



OPEN

DATA DESCRIPTOR

# Multilayer networks describing urban interactions for building the digital twins of five cities in Spain

Jorge P. Rodríguez<sup>1,2</sup> , Alberto Aleta<sup>3,4</sup> & Yamir Moreno<sup>3,4</sup>

Networks specifying who interacts with whom are crucial for mathematical models of epidemic spreading. In the context of emerging diseases, these networks have the potential to encode multiple interaction contexts where non-pharmaceutical interventions can be introduced, allowing for proper comparisons among different intervention strategies in a plethora of contexts. Consequently, a multilayer network describing interactions in a population and detailing their contexts in different layers constitutes an appropriate tool for such descriptions. These approaches however become challenging in large-scale systems such as cities, particularly in a framework where data protection policies are enhanced. In this work, we present a methodology to build such multilayer networks and make those corresponding to five Spanish cities available. Our work uses approaches informed by multiple available datasets to create realistic digital twins of the citizens and their interactions and provides a playground to explore different pandemic scenario in realistic settings for better preparedness.

## Background & Summary

Understanding contact patterns is crucial for modeling the spread of infectious diseases within a population. Over the past decade, mathematical and computational models have advanced from the traditional homogeneous approximation, incorporating complexities that extend beyond mere knowledge of who interacts with whom. These models now aim to account for critical details, including when, where, under what circumstances, and the duration of these interactions<sup>1</sup>. Such nuances can profoundly influence the outcomes of epidemiological models, yet the accessibility to this information often remains limited.

One common method for representing interactions is through the use of contact matrices, which contain interaction rates among different cohorts, typically categorized by criteria such as age<sup>2</sup>. These matrices can be constructed using data from surveys<sup>3</sup>, census data<sup>4</sup>, or a combination of both<sup>5,6</sup>. However, this approach treats all individuals within a given group as equivalent. To capture individual heterogeneities, researchers resort to data-driven approaches<sup>7–9</sup> or to create contact networks in which each node represents an agent, and the links encode their interactions<sup>10</sup>.

There are several methodologies for collecting data and building interaction networks. The simplest approach is to carry out surveys, a technique that is adequate for relatively small populations and situations where contacts are easily identifiable, such as sexual interaction networks<sup>11–13</sup>. However, there are some diseases, like those that are airborne, for which the potential number of contacts is much larger and harder to identify.

For these scenarios, high-resolution data collection can be achieved using wearable sensors<sup>14</sup>, but logistical constraints limit their applicability to specific settings<sup>15</sup> or small areas<sup>16</sup>. During the recent COVID-19 pandemic, digital contact-tracing apps were rapidly deployed, offering the potential to build large-scale networks. However, low participation rates, technical issues, and privacy concerns hindered the success of these apps<sup>17–20</sup>. An alternative approach involves inferring synthetic contact networks from survey data used to create contact matrices, but since most surveys only stratify individuals by age, many socio-economic characteristics of the population remain unaccounted for<sup>21</sup>.

<sup>1</sup>Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC), CSIC-UIB, Palma de Mallorca, Spain. <sup>2</sup>CA UNED Illes Balears, Palma, Spain. <sup>3</sup>Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain. <sup>4</sup>Department of Theoretical Physics, University of Zaragoza, Zaragoza, Spain. ✉e-mail: [jorgeprodriguez@gmail.com](mailto:jorgeprodriguez@gmail.com)

City	$D$	$N$	$N_{NH}$	$N_{EU}$	$N_{TOT}$
Barcelona	10	1,666,530	16,456	129,644	1,812,630
Valencia	19	801,545	75,708	2,193	879,446
Sevilla	11	701,455	2,655	64,371	768,491
Zaragoza	12	681,877	6,905	17,396	706,178
Murcia	8	459,403	1,335	31,800	492,538

**Table 1.** Sizes of the studied cities.  $D$  is the number of districts considered by the census,  $N$  is the population according to the available data,  $N_{NH}$  is the population living in nursing homes,  $N_{EU}$  is the number of external university students and  $N_{TOT} = N + N_{NH} + N_{EU}$ .

In response to these challenges, we use of publicly available, high-resolution data to construct synthetic multilayer networks that faithfully replicate key features of urban areas<sup>22–24</sup>. In multilayer networks, the system is divided into multiple layers, each representing a specific type of interaction<sup>25,26</sup>. This framework allows us to naturally account for different settings in which human interactions occur and also their distinct impacts on pathogen transmission. In this work, we present the multilayer networks that describe individual-level interactions in five Spanish cities, identifying six distinct interaction contexts or layers: households, schools, universities, nursing homes, workplaces, and the broader community. In addition to their utility in studying disease spreading across diverse urban settings and measuring the distinct impacts of characteristic features unique to each city, they are also a versatile platform that may be used to investigate other spreading processes extending beyond epidemics.

## Methods

**Geographical extents.** To limit the geographical scope of our study, we employed the administrative boundaries of each municipality within the respective cities under consideration. Within these municipal limits, we utilized census districts as the primary geographical units assigned to individuals. These census districts were defined by shapefiles available in the cartography of the 2011 census provided by the Spanish “Instituto Nacional de Estadística” (INE).

**Demography.** To construct the synthetic population required for inferring multilayer contact networks, we created a population that accurately mirrored the observed distribution of citizens’ age and sex across various districts within each city (Table 1). We sourced recent demographic information from local census data, primarily obtained from the municipalities’ online open datasets as of January 1, 2020. This data was readily available for Barcelona, Valencia, Sevilla, and Murcia. For Zaragoza, we utilized continuous census data provided by the “Instituto Nacional de Estadística” (INE).

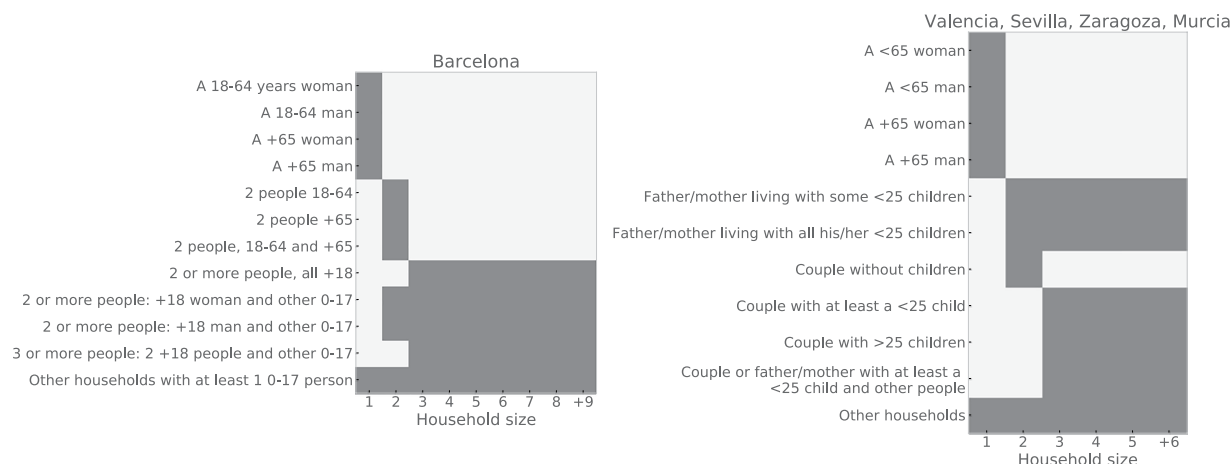
The diverse origins of these datasets, spanning local, regional, and national administrative units, resulted in varying age groupings, typically organized into 5-year cohorts, with variations in the lowest and highest age categories. Specifically, the most elderly age group was designated as +100 for Barcelona and Zaragoza, +90 for Valencia and Sevilla, and +85 for Murcia. Moreover, Murcia subdivided the ‘0–4’ age category into ‘0’ and ‘1–4’. We decided not to standardize the age-groups, aiming to preserve as much information as possible for each city. This decision also allows for assessing the significance of such detailed age information. If necessary, it can be easily aggregated to create a standardized model.

Using the census statistics, we created a synthetic population, specifying the age, sex, and residential district of each individual. For layers that required precise ages (such as university and school layers), we allocated ages with a 1-year resolution for individuals under 30 years old using an interpolation method. This method involved considering an age, say ‘A’, and then calculating the set of ages: A–2, A–1, A, A+1, A+2. We used the number of ages in this set within each binned age-cohort as a weighting factor to estimate the expected number of individuals at that specific age. For instance, for an individual aged 29, the contribution from the 25–29 age cohort would be 3/5, while the contribution from the 30–34 cohort would be 2/5.

Once this synthetic population was created, we compared the total population count within each district with the total population by age and sex in the same district. This validation process matched in Valencia, Sevilla, Zaragoza, and Murcia, and thus, these populations were retained. However, in the case of Barcelona, the total population count, when summed across ages and sexes, did not align with the reported total population in the district. We hypothesize that it was caused by the absence of information on age and sex for some individuals. To address this, we randomly sampled from the synthetic population to maintain the total number of individuals reported, resulting in the final population considered for the synthetic city. Each synthetic individual in the population is represented by a node in the multilayer network of the corresponding city.

**Households layer.** To create the households layer, we used various statistical parameters within each district, including the number of households, average household size, the distribution of household sizes, and the structure of these households. Additionally, for households including adult couples, we considered the average age difference between cohabiting partners.

Our approach began by calculating a scaling factor, representing the ratio between the expected population (calculated from the product of the average household size and the number of households) and the actual population (as detailed in the Demography section). In most cases, this factor was close to 1, needing no further adjustments.



**Fig. 1** Compatibility matrices between the structure of the household and its size. Darker entries represent the combinations that are allowed.

Household sizes were obtained as integer values for smaller sizes but transitioned to categorical sizes for larger households, defined as 9 or more people in Barcelona, and 6 or more people in Valencia, Sevilla, Zaragoza, and Murcia. For these larger households, we estimated their average size, ensuring that the district's overall average household size remained unchanged. This average value was then assigned to these larger households.

To account for the specific household structures, which encompassed broad age cohorts and gender, we associated each structure with household size. We achieved this by using available categories for each city to create a compatibility matrix (see Fig. 1). Note that by using the household structure found in the data, we naturally include inter-generational households, which to some extent capture also relationships between the extended family if they live together. However, to fully account for these interactions one would require reliable information on the structure of extended families - which we were not able to locate.

Using the available information, our household creation algorithm worked as follows. First, for each census district, the average size for unbounded households (e.g., +9 for Barcelona, +6 for Valencia) was calculated and rounded to the nearest integer. In the unlikely case of these sizes being smaller than the minimum boundary, then these households were assigned that boundary. If the average exceeded 20, households were divided into patches of 20 individuals each.

Subsequently, the algorithm generated two vectors for each household: one describing its structure and another detailing its size. Given that not all households in the district had available statistics, we generated household sizes based on the distribution of those with available information while keeping the district's population fixed (see Demography).

For household structures, we allocated individuals according to the profiles present in the data (for Barcelona, sex and age cohorts 0-17, 18-64, +65 and, for Valencia, Sevilla, Murcia, and Zaragoza, sex and age cohorts 0-24, 25-64 and +65). We started with households of size 1; for all the cities, these structures of households corresponded to 4 combinations: woman/man with ages 18-64 or +65. We generated unoccupied households and aimed to fill all of them with synthetic individuals according to their profiles. If, for a given structure, there were no remaining individuals to allocate in the household, we kept that household pending for a secondary iteration of the algorithm performed after assigning the households of size 2.

We then followed the same approach for the households of size 2, keeping the distribution of structures compatible with their sizes (Fig. 1). Additionally, we considered whether the household nuclei were composed of a couple of woman-man or not (statistics also available in the age-difference dataset). If this was the case, we also considered the distribution of age differences. After assigning the households of sizes 1 and 2 to their structures and individuals, some pending households could not be filled with the synthetic population, so they were assigned to the category 'Other' (Valencia, Sevilla, Zaragoza, and Murcia) or 'Other households with at least a 0-17 person' (Barcelona), which in the latter corresponded to households of size 1 and 2 with individuals with 0-17 age.

Lastly, we repeated the same procedure for households with sizes larger than 2. After that, if there were still some unoccupied households that could not be filled according to the distribution of the structures, they were completed randomly with the remaining individuals who did not have an assigned household.

**School layer.** The creation of the school layer requested detailed information about educational institutions, including the levels they offered and their respective districts, as well as data on student groups (such as the number of groups per level and their sizes) and the number of students per level or group. This data was sourced from regional education councils. However, the data was not consistently formatted, and variations in resolution existed among the datasets.

For this layer, we focused on four key educational programs: infant (0-5 years), primary (6-11 years), secondary (which in Spain encompasses compulsory education for ages 12-15 and baccalaureate for ages 16-17), and work training (age 16 and above). Notably, work training programs exhibited a wide range of student ages,

necessitating special consideration. In contrast, for the other educational programs, we assumed that all students fell within the expected age ranges for their respective levels.

Discrepancies emerged in the number of enrolled children and those included in our synthetic population for specific age groups. This issue did not pose a problem when the synthetic population numbers were higher, given the incorporation of recruitment rates below 1. However, in cases where the official enrollment statistics reported higher numbers, we scaled them to match the figures within the synthetic population. This divergence might arise from variations in data collection times between demographic and educational datasets, as well as the registration of children from other municipalities.

To perform this scaling, we also considered that the minimum legal working age in Spain is 16 years old. Accordingly, we extracted the employment rate for individuals aged 16–19 as of January 2019. This information helped us estimate the number of individuals within this age range who were not in the job market and, therefore, were expected to enroll in educational programs. We then compared these estimates with the numbers reported in the statistics, aiding in the alignment of data from different sources and resolving discrepancies.

Then, our goal was to create synthetic educational centers that included students organized into groups according to different educational levels, each overseen by a teacher. This general framework guided our methodology, but it was adapted and refined based on the availability, format, and resolution of data for each city. In the following, we describe the data and processing methods employed for each city.

**Barcelona.** For the case of Barcelona, the statistics were available at the census district level for the academic course 2019–2020. There were available datasets associated with the first (0–2) and second (3–5) cycles of infant education, primary education, secondary education, baccalaureate, and medium and advanced work training programs. The file reports the number of students specified, for each level (generally associated with age), and the number of registered students in the centers of each census district, split amongst different management modes (regional educational council, local authorities, or private).

Additionally, we had data on the centers and groups, which included the number of centers per district with students in each educational program and the number of groups, although not specified by level. Our objective was to create synthetic educational centers with groups assigned to specific levels. To do so, we divided the process in two steps:

1. **Group creation:** We began by looping through different districts and management modes. For each combination, we ensured that the number of groups was at least as large as the number of levels within that educational program. In that case, we created the groups and randomly assigned them a level while making sure that at least every level had a group. If the number of groups was lower than the number of levels, we assumed that groups encompassed different levels, which was more common in the first cycle of the infant program.
2. **Center and group assignment:** We then randomly assigned a center to each group, followed by the distribution of the number of students (scaled from statistics) among these groups and their respective centers.

To populate these centers, we matched individuals in our synthetic population based on their ages and districts to the students' ages and census districts. Initially, we filled the groups with individuals of the same age living in the same district as the center. Subsequently, any remaining vacant student positions were filled with individuals from neighboring districts. We employed a gravity law approach for this purpose, with the probability of transitioning between districts being proportional to the squared inverse of the great circle distance between their centroids. Note that, even though there might be some minor differences across Autonomous Regions, in Spain students are given priority if they apply to public schools or high-schools that are close to their household.

The procedure was similar across different programs, except for the medium and advanced work training programs. For those cases, we utilized an additional dataset that provided the number of individuals registered in each district, by management mode and at each level, with specific ages (measured in years). This information allowed us to create a more accurate representation of the individuals enrolled in these programs.

To complete the educational centers, we assigned a teacher to each group. We selected individuals from the synthetic population aged between 30 and 70 years old who were not already assigned to any school group (note that some individuals older than 30 may be registered in the work training programs). The teacher assignment process mirrored that of students, ensuring that teachers belonged to the same school group and educational center. We followed the same algorithm to assign the teachers to their groups in the rest of the cities. Note that in this approach students only interact with students in the same group, as well as with the teacher, while teachers may interact with any other teacher in the same center.

**Valencia.** For Valencia, our data source provided detailed information for the academic year 2019–2020. Specifically, it outlined the number of groups offered and the number of registered students at each level of the considered programs, for each specific educational center. This high-resolution data greatly facilitated the creation of educational centers and groups within our digital city.

However, since the dataset did not include information about census districts, we turned to another dataset that included geographical coordinates for each educational center. Using this information, we assigned a specific census district to each center based on its geographical location and the boundaries of the districts (as described in the “Geographical Extents” section). A similar approach was applied to Sevilla, Zaragoza, and Murcia.

In the case of work training programs, we relied on national statistics related to the ages of registered individuals for basic, intermediate, and advanced programs. This approach was also applied in Sevilla, Zaragoza, and Murcia.

In scenarios where the number of registered students in the schools of a district exceeded the number of synthetic individuals in that district, we adopted a strategy to balance these figures. Specifically, we selected registered individuals from districts with the opposite scenario, where there were more synthetic individuals than registered students. This process followed the gravity-based approach described earlier for Barcelona, ensuring a balanced representation of individuals within our synthetic population.

**Sevilla.** In the case of Sevilla, the available data had a lower level of resolution compared to other cities. Specifically, we had access to datasets detailing the number of registered students for each program, with further breakdowns by management modes (public, private, and private concerted centers). Additionally, another dataset provided information on the number of groups associated with each program. These datasets pertained to the academic year 2017–2018.

To address the lack of information at the district level, our approach involved iterating over the educational programs, taking into account the programs offered by each school. Note that for infant programs, there was no differentiation between the first (0–2 years) and second (3–5 years) cycles. Therefore, we assumed that students in the first cycle attended centers exclusively offering the infant program.

To populate the synthetic city, we assigned individuals to groups located within their respective census districts. When allocating individuals to groups, we followed the gravity-based approach for centers that included all students from the same district. This method helped ensure that individuals were assigned to groups in a manner consistent with their geographical proximity, considering both program and district-specific factors.

**Zaragoza.** For Zaragoza, our data sources provided information at the municipality level, with data corresponding to the academic year 2019–2020. Specifically, we had access to the number of students enrolled in each program, along with specifications regarding the management mode of the educational centers (public, private, and private concerted centers). Additionally, data on the number of groups were available, detailing the number of groups per age group and management mode. The information also included details on how many centers offered each program, further categorized by management mode. Within this center description, we found information regarding the types of centers that we matched with the educational programs under consideration.

Then, our approach closely mirrored that used for Seville. We iterated over educational programs, ensuring that each center offered at least one group for each level when possible. Any remaining groups were assigned randomly. Following the group assignment, we distributed registered students among these groups and matched them with individuals in our synthetic population. Lastly, we applied the gravity-based approach, as previously described, in cases where the number of registered students in a district exceeded the population of the district at a specific age.

**Murcia.** In the last city, Murcia, we encountered data of a similar nature as in Sevilla and Zaragoza, with information available at the municipality level and data corresponding to the academic year 2016–2017. The available data included details on how many students were registered for each program, categorized by two management modes (public and private). Similarly, the number of groups was also reported at this resolution. Additionally, we had access to data on the number of centers, their geographical coordinates, management mode, and type. The center-type information was particularly useful for associating the different educational programs with each center.

To create groups, we iterated over educational programs, ensuring that each center offered at least one group per age when possible. Following this group assignment, we distributed registered students among these groups and matched them with individuals in our synthetic population. As in previous cases, when the number of registered students of a specific age in a district exceeded the number of individuals of that age in our synthetic population, we applied the gravity-based method to balance these figures.

**University layer.** This layer connects students within the same academic program at both the Degree and Master's levels. We utilized data from the Ministry of Science and Innovation in Spain, which provided information on the number of registered students and the percentage of women in each program at various universities. To ensure geographical relevance, we focused on universities located in the municipality of the city or its metropolitan area. The only exception was the Universidad Politécnica de Cartagena, located in the region of Murcia but outside the capital's metropolitan area.

Next, we incorporated information on the age distribution of students. This data, also obtained from the Ministry of Science and Innovation, categorized students into age and sex groups for the academic year 2019–2020. For Degree students, the age groups were [18–21, 22–25, 26–30, +31], while for Master's students, they were [<25, 23–50, 31–40, +41]. Using this data, we constructed vectors indicating the age group of each female and male student in both Degree and Master's programs.

Not all registered students were likely represented in our synthetic population as university students often have high geographical mobility. We estimated the number of registered students present in our synthetic population using two sets of data. First, we used publicly available statistics describing the province of origin of the students registered in each university. This allowed us to estimate the fraction  $F$  of students who reside in the same province as their university's location. Then, we obtained data from a private request to the Ministry of Science and Innovation which provided the number of registered university students per municipality. We calculated the ratio  $R$  between the number of registered students from the city and the total number of students from any municipality in the province, considering 5 students for municipalities that specified '5 or less'. Consequently, the fraction of registered university students present in our synthetic population was estimated to be  $F \times R$ . The remaining students were treated as external university students who interacted exclusively within the university layer of our multilayer network.



With all this data in hand, we created synthetic university groups as fully-connected patches considering the number of registered students in each Degree or Master's program, their gender distribution, and their age distribution based on regional statistics. For programs with more than 50 registered students, we randomly divided them into multiple groups, each consisting of 50 students, plus any remainder. Within these groups, a fraction ( $FR$ ) represented our synthetic population, and we selected individuals with matching ages and genders to populate these synthetic university groups. We also stored the features (age and gender) of the external university students and assigned them to their respective university groups.

**Nursing homes layer.** In the context of studying infectious disease dynamics, and particularly in light of the recent COVID-19 pandemic, nursing homes are recognized as high-risk locations that warrant inclusion in our synthetic cities<sup>27,28</sup>. However, determining the registration status of residents in the local census proved to be challenging. To address this uncertainty, we decided to treat nursing home residents as an external population separate from our synthetic population. These residents were considered to interact exclusively within the nursing home layer.

We gathered information about nursing homes and their capacities from the program “Envejecimiento en red” (Aging in a network) from the Spanish National Research Council. In particular, for each province, there was available a file detailing the number of nursing homes, their location, and their respective capacities. While the general dataset was updated in 2020, some nursing homes had older updates.

To build this layer, we assumed that the number of residents in each nursing home was equal to its capacity. Additionally, we introduced a caretaker for every 4 residents selected from the synthetic population among those within the legal age limits to work, and who did not have any connections in the school layer (neither as teachers nor students). As previously mentioned, residents were considered as part of the external population, so their age and sex were randomly assigned based on national statistics of residents in nursing homes, obtained from the Instituto Nacional de Estadística (INE).

**Work layer.** This layer represents interactions within the working environment. To realistically model this layer, we obtained data on both employee attributes and company sizes.

For the employee information, we considered the number of individuals affiliated with Social Security in each city  $i$  as of January 2019,  $A_i$ . It is important to note that this number includes various categories, such as unemployed individuals actively seeking employment. To refine this data, we obtained the occupancy rate  $O_i$  which represents the fraction of eligible individuals who were currently employed. Additionally, we calculated the total number  $T_i$  of individuals in each city who were eligible to work, meaning they were older than 16 years. Consequently, the number of workers is  $O_i T_i$ .

With all this information, we estimated the number of workers in each city as follows. First, we obtained the number of freelance workers  $F_i$  from the Social Security dataset. Second, we defined an activity ratio  $r_{Wi}$  as  $O_i T_i / A_i$ , representing the ratio between the expected number of workers ( $O_i T_i$ ) and the number of individuals registered in the Social Security system ( $A_i$ ). Finally, we rescaled the difference between  $A_i$  and  $F_i$  with the activity ratio to obtain the final number of active workers in companies in the city  $i$ ,  $W_i = (A_i - F_i) r_{Wi}$ .

Before proceeding to assign ages and sex to these workers, we excluded those who had already been allocated to other layers, such as teachers in the school layer and caretakers in the nursing homes layer. For the remaining workers, we assigned ages and genders using national statistics from INE for January 2019. These statistics categorized individuals into various age groups: [16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, +65] (with +65 representing ages 65-69). We randomly selected the sex first, based on the estimated distribution, and then randomly assigned an age group to each worker.

Now that we had our “virtual” workers (not yet matched with the synthetic population), we allocated them to companies. Data on company size distributions were available from INE at the province level. This data specified the number of companies within one of the following size ranges: [1-2, 3-5, 6-9, 10-19, 20-49, 50-99, 100-199, 200-499, 500-999, 1000-4999, +5000]. We represented these distributions by assigning the central value to each range, resulting in values [1.5, 4, 7.5, 14.5, 29.5, 69.5, 149.5, 349.5, 749.5, 2999.5]. Notably these distributions followed the same shape across the 50 Spanish provinces, which was described by a power-law distribution  $pdf(S) \sim S^{-2}$ . Thus, we used this information to generate the companies' sizes using the Python package *powerlaw*<sup>29</sup>. The minimum company size was set to 1, and the maximum size was determined by the number of workers who had not yet been allocated.

Finally, after generating the size of each company, we randomly assigned workers from the virtual set. For companies with more than 20 workers, employees were distributed randomly among a number of patches equal to  $P = 1 + S \pmod{20}$  patches.

**Community layer.** All individuals participate also in the community layer, reflecting all the other encounters not considered in the other layers. We further assume that they only interact with individuals located within the same census district. If one wishes to disregard this restriction, this layer could be substituted by any random network generated from the individuals located in the layer.

**Interactions.** In all the layers, as previously discussed, individuals were classified in patches, such as residents in a household, students in a certain class in school or workers in a specific company. We assume that an individual in a certain patch can interact with any other individual in the same patch. In network terminology, we establish a link between each node in a patch and every other node in the same patch. Thus, each layer (except for the community layer) is made of several disconnected patches of nodes, with each of them being a fully-connected network. Note that there cannot be a direct connection from a node located in a patch in layer A and another node located in a patch in layer B. However, since all nodes are matched to a synthetic individual, the same node can participate in two different layers, implicitly establishing connections across layers.

Due to limited data availability for the community layer, we adopted a data-driven approach based on the contact matrices reported in<sup>5</sup> for the category “Other locations” in Spain. These matrices specified contact patterns between individuals grouped into 5-year age brackets up to the age of 80. For older age groups, we extrapolated from the last available age group.

Our modeling approach assumed that community interactions occurred among individuals residing in the same census district. Then, we estimated the total number of links from a specific age group  $i$  to another age group  $j$ ,  $E_{ij}$  as

$$E_{ij} = P_i \cdot \frac{C_{ij} + C_{ji}}{2} \quad (1)$$

Where  $P_i$  represents the population in age group  $i$ , and  $C_{ij}$  is the number of contacts between individuals in age group  $i$  and those in age group  $j$ <sup>30</sup>. Then, we simply extracted  $E_{ij}$  random links from all the possible links between individuals in age groups  $i$  and  $j$  and added them to this layer.

## Data records

The generative process leading to our multilayer networks was informed by datasets which in most cases were publicly available. This section details which datasets were particularly used, together with the details of the final products, *i.e.* the multilayer networks. The final dataset is available in the official repository of the Spanish National Research Council<sup>57</sup>.

**Scientific literature.** We downloaded the contact matrices for 152 countries<sup>5,31</sup>, using the data for Spain and the contacts categorized as ‘other locations’, to inform the generation of the contacts in the community layer.

**Statistics at the national level.** We obtained from INE<sup>32</sup> the following information at the district level: population by age and sex, home sizes, and home structure.

The Ministry of Science, Innovation and Universities (Spain) provided the number of registered students per university and specific degree<sup>33</sup>. This contributed to the generation of the university layer, together with information on the students’ ages per region, obtained from the Ministry of Education, Culture and Sport<sup>34,35</sup>. We also included another dataset after a request to the Ministry of Universities, detailing the number of registered students in BsC and MsC programs per municipality in the academic course 2018–2019.

In the school layer, we considered training program students’ ages, obtained from the Ministry of Education, Culture, and Sport<sup>36</sup>.

To create the work layer, we used the number of companies per region and their number of employees<sup>37</sup>, the results of the national employment survey<sup>38</sup>, both available from INE, together with the statistics of affiliated workers to the Spanish Social Security system by municipality<sup>39</sup>.

Finally, the nursing homes layer was created using data describing the age and gender of residents in nursing homes<sup>40</sup> from INE, as well as the number of places in each of the nursing homes located at each municipality, from the ‘Envejecimiento en Red’ program<sup>41</sup>.

**Barcelona.** To create the synthetic population, we used the local census detailing how many individuals, by age and sex, were registered in each district from Barcelona City Council<sup>42</sup>.

In the school layer, we used a dataset from the Catalan Government describing the number of registered students, units and centers per district<sup>43</sup>.

**Valencia.** We obtained from Valencia City Council the population statistics at the district level, specifying ages and sexes<sup>44</sup>.

We used data on the number of students per level of each program in school, by centers<sup>45</sup>, the same information for job training programs<sup>46</sup>, and the list of scholar centers, together with their geographical coordinates<sup>47</sup>, available from Valencia Regional Government.

**Sevilla.** We considered the population by age and sex, specified at district resolution, available from Sevilla City Council<sup>48</sup>.

To create the school layer, we downloaded data describing the number of students, number of centers, and number of units, per level and municipality<sup>49</sup>, and the list of scholar centers, including their coordinates<sup>50</sup>, provided by Andalucía Regional Government.

**Zaragoza.** We considered, at the level of municipality, the number of students, number of units, and number of centers per level<sup>51</sup>, together with the list of scholar centers that included their coordinates, both available from the Aragón Regional Government.

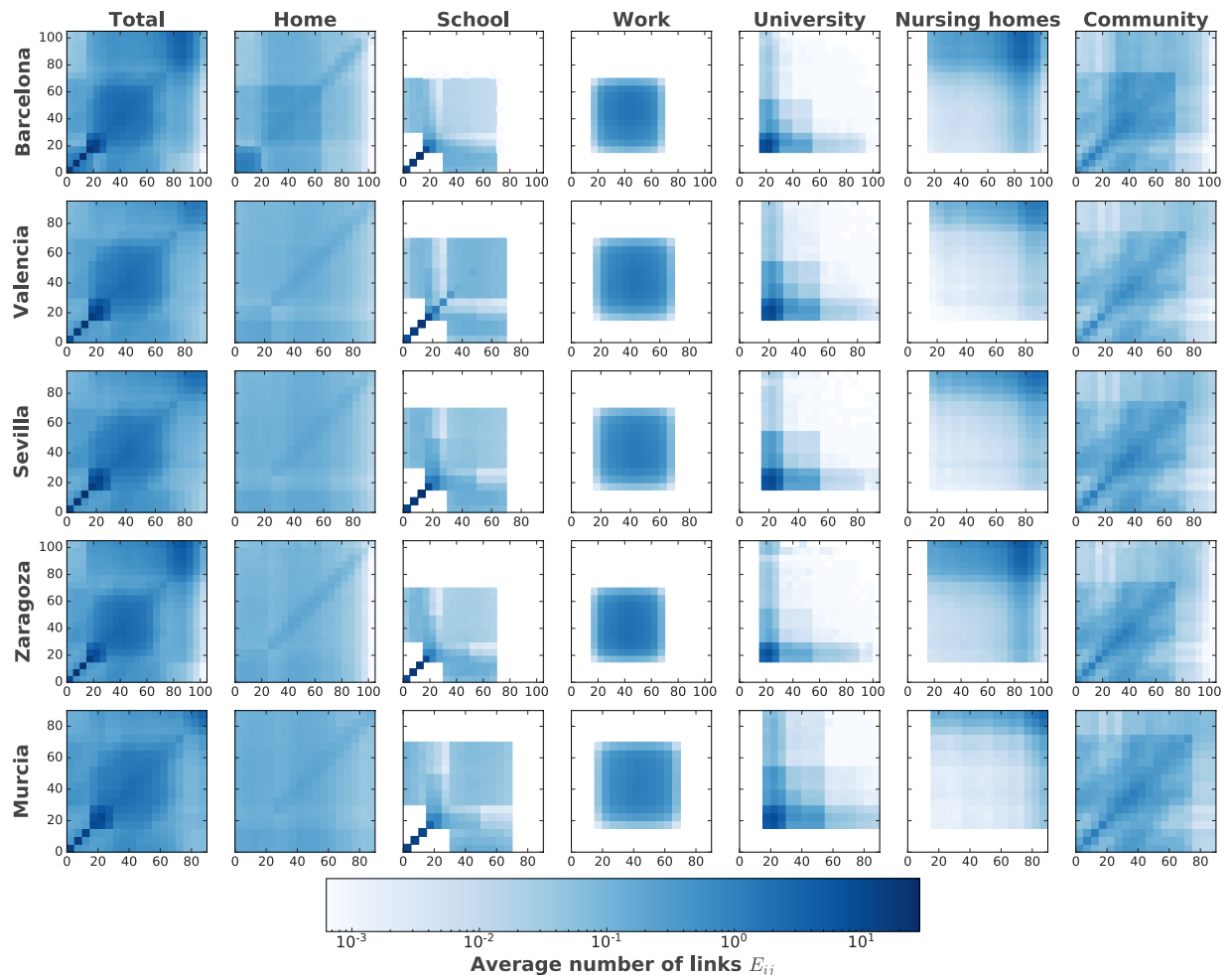
**Murcia.** Considering that districts were divided into sections, we used the population by age, at section resolution for men<sup>52</sup> and women<sup>53</sup>.

To create the school layer, we used a dataset detailing at each level and within each municipality, the number of registered students, the number of units, and the number of centers<sup>54</sup>, together with the list of centers that included their coordinates<sup>55</sup>.

**Multilayer networks and metadata.** We are releasing, together with this article, the multilayer networks of each layer in each city, as a link list in a file in CSV format, where each row represents a link and reports the two individuals that it connects.

City	$\langle k_H \rangle$	$\langle k_S \rangle$	$\langle k_U \rangle$	$\langle k_{NH} \rangle$	$\langle k_W \rangle$	$\langle k_C \rangle$
Barcelona	3.74	26.25	45.43	139.88	14.21	4.31
Valencia	2.05	24.62	45.77	143.73	15.32	3.64
Sevilla	2.26	25.87	46.36	158.32	14.33	3.72
Zaragoza	2.03	23.19	46.56	149.22	18.35	3.64
Murcia	2.48	22.61	46.29	145.19	12.40	3.87

**Table 2.** Average number of contacts in the synthetic cities per layer: households,  $\langle k_H \rangle$ ; schools,  $\langle k_S \rangle$ ; universities,  $\langle k_U \rangle$ ; nursing homes  $\langle k_{NH} \rangle$ ; work,  $\langle k_W \rangle$ ; and community,  $\langle k_C \rangle$ .



**Fig. 2** Contact matrices in each digital city aggregated (Total) and per layer. Each entry  $E_{ij}$  indicates the average number of connections that the individuals of the age represented by row  $i$  will have with the individuals in the age group of column  $j$ .

Additionally, we include three metadata files in CSV format describing the individuals' features: first for the generic population, second for the nursing home residents, and third for the external university students. Each file details the age and sex (male: 0, female: 1) of each individual, together with the census district for the generic population.

### Technical validation

We conducted an analysis of the macroscopic features of the multilayer network that describes the contact within our digital cities to validate our methodology. First, we computed the average degree for individuals who had links in each of the layers, as shown in Table 2. Note that the average number of links in each layer may vary significantly, and thus they should be properly weighted depending on the process being studied<sup>6</sup>.

Second, we analyzed the contact matrices at an aggregated level and for each layer extracted directly from the multilayer network. The results of this analysis are presented in Fig. 2 and are consistent with previous studies<sup>5</sup>.



Lastly, these multilayer networks were used to study the spreading of COVID-19 during the first and second waves in Spain. The calibrated models on these networks showed a good agreement with the data extracted from surveillance reports and offered insights on the mechanisms behind the spreading through the population<sup>56</sup>.

## Data availability

The data is available in the official repository of the Spanish National Research Council<sup>57</sup>.

## Code availability

The custom Python code and raw data processed to create the digital twins is available in the official repository of the Spanish National Research Council<sup>57</sup>.

Received: 3 November 2024; Accepted: 4 July 2025;

Published online: 15 July 2025

## References

1. Mossong, J. *et al.* Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Med.* **5**(3), e74 (2008).
2. Manna, A., Dall'Amico, L., Tizzoni, M., Karsai, M., and Perra, N. "Generalized contact matrices for epidemic modeling," *arXiv*, Jun. 2023.
3. Hoang, T. *et al.* A systematic review of social contact surveys to inform transmission models of close-contact infections. *Epidemiology (Cambridge, Mass.)* **30**(5), 723 (2019).
4. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., and Merler, S., "Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread," *PLoS Computational Biology*, no. 8, e1002673, 2012.
5. Prem, K., Cook, A. R. & Jit, M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS computational biology* **13**(9), e1005697 (2017).
6. Mistry, D. *et al.* Inferring high-resolution human mixing patterns for disease modeling. *Nature Communications* **12**(1), 323 (2021).
7. Iozzi, F. *et al.* Little Italy: An Agent-Based Approach to the Estimation of Contact Patterns- Fitting Predicted Matrices to Serological Data. *PLoS Comput. Biol.* **6**(12), e1001021 (2010).
8. Grefenstette, J. J. *et al.* FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health* **13**(1), 1–14 (2013).
9. Gallagher, S., Richardson, L. F., Ventura, S. L. and Eddy, W. F. "SPEW: Synthetic Populations and Ecosystems of the World," *J. Comput. Graph. Stat.* pp. 773–784, <https://doi.org/10.1080/10618600.2018.1442342> (2018).
10. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**(3), 925–979 (2015).
11. Klovvdahl, A. S. *et al.* Social networks and infectious disease: The Colorado Springs study. *Soc. Sci. Med.* **38**(1), 79–88 (1994).
12. Jolly, A. M., Muth, S. Q., Wylie, J. L. & Potterat, J. J. Sexual networks and sexually transmitted infections: A tale of two cities. *Journal of Urban Health : Bulletin of the New York Academy of Medicine* **78**(3), 433 (2001).
13. Liljeros, F., Edling, C. R. & Amaral, L. A. N. Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes Infect.* **5**(2), 189–196 (2003).
14. Géniois, M. & Barrat, A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.* **7**(1), 1–18 (2018).
15. Smieszek, T. *et al.* Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes. *BMC Infect. Dis.* **16**(1), 1–14 (2016).
16. Ozella, L. *et al.* Using wearable proximity sensors to characterize social contact patterns in a village of rural Malawi. *EPJ Data Sci.* **10**(1), 1–17 (2021).
17. Pandit, J. A., Radin, J. M., Quer, G. & Topol, E. J. Smartphone apps in the COVID-19 pandemic. *Nat. Biotechnol.* **40**(7), 1013–1022 (2022).
18. Barrat, A., Cattuto, C., Kivela, M., Lehmann, S. & Saramäki, J. Effect of manual and digital contact tracing on COVID-19 outbreaks: a study on empirical contact data. *J. R. Soc. Interface* **18**(178), 20201000 (2021).
19. Burdinski, A., Brockmann, D. & Maier, B. F. Understanding the impact of digital contact tracing during the COVID-19 pandemic. *PLoS Digital Health* **1**(12), e0000149 (2022).
20. Rodríguez, P. *et al.* A population-based controlled experiment assessing the epidemiological impact of digital contact tracing. *Nature Communications* **12**(1), 587 (2021).
21. Aleta, A., de Arruda, G. F. & Moreno, Y. Data-driven contact structures: From homogeneous mixing to multilayer networks. *PLoS Comput. Biol.* **16**(7), e1008035 (2020).
22. Backhausz, Á., Kiss, I. Z. & Simon, P. L. The impact of spatial and social structure on an SIR epidemic on a weighted multilayer network. *Period. Math. Hungar.* **85**(2), 343–363 (2022).
23. Bokányi, E., Heemskerck, E. M. & Takes, F. W. The anatomy of a population-scale social network. *Sci. Rep.* **13**(9209), 1–16 (2023).
24. Cremers, J. *et al.* Unveiling the Social Fabric: A Temporal, Nation-Scale Social Network and its Characteristics, *arXiv*, Sep. 2024.
25. Kivela, M. *et al.* Multilayer networks. *Journal of complex networks* **2**(3), 203–271 (2014).
26. Aleta, A. & Moreno, Y. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics* **10**, 45–62 (2019).
27. Ahrenfeldt, L. J., Otavova, M., Christensen, K. & Lindahl-Jacobsen, R. Sex and age differences in COVID-19 mortality in europe. *Wiener klinische Wochenschrift* **133**, 393–398 (2021).
28. Koleva, G. *et al.* "The impact of COVID-19 in nursing homes in Madrid, Spain: a need for assessment," *Lancet Regional Health-Europe*, vol. 11, Dec. 2021.
29. Alstott, J., Bullmore, E. & Plenz, D. powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. *PLoS One* **9**(1), e85777 (2014).
30. Arregui, S., Aleta, A., Sanz, J. & Moreno, Y. Projecting social contact matrices to different demographic structures. *PLoS Comput. Biol.* **14**(12), e1006638 (2018).
31. Age-and-location-specific contact matrices for 152 countries. [Online]. Available: <https://doi.org/10.1371/journal.pcbi.1005697.s002> (2017).
32. Population and housing censuses 2011. [Online]. Available: <https://www.ine.es/censos2011/tablas/Inicio.do> (2011).
33. Serie de estudiantes matriculados por rama de enseñanza, comunidad autónoma, universidad y titulación. [Online]. Available: <https://datos.gob.es/es/catalogo/> (2020).
34. Estadística de estudiantes, grado. [Online]. Available: <https://datos.gob.es/es/catalogo/> (2020).
35. Estadística de Estudiantes, Máster. [Online]. Available: <https://datos.gob.es/es/catalogo/> (2020).
36. Estadística de Enseñanzas no Universitarias, formación Profesional. [Online]. Available: <https://datos.gob.es/es/catalogo/> (2019).
37. Empresas por CCAA, actividad principal (grupos CNAE 2009) y estrato de asalariados (antigua estratificación). [Online]. Available: <https://www.ine.es/jaxiT3/Tabla.htm?t=298&L=0> (2019).
38. Encuesta de Población Activa. [Online]. Available: [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176918&menu=resultados&idp=1254735976595](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&idp=1254735976595) (2019).

39. Total afiliados por régimen. datos municipios. Último día del mes. [Online]. Available: <https://www.seg-social.es/wps/portal/wss/internet/EstadisticasPresupuestosEstudios/Estadisticas/EST8> (2019).
40. Población en establecimientos colectivos por sexo, edad y tipo de establecimiento colectivo. [Online]. Available: <https://www.ine.es/jaxi/Tabla.htm?path=/t20/e244/colectivos/p02/l0/&file=02005.px&L=0> (2011).
41. Envejecimiento en Red. Datos abiertos: residencias. [Online]. Available: <https://envejecimientoenred.csic.es/datos-abiertos/residencias/> (2020).
42. Padrón Municipal de Habitantes, barcelona. [Online]. Available: <https://portaldades.ajuntament.barcelona.cat/es/estad> (2020).
43. Estadística de l'ensenyament. Departament d'Educació, Generalitat de Catalunya. [Online]. Available: <https://educacio.gencat.cat/ca/departament/estadistiques/estadistiques-ensenyament/cursos-anteriors/curs-2019-2020/> (2020).
44. Padrón Municipal de Habitantes. [Online]. Available: <https://www.valencia.es/cas/estadistica/banco-de-datos> (2020).
45. Resumen de escolarización 2020. [Online]. Available: <https://dadesobertes.gva.es/es/dataset/edu-alu-gen-2020> (2020).
46. Alumnado matriculado en centros de Formación Profesional 2020. [Online]. Available: <https://dadesobertes.gva.es/es/dataset/edu-alu-fp-2020> (2020).
47. Guía de centros docentes. [Online]. Available: <http://www.ceice.gva.es/web/centros-docentes/descarga-base-de-datos> (2020).
48. Padrón Municipal de Habitantes. [Online]. Available: <https://www.sevilla.org/servicios/servicio-de-estadistica/datos-estadisticos/> (2020).
49. Sistema de Información Multiterritorial de Andalucía. 3. Sociedad, 3.1 Enseñanza y formación. [Online]. Available: <https://www.juntadeandalucia.es/institutodeestadisticaycartografia/badea/informe/anual/> (2020).
50. Directorio de centros docentes. [Online]. Available: <https://www.juntadeandalucia.es/datosabiertos/portal/dataset/directorio-de-centros-docentes-de-andalucia> (2020).
51. Estadísticas de las enseñanzas de régimen general. [Online]. Available: <https://www.aragon.es/-/ensenanzas-de-regimen-general> (2020).
52. Población según secciones y edad, por sexo. hombres. [Online]. Available: [https://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU\\_padron/p20/sec23\\_sec24\\_31.html](https://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU_padron/p20/sec23_sec24_31.html) (2020).
53. Población según secciones y edad, por sexo. mujeres. [Online]. Available: [https://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU\\_padron/p20/sec24\\_sec25\\_31.html](https://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU_padron/p20/sec24_sec25_31.html) (2020).
54. Estadística de la enseñanza no universitaria. [Online]. Available: [https://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU\\_EDUCNOUNIV/Indice2016.html](https://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU_EDUCNOUNIV/Indice2016.html) (2017).
55. Centros educativos. [Online]. Available: [https://murciaencifras-com.carto.com/tables/centros\\_educativos\\_1/public](https://murciaencifras-com.carto.com/tables/centros_educativos_1/public) (2016).
56. Rodríguez, J. P., Aleta, A. & Moreno, Y. Digital cities and the spread of COVID-19: Characterizing the impact of non-pharmaceutical interventions in five cities in Spain. *Frontiers in Public Health* **11**, 1122230 (2023).
57. Code and data used in the manuscript. [Online]. Available: <https://doi.org/10.20350/digitalCSIC/16527> (2025).

## Acknowledgements

The authors acknowledge the Department of University Statistics of the Spanish Ministry of Universities for facilitating the number of registered students at the university per municipality, which is not currently publicly available online. JR, AA, and YM acknowledge support from the Government of Aragon (FONDO-COVID19-UZ-164255). JR is supported by Govern de les Illes Balears through the Vicens Mut program and the María de Maeztu Excellence Unit 2023-2027 Ref. CEX2021-001164-M, funded by MICIU/AEI/10.13039/501100011033. AA acknowledges support through the grant RYC2021-033226-I funded by MICIU/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. YM was partially supported by the Government of Aragon, Spain, and ERDF “A way of making Europe” through grant E36-23R (FENOL), and by Ministerio de Ciencia, Innovación y Universidades, Agencia Española de Investigación (MICIU/AEI/10.13039/501100011033) Grant No. PID2023-149409NB-I00. The authors acknowledge the use of the computational resources of COSNET Lab at Institute BIFI, funded by Banco Santander through grant Santander-UZ 2020/0274, and by the Government of Aragon (FONDO-COVID19-UZ-164255). JR and AA acknowledge funding from la Caixa Foundation under the project code SR20-00386 (COVID-SHINE). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

## Author contributions

Author contributions: Conceptualization: Y.M., A.A., J.R.; Software: J.R.; Data curation: J.R.; Formal analysis: J.R.; Supervision: Y.M., A.A.; Funding acquisition: Y.M.; Methodology: Y.M., A.A., J.R.

## Competing interests

Authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.P.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025