [28]. For the creation of training data, the Gaussian mixture model probability distribution derived from the data was used. For testing data, the mean and standard deviation of the Gaussian mixture model probability distributions were both increased by 20%, see Figure 15. Some of the Gaussian mixture model probability distributions stretched below zero concentrations for K, U, and T. In sampling from these distributions, if a sample had a concentration less than zero another sample was drawn from the distribution until all concentrations were above zero.

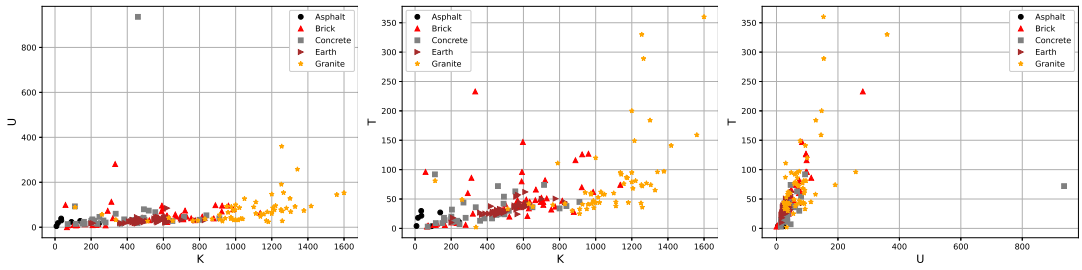


Fig. 14. NORM concentrations for used in this work for each material type compiled through an extensive literature review of material NORM concentrations measured around the globe [23, 22, 24, 25, 26, 27]

Multi-block sections of real cities are often zoned according to purpose (industrial, residential, etc.), constructed by the same contractor, and/or subject to development at similar times, resulting in the use of similar construction materials. As such, each run is broken into NORM concentration zones, as described above. The concentrations of each NORM component for each material in Table III are sampled from the distributions given in Figure 15. The sampled concentration of NORM component c for material ID m is given as $C_{c,m}$ and is measured in units of Bq kg^{-1}: e.g.,

TABLE VII
Correlation coefficients calculated on the NORM datasets for asphalt, brick, concrete, earth, and granite materials compared with those calculated from the Gaussian mixture model probability distributions.

Material		NORM Data			Gaussian Mixture Model		
		K	U	T	K	U	T
Asphalt	K	1	-0.02	-0.08	1	-0.02	-0.09
	U	-	1	0.79	1	1	0.79
	T	-	-	1	-	-	1
Brick	K	1	0.25	0.36	1	0.23	0.35
	U	-	1	0.94	1	1	0.94
	T	-	-	1	-	-	1
Concrete	K	1	0.10	0.42	1	0.09	0.41
	U	-	1	0.44	1	1	0.43
	T	-	-	1	-	-	1
Earth	K	1	0.83	0.95	1	0.83	0.95
	U	-	1	0.61	1	1	0.61
	T	-	-	1	-	-	1
Granite	K	1	0.42	0.48	1	0.42	0.48
	U	-	1	0.67	1	1	0.68
	T	-	-	1	-	-	1

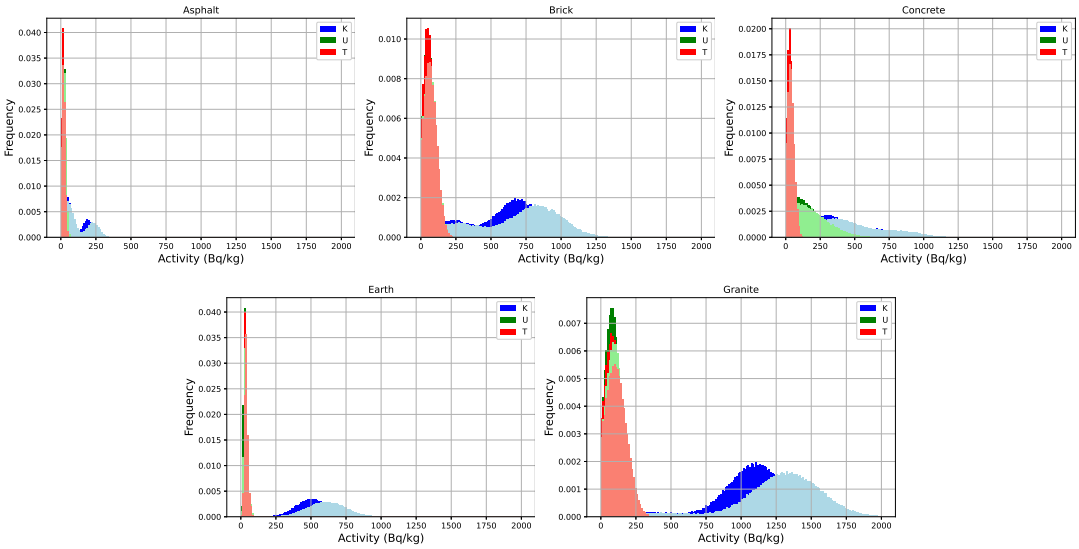


Fig. 15. Training and testing Gaussian mixture model distributions for Asphalt, Brick, Concrete, Earth, and Granite. The testing models have a 20% increased mean and standard deviation.

the concentration of ^{40}K in asphalt would be $C_{\mathrm{K},19}$.

For each block b_i in G_r, the unscaled pulse height spectrum mesh tally for a particular NORM component c and material m in that block is calculated by summing the contributions

from block b_i and the adjacent blocks, b_{i-1} and b_{i+1}. The pulse height spectra are then scaled by $C_{c,m}$ for the current constant-NORM zone to yield scaled per-second energy deposition spectra, $S_{i,m,c}$. Unlike in the *Chameleon Street* dataset, each simulated detected NORM component is kept separate throughout the entire run generation process in order to allow one to analyze the effects of individual NORM components to the detector's signal and algorithm's response. As such, the outputs of this step is a list, V, containing four vectors (one for each NORM component c) of scaled pulse height spectra, S_c, calculated by summing $S_{m,c}$ across each material m containing NORM component c.

III.D.4. Environmental Factors

Including the effects of rain in the dataset involved using the source terms for ^{214}Pb and ^{214}Bi that were modeled using real rain events as described in Section III.C. The time-series surface concentrations of ^{214}Pb and ^{214}Bi as C_{Pb} and C_{Bi}, respectively, were used to scale the simulated count rate tallies in the model to simulate the detector's response to rain events as it moves through the environment. Table VIII lists the parameters that must be set for the rain sampling step of the synthetic run generation framework.

TABLE VIII
User input variables for rain sampling.

Symbol	Description	Units
P_{r}	probability of a rain event occurring in this run	decimal percent
$P_{\mathrm{r,start}}$	If it does rain in this run, the probability of starting the run during the rain event	decimal percent
R	a set of all possible rain events in database	N/A

Each simulated run has a probability that a rain event will occur during the run, P_{r}. Initially, a random value sampled according to $U(0.0, 1.0)$ and compared to P_{r} to determine if the run will contain a rain event. If it does, a single rain event containing the time-series C_{Pb} and C_{Bi} is sampled from R, where each event in R has equal probability of being sampled. C_{Pb} and C_{Bi} are each time series lists where each element represents the surface activity concentration at a given time after the beginning of the rain event. Each rain event has a duration matching that recorded for in the measured data, t_{rain}.

Another value is then sampled according to $U(0.0, 1.0)$ and compared against $P_{\mathrm{r,start}}$, which determines if the run starts *during* a rain event. If the run starts *during* a rain event, the run will

start at $t_{\text{rain,start}}$ seconds into the rain event, which is sampled according to $t_{\text{rain,start}} \sim U(0.0, t_{\text{rain}})$. If the run does not start during a rain event, the amount of time elapsed before the rain event begins is sampled according to $U(0, t_r)$. At each time step in the run during which it is actively raining, the values of C_{Pb} and C_{Bi} are linearly interpolated in time and multiplied by the unscaled count rate tallies to yield two additional lists, S_{P} and S_{B}.

Another environmental contributor to the gamma-ray background is the cosmogenic component. As described in Section III.B, this component was not simulated in a way that requires it to be scaled. As such, the spectral detector response for the cosmic components, S_{COS}, is directly loaded in from the simulations database for the street's geometry.

Just like for the other NORM components, the rain and cosmic spectral components are stored separately, with V now containing S_{K}, S_{U}, S_{T}, S_{C}, S_{P}, S_{B}, and S_{COS}.

III.D.5. Nuisance/Threat Sources

The nuisance/threat sources were modeled at an activity of 1 Bq for the point and medical sources and at a variety of masses for the nuclear material sources, as given in Table IV. When included in the dataset, the unscaled pulse height spectrum mesh tally for a particular source is given as $\hat{S}_{\text{src}}$. $\hat{S}_{\text{src}}$ can then be multiplied by a linear scaling factor f to yield the scaled pulse height spectra, S_{src}. The scaling factor is generally calculated based on the desired peak signal-to-noise ratio (SNR) SNR_{peak} for a particular source. SNR_{peak} is defined in Equation 1 where S is the unscaled voxelized count rate from the source and B is the sum of the voxelized count rate from all background components.

$$\text{SNR}_{\text{peak}} = \max\left(\frac{fS}{\sqrt{fS+B}}\right) \tag{1}$$

A given run can contain more than one source. For each source in a given run, the scaled voxelized series of time-normalized spectra for the source, $S_{\text{src}} = f\hat{S}_{\text{src}}$, is appended to V.

III.D.6. Spectrum Processing

At this step in the process, V now contains the voxelized pulse height spectra for the NORM background components, the environmental background components, and the nuisance/threat source(s). Gaussian energy broadening (GEB) is then applied to the pulse height spectra by

using the method employed by GADRAS [29]. The functional form describing energy broadening is a piecewise function, defined as:

$$\text{FWHM} = 6.61 \cdot P_7 \left(\frac{E}{661} \right)^{P_8} \tag{2}$$

where parameters (P_6, P_7, P_8) = (0.0, 7.5, 0.7). After GEB, the pulse height spectra for each voxel are scaled by voxel duration, which is provided by the kinematics simulation described in Section III.D.2. The spectrum is then Poisson sampled to generate statistical noise that a real detector would experience while taking data in each voxel. This is done for each isotope separately so that photons of individual source and background components can be tracked.

III.D.7. Listmode Sampling

The final dataset is recorded in listmode.Listmode is used to allow for algorithm developers to use any integration time they see fit, and significantly reduces data size relative to spectra with small integration times, which are the alternative if relevant temporal information is not to be lost. For each voxel, each energy bin in each GEB spectrum component (nuisance/threat sources, environmental background components, NORM background components) is scaled by time within the voxel, Poisson-sampled to determine the number of events within that discrete energy and time interval and then each event is sampled uniformly over that energy range and time to create discrete photon arrival times and energies. The list-mode data are then sorted in time and the difference in time between events is calculated. The global timestamp is not used in the final dataset, rather the time difference, to reduce the overall size of the data by storing the interval in milliseconds as a 16-bit integer. The time difference between the current event and the last, the photon energy, and an identifier for origin of that photon are all saved.

III.E. Dataset Design

The overarching objective in creating the RADAI dataset was to supply algorithm researchers with data that are realistic enough to exercise relevant data analyses methods within the urban-search context. Ideally, the data would be calibrated in difficulty so that performance gains from new algorithms are both measurable and meaningful. To that end we adopted an iterative, data-driven design loop where a large space of background datasets were generated with parameterized

background strengths and initially selected background parameters were used to created preliminary training and testing datasets were generated through the stochastic pipeline described in Section III.D, then a suite of reference algorithms was deployed on these runs to estimate likely algorithm performance on source detection and identification tasks for the datasets. Finally, a subset of generated datasets was selected and source activity difficulty levels were determined through a formal *Design of Experiments* until those baselines spanned the “Goldilocks" regime where easy, moderate, and difficult encounters (as determined by baseline algorithm performance) were all represented. This section covers this process in more detail.

III.E.1. Baseline Algorithm Performance

Multiple algorithms - Nonnegative matrix factorization (NMF) [30], nuisance-rejecting spectral comparison ratios for anomaly detection (NSCRAD) [31], multiplexed censored energy window (mCEW) [32], and gross counts k-sigma [33] - were used to assess the suitability of the dataset, as their performance on spectroscopic anomaly detection and source identification tasks is well understood. The k-sigma algorithm analyzes gross counts to detect anomalies. Therefore its performance on this dataset is expected to be relatively poor because the background gross count rate can vary significantly. NSCRAD and mCEW are spectroscopic region-of-interest algorithms which perform well on sources with well-defined spectral features that stand out from the background. By contrast, their performance on NORM and other sources whose spectra are broad is expected to be worse because the regions of interest are wide and ill-defined. An example of some of the windows found by the mCEW algorithm is shown in Figure 16.

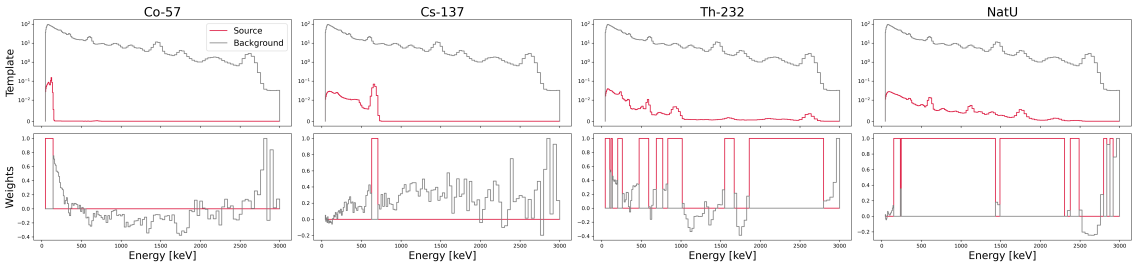
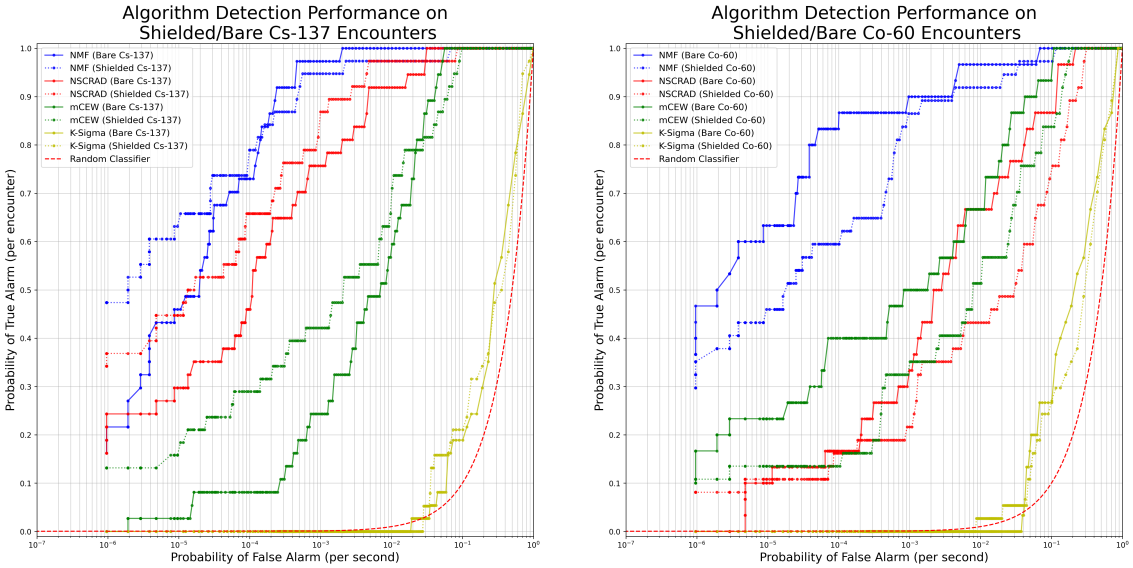
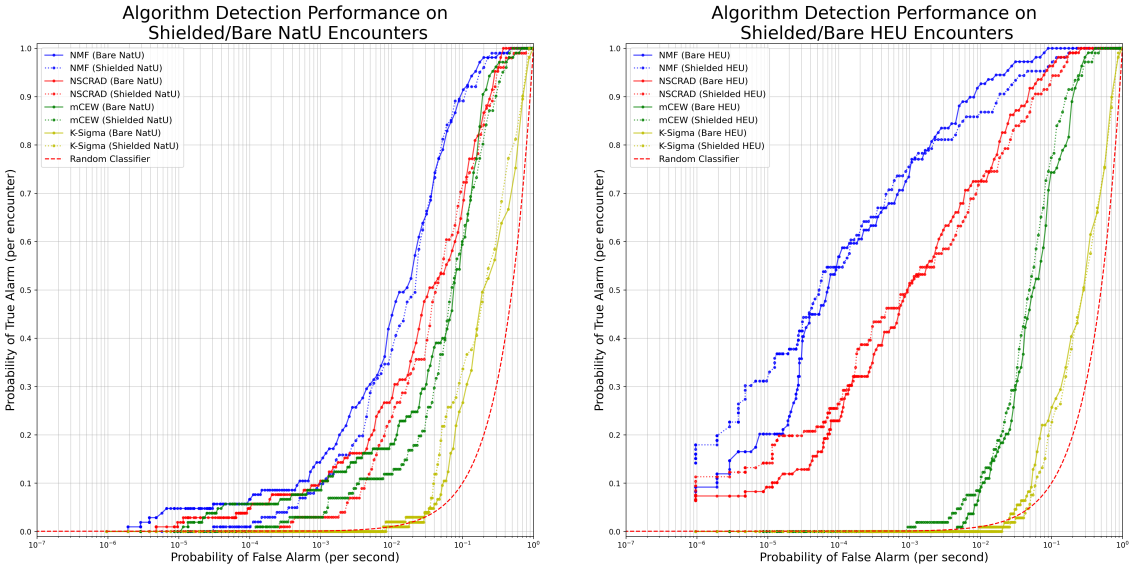


Fig. 16. Examples of windows — determined by assigning weights to regions of the spectrum that are associated with sources (red) or background (black) — found by mCEW. The windows for ^{137}Cs and ^{57}Co are narrow and well-defined whereas those found for ^{232}Th and natural Uranium span nearly the entire range of energies.



(a) ^{137}Cs (b) ^{60}Co



(c) NatU (d) HEU

Fig. 17. ROC curves for NMF, NSCRAD, mCEW, and K-Sigma on ^{137}Cs, ^{60}Co, NatU, and HEU.

NMF is a dimensionality reduction technique which has been shown to be a useful framework for modeling background gamma radiation. Here, the NMF-based approach to spectroscopic anomaly detection and source identification laid out in Ref. [34] was implemented. This method is expected to outperform the other previously described algorithms because it is performing full-spectrum analysis and leveraging all of the available spectroscopic information rather than looking

only at windows of interest or gross counts. Similarly to NSCRAD and mCEW, though, NMF generally performs well on detection tasks where the spectral features of the source are distinct from the background. Detection performance suffers when attempting to detect NORM sources or other isotopes whose spectral features are similar to background.

Each algorithm was trained and source templates were generated using the perfect-knowledge spectroscopic training data. Each algorithm was then tested on all 300 runs in the testing dataset. The receiver operating characteristic (ROC) curves shown in Figure 17 show the detection performance of each algorithm on individual sources. Most sources in the dataset have at least two different shielding configurations, namely "bare" and "shielded". The results are broken out by shielding configuration to demonstrate the difference in algorithm performance in each case.

After training each algorithm on the training dataset and evaluating its performance on the testing dataset, SNR ranges of the individual source encounters were adjusted to promote a probability of detection of 50% at a FAR of 1 per 8 hours. SNR was adjusted based on shielding configuration to ensure shielded source encounters were not too difficult. The results in Figure 17 bear this out, as the difference in detection performance between shielded and bare source encounters for a single algorithm is generally not drastic. The performance of each baseline algorithm tested on the dataset reflects the relative performance expectations set forth above (in particular, NMF $>$ NSCRAD $>$ mCEW $>$ K-Sigma). Additionally, the gradual increase in probability of detection as a function of false-alarm rate for each algorithm indicates that the overall difficulty of the dataset is reasonable, comprising easy, medium, and hard detection tasks.

III.E.2. Design of Experiments

Strategies from the field of Design of Experiments were leveraged to ensure that the dataset would be adequately robust to serve as an effective standard for developing, testing, and comparing detection algorithms. To benefit radiation detection algorithm development, the dataset must include: (1) complex backgrounds and source types, to represent the wide variety of urban environments that could be encountered; (2) sufficient source encounters to estimate the performance of an algorithm under development; and (3) varied difficulty, to distinguish between the performance of different algorithms. Methodologies from the field of Design of Experiments, including the use of space-filling designs and nonparametric strategic subdata selection, were applied to meet these

objectives[3, 35, 36, 37, 38, 39].

The street geometry, rain and cosmics, and vehicle speeds combine to define the radiological backgrounds. The dataset design objective was twofold: (1) identify locations with sufficiently varied background characteristics to host a source; and (2) assign source types and strengths at those locations. Source type encompassed shielding options, as defined in Table IV. Source strength was defined by ranges of $\mathrm{SNR}_{\mathrm{peak}}$ defined in Equation 1. Probabilistic sampling, as detailed in previous sections, was used to generate a large collection of runs that spanned the desired variety of backgrounds. Statistical analyses were conducted to ensure desired variety and coverage of street geometries, rain and cosmics, and vehicle speed.

A successful detection algorithm must be able to separate the background signal from the source under a wide variety of possible background-source combinations. Thus, a key component of the dataset design involved developing a strategy to ensure a diversity of background-source combinations, while limiting the rate of source encounters to retain realism, and simultaneously avoid the need for excess simulation. The field of Design of Experiments provides methods for strategically selecting input combinations that span a defined space of interest, while adhering to a fixed budget. Space-filling designs are well-suited to the dataset design problem because they seek to spread design points out nearly evenly or uniformly throughout the space of interest [40]. A wide variety of space-filling designs exist in the Design of Experiments literature [41, 35, 36], including Non-Uniform Space-Filling Designs [42], which were successfully used for dataset design in the previous urban search data competition [1]. However, the size and complexity of the current dataset required a different approach.

In the current application, the background space of interest was defined by the large collection of probabilistically generated runs. This posed two design challenges. First, the region of interest was complicated and large, in terms of both number of observations and dimension. To address this issue, the background inputs were reparametrized as spectral count rate variation and gross count rate variation, for the purpose of dimension reduction. These two new variables served as proxies for the background variability created by changes in street geometry, rain and cosmics, and vehicle speed. The problem was further simplified to focus solely on variability at potential source encounters. Thus, summary metrics defined by maximum spectral count rate variation and maximum gross count rate variation over the 6-second interval around the time of a potential

source encounter were used. With this parametrization, the problem size was reduced to six factors of interest for source placement decision-making: source type, source strength, block type, run, and the background summary metrics maximum spectral count rate variation and maximum gross count rate variation.

The second challenge posed was that the background space of interest generated via the probabilistic sampling was fixed once the sampling step had been completed. This differs from the standard Design of Experiments framework, wherein the experimenter is free to select the desired settings of all inputs of interest. In the dataset generation scenario, only source type and source strength could be selected freely; source location, which includes block type, run, maximum spectral count rate variation, and maximum gross count rate variation, were required to be selected strategically from the large collection of existing options generated via the synthetic data generation step (Sec. III.D).

To address this issue, strategic subdata selection methodology was employed. Strategic subdata selection methods, which constitute a growing area of research within the field of Design of Experiments, approach the data reduction problem from a Design of Experiments perspective. The goal of data reduction is to select s observations from a dataset consisting of n observations, where $s \ll$n. Strategic subdata selection methods require the subdata to be judiciously selected from the full data in such a way as to meet specified objectives based on principles from the field of Design of Experiments. For instance, the Information-Based Optimal Subdata Selection method proposes a subdata selection strategy to achieve desired theoretical conditions of the subdata under the D-optimality criterion, a criterion commonly used for design generation in Design of Experiments [38]. Similarly, the Optimal Design Based subdata selection scheme provides a framework that adapts the D-, A-, or I-optimality criteria from the Design of Experiments literature to the subdata selection problem [39]. In Design of Experiments-based approaches to subdata selection, the subdata is treated as an s-point design, and the full data is taken to be the design space of interest [38].

Nonparametric strategic subdata selection methods have been proposed that take inspiration from space-filling design methodology. One of these approaches, called the nearest-neighbor space-filling design method [37], works by first generating a space-filling design with s design points that spans the space of the full data. Then the nearest neighbor to each design point is identified from

the full data, using the Euclidean distance measure. These nearest neighbors are selected to make up the subdata. This approach was leveraged for the RADAI dataset generation problem. Testing and training sets were generated separately for the RADAI dataset. In both cases, s was taken to be the desired number of source encounters and the collection of probabilistically generated backgrounds was treated as the full data.

The s-point space-filling design was constructed across the six-dimensional space that included the four background inputs as well as source type and source strength. The probabilistic sampling resulted in a nonrectangular design space. To accommodate the irregular region, a Fast Flexible Space-Filling Design was used in the space-filling design step. In addition to accommodating linear constraints in the design space, these designs allow for categorical or discrete numeric inputs. This was an important advantage because source type needed to be treated as categorical inputs and we elected to treat source strength categorically as well.

Source encounters with very high SNR are expected to be correctly identified by algorithms with high probability, while those with very low SNR are likely to be correctly identified with low probability, across a large collection of potential source identification algorithms. Thus, these cases do not aid in distinguishing performance between different algorithms under development and are, therefore, less desirable than moderately difficult cases with midrange SNR. For that reason, it was not desired to obtain a source-placement design with SNR approximately uniformly distributed; rather, midrange values should appear more often. Further, the distribution of SNR should differ between the dataset used for training and the dataset used for testing: the testing set should include more challenging cases (with lower SNRs). This was accomplished by manipulating the Fast Flexible Space-Filling design methodology to enable non-uniformity in the source strength input by artificially adding levels to the categorical input and then later collapsing those levels. For instance, if SNR consisted of two categorical levels: Hard (low SNR values) and Easy (high SNR values) and it was desired to obtain a design with approximately twice as many Easy encounters as Hard encounters, the designer could simply artificially expand the "Easy" category into two separate categories labeled "Easy.1" and "Easy.2." With that, the Fast Flexible Space-filling design would treat SNR as a three-level categorical factor, with levels "Easy.1," "Easy.2," and "Hard," resulting in a design in which each occurs in approximately one-third of the cases. Once the design has been generated, the designer can collapse all the "Easy.1" and "Easy.2" cases into a

single “Easy” category, thus obtaining a design with approximately twice as many Easy cases as Hard cases. While the source-placement design generation problem was more complicated than this example, this general approach was taken to obtain the desired (non-uniform) distributions of SNR in both the training and the testing sets.

Several other considerations were taken in the Fast Flexible Space-Filling design generation step to further distinguish the training set from the testing set. The reasons for these differences were twofold: (1) to guard against overfitting in the training stage and (2) to ensure that the testing set could be used to effectively assess performance of the algorithms. To meet these goals, some purposeful gaps were included in the training set, relative to the testing set, to allow the dataset user to assess how well their algorithm interpolates. Additionally, the scope of the testing set was designed to surpass that of the training set in some instances, to allow the user to assess how well their algorithm extrapolates. Gaps were left in the training set to test interpolation by excluding certain blocks from the training set. The excluded blocks shared characteristics with included blocks to ensure a test of interpolation rather than extrapolation. To test extrapolation, certain challenging source strength settings (low SNRs) were excluded entirely from the training set.

Once the Fast Flexible Space-Filling design was generated, the nearest neighbor to each design point was identified from the probabilistically sampled background data, with respect to the four background inputs. These instances were included in the final data sets. For each instance, the two inputs that could be set freely, source type and source strength, were assigned according to the results of the Fast Flexible Space-Filling design, without modification.

Thus, using methodologies from the field of Design of Experiments, including the use of space-filling designs and nonparametric strategic subdata selection, the training and testing datasets were judiciously designed and generated to span a large and varied space of backgrounds and source types, providing a robust representation of the wide array of urban environments of interest, while offering diverse levels of difficulty to adequately assess algorithm performance and distinguish between different algorithms under development.

IV. RELEASED DATASETS AND CODE

Using the methods described herein, we have generated and released three datasets: **Training**, **Developer**, and **Testing**. These datasets can be accessed by registering an account through the Berkeley Data Cloud (BDC) hosted at this link (https://bdc.lbl.gov/), which is operated by LBNL. Table IX outlines the values of all relevant hyperparameters used to generate each dataset.

- The **Training Dataset** contains full simulation-level provenance information for each detected gamma-ray. It is intended to support algorithm development, supervised learning, and data augmentation.
- The **Testing Dataset** is structurally identical to the training dataset but does not include any ground truth information. It is intended for final algorithm evaluation. Performance scoring against the testing dataset is available through the RADAI Online Scoring Portal, which can also be accessed through BDC. Example submissions and an official submission template are provided through BDC to support the community in preparing for online evaluation.
- The **Developer Dataset** comprises a smaller number of shorter-duration runs with stronger anomaly events. Each source is simulated at a high SNR of 60, so that users can augment the source activity as desired for their experiments. The RADAI software toolkit repository provides scripts to aid users in data augmentation using the developer dataset.

Additionally, we have open-sourced the RADAI software toolkit, which provides users with Python tools for algorithm development and evaluation with the RADAI datasets. This code can be accessed here.

V. CONCLUSION

This work presents the RADAI dataset, a large-scale, realistic, and publicly available synthetic dataset designed to advance the development and evaluation of radiation detection algorithms, including those based on AI/ML. Building upon previous efforts such as *Chameleon Street*, RADAI introduces substantial improvements in physical realism, environmental complexity, and

TABLE IX
Dataset-generation hyper-parameters for the training, testing, and developer sets.

Parameter	**Training**	**Testing**	**Developer**
Number of runs	300	300	10
Target run duration t_r (s)	3600	3600	3600
Min. detector velocity $v_{\min}$ ($\mathrm{m\,s^{-1}}$)	2.2	2.2	2.2
Max. detector velocity $v_{\max}$ ($\mathrm{m\,s^{-1}}$)	8.0	13.4	8.0
Min. speed-zone length $l_{\mathrm{zone,min}}$ (blocks)	2	3	2
Max. speed-zone length $l_{\mathrm{zone,max}}$ (blocks)	6	10	6
Stop-light probability p_{stop}	0.1	0.1	0.1
Min. stop duration $t_{\mathrm{stop,min}}$ (s)	1.0	1.0	1.0
Max. stop duration $t_{\mathrm{stop,max}}$ (s)	10.0	10.0	10.0
Max. stop distance $d_{\mathrm{stop,max}}$ (m)	30	30	30
Min. NORM-zone length $l_{\mathrm{NORM,min}}$ (blocks)	4	4	8
Max. NORM-zone length $l_{\mathrm{NORM,max}}$ (blocks)	12	12	16
Rain probability P_r	0.1	0.1	0.1
Rain-start-during-run probability $P_{r,\mathrm{start}}$	0.5	0.5	0.5
Excluded sources	^{60}Co‡, ^{133}Ba‡, ^{67}Cu†,‡, ^{90}Sr§, ^{192}Ir$^{\ddagger,\parallel}$, ^{177}Lu*	None	^{133}Ba‡, ^{67}Cu†,‡, ^{90}Sr§, ^{192}Ir$^{\ddagger,\parallel}$, ^{177}Lu*

Symbol legend: † bare, ‡ 1 cm steel, § 2 cm Al, $^{\parallel}$ 5 cm steel, * 8 cm PMMA

operational variability, including dynamic kinematics, diverse urban topographies, cosmic and rain-induced background contributions, and a broader range of threat and nuisance sources.

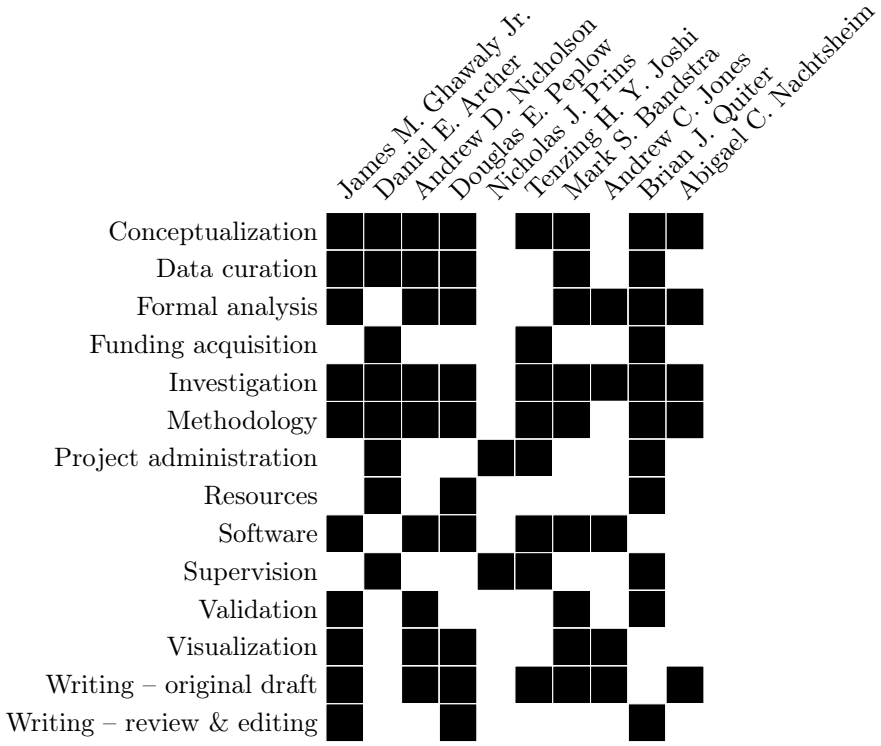
The dataset is generated through a modular Monte Carlo simulation framework that enables arbitrarily long, unique detector traversals through urban environments while maintaining accurate physics-based signal generation. Extensive validation and testing of simulation components, including comparisons against measured spectra, ensure that the synthetic data remain representative of real-world observations.

By releasing training, developer, and testing datasets with varying levels of difficulty and ground truth access, RADAI supports the full spectrum of algorithm research, from exploratory development to rigorous benchmarking. In doing so, it aims to foster reproducible comparisons, accelerate algorithmic innovation, and ultimately contribute to more robust and reliable systems for radiological search in complex environments.

ACKNOWLEDGEMENTS

This research was supported by the U.S. National Nuclear Security Administration (NNSA) Office of Defense Nuclear Nonproliferation Research and Development within the U.S. Department of Energy at ORNL under Contract AC05-00OR2272, at LBNL under Contract DE-AC02-05CH11231 and at LANL under Contract 89233218CNA000001.

VI. AUTHOR CONTRIBUTIONS STATEMENT



VII. DISCLOSURE STATEMENT

The authors report there are no competing interests to declare.

REFERENCES

[1] J. M. Ghawaly Jr., A. D. Nicholson, D. E. Peplow, C. M. Anderson-Cook, K. L. Myers, D. E. Archer, M. J. Willis, and B. J. Quiter, "Data for Training and Testing Radiation Detection Algorithms in an Urban Environment," *Scientific Data*, **7** (2020); 10.1038/s41597-020-00672-2.

[2] A. D. Nicholson, D. E. Peplow, J. M. Ghawaly Jr., M. J. Willis, and D. E. Archer, "Generation of Synthetic Data for a Radiation Detection Algorithm Competition," *IEEE Transactions on Nuclear Science*, **67**, *8*, 1968 (2020); 10.1109/TNS.2020.3001754.

[3] C. M. Anderson-Cook, D. Archer, M. S. Bandstra, J. C. Curtis, J. M. Ghawaly, T. H. Joshi, K. L. Myers, A. D. Nicholson, and B. J. Quiter, "Radiation Detection Data Competition Report," LA-UR-20-27292, Los Alamos National Laboratory, Los Alamos, NM (2021); 10.2172/1778748.

[4] B. T. Rearden and E. Jessee, M. A., "SCALE Code System," ORNL/TM-2005/39 Version 6.2.3, Oak Ridge National Laboratory, Oak Ridge, TN (2018); 10.2172/1426571.

[5] R. J. McConn Jr., C. J. Gesh, R. T. Pagh, R. A. Rucker, and R. G. Williams III, "Compendium of Material Composition Data for Radiation Transport Modeling," PIET-43741-TM-963, PNNL-15870, Rev. 1, Pacific Northwest National Laboratory, Richland, WA (2011); 10.2172/1023125.

[6] M. Cristy and K. F. Eckerman, "Specific Absorbed Fractions of Energy at Various Ages from Internal Photon Sources 1. Methods," ORNL/TM-8381/V1, Oak Ridge National Laboratory, Oak Ridge, TN (1987); 10.2172/6233735.

[7] J. C. Curtis, R. J. Cooper, T. H. Joshi, B. Cosofret, T. Schmit, J. Wright, J. Rameau, D. Konno, D. Brown, F. Otsuka, E. Rappeport, M. Marshall, and J. Speicher, "Simulation and validation of the Mobile Urban Radiation Search (MURS) gamma-ray detector response," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **954**, 161128 (2020); 10.1016/j.nima.2018.08.087.

[8] A. Zoglauer, R. Andritschke, and F. Schopper, "MEGAlib – The Medium Energy Gamma-ray Astronomy Library," *New Astronomy Reviews*, **50**, 7, 629 (2006); 10.1016/j.newar.2006.06.049.

[9] S. Agostinelli, J. Allison, K. Amako et al., "Geant4—a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **506**, 3, 250 (2003); 10.1016/S0168-9002(03)01368-8.

[10] M. S. Bandstra, J. M. Ghawaly, D. E. Peplow, D. E. Archer, B. J. Quiter, T. H. Y. Joshi, A. D. Nicholson, M. J. Willis, I. Garishvili, A. J. Rowe, B. R. Longmire, and J. T. Natress, "Full Spectrum Modeling of *In Situ* Gamma-ray Detector Measurements with a Focus on Precipitation-Induced Transients," *JJJJJJJJJJ*, **V**, P– (2025); 10.000123/000123.

[11] ANSI, "American National Standard Performance Criteria for Handheld Instruments for the Detection and Identification of Radionuclides," ANSI N42.34-2015 (Revision of ANSI N42.34-2006), American National Standards Institute (2016); 10.1109/IEEESTD.2016.7551091.

[12] ANSI, "American National Standard for Evaluation and Performance of Radiation Detection Portal Monitors for Use in Homeland Security," ANSI N42.35-2016 (Revision of ANSI N42.35-2006), American National Standards Institute (2016); 10.1109/IEEESTD.2016.7551097.

[13] ANSI, "American National Standard for Performance Criteria for Spectroscopy-Based Portal Monitors Used in Homeland Security," ANSI N42.38-2015 (Revision of ANSI N42.38-2006), American National Standards Institute (2016); 10.1109/IEEESTD.2016.7394937.

[14] ANSI, "American National Standard Performance Criteria for Mobile and Transportable Radiation Monitors Used for Homeland Security," ANSI N42.43-2016 (Revision of ANSI N42.43-2006), American National Standards Institute (2016); 10.1109/IEEESTD.2016.7488155.

[15] ANSI, "American National Standard Performance Criteria for Backpack-Based Radiation-Detection Systems Used for Homeland Security," ANSI N42.53-2013, American National Standards Institute (2013); 10.1109/IEEESTD.2013.6587250.

[16] M. W. Enghauser, "FRMAC Gamma Spectroscopist Knowledge Guide," SAND2019-9768 R, Sandia National Laboratories, Albuquerque, New Mexico (2019); 10.2172/1763003.

[17] "Stabilization, Packaging and Storage of Plutonium-Bearing Materials," DOE-STD-3013-2018, US Department of Energy (2018)URL https://www.standards.doe.gov/standards-documents/3000/3013-astd-2018/@@images/file.

[18] G. G. Thoreson, "Detecting Radiation Algorithms Group (DRAG)," SAND2021-3122PE-694854, Albuquerque, NM (2021)URL www.osti.gov/servlets/purl/1856089.

[19] M. Bandstra, C. Britt, J. Ghawaly, T. Grimes, T. Haard, P. Heimberg, T. Joshi, H. Komkov, S. Labov, N. McFerran, T. Morrow, A. Nicholson, M. Paff, B. Quiter, M. Reed, and G. Thoreson, "Metrics and Methods for Radiation Detection Algorithm Characterization for Nuclear/Radiological Source Search," ORNL/TM-2023/2906, Oak Ridge National Laboratory, Oak Ridge, Tennessee (2024); 10.2172/2439899.

[20] D. E. Peplow, D. E. Archer, J. M. Ghawaly, Jr., T. H. Y. Joshi, M. S. Bandstra, and B. J. Quiter, "Threat Sources for Creating Synthetic Urban Search Data," ORNL/TM-2021/2076, Oak Ridge National Laboratory, Oak Ridge, Tennessee (2021).

[21] R. Akcelik and D. C. Biggs, "Acceleration Profile Models for Vehicles in Road Traffic," *Transportation Science*, **21**, *1*, 36 (1987); 10.1287/trsc.21.1.36.

[22] M. W. Swinney, D. E. Peplow, B. W. Patton, A. D. Nicholson, D. E. Acher, and M. J. Willis, "Development of the NaI Detector Pulse-height Tally Spectra on Five-degree Intervals from the Cosmic Photon Background," *Nuclear Technology*, **203**, *3*, 325 (2018); 10.1080/00295450.2018.1458558.

[23] R. Trevisi, S. Risica, M. D'alessandro, D. Paradiso, and C. Nuccetelli, "Natural radioactivity in building materials in the European Union: a database and an estimate of radiological significance," *Journal of Environmental Radioactivity*, **105**, 11 (2012); 10.1016/j.jenvrad.2011.10.001.

[24] A. Baba, A. Bassari, F. Erees, and S. Cam, "Natural radioactivity and metal concentrations in soil samples taken along the Izmir-Ankara E-023 highway, Turkey," *Proceedings of the INSINUME Workshop. In situ nuclear metrology as a tool of radioecology* (2004)URL inis.iaea.org/records/3ttxv-ma432.

[25] A. Y. Ahmad, M. A. Al-Ghouti, I. AlSadig, and M. Abu-Dieyeh, "Vertical distribution and radiological risk assessment of ^{137}Cs and natural radionuclides in soil samples," *Scientific Reports*, **9**, *1*, 1 (2019); 10.1038/s41598-019-48500-x.

[26] M. A. Hannan, K. A. Wahid, and N. Nguyen, "Assessment of natural and artificial radionuclides in Mission (Texas) surface soils," *Journal of Radioanalytical and Nuclear Chemistry*, **305**, 573 (2015); 10.1007/s10967-015-4018-4.

[27] M. O. Isinkaye and Y. Ajiboye, "Natural radioactivity in surface soil of urban settlements in Ekiti State, Nigeria: baseline mapping and the estimation of radiological risks," *Arabian Journal of Geosciences*, **15**, *6*, 1 (2022); 10.1007/s12517-022-09835-4.

[28] M. I. Walker and K. S. B. Rose, "The radioactivity of the sea," *Nuclear Energy*, **29**, *4*, 267 (1990)URL inis.iaea.org/records/72hh6-cc938.

[29] D. J. Mitchell, L. T. Harding, G. G. Thoreson, and S. M. Horne, "GADRAS Detector Response Function," SAND2014-19465, Sandia National Laboratory (2014); 10.2172/1163837.

[30] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, **401**, 788– (1999); 10.1038/44565.

[31] K. K. Anderson, K. D. Jarman, M. L. Mann, D. M. Pfund, and R. C. Runkle, "Discriminating nuclear threats from benign sources in gamma-ray spectra using a spectral comparison ratio method," *Journal of Radioanalytical and Nuclear Chemistry*, **276**, *3*, 713 (2008); 10.1007/s10967-008-0622-x.

[32] E. Lei, "Robust Detection of Radiation Threat," (2017); SEE NOTES IN reference.bib FILE.

[33] D. E. Archer, D. E. Hornback, J. O. Johnson, T. M. Miller, A. D. Nicholson, B. W. Patton, D. E. Peplow, and B. Ayaz-Maierhaferm, "Systematic Assessment of Neutron and Gamma Backgrounds Relevant to Operational Modeling and Detection Technology Implementation," ORNL/TM-2014/687, Oak Ridge National Laboratory, Oak Ridge, Tennessee (2015); 10.2172/1185844.

[34] K. J. Bilton, T. H. Joshi, M. S. Bandstra, J. C. Curtis, B. J. Quiter, R. J. Cooper, and K. Vetter, "Non-negative Matrix Factorization of Gamma-Ray Spectra for Background

Modeling, Detection, and Source Identification," *IEEE Transactions on Nuclear Science*, **66**, *5*, 827 (2019); 10.1109/TNS.2019.2907267.

[35] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of Statistical Planning and Inference*, **26**, *2*, 131 (1990); 10.1016/0378-3758(90)90122-B.

[36] V. R. Joseph, E. Gul, and S. Ba, "Maximum projection designs for computer experiments," *Biometrika*, **102**, *2*, 371 (2015); 10.1093/biomet/asv002.

[37] A. Nachtsheim, "Contribution to optimal experimental design and strategic subdata selection for big data," PhD Thesis, Arizona State University (2020).

[38] H. Wang, M. Yang, and J. Stufken, "Information-based optimal subdata selection for big data linear regression," *Journal of the American Statistical Association*, **114**, *525*, 393 (2019).

[39] L. Deldossi and C. Tommasi, "Optimal design subsampling from big datasets," *Journal of Quality Technology*, **54**, *1*, 93 (2021).

[40] D. Montgomery, *Design and analysis of experiments*, Hoboken, NJ: John Wiley and Sons, Inc. (2013).

[41] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, **21**, *2*, 239 (1979); 10.2307/1268522.

[42] L. Lu, C. M. Anderson-Cook, and T. Ahmed, "Non-uniform space filling (NUSF) designs," *Journal of Quality Technology*, **53**, *3*, 309 (2021); 10.1080/00224065.2020.1727801.