










# Machine learning and metagenomics reveal shared antimicrobial resistance profiles across multiple chicken farms and abattoirs in China

Received: 9 January 2023

Accepted: 7 July 2023

Published online: 10 August 2023

 Check for updates

Michelle Baker <sup>1,12</sup>, Xibin Zhang<sup>2,12</sup>, Alexandre Maciel-Guerra<sup>1,12</sup>, Yingping Dong<sup>3</sup>, Wei Wang<sup>3</sup>, Yujie Hu <sup>3</sup>, David Renney<sup>4</sup>, Yue Hu<sup>1</sup>, Longhai Liu<sup>5</sup>, Hui Li<sup>6</sup>, Zhiqin Tong<sup>6</sup>, Meimei Zhang<sup>7</sup>, Yingzhi Geng<sup>7</sup>, Li Zhao<sup>8</sup>, Zhihui Hao<sup>9</sup>, Nicola Senin <sup>10</sup>, Junshi Chen<sup>3</sup>, Zixin Peng <sup>3,13</sup> , Fengqin Li <sup>3,13</sup>  & Tania Dottorini <sup>1,11</sup> 

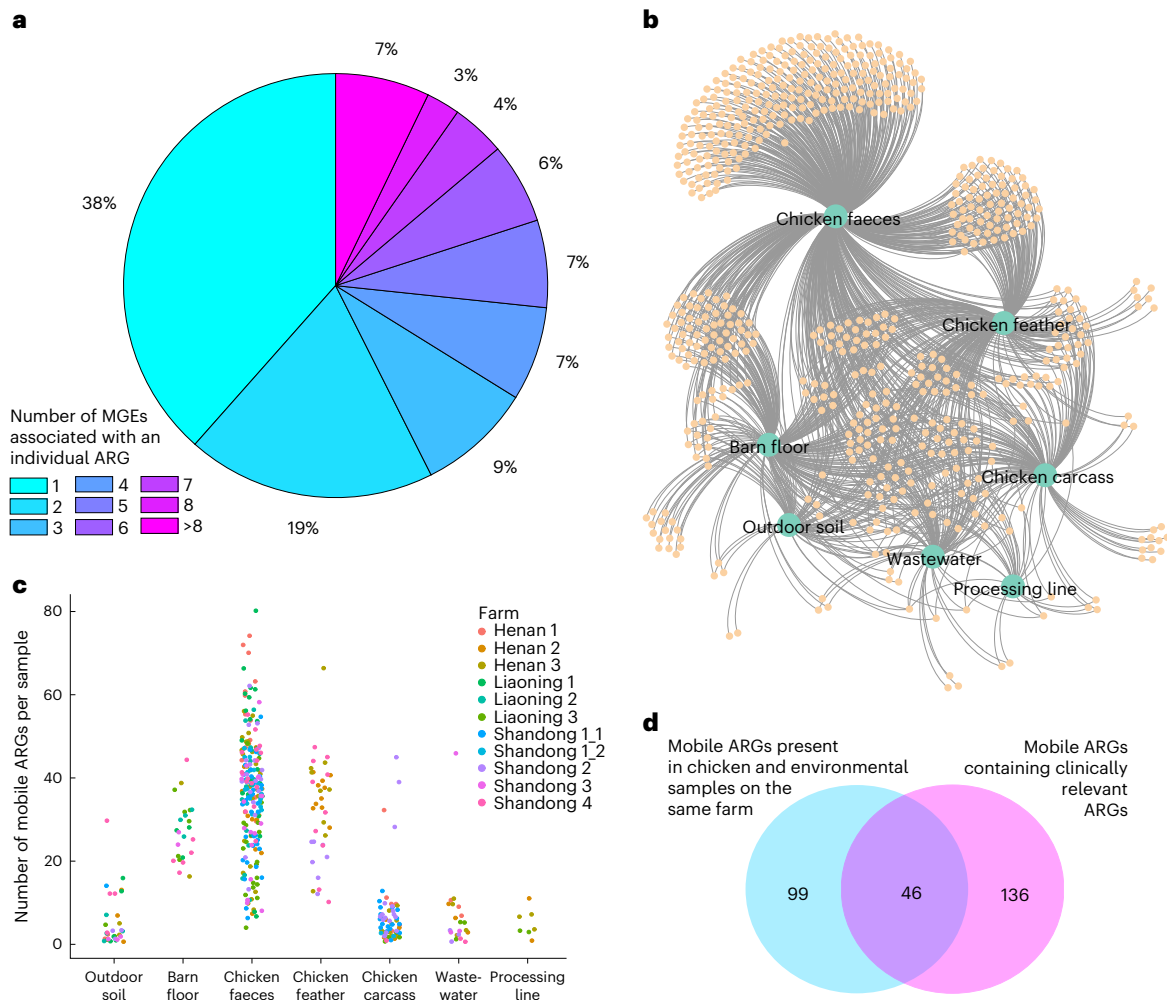
China is the largest global consumer of antimicrobials and improving surveillance methods could help to reduce antimicrobial resistance (AMR) spread. Here we report the surveillance of ten large-scale chicken farms and four connected abattoirs in three Chinese provinces over 2.5 years. Using a data mining approach based on machine learning, we analysed 461 microbiomes from birds, carcasses and environments, identifying 145 potentially mobile antibiotic resistance genes (ARGs) shared between chickens and environments across all farms. A core set of 233 ARGs and 186 microbial species extracted from the chicken gut microbiome correlated with the AMR profiles of *Escherichia coli* colonizing the same gut, including *Arcobacter*, *Acinetobacter* and *Sphingobacterium*, clinically relevant for humans, and 38 clinically relevant ARGs. Temperature and humidity in the barns were also correlated with ARG presence. We reveal an intricate network of correlations between environments, microbial communities and AMR, suggesting multiple routes to improving AMR surveillance in livestock production.

Antimicrobial use in poultry production in China is five times higher than the international average<sup>1</sup>. Antibiotic use, even at low levels, alters and expands the gut resistome in livestock<sup>2</sup>, and the microbial community can shape antimicrobial resistance (AMR) phenotypes<sup>3</sup>. External events such as changes in diet, temperature and stress<sup>4,5</sup> may result in the colonization of new resident species or AMR transfer between species<sup>6</sup>. Temperature, humidity and both bacterial species abundance and the presence of antibiotic resistance genes (ARGs)<sup>7–9</sup> can influence bacterial infection in broilers<sup>10</sup>. Links between environmental conditions and AMR are particularly

relevant for China and low- and middle-income countries (LMICs), where maintaining stable environmental conditions in industrial-scale farming may be challenging compared with in high-income countries<sup>11</sup>.

AMR surveillance in non-healthcare domains has not been widely adopted<sup>12</sup>, but is key to understanding how food production systems contribute to the selection and dissemination of antibiotic-resistant bacteria (ARB) and ARGs. Machine learning (ML) and big data mining offer tools to advance precision poultry farming<sup>13,14</sup>. Culture-based approaches involving whole genome sequencing (WGS) of individual

A full list of affiliations appears at the end of the paper.  e-mail: [pengzixin@cfsa.net.cn](mailto:pengzixin@cfsa.net.cn); [lifengqin@cfsa.net.cn](mailto:lifengqin@cfsa.net.cn); [tania.dottorini@nottingham.ac.uk](mailto:tania.dottorini@nottingham.ac.uk)



**Fig. 1 | Analysis of potentially mobile ARGs. a**, Pie chart showing the proportion of ARGs (out of the 195 found) associated with one or multiple MGEs. **b**, Undirected network graph showing potentially mobile ARGs (small orange circles) associated with different sample sources (large green circles). The edges in the graph link the potentially mobile ARGs to the sources in which they were found. **c**, Number of potentially mobile ARGs per sample per source. Each circle represents a single sample, with circles coloured by farm. **d**, Venn diagram

showing that 145 (out of 661) potentially mobile ARGs were found to be present in both chicken and environmental samples from the same farm, and 182 potentially mobile ARGs contained clinically relevant ARGs. An overlap of 46 clinically relevant<sup>30</sup> potentially mobile ARGs was found in chicken and environmental sources obtained from the same farm. Note that in this analysis, samples from the same source collected at  $t_1$  (week 3) and  $t_2$  (week 6) were aggregated together, leading to a total of seven sources considered for each farm.

pathogens, antibiotic susceptibility testing and ML techniques are effective predictors of genomic characteristics linked to AMR for both *Escherichia coli* isolates<sup>15–18</sup> and other bacteria<sup>19–24</sup>. However, surveillance approaches focusing solely on WGS of individual pathogens may not capture the diversity of the microbial communities and resistomes within livestock production and ARG data may be missed<sup>25</sup>. In a recent proof-of-concept study, we observed that several ARGs present in the chicken faecal resistome were found to correlate with the resistance/susceptibility profiles of *E. coli* isolates cultured from the same samples<sup>26</sup>.

In this study, we developed a reference method for metagenomic-based surveillance targeting Chinese livestock farming, where AMR surveillance is particularly challenging, using an approach that takes into consideration the lack of laboratory resources commonly experienced in China and LMICs<sup>27,28</sup>. We used *E. coli* as an indicator species for AMR within the wider context of the microbial community populating the chicken gut. To address wider contexts, we explored the impacts on the microbiomes of the surrounding and connected farm environments, barn temperature and humidity, and adopted antimicrobial administration protocols.

## Results

### Birds and environment share clinically relevant mobile ARGs

Biological samples were collected from ten large-scale commercial poultry farms (see Methods, Supplementary Information, Supplementary Fig. 1 and Supplementary Tables 1 and 2). Microbial communities and ARGs were differentiated across farm sources and between farms and abattoir (Supplementary Information, Supplementary Figs. 2–5 and Supplementary Tables 3–5). As gene mobility may influence ARG presence across sources and because of the potential importance of mobile genetic elements (MGEs) in the development of effective surveillance systems<sup>29</sup>, we looked for ARGs that were within 5 kilobases (kb) of an MGE<sup>26</sup> and considered these MGE–ARG combinations to be potentially mobile ARGs. In total, 661 different MGE–ARG combinations (potentially mobile ARGs) were found, featuring 195 unique ARGs (Supplementary Table 6). Of these, 75 ARGs (38%) were found in only one MGE–ARG combination, while the remaining 120 (62%) were found in multiple combinations (2 to 22; Fig. 1a). Over half (56%) of the 661 potentially mobile ARGs were present in more than one source (Fig. 1b), with three MGE–ARG combinations (*IS1216-poxTA*, *IS15-APH(3′)-Ia* and *ISCfr1-AAC(3)-IId*) present in all sources except

feathers. Chicken faeces had the highest number of potentially mobile ARGs, but also the greatest variance (Fig. 1c). Feathers and barn floor also carried many potentially mobile ARGs, the mean number statistically equivalent to faeces (Dunn's test adjusted  $P > 0.05$ ). Outdoor soil, carcasses, processing line and wastewater generally had lower numbers of potentially mobile ARG patterns per sample, with these numbers differing significantly (Dunn's test adjusted  $P < 0.01$ ) from faeces and feather, but not from each other. In total, across all 10 farms, 145 different MGE-ARG combinations were found in bird and environmental sources on the same farm, with some of these appearing on multiple farms. Of these, 46 contained clinically relevant ARGs<sup>30</sup> (Fig. 1d). Notably, we found *bla*<sub>NDM-5</sub> in chicken faeces, feathers and environmental barn floor samples. This gene is commonly found on the IncX3 plasmid, which can be disseminated among humans, animals, food and environment<sup>31</sup>, although we did not confirm plasmid presence in our short-read metagenomic sequencing (MGS) data. Another important clinically relevant gene, *qnrS1*, was found in chicken faeces, feather, environmental barn floor and wastewater samples. This plasmid-mediated quinolone resistance gene is known to be present in the chicken supply chain and is capable of being transferred to different bacteria<sup>32</sup>.

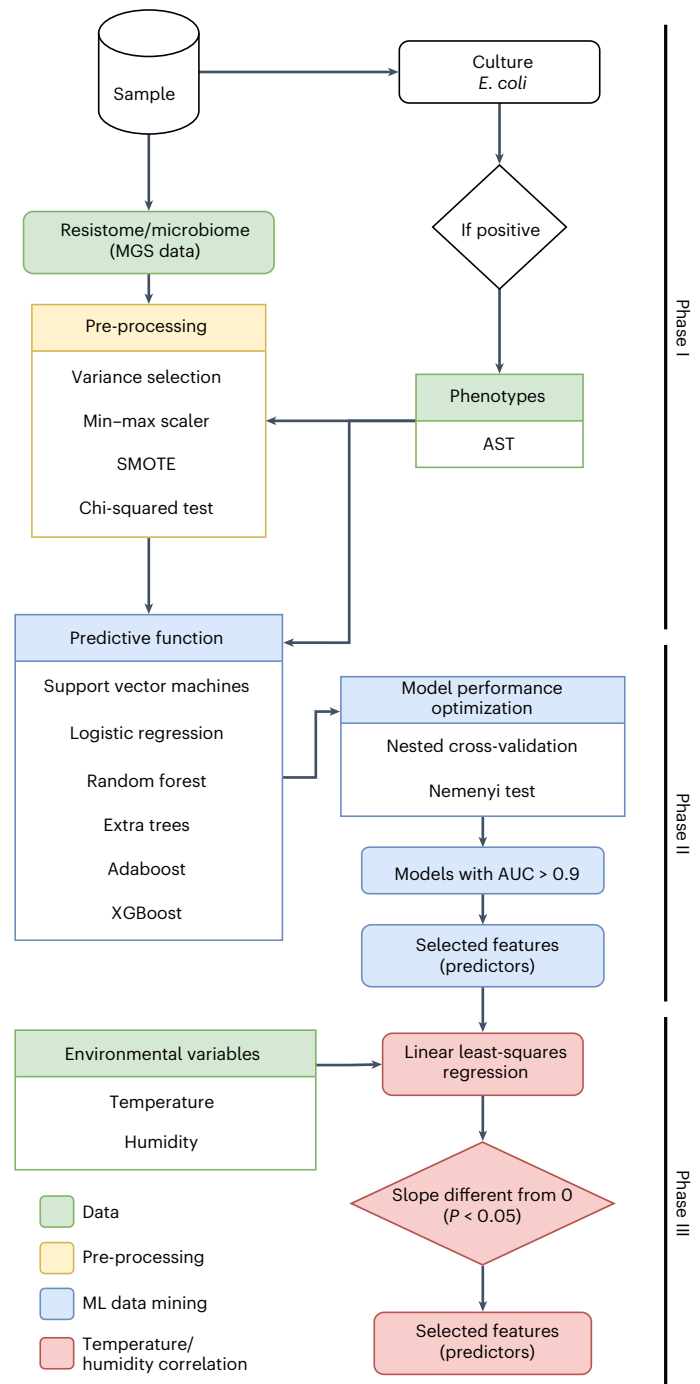
### *E. coli* AMR correlates with the gut microbiome it inhabits

We further investigated whether there was a correlation between the bacterial species found in the chicken gut, the resistome (that is, the ARGs from all species) and the AMR profiles of *E. coli* isolates taken from the same samples as the metagenome data. We cultured *E. coli* isolates from 170 chicken faecal samples (a subset of the samples that had been used for metagenomics) and characterized their AMR profiles against a panel of 26 antibiotics. The proportion of isolates resistant to each antibiotic ranged from 1% to 98% (Supplementary Table 7). All isolates were resistant to at least one antibiotic, with 169 resistant to at least three.

To investigate the correlations between antibiotic resistance in *E. coli* and gut microbiome, we developed a bespoke data mining method based on ML (Fig. 2). The method consists of building an ML-powered 'predictive function' whose input is the aggregation of information from the gut microbial community (relative abundances of microbial species) and gut resistome (ARG count) and whose output is the resistance of *E. coli* to a specific antibiotic (true or false) from antimicrobial susceptibility testing (AST). The predictive function was trained by using experimental data (supervised learning) and swapping different underlying ML technologies until optimal prediction performance was achieved. A set of the most informative features, also referred to as 'predictors', was extracted from the ML models. The set was then refined by analysis of the correlation with temperature and humidity (see later).

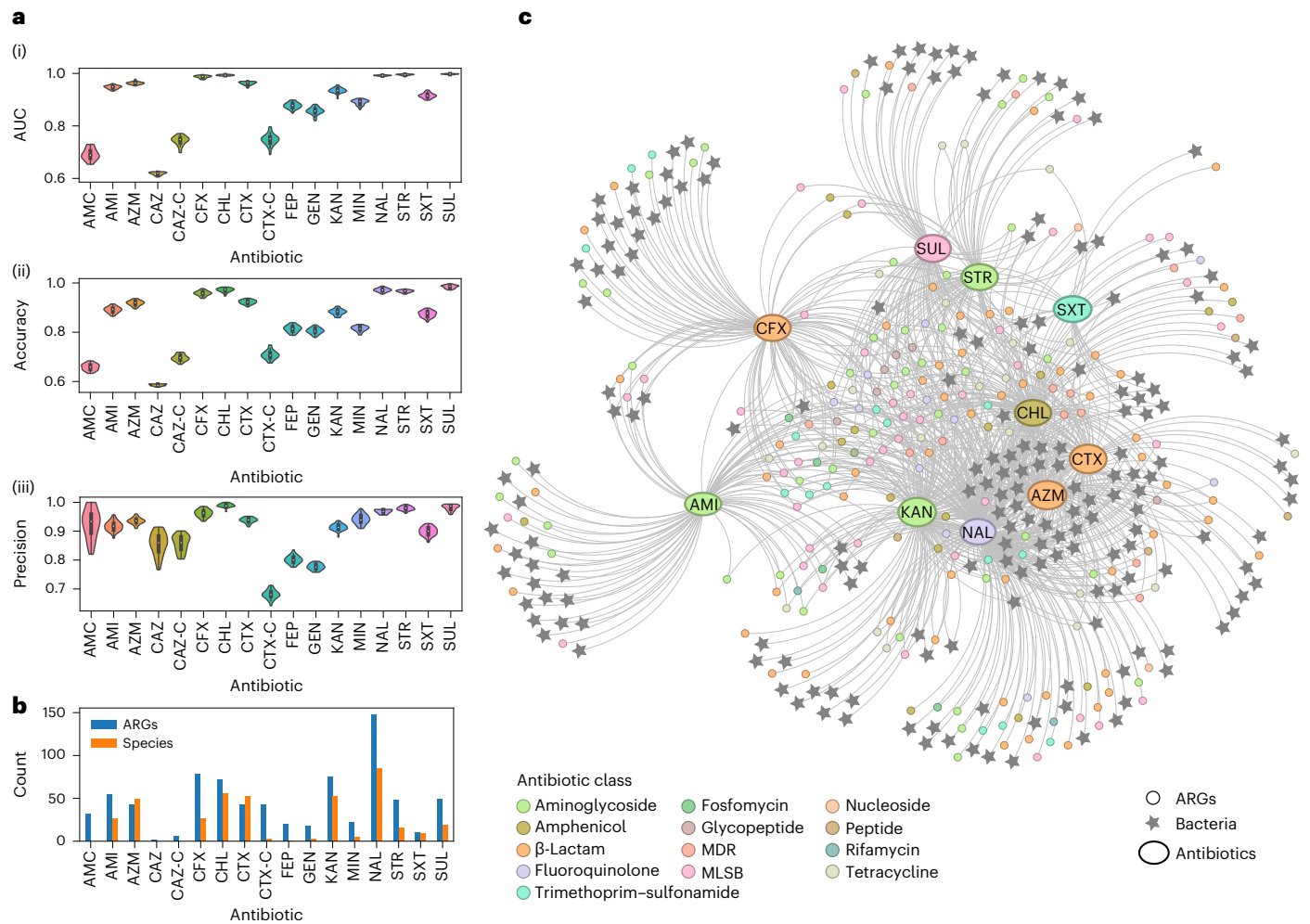
Out of the 26 antibiotics, only 17 had sufficient data (resistance and susceptibility cases) to allow proper ML training: amikacin, amoxicillin-clavulanic acid, aztreonam, cefepime, ceftazidime, ceftazidime-clavulanic acid, chloramphenicol, cefotaxime, gentamycin, kanamycin, minocycline, nalidixic acid, streptomycin, sulfafurazole and trimethoprim-sulfamethoxazole. For all, the best prediction performance (Nemenyi test) was observed with the extra tree classifier (ML technology; Supplementary Table 8 and Supplementary Fig. 6). The prediction performance indicators computed using the extra tree method are reported in Fig. 3a and Supplementary Fig. 7. Ten predictive models (amikacin, aztreonam, ceftazidime, chloramphenicol, cefotaxime, kanamycin, nalidixic acid, streptomycin, sulfafurazole and trimethoprim-sulfamethoxazole) achieved performances exceeding AUC > 0.90.

Data mining showed that a core subset of the chicken gut resistome (all detectable ARGs from the chicken faeces metagenomic data) and microbial species (all bacterial species from the chicken faeces metagenomic data) exhibited strong predictive power for *E. coli* resistance. This core consisted of 419 features



**Fig. 2 | Data mining pipeline to find correlations between gut microbiome, antibiotic resistance in *E. coli*, temperature and humidity.** The full data analysis workflow of the bespoke data mining method based on ML. Input data are shown in green. Phase I involves metagenome data pre-processing (in yellow). The steps are described in detail in the Methods section. Phase II involves the training and testing of ML-powered predictive functions to isolate metagenomic features (that is, the ARG count and relative abundances of microbial species present in the sample) correlated with phenotypic resistance (in blue). Phase III involves fitting regression models (discussed in the next section) to isolate metagenomic features that better correlate with variations of temperature and humidity (in red). AUC, area under the curve.

(186 microbial species and 233 ARGs) acting as strong predictors of *E. coli* resistance/susceptibility to 10 antibiotics (Fig. 3b,c and Supplementary Table 9) with an AUC of over 0.90. The 233 ARGs



**Fig. 3 | ML performance and feature selection from correlations between gut microbial species, resistome and antibiotic resistance in *E. coli*.**

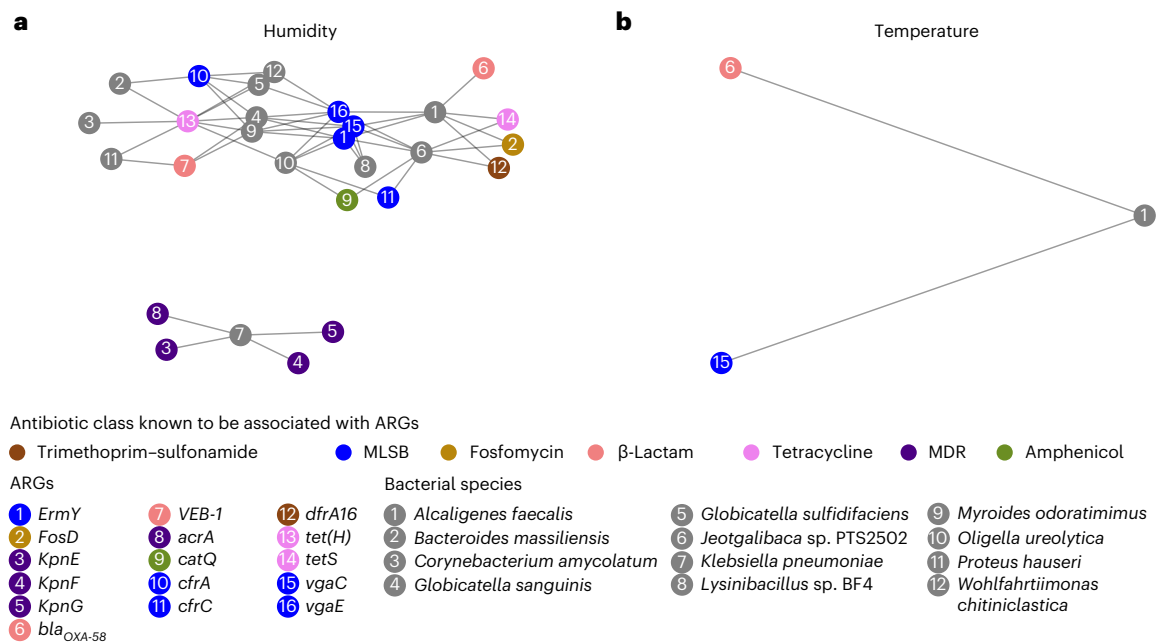
**a**, Performance of the ML-powered predictive functions of *E. coli* resistance to specific antibiotics (ML technology: extra tree classifier; see Methods). Performance indicators (AUC, accuracy and precision) were computed as the average of 30 iterations of nested cross-validation (see Methods). See Supplementary Fig. 2 for performance indicator sensitivity, specificity and Cohen's kappa score. The violin plots show the distribution of the data, with each data point representing one antibiotic model. Inside each violin plot is a box plot, with the box showing the interquartile range (IQR), the whiskers showing the rest of the distribution as a proportion of 1.5 x IQR and the white circle representing the median value. **b**, Counts of metagenomic features (ARGs and microbial species) found as the strongest predictors of *E. coli* resistance/susceptibility

profiles to each antibiotic. **c**, Undirected graph showing the strongest predictors (metagenomic features in the chicken gut) for each antibiotic model. The edges of the graph link ARG or bacteria species nodes (predictor variables) to the antibiotic model in which they were found to be predictive. Both the ARG and antibiotic model nodes are colour coded according to the antibiotic/ARG is known to be associated with. The ML models were run for the following antibiotics: amoxicillin–clavulanic acid (AMC), amikacin (AMI), aztreonam (AZM), ceftazidime (CAZ), ceftazidime–clavulanic acid (CAZ-C), cefotaxime (CTX), cefotaxime–clavulanic acid (CTX-C), ceftiofur (CFX), chloramphenicol (CHL), cefepime (FEP), gentamycin (GEN), kanamycin (KAN), minocycline (MIN), nalidixic acid (NAL), streptomycin (STR), sulfafurazole (SUL) and trimethoprim–sulfamethoxazole (SXT). MDR, multidrug resistant.

from the top 10 antibiotic models belonged to  $\beta$ -lactams (24% of the ARGs), aminoglycosides (18%), and macrolides, lincosamides and streptogramin B (MLSb; 18%), with other antibiotic classes accounting for less than 10% each. Of these 233 ARGs, based on the correlation of ARG read depth with species abundance (see Methods)<sup>33</sup>, 46 were found to be present in contigs identified as originating from *E. coli*. A further 16 ARGs (of the 233) were present only in contigs identified as other bacterial species (that is, they did not originate from *E. coli*). To further explore the relationship between core gut features and antibiotic resistance, the 419 features and 10 antibiotic resistances were visualized as nodes of a graph, with edges only connecting predictors to predicted resistances (Fig. 3c). This analysis highlighted a core of 66 ARGs (15 clinically relevant, including *bla*<sub>NDM-5</sub>, *bla*<sub>CTX-M-15</sub>, *dfrA15* and *dfrA5*) acting as predictors of more than

three antibiotic resistances. Three ARGs (*aphA6*, *vatA*) and *vgb(A)* were found to be predictors of eight antibiotic resistances. The same analysis revealed 28 microbial species in the gut acting as predictors of 5 antibiotic resistances (aztreonam, chloramphenicol, cefotaxime, kanamycin and nalidixic acid). These 28 species included the bacterial genera *Arcobacter*, *Acinetobacter* and *Sphingobacterium* in addition to other commensal bacteria.

Shapley additive explanation values were used to explain the AMR-related features selected by ML (Supplementary Fig. 8). The top ten most important features found to predict resistance for each antibiotic model indicated that 41% of the features had their presence positively associated with the prediction of resistant phenotypes, while 59% had their absence positively associated with the prediction of the resistant phenotype, most notably for the antibiotic models nalidixic



**Fig. 4 | Gut features identified as predictors of *E. coli* resistance. a, b,** Microbial species and ARGs correlated with humidity (a) and temperature (b). Microbial species and ARGs are correlated with humidity or temperature, and also with each other, indicating that the ARGs are likely to be present in the species. Features were considered correlated if the slopes of the linear regression lines

were significantly different from zero ( $P < 0.05$  using a two-sided  $t$ -test). Nodes indicate ARGs or microbial species; edges connect species to ARGs likely present in the species. ARG nodes are colour-coded according to the antibiotic class known to be associated with the ARG; microbial species nodes are shown in grey.

acid and streptomycin. Conversely, eight of the top ten features in each model were negatively associated with the prediction of resistance. The chloramphenicol antibiotic model had the highest number of ARGs known to confer resistance to the same antibiotic class (phenicol; *optrA* (ref. 34), *lsaE* (ref. 35) and *mel* (ref. 35)) or known to facilitate resistance to it (*oqxA* (ref. 36)).

### Temperature and humidity shape the gut microbiome linked to AMR

For the top ten antibiotic models, we developed bespoke regression models using individual gut features as independent variables (one model per variable) and temperature or humidity as dependent variables to ascertain whether model fitting would highlight a correlation (see phase III in Fig. 2 and Methods). Temperature and humidity were measured in all farms except Liaoning 1 (LN1) over a full chicken production cycle (Supplementary Table 10 and Supplementary Fig. 9). Amongst the original 419 features, 130 ARGs and 48 microbial species correlated with humidity, whilst 39 ARGs and 20 microbial species correlated with temperature (Supplementary Fig. 10 and Supplementary Table 11). The correlation with humidity was on average stronger (higher  $R^2$  values in the regression analysis, Supplementary Fig. 10). Of the 130 ARGs correlated with humidity, 22% were MLSb, 18% were  $\beta$ -lactams, 17% were aminoglycosides and 11% were tetracyclines. Of the 39 ARGs correlated with temperature, 23% were  $\beta$ -lactams, 18% were MLSb, 15% were aminoglycosides and 13% were glycopeptides. Nineteen ARGs correlated with both temperature and humidity, four of them clinically relevant (*qnrA1*, *qnrS2*, *bla<sub>NDM-1</sub>* and *catA8*). Four microbial species from the phyla Proteobacteria (*Helicobacter pullorum* and *Alcaligenes faecalis*), Firmicutes (*Bacillus cereus* group) and Bacteroidetes (*Bacteroides stercoris*) correlated with both temperature and humidity. One species from Tenericutes (*Mycoplasma yeatsii*) correlated with temperature only, while other species from Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria correlated with either temperature or humidity (Supplementary Table 11).

We tested for the possibility that some ARGs found to be correlated with temperature or humidity might belong to microbial species that are also correlated with temperature and humidity. This was done by correlating ARG read depth with microbial species read depth as proposed by Tong et al.<sup>33</sup>. The analysis highlighted two distinct subgraphs correlated with humidity (Fig. 4a) and one correlated with temperature (Fig. 4b). Notably, one of the subgraphs correlated with humidity contained *Klebsiella pneumoniae* and four related ARGs (*kpnE*, *kpnF*, *kpnG* and *acrA*). The subgraph containing *A. faecalis* and ARGs *vgaC* and *bla<sub>OXA-58</sub>* was found in both analyses (that is, it correlated with both temperature and humidity).

We then investigated whether the gut ARG features identified as predictors of resistance in *E. coli*, and further identified as correlated with humidity or temperature, were in close proximity to MGEs. Ten ARGs were found located in close proximity to MGEs (MLSb: *optrA*, *mph(F)* and *erm(X)*;  $\beta$ -lactams: *bla<sub>NDM-1</sub>* and *bla<sub>OXA-58</sub>*; amphenicol: *catA8* and *catB2*; aminoglycoside: *aadA1*; fluoroquinolone: *qnrS2* and *qnrA1*). Three of the ten ARGs were found to be associated with only one MGE (*catB2* with *ISPa25*, *mph(F)* with *IS15* and *qnrA1* with *IS15*), whilst the other seven were associated with two to nine different MGEs. All the MGE-ARG pairs were investigated for conserved structure across farms or sources. For example, the clinically important *bla<sub>NDM-1</sub>* was found in close proximity to *IS15* in four samples (three chicken faeces from LN1 and one barn floor sample from Liaoning 3 (LN3); Supplementary Fig. 11). In 19 samples from chicken faeces and feather samples from LN1, LN3, Shandong 2 (SD2) and Shandong 4 (SD4), *bla<sub>NDM-1</sub>* was found in proximity to MGE *ISAbA125* and located next to another ARG, *ble*, which is a known association for plasmid-borne *bla<sub>NDM-1</sub>* in *Enterobacteriaceae* species from Asian regions<sup>37</sup>. Despite having found the same *bla<sub>NDM-1</sub>*-*ISAbA125* pattern in several farms (LN1, LN3, SD2 and SD4), there was no evidence of transmission between farms (Fig. 5). Instead, evolutionary analysis of the contigs (using a molecular clock model to predict the rate of molecular evolution on each branch of the phylogenetic tree<sup>38</sup>) suggested recent branching of isolates within individual



isolation in combination with statistical and ML methods to draw out complex correlations showing AMR trends and patterns. *E. coli* has an established role as a reference indicator of AMR<sup>26</sup>. We found that 38 clinically relevant ARGs correlated with resistance to multiple antibiotics. Some of these antibiotics had no previously known association with these ARGs. In particular, 14 ARGs (*aadA16*, *aph(3')-Ia*, *aph(3')-VIa*, *bla<sub>CARB-16</sub>*, *catQ*, *dfrA15*, *dfrA16*, *dfrA27*, *bla<sub>OXA-58</sub>*, *bla<sub>PER-1</sub>*, *qnrD1*, *tet(Z)*, *tet(39)* and *bla<sub>SHV-110</sub>*) found to be associated with resistance to the highest number of antibiotics had previously been found in earlier studies on poultry in China<sup>40–42</sup>, confirming our method. However, we found a cluster of gut bacteria that correlated well with *E. coli* resistance to five different antibiotics. These included *Arcobacter* (an emerging waterborne and foodborne zoonotic pathogen, responsible for gastroenteritis in humans<sup>43</sup>), *Acinetobacter* (commensal in the poultry gut, but capable of causing extraintestinal diseases in both humans and poultry<sup>44</sup>) and *Shingobacterium* (clinically relevant in humans and animals<sup>45</sup>). This result suggests that, in agreement with previous studies<sup>12,29,46–50</sup>, focusing exclusively on *E. coli* within the farm for surveillance purposes may not be as effective as monitoring a larger number of pathogens.

In our study, the farms that used tetracyclines, lincosamides and polypeptides were positively correlated with the presence of ARGs from a wide range of classes, beyond those specific to the selected antibiotics. This appears to be consistent with previous findings<sup>51,52</sup>, but contrasts with a recent study from the United States<sup>53</sup>. It is possible that the co-localization of AMR genes is playing an important role in AMR selection in our farms. Indeed, the co-localization of AMR genes in bacterial genomes in food animals has previously been observed and recognized as a food safety concern in China<sup>54</sup> as well as elsewhere<sup>55,56</sup>.

The chickens in our study were housed in sheds that did not have an effective climate control system, and therefore experienced substantial temperature and humidity variations. Our results indicate that the core features of the gut microbial community and resistome, found to be correlated with resistance in *E. coli*, are also correlated with changes in temperature and humidity in chicken housing. Our results confirm and expand findings of previous studies<sup>7–9,26,37</sup>. Of note, the relative abundance of *A. faecalis* and the ARGs *vga(C)* and *bla<sub>OXA-58</sub>* originating from this species (via analysis of ARG and species read depths) were found to be correlated with changes in both temperature and humidity. A greater abundance of *A. faecalis* and more severe clinical symptoms in higher humidity conditions have been observed previously in a case–control study of turkeys kept at different humidity levels and inoculated with *A. faecalis*<sup>58</sup>. This bacterium is commonly found in birds<sup>59</sup> and would not typically be monitored by conventional surveillance. However, it is considered an emerging pathogen, has been associated with infections in humans and is considered difficult to treat due to its capacity to become extensively drug resistant<sup>60</sup>. Similarly, the important opportunistic pathogen *K. pneumoniae* and four ARGs (*kpnE*, *kpnF*, *kpnG* and *acrA*) originating from this bacterium and important for *K. pneumoniae* resistance<sup>61</sup> were found to be correlated with changes in humidity. *K. pneumoniae* can be transmitted via airborne contamination and has previously been found to have increased survival in indoor high-humidity conditions, highlighting the importance of studying this bacterium in indoor environments<sup>62</sup>. The associations between environmental variables (easily monitorable and controllable) and the species and genes associated with AMR present opportunities for the development of novel AMR monitoring solutions, especially in LMICs where these variables are not controlled and pose a risk to the animals that are exposed to changes in them.

Ten potentially mobile ARGs in the gut resistome were found to correlate with *E. coli* resistance and with temperature and humidity. In addition, 67 potentially mobile ARGs were found to correlate with *E. coli* resistance and humidity. One of these, the gene *cfp(C)*, encoding a ribosomal RNA methyltransferase conferring resistance to linezolid and phenicol antibiotics, was found near *ISEc9* (within 5 kb) and is

associated with CTX-M genes<sup>63,64</sup> (as we also found). The association of *drfA16* with the transposase *IS6100* has previously been reported in only a single study with an association with *Corynebacterium diphtheriae*, the causative agent of cutaneous diphtheria<sup>65</sup>. These associations potentially indicate an environment-specific evolution of these MGEs, as has been hinted at in previous work on pig farms that showed the importance of MGEs for AMR varied according to the season<sup>66</sup>.

Even though our analysis relied on a large set of samples from many heterogeneous sources with geographical and seasonal variations, our scope was limited to *E. coli*, did not consider human samples and would benefit from extending the analysis to other indicator species such as *Enterococcus*<sup>67</sup>. Spatial and temporal variations in farm/slaughterhouse microbial communities and resistomes are mirrored in human faecal samples, as our previous work and that of others have shown<sup>26,68</sup>, but whether these observations would be generalizable and globally true is currently unknown.

Metagenomic sequencing has the potential to broaden our knowledge of the factors driving resistance and improve AMR surveillance<sup>69</sup>. Metagenomic sequencing data are essential for developing new infection and resistance control policies, raising awareness of AMR and allowing the optimal use of antibiotics by veterinary professionals<sup>12,70</sup>. MGS shows great promise for AMR surveillance in environmental sectors, but methodologies need to be standardized and data gaps filled<sup>71,72</sup>, with few laboratories and countries at present having both the resources and expertise to use MGS for AMR surveillance<sup>73</sup>. With further development, metagenomic and ML approaches could be deployed to provide fast and reliable predictions of AMR outbreaks, emerging pathogens and transmission routes<sup>69</sup>.

Despite the increasing availability of low-cost precision farming technologies and metagenomics<sup>29,74</sup>, innovation and methodological advances must further enable the development of surveillance solutions capable of monitoring AMR dynamics<sup>12,29</sup>. Drug resistance arises from complex interactions across ARBs, microbial communities, geographical niches and environments, evolutionary forces, climate and human practices. We have demonstrated how methodologies can be developed that are capable of associating a wide array of microbial species and genes with observable AMR, and further assessed how those are associated with the environmental variables of temperature and humidity. Consideration of all relevant and interconnected AMR datasets in a 360° approach will drive forward our understanding and control of AMR spread.

## Methods

### Ethics statement

This study complied with all relevant ethical regulations and was specifically performed in accordance with protocols approved by the Ethics Committee of the State Key Laboratory of the China National Center for Food Safety Risk Assessment (ethical approval number: 2018018). Ethical approval was also obtained from the Research Ethics Committee of the School of Veterinary Medicine and Science at the University of Nottingham (application identification number: 2340180613).

### Collection of biological samples and environmental sensor data

For this study, we selected ten large-scale commercial poultry farms in three different provinces in China (Shandong, Henan and Liaoning; hereafter, the farms are denoted SDx, HNx and LNx, respectively), covering an area of 472,500 km<sup>2</sup>, each farm feeding into one of four regional abattoirs (two in Henan, one in Liaoning and one in Shandong). Each farm featured multiple barns, each barn containing between 12,000 and 32,800 birds, leading to a total production capacity of 110,730 to 380,000 birds per breeding cycle (depending on farm). Broiler production was based on self-breeding with broilers bred on the farm and moved to barns in same-aged batches. Of the ten selected farms, four (three in Liaoning and one in Shandong) used net housing

systems, whilst the other six used cage housing systems. During collection, the number of birds per barn did not significantly differ between the two housing systems ( $t$ -test,  $P = 0.07$ ).

Sampling followed the same pooled birds over one breeding cycle, except for one farm in the Shandong province (Shandong 1), which was sampled over two cycles to conduct a pilot study to fine-tune the collection campaign and data analysis protocols<sup>26</sup>. Biological samples were collected at the same three time points in every breeding cycle (including both cycles in Shandong 1):  $t_1$  (week 3),  $t_2$  (week 6) and  $t_3$  (1–5 days after week 6). Biological samples consisted of pooled faeces and feathers (not necessarily from the same animals) from the droppings of live birds in the barns collected from the barn floor immediately after excretion at mid-life ( $t_1$ ) and at the end of life ( $t_2$ ) of the animals, as well as barn floor samples (litter) collected at the same time points. In the abattoirs, samples were collected on slaughtering day ( $t_3$ ) from carcasses, meat processing surfaces (referred to as the processing line) and wastewater. Soil samples were collected from outside areas surrounding the farms at  $t_1$  and  $t_2$ . Details of the collection methods are available in the Supplementary Information.

All the farms involved in this study were equipped with heating/air conditioning systems. Environmental sensor data (temperature and humidity) were collected at intervals of 5 min using the automated sensors and data loggers available in most farms (HN1, HN2, HN3, SD2, SD3 and SD4). Three farms (SD1, LN2 and LN3) were unequipped with automated solutions and manual measurements were performed using SMART SENSOR AS837 temperature/humidity devices either daily or every 6 h. Farm LN1 had technical issues with the sensor and did not acquire any measurements. In all cases, the temperature and humidity data were averaged over three measurements taken at different locations within the barn.

### DNA library construction and sequencing

DNA extraction was performed on faeces, barn floor and outdoor soil samples using a magnetic bead genomic DNA extraction kit (DOP336-T3, TIANGEN Biotech). For carcass samples, the cetyltrimethylammonium bromide method<sup>75</sup> was used. Samples with DNA content above 1  $\mu\text{g}$  were used to construct the DNA library. The DNA concentration was measured using a Qubit dsDNA Assay Kit and Qubit 2.0 fluorometer (Life Technologies), and the integrity was measured using 1% agarose gel electrophoresis. A total amount of 1  $\mu\text{g}$  DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using NEBNext Ultra DNA Library Prep Kit for Illumina (NEB). The DNA sample was fragmented into 350 bp, and then DNA fragments were end-polished, A-tailed and ligated with the full-length adaptor for Illumina sequencing with further PCR analysis. Finally, the PCR products were purified (AMPureXP system) and the libraries analysed for size distribution using an Agilent 2100 Bioanalyzer (Agilent Technologies) and quantified using real-time PCR. After cluster generation, the library preparations were sequenced on an Illumina Novaseq 6000 platform and 150 bp paired-end reads were produced.

### Bioinformatics analysis

The raw sequence reads, obtained from the Illumina HiSeq sequencing platform, were pre-processed and filtered using Readfq (V8, <https://github.com/cjfields/readfq>) to acquire high-quality data for subsequent analysis. Host DNA was filtered using Bowtie 2 (v2.3.4.1)<sup>76</sup> and SAMtools (v1.9 (ref. 77)); reference genome accession code: GCF\_000002315.6). The microbiomes of samples were constructed by assembling the metagenome sequencing data for the different sample sources (chicken faeces, chicken feather, chicken carcass, barn floor, outdoor soil, wastewater and processing line) separately using binning and dereplication pipelines<sup>26,78</sup>. MEGAHIT (v1.1.2)<sup>79</sup> software was used to assemble the sequences. Single sample assemblies were generated for all samples with MEGAHIT default parameters.

Co-assemblies were generated for each sample source group (chicken faeces, chicken feather, chicken carcass, barn floor, outdoor soil, wastewater and processing line), each with the MEGAHIT setting parameters “--continue --kmin-1pass --min-contig-len 1000” as previously used for co-assemblies<sup>78</sup>. Filtered contigs (>2,000 bp) were mapped to single assemblies and co-assemblies using Burrows–Wheeler Aligner–Maximal Exact Match (BWA-MEM v2.2.1)<sup>80</sup> and SAMtools (v1.9)<sup>77</sup> to produce the Binary Alignment Map (BAM) files. METABAT2 (v2.15)<sup>81</sup> was used to obtain the depth of coverage. The taxonomic classification and composition (relative species abundances) of the metagenome reads were profiled using MetaPhlan (v3.0)<sup>82</sup> with Bowtie 2 (v2.3.4.1)<sup>76</sup> using the default settings –bowtie2out –input\_type fastq. Nonmetric multidimensional scaling (NMDS) of the relative species abundance was performed in R (v3.6.2) using the vegan<sup>83</sup> package with Bray–Curtis dissimilarity. Analysis of variance was performed in R using PERMANOVA from the vegan package<sup>83</sup> with pairwise testing using the pairwise adonis function<sup>84</sup> with Holm correction for multiple comparisons. Relative abundances were visually analysed by combining violin plots and categorical scatter plots, and differences were assessed by Wilcoxon rank sum test with Holm correction (adjusted  $P = 0.05$ ).

As sequencing depth can affect the observed diversity in genomic sequencing, rarefaction is widely used to normalize samples before analysis across different sample types<sup>85</sup>. However, the use of rarefaction is controversial as the subsampling leads to the loss of information available in the non-rarefied sample<sup>86</sup>. Hence, in this study, we used rarefied data only where necessary (to compare different sample types) and used non-rarefied data where only a single sample type was being considered. Host-removed reads were rarefied using the minimum sample depth using seqtk (<https://github.com/lh3/seqtk>), with the random seed fixed for each pair of reads.

### Analysis of resistome and MGEs

Assembled genomes were searched for sequence similarity to annotated ARGs present in the Comprehensive Antibiotic Resistance Database (CARD)<sup>61</sup> using Basic Local Alignment Search Tool–nucleotide (BLASTn)<sup>87</sup> with an identity threshold of 95% and coverage threshold of 95% (stricter thresholds with respect to our previous study<sup>26</sup>) to minimize the likelihood of mislabelled ARG variants. NMDS analysis was performed on the resulting gene count matrix in the vegan R package<sup>83</sup> using Bray–Curtis dissimilarity. Comparisons were made using (1) the total number of ARGs present per sample, (2) the actual count of individual ARGs per sample and (3) the relative ARG abundance per antibiotic class according to CARD (the number of ARGs present in the sample divided by the total number of ARGs in that class). These three approaches were visually analysed by combining violin plots and categorical scatter plots, and differences were assessed by Wilcoxon rank sum test with Holm correction (adjusted  $P = 0.05$ ).

To identify the source bacteria from which the ARGs originated, in accord with a previous study<sup>33</sup>, rarefied reads from each metagenome sample were mapped to their single assemblies using BWA-MEM (v2.2.1)<sup>80</sup> and SAMtools (v1.9)<sup>77</sup>. The average depths were assigned to the ARG-carrying contigs and ARGs. The coverage of ARGs, normalized by gene/contig length<sup>33</sup>, was then used to correlate with species abundance through the Spearman correlation test. ARG–species pairs were considered significantly correlated if  $P < 0.05$  and the Pearson correlation coefficient  $\geq 0.6$ .

To look for the presence of potentially mobile ARGs shared across different sources, ARGs carried by both the environment and chickens were considered. Filtered contigs (>500 bp) in each assembly were searched for ARGs and MGEs by a BLASTn search against the CARD<sup>61</sup> and ISfinder (<https://isfinder.biotoul.fr/>) databases using an identity threshold of 95% and coverage threshold of 95% to prevent false positives and variant uncertainty<sup>88</sup>. The distance between an ARG and MGE was calculated from the positions of the ARG and MGE in the contig<sup>26</sup>. ARG-carrying contigs with a distance of more than 5 kb between ARG



and MGE were discarded<sup>68,89–91</sup>, with the remaining contigs classed as potentially mobile ARGs. Contigs were annotated using Prokka (v1.14.6)<sup>92</sup>. Potentially mobile ARG patterns found in only a single sample were discounted in the analysis. ARGs were further classified as clinically important if the ARG was included in the Risk I category (clinically important ARGs dataset) according to Zhang et al.<sup>30</sup>. These genes were classed as Risk I if they were (1) present in human-associated environments, (2) potentially mobile genes and (3) present in ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species). The structures of the potentially mobile ARG patterns (MGE type, ARG carried, MGE carried, sample source, farm, number of samples carrying potentially mobile ARG and distance) are summarized in Supplementary Table 8. For IS*Aba125*–*bla*<sub>NDM-1</sub>, the gene structure was visualized using EasyFig<sup>93</sup>.

Evolutionary phylogeny was reconstructed for contigs carrying the potentially mobile ARG IS*Aba125*–*NDM-1* using BEAST (v1.10.4)<sup>38</sup>. All combinations of three clock models (strict, uncorrelated log normal and uncorrelated exponential) and three tree priors (constant coalescent, logistic growth and Bayesian skyline) were tested using stepping-stone sampling on the contigs to identify the best model. The best model was found to be a random uncorrelated log-normal clock model with a Bayesian skyline growth model. The GTR-gamma nucleotide substitution model was used, as selected by a maximum likelihood tree analysis in IQ-tree2 (v2.0.6) using automated model selection<sup>94</sup>. The analysis was conducted for three independent chains until the effective sample size, that is, the effective number of independent draws from the posterior distribution, for all parameters was greater than 200 per chain. This entailed each chain running for 100 million steps. Convergence was assessed in Tracer (v1.7.1)<sup>95</sup>, and chains were subsequently combined using LogCombiner (v1.10.4)<sup>96</sup>. The maximum clade credibility tree was selected using TreeAnnotator (v1.10.4) and then visualized in iTOL (v5)<sup>97</sup>.

### Investigation of correlations between faecal metagenomic features, antibacterial resistance and temperature/humidity

*E. coli* strains were taken from the same samples as the chicken gut metagenome data and cultured and used as indicator species for AMR<sup>98</sup> for each chicken faeces sample (see Supplementary Information for details of the culture and AST methodology). Only 191 of the 223 samples were positive for an *E. coli* isolate. Of these 191 samples, a further 21 (from LN1) were discarded from this analysis as technical issues with the environmental sensors resulted in these samples not having the necessary temperature and humidity data needed for the ML pipeline. Therefore, 170 samples remained to be analysed by the ML pipeline.

The antibiotic susceptibility/resistance profiles of the *E. coli* strains were evaluated against a panel of 26 antibiotics (Supplementary Table 1) using broth microdilution and interpreted according to the criteria of the Clinical and Laboratory Standards Institute<sup>99</sup>. The overall data analysis pipeline, implemented in Python (v3.9.15)<sup>100</sup> and SciPy (v1.9.3)<sup>101</sup> consisted of three phases (Fig. 2):

- Phase I: pre-selection of metagenomic features. For each antibiotic, isolation of a first set of faecal metagenome features (that is, ARG counts and relative abundance of microbial species) showing correlation with the resistance/susceptibility profiles of *E. coli* based on a chi-squared test.
- Phase II: assessment of the feature predicting power through the development of ML-powered predictive functions. Development of ML-based predictive functions of resistance/susceptibility (one predictive function per antibiotic) that operate from the pre-selected features (see below for more details), supervised training with available samples and then inspection of the best-fit state of each predictive function to retrieve the predictive

influence of each feature, that is, the relative weight of the feature in driving the prediction result.

- Phase III: assessment of feature dependency on temperature/humidity through the development of ML-powered regressors. Development of ML-based regressors to identify correlations between the set of faecal metagenome features identified in phase II and temperature/humidity conditions.

The three phases are described in detail below.

#### Phase I

An initial set of features was considered for each of the 26 antibiotics and comprised all data on ARG count and microbial species abundance in the faecal metagenome. The following steps were applied to process and reduce such sets using the Python package Scikit-learn<sup>102</sup>:

1. Abundances were turned into relative abundances (0–1 range) using min–max normalization.
2. For each specific antibiotic, imbalances in sample size between resistance and susceptibility observations were compensated with synthetically generated data using the synthetic minority oversampling technique (SMOTE)<sup>103</sup>, adopting five-nearest neighbours as the default parameter.
3. Features (ARG count and relative abundance of species) with a variance equal to zero (that is, features that had the same value in all samples) were removed as redundant (incapable of acting as effective predictors).
4. Features that did not show strong association with the prediction result (resistance/susceptibility profile), according to a chi-squared test, were removed (all the features with a *P* value greater than 0.01 were removed). No multiple comparison correction was used as we were looking to assess each feature in its own right<sup>104</sup>.
5. The remaining set of features were subjected to visual inspection via a graph representation designed to create spatial clusters that highlight correlation. The analysis was performed using the NetworkX<sup>105</sup> library in python. In the resulting graph, nodes representing features (ARG count or relative abundance of species) are connected to nodes representing resistance/susceptibility to a specific antibiotic if the existence of a correlation had been demonstrated by the chi-squared test (see the previous step). The nodes were spatially arranged using the Kamada–Kawai path-length cost function<sup>106</sup>.

#### Phase II

Predictive functions based on multiple underlying ML technologies were developed and tested, each trained to predict resistance/susceptibility to a specific antibiotic, using the features pre-selected in Phase I as input of supervised learning. A predictive function was trained and validated for each of the 26 antibiotics tested. Upon successful training and validation, inspection of the best-fit state of each predictive function allowed retrieval of the quantitative influence of each feature (that is, relative weight) in relation to predicting resistance/susceptibility to each antibiotic.

The following ML technologies were tested for implementation of the predictive functions: logistic regression, linear support vector machine, radial basis function support vector machine, extra tree classifier, random forest, Adaboost and XGBoost, all implemented using the Python package Scikit-learn<sup>102</sup>. Nested cross-validation (NCV)<sup>107</sup> was used to assess the performance and select the optimal hyperparameters for each technology. NCV is an iterative procedure in which different configurations of the predictive function (that is, different hyperparameters driving the selected technology) are repeatedly tested for performance whilst reshuffling the training and testing sets. NCV consists of an outer loop dedicated to randomly reallocating

observations into new training and testing sets, and an inner loop where different configurations (sets of hyperparameters) for the predictive function are tested with the current training and testing set. In our analysis, we ran an NCV with a fivefold outer loop (five reshuffles of the training and testing sets) and a threefold inner loop (three reshuffles of the training set) for each different ML technology. Prediction performance was measured via the receiver operating characteristic area under the curve (ROC-AUC, referred to simply as AUC in the following), accuracy, sensitivity, specificity and precision, all computed at each iteration of the outer loop<sup>108</sup>. Thirty iterations of the NCV assessments were completed for each ML technology. The technologies were then compared by running an *F*-test on the mean quantitative results for each using the AUC metric. A minimum of 12 samples in the minority class were required for the classification for SMOTE and NCV. Nine antibiotics (ampicillin, ampicillin–sulbactam, cefazolin, ciprofloxacin, doxycycline, imipenem, levofloxacin, meropenem and tetracycline) lacked sufficient samples in one class to allow cross-validation and SMOTE, and so were not taken further. We compared the seven ML architectures to avoid bias in the analysis related to choosing a specific ML technology. Prediction performance was measured using 30 NCV iterations, with the final performance score defined as the mean of all runs. The Nemenyi test was used to verify which predictive function performed best out of the seven ML methods. The extra tree predictive functions ranked best according to all studied performance indicators apart from sensitivity (where all the predictive functions were considered statistically equivalent) and were finally selected to produce the correlation results. As the extra tree method had been selected to power the final predictive functions, Gini importance was used to extract the strongest predictors from the final, trained models.

### Phase III

The last phase of the analysis consisted of the development of regression models to identify correlations between the set of faecal metagenome features identified in Phase II (predictors) and temperature/humidity conditions. Only the predictors extracted from ML-powered models with AUC > 0.9 were considered.

A separate regression model was created to represent the relationship of each predictor (considered as the input/explanatory variable) with either temperature or humidity (considered as the dependent variable). The predictor was treated as continuous if related to either a relative microbial abundance or ARG count. Temperature and humidity values were collected at each farm and averaged from the 7 days before the two time points  $t_1$  and  $t_2$ .

Each regression model was developed using linear least-squares fitting (using the Python package SciPy<sup>101</sup>) using the coefficient of determination ( $r^2$ ) to assess the goodness of fit. Metagenome features were considered to be significantly correlated with temperature or humidity if the slope of the regression line statistically differed from zero ( $P < 0.05$  using the Wald test with *t*-distribution of the test statistic). We looked for correlations between the ARG read depth and species read depth, which would indicate the likelihood of ARGs originating from a particular species, as proposed by Tong et al.<sup>33</sup>. An undirected graph was created using NetworkX (v2.8.4)<sup>105</sup> to visualize the interconnected ARGs and species selected by the regression framework for humidity and temperature.

### Analysis of antibiotic use bias

The correlations observed between the metagenomic data in chicken faeces and the resistance profiles observed in *E. coli* may be influenced by the different antibiotic protocols that each farm adopted (Supplementary Table 13). To identify whether the differences in antibiotic treatment in each farm led to bias in the selected metagenomic features, we calculated the relative abundance of ARGs expressed by first grouping ARGs by relationship to each specific antibiotic, and then by computing ratios of ARGs present in the sample, divided by the total

number of ARGs for each antibiotic, and then calculated the relative abundance of the microbial species. For these three cases, we used the Wilcoxon rank sum test to verify whether there was a difference between the samples from farms that received an antibiotic against the samples that did not receive that antibiotic.

### Statistical analysis

For details of all statistical analyses, see the Supplementary Information.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The metagenomic sequencing data supporting the conclusions of this article are available in the NCBI database under Bioproject accession numbers [PRJNA678871](https://doi.org/10.1038/s43016-023-00814-w) (for Shandong 1\_1 and 1\_2) and [PRJNA841806](https://doi.org/10.1038/s43016-023-00814-w) (for all other farms). In addition, the reference genome used for filtering host DNA is available in the NCBI database under accession code GCF\_000002315.6. All source data needed to recreate the figures are provided in the Supplementary Data 1.

### Code availability

The code is available via Github at [https://github.com/tan0101/Commercial\\_MGS2023](https://github.com/tan0101/Commercial_MGS2023). The code was used for the machine learning classification, regression analysis and network analysis.

### References

1. Wu, Z. *Antibiotic Use and Antibiotic Resistance in Food-Producing Animals in China* OECD Food, Agriculture and Fisheries Paper No. 134 (OECD, 2019); <https://doi.org/10.1787/4adba8c1-en>
2. Looft, T. et al. In-feed antibiotic effects on the swine intestinal microbiome. *Proc. Natl Acad. Sci. USA* **109**, 1691–1696 (2012).
3. Vega, N. M., Allison, K. R., Samuels, A. N., Klempner, M. S. & Collins, J. J. *Salmonella typhimurium* intercepts *Escherichia coli* signaling to enhance antibiotic tolerance. *Proc. Natl Acad. Sci. USA* **110**, 14420–14425 (2013).
4. Sommer, F., Anderson, J. M., Bharti, R., Raes, J. & Rosenstiel, P. The resilience of the intestinal microbiota influences health and disease. *Nat. Rev. Microbiol.* **15**, 630–638 (2017).
5. Pan, D. & Yu, Z. Intestinal microbiome of poultry and its interaction with host and diet. *Gut Microbes* **5**, 108–119 (2014).
6. Baron, S. A., Diene, S. M. & Rolain, J.-M. Human microbiomes and antibiotic resistance. *Hum. Microb. J.* **10**, 43–52 (2018).
7. Gautam, R. et al. Modeling the effect of seasonal variation in ambient temperature on the transmission dynamics of a pathogen with a free-living stage: example of *Escherichia coli* O157:H7 in a dairy herd. *Prev. Vet. Med.* **102**, 10–21 (2011).
8. Oakley, B. B. et al. The cecal microbiome of commercial broiler chickens varies significantly by season. *Poult. Sci.* **97**, 3635–3644 (2018).
9. Wang, X. et al. Effects of high ambient temperature on the community structure and composition of ileal microbiome of broilers. *Poult. Sci.* **97**, 2153–2158 (2018).
10. Sohsuebgarm, D., Kongpechr, S. & Sukon, P. Microclimate, body weight uniformity, body temperature, and footpad dermatitis in broiler chickens reared in commercial poultry houses in hot and humid tropical climates. *World Vet. J.* **9**, 241–248 (2019).
11. Thornton, P. K., van de Steeg, J., Notenbaert, A. & Herrero, M. The impacts of climate change on livestock and livestock systems in developing countries: a review of what we know and what we need to know. *Agric. Syst.* **101**, 113–127 (2009).
12. Ko, K. K. K., Chng, K. R. & Nagarajan, N. Metagenomics-enabled microbial surveillance. *Nat. Microbiol.* **7**, 486–496 (2022).

13. Astill, J., Dara, R. A., Fraser, E. D. G. & Sharif, S. Detecting and predicting emerging disease in poultry with the implementation of new technologies and big data: a focus on avian influenza virus. *Front. Vet. Sci.* <https://doi.org/10.3389/fvets.2018.00263> (2018).
14. Ahmed, G. et al. An approach towards IoT-based predictive service for early detection of diseases in poultry chickens. *Sustainability* **13**, 13396 (2021).
15. Her, H.-L. & Wu, Y.-W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. *Bioinformatics* **34**, i89–i95 (2018).
16. Hyun, J. C., Kavvas, E. S., Monk, J. M. & Palsson, B. O. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput. Biol.* **16**, e1007608 (2020).
17. Pearcy, N. et al. Genome-scale metabolic models and machine learning reveal genetic determinants of antibiotic resistance in *Escherichia coli* and unravel the underlying metabolic adaptation mechanisms. *mSystems* **6**, e00913–e00920 (2021).
18. Peng, Z. et al. Whole-genome sequencing and gene sharing network analysis powered by machine learning identifies antibiotic resistance sharing between animals, humans and environment in livestock farming. *PLoS Comput. Biol.* **18**, e1010018 (2022).
19. Kavvas, E. S. et al. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, 4306 (2018).
20. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D. & Palsson, B. O. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.* **11**, 2580 (2020).
21. Liu, Z. et al. Evaluation of machine Learning models for predicting antimicrobial resistance of *Actinobacillus pleuropneumoniae* from whole genome sequences. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2020.00048> (2020).
22. ValizadehAslani, T., Zhao, Z., Sokhansanj, B. A. & Rosen, G. L. Amino acid *k*-mer feature extraction for quantitative antimicrobial resistance (AMR) prediction by machine learning and model interpretation for biological insights. *Biology* **9**, 365 (2020).
23. Wang, W. et al. Novel SCCmec type XV (7A) and two pseudo-SCCmec variants in foodborne MRSA in China. *J. Antimicrob. Chemother.* **77**, 903–909 (2022).
24. Wang, W. et al. Whole-genome sequencing and machine learning analysis of *Staphylococcus aureus* from multiple heterogeneous sources in China reveals common genetic traits of antimicrobial resistance. *mSystems* **6**, e01185–01120 (2021).
25. Hendriksen, R. S. et al. Using genomics to track global antimicrobial resistance. *Public Health Front.* <https://doi.org/10.3389/fpubh.2019.00242> (2019).
26. Maciel-Guerra, A. et al. Dissecting microbial communities and resistomes for interconnected humans, soil, and livestock. *ISME J.* **17**, 21–35 (2022).
27. Okeke, I. N. et al. Leapfrogging laboratories: the promise and pitfalls of high-tech solutions for antimicrobial resistance surveillance in low-income settings. *BMJ Glob. Health* **5**, e003622 (2020).
28. Iskandar, K. et al. Surveillance of antimicrobial resistance in low- and middle-income countries: a scattered picture. *Antimicrob. Resist. Infect. Control* **10**, 63 (2021).
29. Ikhimiukor, O. O., Odih, E. E., Donado-Godoy, P. & Okeke, I. N. A bottom-up view of antimicrobial resistance transmission in developing countries. *Nat. Microbiol.* **7**, 757–765 (2022).
30. Zhang, A.-N. et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat. Commun.* **12**, 4765 (2021).
31. Tang, B. et al. Characterization of an NDM-5 carbapenemase-producing *Escherichia coli* ST156 isolate from a poultry farm in Zhejiang, China. *BMC Microbiol.* **19**, 82 (2019).
32. Cui, M. et al. Prevalence and characterization of fluoroquinolone resistant *Salmonella* isolated from an integrated broiler chicken supply chain. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2019.01865> (2019).
33. Tong, C. et al. Swine manure facilitates the spread of antibiotic resistome including tigeicycline-resistant tet(X) variants to farm workers and receiving environment. *Sci. Total Environ.* **808**, 152157 (2022).
34. Wang, Y. et al. A novel gene, *optrA*, that confers transferable resistance to oxazolidinones and phenicols and its presence in *Enterococcus faecalis* and *Enterococcus faecium* of human and animal origin. *J. Antimicrob. Chemother.* **70**, 2182–2190 (2015).
35. Aradas, M., Poljak, Z., Fittipaldi, N., Ricker, N. & Farzan, A. Serotypes, virulence-associated factors, and antimicrobial resistance of *Streptococcus suis* isolates recovered from sick and healthy pigs determined by whole-genome sequencing. *Front. Vet. Sci.* **8**, 742345 (2021).
36. Hansen, L. H., Sørensen, S. J., Jørgensen, H. S. & Jensen, L. B. The prevalence of the OqxAB multidrug efflux pump amongst olaquinox-resistant *Escherichia coli* in pigs. *Microb. Drug Resist.* **11**, 378–382 (2005).
37. Dortet, L., Nordmann, P. & Poirel, L. Association of the emerging carbapenemase NDM-1 with a bleomycin resistance protein in *Enterobacteriaceae* and *Acinetobacter baumannii*. *Antimicrob. Agents Chemother.* **56**, 1693–1697 (2012).
38. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
39. Laird, T. J. et al. Diversity detected in commensals at host and farm level reveals implications for national antimicrobial resistance surveillance programmes. *J. Antimicrob. Chemother.* **77**, 400–408 (2022).
40. Zhou, W. et al. Antimicrobial resistance and genomic characterization of *Escherichia coli* from pigs and chickens in Zhejiang, China. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2022.1018682> (2022).
41. He, D. et al. CTX-M-123, a novel hybrid of the CTX-M-1 and CTX-M-9 group  $\beta$ -lactamases recovered from *Escherichia coli* isolates in China. *Antimicrob. Agents Chemother.* **57**, 4068–4071 (2013).
42. Wang, Y. et al. Antibiotic resistance gene reservoir in live poultry markets. *J. Infect.* **78**, 445–453 (2019).
43. Sciortino, S. et al. Occurrence and antimicrobial resistance of *Arcobacter* spp. recovered from aquatic environments. *Antibiotics* **10**, 288 (2021).
44. Jochum, J. M., Redweik, G. A. J., Ott, L. C. & Mellata, M. Bacteria broadly-resistant to last resort antibiotics detected in commercial chicken farms. *Microorganisms* <https://doi.org/10.3390/microorganisms9010141> (2021).
45. Błażejewska, A., Zalewska, M., Grudniak, A. & Popowska, M. A comprehensive study of the microbiome, resistome, and physical and chemical characteristics of chicken waste from intensive farms. *Biomolecules* <https://doi.org/10.3390/biom12081132> (2022).
46. de Mesquita Souza Saraiva, M. et al. Antimicrobial resistance in the globalized food chain: a One Health perspective applied to the poultry industry. *Braz. J. Microbiol.* **53**, 465–486 (2022).
47. Surveillance and One Health in food production key to halting antimicrobial resistance. *World Health Organisation* (7 June 2021); <https://www.who.int/europe/news/item/07-06-2021-surveillance-and-one-health-in-food-production-key-to-halting-antimicrobial-resistance>

48. Davies, N., Jørgensen, F., Willis, C., McLauchlin, J. & Chattaway, M. A. Whole genome sequencing reveals antimicrobial resistance determinants (AMR genes) of *Salmonella enterica* recovered from raw chicken and ready-to-eat leaves imported into England between 2014 and 2019. *J. Appl. Microbiol.* **133**, 2569–2582 (2022).
49. Conesa, A., Garofolo, G., Di Pasquale, A. & Cammà, C. Monitoring AMR in *Campylobacter jejuni* from Italy in the last 10 years (2011–2021): microbiological and WGS data risk assessment. *EFSA J.* **20**, e200406 (2022).
50. Rohr, J. R. et al. Emerging human infectious diseases and the links to global food production. *Nat. Sustain.* **2**, 445–456 (2019).
51. Xiong, W. et al. Antibiotic-mediated changes in the fecal microbiome of broiler chickens define the incidence of antibiotic resistance genes. *Microbiome* **6**, 34 (2018).
52. Zhou, Y. et al. Antibiotic administration routes and oral exposure to antibiotic resistant bacteria as key drivers for gut microbiota disruption and resistome in poultry. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2020.01319> (2020).
53. Noyes, N. R. et al. Resistome diversity in cattle and the environment decreases during beef production. *Elife* **5**, e13195 (2016).
54. Zhang, C. Z. et al. The emergence of chromosomally located *bla*<sub>CTX-M-55</sub> in *Salmonella* from foodborne animals in China. *Front. Microbiol.* **10**, 1268 (2019).
55. Storey, N. et al. Use of genomics to explore AMR persistence in an outdoor pig farm with low antimicrobial usage. *Microb. Genom.* <https://doi.org/10.1099/mgen.0.000782> (2022).
56. Thu, W. P. et al. Prevalence, antimicrobial resistance, virulence gene, and class 1 integrons of *Enterococcus faecium* and *Enterococcus faecalis* from pigs, pork and humans in Thai–Laos border provinces. *J. Glob. Antimicrob. Resist.* **18**, 130–138 (2019).
57. Yang, Y., Liu, G., Ye, C. & Liu, W. Bacterial community and climate change implication affected the diversity and abundance of antibiotic resistance genes in wetlands on the Qinghai–Tibetan Plateau. *J. Hazard. Mater.* **361**, 283–293 (2019).
58. Slavik, M. F. et al. Effect of humidity on infection of turkeys with *Alcaligenes faecalis*. *Avian Dis.* **25**, 936–942 (1981).
59. Filipe, M. et al. Fluoroquinolone-resistant *Alcaligenes faecalis* related to chronic suppurative otitis media, Angola. *Emerg. Infect. Dis.* **23**, 1740–1742 (2017).
60. Huang, C. Extensively drug-resistant *Alcaligenes faecalis* infection. *BMC Infect. Dis.* **20**, 833 (2020).
61. Alcock, B. P. et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
62. Barnes, N. M. & Wu, H. Mechanisms regulating the airborne survival of *Klebsiella pneumoniae* under different relative humidity and temperature levels. *Indoor Air* **32**, e12991 (2022).
63. Zheng, W., Yue, M., Zhang, J. & Ruan, Z. Coexistence of two *bla*<sub>CTX-M-14</sub> genes in a *bla*<sub>NDM-5</sub>-carrying multidrug-resistant *Escherichia coli* strain recovered from a bloodstream infection in China. *J. Glob. Antimicrob. Resist.* **26**, 11–14 (2021).
64. Hernández, M. et al. First report of an extensively drug-resistant ST23 *Klebsiella pneumoniae* of capsular serotype K1 co-producing CTX-M-15, OXA-48 and ArmA in Spain. *Antibiotics* <https://doi.org/10.3390/antibiotics10020157> (2021).
65. Barraud, O., Badell, E., Denis, F., Guiso, N. & Ploy, M. C. Antimicrobial drug resistance in *Corynebacterium diphtheriae mitis*. *Emerg. Infect. Dis.* **17**, 2078–2080 (2011).
66. Song, L. et al. Bioaerosol is an important transmission route of antibiotic resistance genes in pig farms. *Environ. Int.* **154**, 106559 (2021).
67. Aarestrup, F. M. et al. Resistance to antimicrobial agents used for animal therapy in pathogenic-, zoonotic- and indicator bacteria isolated from different food animals in Denmark: a baseline study for the Danish Integrated Antimicrobial Resistance Monitoring Programme (DANMAP). *APMIS* **106**, 745–770 (1998).
68. Sun, J. et al. Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. *Nat. Commun.* **11**, 1427 (2020).
69. Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F. & Reimer, A. Metagenomics: the next culture-independent game changer. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2017.01069> (2017).
70. Yadav, S. & Kapley, A. Antibiotic resistance: global health crisis and metagenomics. *Biotechnol. Rep.* **29**, e00604 (2021).
71. Yang, X. et al. Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain. *Appl. Environ. Microbiol.* **82**, 2433–2443 (2016).
72. Duarte, A. S. R. et al. Addressing learning needs on the use of metagenomics in antimicrobial resistance surveillance. *Public Health Front.* <https://doi.org/10.3389/fpubh.2020.00038> (2020).
73. Pillay, S., Calderón-Franco, D., Urhan, A. & Abeel, T. Metagenomic-based surveillance systems for antibiotic resistance in non-clinical settings. *Front. Microbiol.* **13**, 1066995 (2022).
74. Li, N., Ren, Z., Li, D. & Zeng, L. Review: automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming. *Animal* **14**, 617–625 (2020).
75. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
76. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
77. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
78. Glendinning, L., Stewart, R. D., Pallen, M. J., Watson, K. A. & Watson, M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. *Genome Biol.* **21**, 34 (2020).
79. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
81. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
82. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
83. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
84. Arbizu, P. M. pairwiseAdonis: pairwise multilevel comparison using adonis. R version 0.4 <https://github.com/pmartinezarbizu/pairwiseAdonis> (2020).
85. Cameron, E. S., Schmidt, P. J., Tremblay, B. J. M., Emelko, M. B. & Müller, K. M. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Sci. Rep.* **11**, 22302 (2021).
86. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2017.02224> (2017).
87. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

88. Schmidt, K. et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* **72**, 104–114 (2016).
89. Che, Y. et al. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc. Natl Acad. Sci. USA* **118**, e2008731118 (2021).
90. Ellabaan, M. M. H., Munck, C., Porse, A., Imamovic, L. & Sommer, M. O. A. Forecasting the dissemination of antibiotic resistance genes across bacterial genomes. *Nat. Commun.* **12**, 2435 (2021).
91. Hua, X. et al. BacAnt: a combination annotation server for bacterial DNA sequences to identify antibiotic resistance genes, integrons, and transposable elements. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2021.649969> (2021).
92. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
93. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
94. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
95. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
96. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
97. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, 293–296 (2021).
98. Anjum, M. F. et al. The potential of using *E. coli* as an indicator for the surveillance of antimicrobial resistance (AMR) in the environment. *Curr. Opin. Microbiol.* **64**, 152–158 (2021).
99. CLSI. *Performance Standards for Antimicrobial Susceptibility Testing*, 31st ed. (Clinical Laboratory Standards Institute, 2021).
100. Python v3.9.15 (Python Software Foundation, 2023); <https://docs.python.org/3/index.html>
101. Jones, E., Oliphant, T. & Peterson, P. SciPy: open source scientific tools for Python. (2001).
102. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
103. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
104. Perneger, T. V. What's wrong with Bonferroni adjustments. *Br. Med. J.* **316**, 1236–1238 (1998).
105. Hagberg, A., Swart, P. & Chult, D. S. Exploring Network Structure, Dynamics, and Function Using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)* (eds Varoquaux, G., Vaught, T. & Millman, J.) (Los Alamos National Laboratory, 2008).
106. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31**, 7–15 (1989).
107. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
108. Wainer, J. & Cawley, G. Empirical evaluation of resampling procedures for optimising SVM hyperparameters. *J. Mach. Learn. Res.* **18**, 1–35 (2017).

## Acknowledgements

This work was supported by the InnovateUK (grant no. 104986), FARMWATCH: Fight AbR with Machine learning and a Wide Array of sensing TeChnologies (T.D. and D.R.) and the Ministry of Science and Technology of the People's Republic of China through the Grant Key Project of International Scientific and Technological Innovation Cooperation Between Governments (number 2018YFE0101500, Z.P.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors gratefully acknowledge the support received from the University of Nottingham Research Beacon of Excellence: Future Food and the Innovate UK monitoring officer L. Viatge for support.

## Author contributions

J.C., F.L., Z.P., X.Z., L.L. and T.D. designed and supervised the study; Z.P., X.Z., L.L., F.L., J.C. and T.D. planned the methodology; M.B., A.M.G., N.S. and T.D. wrote the draft; Z.P., J.C., F.L., M.B., A.M.G., N.S. and T.D. edited and reviewed the draft and provided critical comments; Z.P., W.W., Y.D., Yujie Hu, H.L., Z.T., M.Z., Y.G., L.Z., Z.H. and X.Z. carried out the experiments and collected the animal and environmental samples; A.M.G., M.B. and Yue Hu performed the data analysis and visualization of the analysed data with critical comments from N.S. and T.D.; Z.P., D.R. and T.D. acquired the funding. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43016-023-00814-w>.

**Correspondence and requests for materials** should be addressed to Zixin Peng, Fengqin Li or Tania Dottorini.

**Peer review information** *Nature Food* thanks the anonymous reviewers for their contribution to the peer review of this work.

**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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

<sup>1</sup>School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, UK. <sup>2</sup>Shandong New Hope Liuhe Group Co. Ltd and Qingdao Key Laboratory of Animal Feed Safety, Qingdao, People's Republic of China. <sup>3</sup>NHC Key Laboratory of Food Safety Risk Assessment, China National Center for Food Safety Risk Assessment, Beijing, People's Republic of China. <sup>4</sup>Nimrod Veterinary Products Ltd., Moreton-in-Marsh, UK. <sup>5</sup>Shandong Kaijia Food Co., Weifang, People's Republic of China. <sup>6</sup>Luoyang Center for Disease Control and Prevention, Luoyang City, People's Republic of China. <sup>7</sup>Liaoning Provincial Center for Disease Control and Prevention, Shenyang City, People's Republic of China. <sup>8</sup>Agricultural Biopharmaceutical Laboratory, College of Chemistry and Pharmaceutical Sciences, Qingdao Agricultural University, Qingdao City, People's Republic of China. <sup>9</sup>Chinese Veterinary Medicine Innovation Center, College of Veterinary Medicine, China Agricultural University, Beijing City, People's Republic of China. <sup>10</sup>Department of Engineering, University of Perugia, Perugia, Italy. <sup>11</sup>Centre for Smart Food Research, Nottingham Ningbo China Beacons of Excellence Research and Innovation Institute, University of Nottingham Ningbo China, Ningbo, People's Republic of China. <sup>12</sup>These authors contributed equally: Michelle Baker, Xibin Zhang, Alexandre Maciel-Guerra. <sup>13</sup>These authors jointly supervised this work: Zixin Peng, Fengqin Li, Tania Dottorini. ✉e-mail: [pengzixin@cfsa.net.cn](mailto:pengzixin@cfsa.net.cn); [lifengqin@cfsa.net.cn](mailto:lifengqin@cfsa.net.cn); [tania.dottorini@nottingham.ac.uk](mailto:tania.dottorini@nottingham.ac.uk)

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

## Data analysis

Our code is available on Github: [https://github.com/tan0101/Commercial\\_MGS2023](https://github.com/tan0101/Commercial_MGS2023)  
 For the bioinformatics analysis the following software were used:  
 Readfq (v8, <https://github.com/cjfields/readfq>)  
 Bowtie2 v2.3.4.1  
 SAMtools v1.9  
 MEGAHIT software v1.1.2  
 BWA MEM v2-2.1  
 METABAT2v2.15  
 MetaPhlan v3.0  
 Rv3.6.2  
 To perform the ML and data analysis analysis the following software were used:  
 IQ-tree2v2.0.6  
 BEASTv1.10.4  
 Tracer v1.7.1  
 python (v3.9.15)  
 scikit-learn(v1.0.2),  
 scipy (v1.9.3)  
 networkx (v2.8.4)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The metagenomic sequencing data supporting the conclusions of this article are available in the NCBI database under Bioproject accession numbers PRJNA678871 (for Shandong 1\_1 and 1\_2) and PRJNA841806 (for all other farms) available on: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA678871> and <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA841806>. In addition the reference genome used for filtering host DNA is available in NBCI database under accession GCF\_000002315.6 [https://www.ncbi.nlm.nih.gov/datasets/genome/GCF\\_000002315.6/](https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002315.6/).

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

For the analysis of how microbial communities and resistomes are differentiated across farm sources and between farm and abattoir, we did not perform sample size calculation as this was an observational/exploratory study.  
 For the results demonstrating the proposed approach to the AMR surveillance (based on ML prediction of resistant phenotypes), sample size



is based on achieving desired power in the predictor. For binary classifiers, power is the sensitivity (true positive rate, defined as  $1 - \beta$ , where  $\beta$  is the false negative rate, i.e. type II error (Banerjee, A. et al. 2009, Industrial psychiatry journal). Note that the type II error is particularly relevant for resistance, as it implies a resistant phenotype escaping detection (Mahfouz, N. et al. 2020, Journal of Antimicrobial Chemotherapy). Using 191 samples, we achieved an average power of 92% for the 11 antibiotic models studied. We also wanted to identify the minimum number of samples required to achieve at least 80% sensitivity (power). Because for classifiers based on ML (e.g. SVMs, decision trees, random forest, adaboost, neural networks), sample size calculation to achieve power is not directly possible using conventional analytical methods (Li. J. et al. 2020, Patterns), we applied a bespoke iterative method (wrapper backward selection - WBS, Figueroa, R et al. 2012, BMC Medical Informatics and Decision Making) as done in our previous paper (Maciel-Guerra, A. et al. 2022, The ISME Journal). The method estimates how power decreases with smaller sample sizes. In our case, WBS estimated the need of 160 samples on average, to achieve 80% power, which is less to what we used (191).

Data exclusions	No data were excluded.
Replication	At least three biological replicates per sample were taken, all were successful.
Randomization	Biological samples were collected randomly without knowing the AMR phenotypes. For the analysis of how microbial communities and resistomes are differentiated across farm sources and between farm and abattoir, random assignment to groups was not performed as this is an exploratory/observational study. For the ML classification the samples were randomly assigned to training and testing groups using a nested cross validation procedure (30 iterations per classifier).
Blinding	Biological samples were collected randomly without knowing the AMR phenotypes.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging