RESEARCH ARTICLE SUMMARY

DNA DAMAGE

The specificity and structure of DNA cross-linking by the gut bacterial genotoxin colibactin

Erik S. Carlson†, Raphael Haslecker†, Chiara Lecchi, Miguel A. Aguilar Ramos, Vyshnavi Vennelakanti, Linda Honaker, Alessia Stornetta, Estela S. Millán, Bruce A. Johnson, Heather J. Kulik, Silvia Balbo*, Peter W. Villalta*, Victoria M. D'Souza*, Emily P. Balskus*

Full article and list of author affiliations: https://doi.org/10.1126/science.ady3571

INTRODUCTION: Colibactin is a chemically unstable genotoxic gut bacterial natural product that is linked to colorectal cancer (CRC). Though it has eluded isolation and structural characterization, colibactin is proposed to contain two cyclopropane "warheads" capable of forming DNA interstrand cross-links (ICLs) connected by a reactive central scaffold of unresolved structure. The discovery of distinctive mutational signatures arising from colibactin exposure and their detection in cancer genomes suggest that colibactin influences CRC. However, we lack direct information regarding the specificity and structure of the colibactin-DNA ICL, limiting our understanding of how this natural product targets DNA and the origins of mutations arising from this DNA damage.

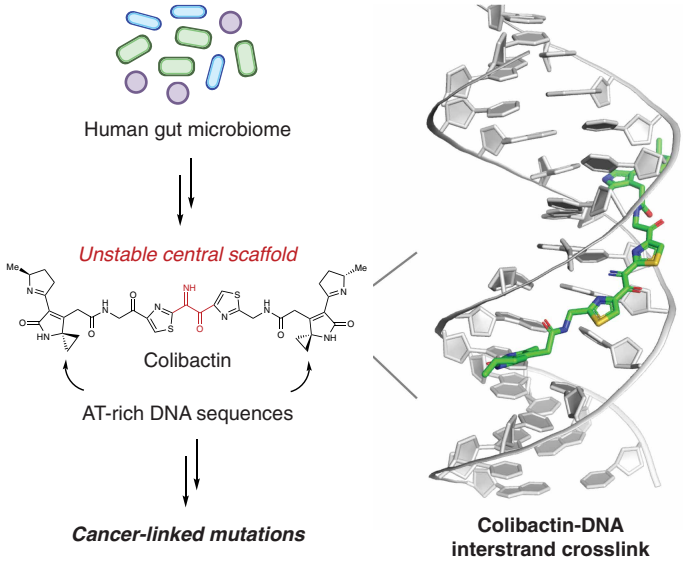
RATIONALE: Though prior studies had revealed colibactin's DNA alkylating activity and implicated adenine (A)– and thymine (T)–rich sequences as likely sites for ICL formation, the precise nature of colibactin's interactions with DNA, the exact sites of alkylation, and its sequence specificity were unknown. To address these gaps in knowledge, we sought to experimentally elucidate the specificity and structure of the colibactin-DNA ICL using biochemical assays, advanced mass spectrometry (MS), and nuclear magnetic resonance (NMR) spectroscopy approaches.

RESULTS: We first investigated the reactivity of colibactin toward DNA oligonucleotides in vitro using a newly developed MS-based assay, overcoming the challenge of its chemical instability by leveraging in situ bacterial production. We observed ICL formation of bis-N3-adenine ICLs within a preferred motif of 5′-W**A**WWTW-3′ (where the adenines bolded and opposite the underlined thymine are alkylated, and W represents either A or T). This preference for AT-rich sequences is consistent with the locations of colibactin-derived mutational signatures. Additional experiments suggested that colibactin binds and alkylates in the minor groove.

To gain initial insights into the structure of the colibactin-DNA ICL, we further applied MS to characterize the intact lesion. Unexpectedly, we observed a mass consistent with ICL formation arising from a proposed colibactin structure containing a chemically unstable central α-ketoimine.

To obtain more detailed structural information, we produced the colibactin-DNA ICL on a large enough scale to enable solution-state NMR studies. The structure that we obtained verifies the sites and locations of colibactin DNA alkylation identified in our in vitro assays. Analysis of the structure revealed chemical features of colibactin that are important for DNA binding and alkylation and explain its sequence specificity. Most notably, the positively charged central α-ketoiminium of colibactin makes extensive electrostatic and hydrogen bonding interactions with the floor of the minor groove. The results of calculations and experiments with a synthetic colibactin analog further support the importance of this unstable central functional group to the specificity of colibactin-DNA ICL formation.

CONCLUSION: Our study reveals the specificity and structure of the colibactin-DNA ICL by combining MS and NMR. Colibactin's preference for alkylating AT-rich sequences sheds light on the origins of mutational signatures. These results also help resolve the structure of colibactin's unstable central region and implicate it as a key determinant of sequence specificity. Together, our findings reveal a strategy for DNA alkylation distinctive among natural products, enhancing our understanding of colibactin's chemical structure, its recognition of and reaction with DNA, and its downstream effects on the host genome. □



Structure and specificity of the colibactin-DNA interstrand cross-link. Characterization of DNA damage by the human gut bacterial genotoxin colibactin reveals a preference for alkylation at AT-rich DNA sequences and the importance of an unstable central α-ketoimine in mediating this specificity. These results help to explain the locations of cancer-linked mutations derived from colibactin exposure.

*Corresponding author. Email: balskus@chemistry.harvard.edu (E.P.B.); dsouza@g.harvard.edu (V.M.D.); villa001@umn.edu (P.W.V.); balbo006@umn.edu (S.B.) †These authors contributed equally to this work. **Cite this article as E. S. Carlson *et al.*, *Science* 390, eady3571 (2025). DOI: 10.1126/science.ady3571**

DNA DAMAGE

The specificity and structure of DNA cross-linking by the gut bacterial genotoxin colibactin

Erik S. Carlson[1]†, Raphael Haslecker[2]†, Chiara Lecchi[3], Miguel A. Aguilar Ramos[1], Vyshnavi Vennelakanti[4,5], Linda Honaker[2], Alessia Stornetta[3], Estela S. Millán[1]‡, Bruce A. Johnson[6], Heather J. Kulik[4,5], Silvia Balbo[3]*, Peter W. Villalta[3,7]*, Victoria M. D'Souza[2]*, Emily P. Balskus[1,8,9]*

Accumulating evidence has connected the chemically unstable, DNA-damaging gut bacterial natural product colibactin to colorectal cancer, including the identification of mutational signatures that are thought to arise from colibactin-DNA interstrand cross-links (ICLs). However, we currently lack direct information regarding the structure of this lesion. In this work, we combined mass spectrometry and nuclear magnetic resonance spectroscopy to elucidate the specificity and structure of the colibactin-DNA ICL. We found that colibactin alkylates within the minor groove of adenine- and thymine-rich DNA, explaining the origins of mutational signatures. Unexpectedly, we discovered that the chemically unstable central motif of colibactin mediates the sequence specificity of cross-linking. By directly elucidating colibactin's interactions with DNA, this work enhances our understanding of the structure and genotoxic mechanisms of this cancer-linked gut bacterial natural product.

The human gut microbiome has been increasingly linked to the development of colorectal cancer (CRC) (*1*, *2*). Particularly prominent potential contributors to this disease are gut bacteria that produce colibactin (*3*, *4*). Colibactin is a complex, chemically unstable genotoxic natural product (Fig. 1A and fig. S1A) produced by commensal Enterobacteriaceae, including strains of *Escherichia coli*, that harbor the *pks* (or *clb*) gene cluster (*5*). This gene cluster encodes a biosynthetic pathway that uses a nonribosomal peptide synthetase–polyketide synthase assembly line. Exposure of human cells to *pks*$^+$ bacteria results in DNA damage, including double-strand breaks (DSBs) (*5*). This gives rise to various phenotypes in vitro and in vivo, including genomic instability, megalocytosis, G2/M cell cycle arrest, cellular senescence, and increased tumor formation in mouse models of CRC (*5*–*11*). Notably, *pks*$^+$ *E. coli* are detected more frequently in CRC patients (*8*, *9*, *11*–*14*), fueling the hypothesis that colibactin exposure may play a role in the initiation and/or progression of cancer.

Understanding the molecular basis of colibactin's genotoxic activity has been particularly challenging because this natural product has been recalcitrant to traditional isolation and structure elucidation approaches. Studying biosynthetic enzymes and identifying shunt products from *pks* mutant strains revealed structural information, including the unexpected incorporation of cyclopropane rings into colibactin, leading to the proposal that it directly alkylates DNA (*15*–*17*). Subsequent discovery of colibactin-derived DNA adducts and the observation that *pks*$^+$ *E. coli* generate DNA interstrand cross-links (ICLs) in vitro and in cell lines further supported this hypothesis (*18*–*20*). Additional biochemical studies, chemical synthesis, and isolation attempts ultimately led to the proposal that colibactin is a pseudosymmetric molecule containing two electrophilic cyclopropane "warheads" capable of alkylating DNA at adenine (Fig. 1A and fig. S1A) (*21*, *22*). These warheads are connected by a central scaffold of unresolved structure that is predicted biosynthetically to be an α-aminoketone (or its enolamine tautomer); however, this motif likely undergoes rapid oxidation to the corresponding α-ketoimine, followed by hydrolysis to a 1,2-diketone that is susceptible to oxidative C–C bond cleavage (fig. S1A) (*23*). This highly unstable central structural motif has been replaced by two methylene groups in a "stable" synthetic colibactin analog (fig. S1B) (*24*). However, this analog is a minor component of an inseparable product mixture, with the major components being two β-hydroxy lactam ring diastereomers of unknown biological significance. Though the proposed structure(s) of colibactin account for the activities of all essential biosynthetic enzymes and explain ICL formation (Fig. 1A and fig. S1A), notable gaps in our understanding of its structure and activity remain, including the identity and function of its central scaffold.

The molecular details underlying colibactin's interaction with DNA are also unclear. The discovery of mutational signatures arising from exposure of human cells and organoids to *pks*$^+$ *E. coli* has provided intriguing indirect insights (*25*, *26*). These mutational signatures are primarily thymine (T)–to–cytosine (C) single-base substitutions and indels that occur within adenine (A)– and T-rich sequence motifs (i.e., 5′-AAWWTT-3′, where W represents A or T), and their transcriptional strand bias is consistent with ICL formation between two adenines. Colibactin-induced DSBs occur within identical AT-rich sequences and are suggested to arise from degradation of ICLs (*26*). Additionally, molecular modeling suggests that the colibactin-DNA ICL spans 4 to 5.5 Å [or 3 to 4 base pairs (bp)] but could not elucidate the specific adenines alkylated or the molecular details of DNA binding (*26*). Colibactin mutational signatures have been detected in many cancer genomes, including 5 to 20% of CRC genomes; occur in driver genes, such as *APC*; and are correlated with early-onset CRC (*25*–*33*). We also recently identified colibactin-DNA adducts in human colonoscopy samples (*34*). Together, these data indicate that humans are exposed to colibactin, strengthening its connection to cancer development. They also provide indirect evidence that colibactin cross-links DNA in a highly specific manner. However, we currently lack direct information regarding the specificity and structure of the colibactin-DNA ICL, limiting our understanding of how this natural product targets DNA and the origins of mutational signatures arising from this DNA damage.

Results

To study colibactin-DNA ICL formation in detail, we initially identified a short, double-stranded oligodeoxynucleotide (dsODN) that was cross-linked upon exposure to *pks*$^+$ *E. coli*. We generated colibactin in situ from bacterial metabolism because of the prominent gaps in our understanding of this metabolite and the concern that synthetic analogs may differ in activity and/or selectivity. Examining colibactin's reactivity toward plasmid fragments identified a 500-bp region of pET28(a) that was cross-linked efficiently. Systematically truncating this 500-nucleotide oligomer identified a 50-nucleotide oligomer that is cross-linked (71% ICL formation) after a 5-hour incubation at 37°C with *pks*$^+$ *E. coli* in Dulbecco's minimum essential medium (DMEM)–HEPES medium (Fig. 1B-C). No ICL formation was observed when incubating the 50-nucleotide oligomer with an isogenic *pks*$^-$ *E. coli* strain. We noted that

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [3]Masonic Cancer Center, University of Minnesota, Minneapolis, MN, USA. [4]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Structural Biology Initiative, City University of New York (CUNY) Advanced Science Research Center, New York, NY, USA. [7]Department of Medicinal Chemistry, University of Minnesota, Minneapolis, MN, USA. [8]Howard Hughes Medical Institute, Harvard University, Cambridge, MA, USA. [9]Broad Institute of MIT and Harvard, Cambridge, MA, USA. ***Corresponding author. Email: balskus@chemistry.harvard.edu (E.P.B.); dsouza@g.harvard.edu (V.M.D.); villa001@umn.edu (P.W.V.); balbo006@umn.edu (S.B.)** †These authors contributed equally to this work. ‡Present address: Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA.

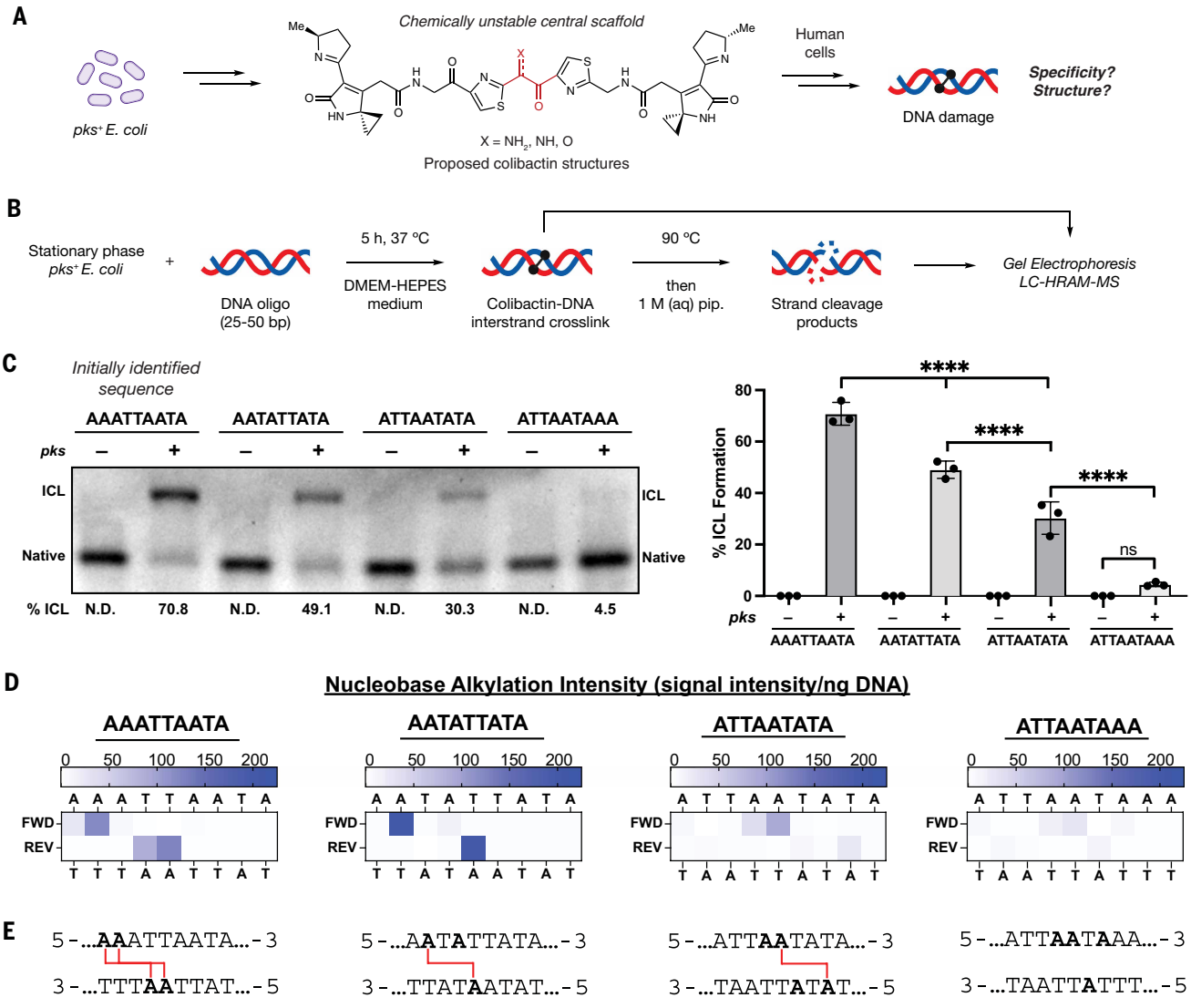


Fig. 1. The chemically unstable gut bacterial genotoxin colibactin forms DNA ICLs in a sequence-specific manner. (**A**) The chemical structure of colibactin and the structure and specificity of the colibactin-DNA ICL are unresolved. (**B**) Experimental workflow for characterizing ICL formation on 25- to 50-bp double-stranded oligonucleotides using colibactin produced in situ by *pks*$^{+}$ *E. coli*. (**C**) Extent of ICL formation within 50-nucleotide oligomers possessing the indicated sequence motifs. Cross-link formation was identified by denaturing gel electrophoresis and quantified by densitometry. Quantified values are displayed as a column plot. ns, not significant; ****$P < 0.0001$; one-way analysis of variance and Tukey's multiple comparison test. (**D**) Residue-specific alkylation of 25-nucleotide oligomers containing the indicated sequence motifs. Alkylation intensities were determined through LC-HRAM-MS analysis and normalized to total DNA injected. The intensity reported is the difference between the average detected in assays with *pks*$^{+}$ *E. coli* and the average detected in assays with *pks*$^{-}$ *E. coli*. (**E**) Inferred ICL locations within tested sequence motifs. All alkylated residues are bolded, and those cross-linked are connected by red lines. All data are mean ± SD with $n = 3$ biological replicates. Full 50- and 25-bp sequences are available in the supplementary materials. N.D., not determined.

this 50-nucleotide oligomer contained an AT-rich motif (5′-AAATTAATA-3′) that is consistent with previously reported colibactin mutational signatures (*25*, *26*), leading us to hypothesize that colibactin directly forms ICLs within this region.

To determine the site(s) of alkylation within this sequence motif, we developed a strand cleavage assay utilizing liquid chromatography–high-resolution accurate mass–mass spectrometry (LC-HRAM-MS) for detection (Fig. 1B and fig. S2). Colibactin-ICLs are thermally unstable and spontaneously depurinate to form abasic sites when heated to 90°C (*35*, *36*). Subsequent base treatment induces strand cleavage at the colibactin-specific abasic sites, and the masses of the resulting strand cleavage products are determined by isotopically resolved mass deconvolution of the LC-HRAM-MS data to identify the location and measure the relative abundance of DNA alkylation (Fig. 1, D and E).

We applied this assay to a *pks*$^{+}$ *E. coli*–exposed 25-nucleotide oligomer containing the AT-rich motif identified from the 50-nucleotide oligomer. This shortened dsODN was chosen to increase the analytical capability of the LC-MS analysis. The results indicate site-specific alkylation of this dsODN by colibactin at 5′-**AA**ATTAATA-3′ (where alkylated nucleotides are bolded or opposite to underlined nucleotides) (Fig. 1, D and E, and figs. S2 and S3). The accuracy of our LC-MS approach was confirmed by subjecting a 5′-6-carboxyfluorescein–labeled 50-nucleotide oligomer to our cross-linking and strand cleavage protocol and determining alkylation site locations using traditional Maxam-Gilbert gel sequencing (figs. S4 to S6) (*37*).

This result suggests that the colibactin-DNA ICL spans 3 to 5 bp. To determine the precise length of the ICL, we repeated our cross-linking and strand cleavage assays with a 50- and 25-nucleotide oligomer

containing the sequence 5′-AATATTATA-3′, which has the potential to form ICLs between pairs of adenines 2 to 8 bp apart. The results of this assay indicated that colibactin forms a single ICL between two adenines spanning 4 bp (5′-A**A**TATTATA-3′). This is consistent with previous studies showing that colibactin-induced DSBs have 2-bp overhangs (*26*) and suggests that the four sites of alkylation within the 5′-AAATTAATA-3′ motif that we initially tested likely correspond to a mixture of two colibactin-DNA ICLs (Fig. 1, D and E, and fig. S4). The extent of ICL formation was reduced compared with that of the initial sequence, perhaps owing to this sequence possessing only one recognition site (Fig. 1, C to E).

We further explored the position of colibactin-DNA ICL formation by performing incubations with dsODNs containing A to T transversions at the previously observed sites of alkylation (5′-ATTAATATA-3′). If colibactin can alkylate at the same position but on the opposite strand, we would expect ICL formation in the same location (i.e., 5′-ATTA**A**TATA-3′); if not, we would expect the ICL to shift to a new location (5′-ATTA**A**TATA-3′). ICL formation and strand cleavage analysis revealed the ICL in the new location, suggesting that colibactin preferentially cross-links at a TA base pair three positions downstream from the initial A (Fig. 1, C to E). Lastly, we detected little to no alkylation of dsODNs containing 5′-ATTAATAAA-3′, which has no 5′-AWWT-3′ motifs (Fig. 1, C to E). Taken together, these data suggest that colibactin-DNA ICLs form at 5′-W**A**WWTW-3′ motifs.

We next introduced single GC base pairs into the 5′-AATATT-3′ motif to test colibactin's specificity for adenine alkylation (fig. S7A) and tolerance for changes to the surrounding sequence. It has been postulated that colibactin may not recognize GC-containing sequences due to their altered DNA groove widths (*26*). Single GC base pair–containing sequence variants were largely cross-linked at comparable levels to that of the parent sequence except for 5′-AGTATT-3′ and 5′-AATACT-3′, which contain a substitution at an alkylation site, and 5′-AATATC-3′, which contains a substitution at an outer base pair flanking this site (fig. S7B). Strand cleavage analysis revealed sites of alkylation identical to those of the parent sequence, except for 5′-AGTATT-3′ and 5′-AATACT-3′. Though no alkylation was observed at the newly incorporated guanine in these sequences, formation of colibactin-DNA monoadducts was observed at the adenine on the other strand (fig. S7, C and D). Altogether, these studies directly support colibactin forming bis-adenine ICLs within sequence motifs matching previously reported sites of DSBs and mutational signatures, strengthening the evidence that colibactin ICLs are the inciting incidents to these downstream events. Our results also reveal that colibactin can react within a broader sequence context than suggested by the mutational signatures.

With direct evidence that colibactin alkylates AT-rich DNA in a sequence-specific manner, we next sought to elucidate its groove specificity. On the basis of the structural characterization of colibactin-DNA adducts, colibactin is thought to alkylate N3 of adenine, which is accessible through the minor groove of DNA (*18*). Prior calculations also suggested that shape complementarity and electrostatics could lead to a preference for minor groove binding (*26*). We directly tested this proposal by incubating the 50-nucleotide oligomer containing the 5′-AAATTAATA-3′ motif with *pks*$^+$ *E. coli* in the presence of DNA-binding small molecules possessing differing topological preferences (fig. S8) after confirming that these molecules had no effect on bacterial growth or colibactin production (figs. S9 and S10). AT-rich minor groove binders [netropsin and 4′,6-diamidino-2-phenylindole (DAPI)] (*38*, *39*) inhibited ICL formation in a dose-dependent manner (Fig. 2). By contrast, a major groove binder (methyl green) (*40*) and a GC-rich minor groove binder (actinomycin D) (*41*) showed no effect. These results provide strong experimental evidence that colibactin specifically alkylates AT-rich minor grooves.

We next investigated the site of colibactin DNA alkylation on adenine. As highlighted above, we previously characterized an N3-adenine colibactin-DNA adduct (*18*). Additional adenine adducts containing the other "half" of colibactin were identified by using mass spectrometry but were not characterized by nuclear magnetic resonance (NMR) spectroscopy, leaving their connectivity unknown (*21*, *22*). Hypothesizing that both colibactin cyclopropanes alkylate the N3 position of adenine, which would be consistent with selective minor groove binding, we incubated *pks*$^+$ *E. coli* with 50-nucleotide oligomers containing a 5′-A**A**TATT-3′ sequence motif in which the bolded residue was replaced with N3-deaza-deoxyadenosine (dAdo) (*42*), eliminating the possibility for N3-dAdo adducts to form. ICL formation was abolished when N3-deaza-dAdo was incorporated into the forward, reverse, or both strands (fig. S11). Strand cleavage analysis of assays with 25-nucleotide oligomers containing these sequence variants also showed no alkylation of N3-deaza-dAdo, instead revealing monoadduct formation for the singly substituted variants. The presence of monoadducts was confirmed using the LC-HRAM-MS cleavage assay (fig. S11, C and D). Taken together, these data indicate that colibactin exclusively forms bis-N3-dAdo ICLs within the minor groove of AT-rich DNA.

To gain initial insights into the structure of the colibactin-DNA ICL, we characterized 14- and 25-bp oligo substrates containing previously examined sequence motifs with LC-HRAM-MS after *pks*$^+$ *E. coli* incubation (Fig. 3). For each dsODN that became cross-linked, we observed a mixture of native and modified dsODNs. After direct deconvolution analysis of the 14-bp dsODN samples, we measured a mass difference of mass/charge ratio (*m/z*) 771.26 (Fig. 3). The 25-bp dsODN samples required additional analysis and provided mass differences consistent with the value measured for the 14-bp samples (fig. S12; see supplementary discussion). We did not observe any other species in these assays. All dsODNs incubated with *pks*$^-$ *E. coli* showed no mass shift. The observed mass is consistent with formation of a single ICL arising from the proposed colibactin structure containing an α-ketoimine. This unexpected observation sharply contrasts with the structures proposed for colibactin detected in aqueous solution and in putative colibactin and ICL degradation products, which all contain a 1,2-diketone (*21*, *22*, *43*). Additionally, the observed mass of colibactin indicates no oxidation of the ring-opened electrophilic warheads, unlike in previously characterized colibactin-DNA adducts (*18*–*21*). This direct characterization of the colibactin-DNA ICL through MS helps to resolve important aspects of colibactin's structure.

To understand the molecular details of how colibactin binds and cross-links DNA, we used solution-state NMR. This required generating the colibactin-DNA ICL on a large scale using *pks*$^+$ *E. coli*. We chose a 14-bp dsODN with a palindromic sequence (5′-CGCGA**A**TATTCGCG-3′) and substituted A6 with 2′-fluoro-deoxyadenosine to increase the glycolytic bond strength and minimize depurination (*44*). To access sufficient quantities of the colibactin-DNA ICL (~15 nmol), hundreds of small-scale incubations were performed in 96-well plates, combined, and purified to provide a mixture of cross-linked (~55 to 65%)

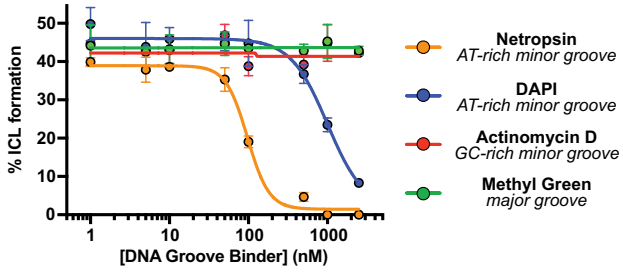


Fig. 2. Colibactin binds AT-rich regions of the minor groove. ICL formation by *pks*$^+$ *E. coli* in the presence of select DNA groove–binding small molecules. Dose-dependent inhibition of ICL formation by AT-rich minor groove binders was observed by denaturing gel electrophoresis and quantified by densitometry. Data are mean ± SD with $n = 3$ biological replicates.

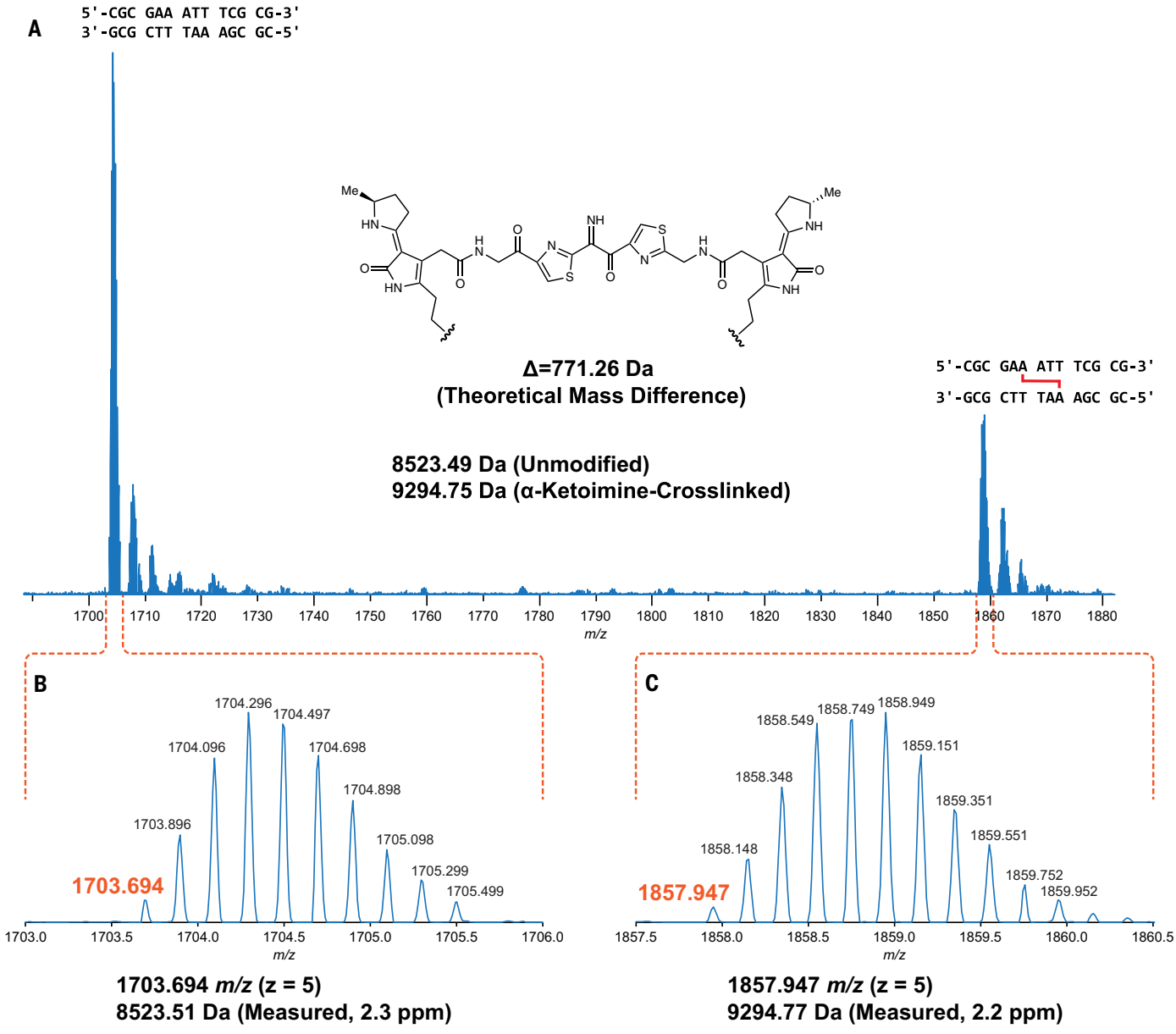


Fig. 3. MS analysis of an intact ICL reveals a central α-ketoimine in colibactin. HRAM LC analysis of cross-linked 14--nucleotide oligomer DNA double-strand oligonucleotide exposed to *pks*$^+$ *E. coli*. (**A**) Full scan spectrum containing −5 charge state signal of unmodified and modified 5′-CGCGAAATTTCGCG-3′ double-strand oligonucleotides. (**B**) Expanded region of spectrum corresponding to the unmodified oligonucleotide. (**C**) Expanded region of the spectrum corresponding to the modified oligonucleotide.

and free DNA. Thus, we first assigned the spectra of free DNA (with and without the 2′-fluoro-A6 modification), which allowed for comparative, unambiguous assignments of DNA chemical shifts that are perturbed upon cross-linking to colibactin. Overall, we did not observe any substantial changes in base-to-ribose nuclear Overhauser effect (NOE) walks throughout the DNA molecule, indicating that the overall B-form groove parameters are maintained upon colibactin cross-linking. This observation suggests that the shape of colibactin is complementary to that of the minor groove.

The structural data verified the sites of alkylation by colibactin identified in our earlier experiments. Compared with the spectra of free DNA, the aromatic ring protons of $A6_a$ and the equivalent $A6_b$ (complementary strand) residues had the largest chemical shift change, with the H2 protons moving downfield by ~0.36 and ~0.26 parts per million (ppm), and the H8, by ~0.83 and ~0.78 ppm, respectively (fig. S14A). By contrast, the average change for the other 2′-deoxyadenosines in the DNA was <0.04 ppm. As expected, the majority of interactions between colibactin and DNA are located at the ATAT sequence where colibactin alkylates, but interactions were also observed from the flanking base pairs, $A5_a$–$T10_b$ and $T10_a$–$A5_b$, indicating an expanded recognition motif.

The pseudosymmetric nature of the colibactin-DNA ICL was also readily discernible. First, equivalent protons from both the colibactin and the palindromic DNA have distinct chemical shifts; for example, the equivalent $A6_a$ and $A6_b$ H8s differ by 0.05 ppm, indicating that they are in slightly different environments (fig. S14A). Second, although such equivalent protons gave rise to similar NOE connectivities with protons in close proximities, the NOEs have slightly different intensities, indicating the same interactions but at different distances (fig. S15). Overall, with the exception of the pyrrolidine rings at either

end of colibactin, we observed NOEs from almost all of its protons to the extended AATATT sequence. The observations that define the orientation of the various colibactin rings with respect to the DNA groove are: (i) no connections from the terminal pyrrolidine rings; (ii) the thiazole hydrogens C34H and C39H gave connections to the outer deoxyribose hydrogens of $T9_a$ and $T10_b$ (H4′/H5′), respectively; and (iii) the N3H and N3′H amides and hydrogens of C27 and C28 (the carbon attached to the N3 of A6), which flank the pyrrolidinone rings, show connections to each other, thus confining the ring inside the groove (Fig. 4, B and D, and fig. S15). Consequently, the structures show that, whereas the terminal pyrrolidine rings point out of the minor groove, the thiazole rings are wedged in, aligning with the phosphodiester backbone with their bulky sulfur atoms pointing outward. Likewise, the pyrrolidinone rings of the warheads are stacked parallel to the groove, slightly outside of the AATATT sequence. This results in a tightly packed DNA-colibactin ICL with colibactin in an extended, concave conformation spanning along half a turn of the minor groove (Fig. 4, B and D, and fig. S14C). The curved shape and close contacts between its heterocyclic aromatic rings and the hydrophobic surfaces of the minor groove walls suggest that shape complementarity as well as van der Waals and hydrophobic interactions all contribute to colibactin-DNA binding. As discussed below, similar features are

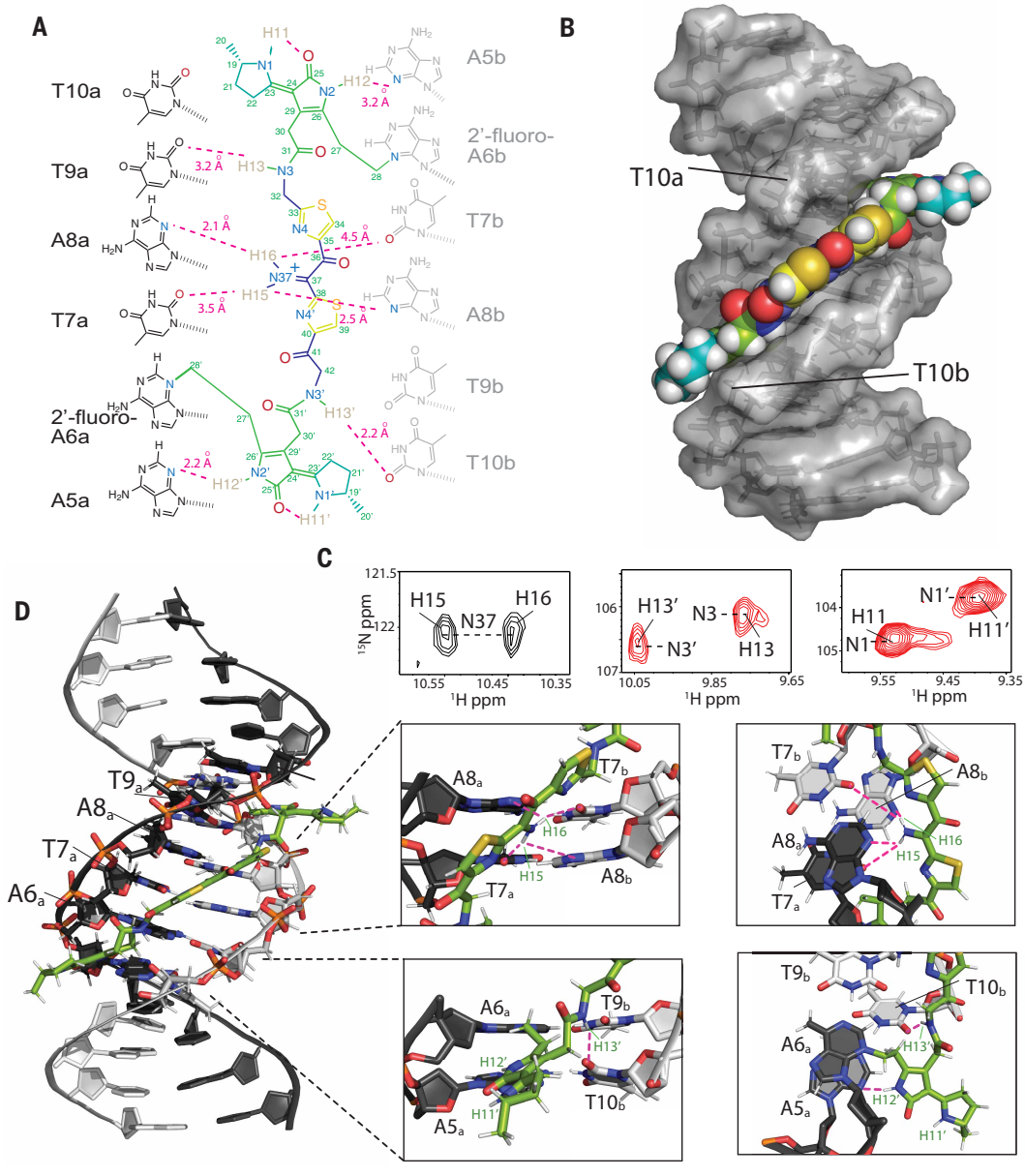


Fig. 4. Structure of the colibactin-DNA ICL. (**A**) Schematic and atom numbering of colibactin and its hydrogen bonding and electrostatic interaction with nucleotide bases, as denoted by magenta dashed lines. (**B**) Surface representation of the structure showing the pyrrolidine, thiazole, and pyrrolidinone rings stacked in between the minor groove edges of the DNA parallel to the long axis. (**C**) Portions of a two-dimensional HSQC experiment with a [^{15}N,^{13}C]-colibactin-DNA ICL sample showing two hydrogen atoms correlated with N37, indicating protonation of the nitrogen (left). The equivalent N3/3′ (center) and N1/1′ (right) nitrogens and their associated protons have different chemical shifts, indicating different environments and the pseudosymmetric nature of the interaction. (**D**) Cartoon representation of the interaction and zoomed-in views showing the placement of (i) colibactin iminium moiety with the electronegative atoms of the center T7A8 residues of the DNA (top) and (ii) the amide hydrogen (H13) and the pyrrolidinone ring hydrogen (H12) interacting with $T10_b$ and $A5_a$ nucleobase (bottom). For clarity, only the inner motif residues (A6 to T9) on the a-strand have been labeled.

found in other minor groove–binding natural products and synthetic compounds.

We also observed hydrogen bonds and/or electrostatic interactions from all nitrogen protons in colibactin, which may explain its specificity for the preferred sequence motif. First, although the NMR data confirmed the presence of a nitrogen in the central scaffold of colibactin, we found this nitrogen within an iminium functional group. Two distinct proton peaks are correlated with N37 in the ^{15}N–heteronuclear single quantum coherence (HSQC) experiment, indicating a protonation event at this site (Fig. 4C). Iminium formation may be explained by the central region of the AATATT binding pocket, which comprises the floor of the minor groove and provides a highly electronegative environment for colibactin, with N37 being near O2 of T7 and N3 of A8 of both strands. Although each iminium proton gives NOEs to A8H2/H1' protons of both strands, H16 had comparatively stronger NOEs, thus placing the iminium slightly closer to the a-strand of the DNA. The NOE correlations from these protons are thus in line with the iminium forming a hydrogen bond with N3 of A8$_a$ and electrostatic interactions with O2 of T7$_b$, O2 of T7, and N3 of A8$_b$. Second, the equivalent colibactin H13 and H13' amide protons showed connectivities to the flanking T9$_a$ and T10$_b$ H1', respectively, suggesting they are positioned to form hydrogen-bonding and electrostatic interactions with the O2 position of these thymines. Third, the H12 and H12' protons of colibactin's pyrrolidinone rings, which are the sites of alkylation, also showed connectivities to A5$_b$H1' and C11$_a$H2' as well as A5$_a$H1' and T10$_b$H2', respectively, indicating that these nitrogens are situated in the center of the groove, electrostatically interacting with the A5 N3 position on either end of the sequence motif (Fig. 4C and fig. S15). This orientation would position the cyclopropane rings of the colibactin warheads close to N3 of A6. Lastly, although H11 and H11' gave no connections to DNA, they were still in slow exchange and gave strong connectivities to the hydrogens in the pyrrolidine rings, indicating that they are most likely forming internal hydrogen bonds to the C25 and C25' carbonyl groups in the adjacent pyrrolidinone rings (fig. S15). These intramolecular hydrogen bonds may form prior to alkylation and, if so, would place the heterocyclic rings of the electrophilic warheads coplanar, enhancing their electrophilicity. Notably, we saw no evidence for intramolecular cyclization of colibactin (*24*), indicating that the ICL derives from a linear pseudosymmetric structure.

This structural information allows us to formulate a model for colibactin-DNA ICL formation (fig. S16). We propose that the central region of colibactin, with its concave shape and heterocyclic aromatic rings, facilitates binding within the minor groove of AT-rich DNA sequences. We also predict that colibactin-DNA binding will be enhanced by multiple positively charged functional groups. In addition to the positively charged pyrrolinium rings of the two electrophilic warheads, the positively charged nitrogen atom of the central aminoketone, enolamine, or α-ketoiminium of colibactin likely plays a critical role in its selectivity for AT-rich minor groove binding by providing favorable electrostatic interactions and hydrogen bonding to the central base pairs in the 5′-WAWWTW-3′ motif, which form the floor of the minor groove. Lastly, the structure also suggests that binding within the minor groove promotes inter- and intramolecular hydrogen bonding and electrostatic interactions involving the colibactin warheads that enhance their electrophilicity and orient the spirocyclopropanes and π-systems of the adjacent pyrrolinium heterocycles with the appropriate stereoelectronic alignment to trigger ring opening by N3 of adenine.

To test this model and, specifically, the importance of colibactin's central structural motif to the specificity of ICL formation, we compared its sequence specificity to that of the stable synthetic colibactin analog possessing two central methylene groups (Fig. 5A) (*24*). We posited that colibactin and the synthetic analog would have different preferences for alkylation of sequences containing multiple CG base pairs within the 5′-WAWWTW-3′ motif. If hydrogen bonding and electrostatic interactions with colibactin's central nitrogen atom are critical for sequence specificity, then exchanging the two central TA base pairs for CG base pairs should diminish alkylation owing to the resulting steric hindrance and electrostatic repulsion from the exocyclic amines of the two guanines, which alter the surface of the minor groove floor. By contrast, the stable colibactin analog lacking the central structural motif should tolerate this substitution.

We incubated either *pks*$^+$ *E. coli* or the synthetic analog with a series of 50-bp dsODNs that contained two GC base pairs at different positions within a 5′-AATATT-3′ motif and measured ICL formation using gel electrophoresis (Fig. 5B). As predicted, the introduction of two GC base pairs at the center of this motif almost completely abolished alkylation by colibactin. ICL formation was also abolished when GC base pairs were installed at the two sites of alkylation and greatly diminished when GC base pairs were introduced at the two outer positions flanking the alkylation site. Other substitutions had minimal effects on ICL formation. Notably, the specificity of the synthetic analog differs from that of natural colibactin (Fig. 5C). We observed a preference for alkylation at the 5′-AACGTT-3′ sequence, which is not targeted by natural colibactin, and minimal alkylation at other sequences. Lastly, we compared the reactivity of colibactin and the synthetic analog toward additional oligonucleotides containing two GC base pairs in the central positions (5′-AAGCTT-3′, 5′-AACCTT-3′, and 5′-AAGGTT-3′) (fig. S17). Again, colibactin did not alkylate these motifs, but the analog displayed robust ICL formation. These results further support a critical role for the central structural motif and its positively charged nitrogen atom in the selectivity of colibactin ICL formation.

To gain additional insight into the interactions influencing the specificity of colibactin-DNA ICL formation, we calculated the electrostatic potential (ESP) values centered on atoms in proposed colibactin structures bearing different central functional groups, both free and cross-linked to DNA (Fig. 6A and figs. S18 and S19). We observed the highest ESP on the iminium nitrogen atom of colibactin bearing a central α-ketoiminium (~664 kJ·mol^{-1}·e^{-1}), consistent with the α-ketoiminium having a positive charge. We calculated a lower ESP on the iminium nitrogen of the α-ketoiminium when this proposed colibactin was cross-linked to DNA compared with that of the corresponding free colibactin (Fig. 6, B and C, and figs. S20 and S21). This decrease could indicate stabilization by electrostatic or other noncovalent (e.g., hydrogen-bonding) interactions with the DNA. However, we did not observe a substantial decrease in ESP of the heavy atoms of the central functional groups in the other proposed colibactin structures when they were cross-linked to DNA, suggesting that the electrostatic interactions between DNA and colibactin are much stronger for the proposed structure bearing a central α-ketoiminium.

Additional calculations focused on DNA further support the importance of electrostatic interactions to colibactin-DNA recognition. Using DNAPhi to predict the ESP of the minor groove of the dsODN used for our structural studies, we found that, as expected, the center of the 5′-AATATT-3′ motif is substantially more electronegative than the other positions, accounting for the increased pK_a (where K_a is the acid dissociation constant) of N37 of colibactin and its protonation (Fig. 6D) (*45*). Substitution of a single central base pair with a CG base pair increased ESP at this position, whereas substituting both central base pairs, which abolished ICL formation by colibactin, resulted in a loss of this ESP differential (Fig. 6, E and F). Modeling interactions between colibactin and these different sequences showed distortions when guanines were introduced owing to repulsions from the large amino substituents of these central base pairs (Fig. 6, G to I). The outer base pairs flanking the alkylation site (A5$_a$ toT10$_b$ and T10$_a$ to A5$_b$) also contribute to the negative ESP, and modeling GC base pair substitutions at these positions also resulted in distortions and repulsions, potentially disrupting key hydrogen bonding and electrostatic interactions with the warheads and accounting for the reduced ICL formation in these sequences (fig. S22). This structure, experimental data, and computations reveal and explain colibactin's specificity for DNA alkylation, highlighting an

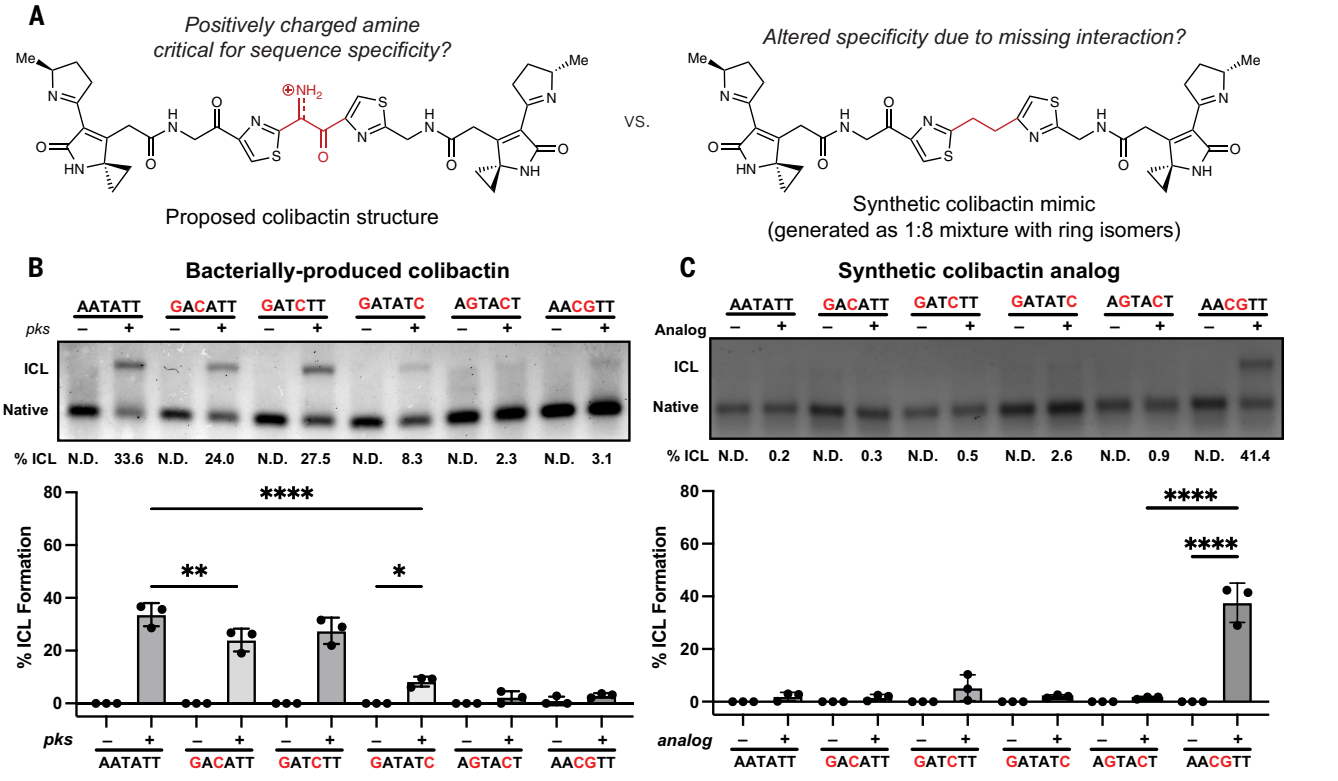


Fig. 5. The unstable central motif of colibactin influences its sequence specificity for ICL formation. (**A**) Structural comparison of bacterially produced colibactin and a synthetic "stable" colibactin analog suggests that the analog may have altered specificity for ICL formation. (**B**) Quantification of ICL formation by coincubation with bacteria through denaturing gel analysis for 50-nucleotide oligomers containing a double GC substitution within the sequence 5′-AATATT-3′. Results were quantified through densitometry and shown as bar plots. (**C**) Quantification of ICL formation by treatment with a colibactin analog through denaturing gel analysis for 50-nucleotide oligomers containing a double GC substitution within the sequence 5′-AATATT-3′. Results were quantified through densitometry and shown as bar plots. All data are mean ± SD with $n = 3$ biological replicates. ****$P < 0.0001$; **$P < 0.01$; *$P < 0.05$; not significant (ns), $P > 0.05$; one-way analysis of variance and Tukey's multiple comparison test. N.D., not determined.

especially notable role for electrostatic interactions involving its chemically unstable central motif.

Discussion

Gaining a molecular understanding of DNA-damaging agents is critical for advancing our understanding of their connections to cancer and use as therapeutic agents. As mutational signatures are increasingly identified in cancer genomes (*46*, *47*), it is imperative to elucidate their origins. Characterizing the DNA lesions that give rise to specific mutational signatures can provide starting places for investigating repair pathways and processes that lead to misrepair. DNA-targeting small molecules are also an important class of drugs (*48*). Studying the specificities and structures of DNA-binding and -alkylating natural products and synthetic compounds has revealed important chemical principles underlying small molecule–DNA recognition and reactivity (*49*, *50*). The high prevalence of *pks*$^+$ *E. coli* in many geographic locations (*34*), the accumulating evidence linking colibactin to CRC development (*5*–*14*, *25*–*33*), and the chemical instability of this genotoxin provide particularly strong motivation for understanding its DNA-damaging activity.

Leveraging advanced MS and NMR approaches, our studies of the colibactin-DNA ICL explain the locations of mutational signatures and enhance our understanding of this unstable natural product. We hypothesize that the α-ketoimine found in the colibactin-DNA ICL is derived from oxidation of an initial α-aminoketone or its enolamine tautomer (fig. S1A) (*23*, *51*), though the timing of this oxidation relative to DNA alkylation is unclear. The presence of this central functional group was unexpected owing to the reactivity of this structural motif and the structures proposed using MS for colibactin and related compounds detected in culture supernatants, which are all suggested to contain a 1,2-diketone. We also see no evidence for DNA alkylation by ring isomers, suggesting that these species, which are generated in the synthesis of stable colibactin mimics and have been observed in decomposition products (*24*, *43*), are likely not relevant for colibactin's DNA-damaging activity. Although the precise identity of the central functional group initially generated by the colibactin biosynthetic pathway remains elusive, our discovery highlights the relevance of proposed structures containing central nitrogen atoms, specifically the α-ketoimine, to colibactin-DNA alkylation. It has also been proposed on the basis of the identification of shunt products that the colibactin biosynthetic pathway produces multiple metabolites with the potential to cause DNA damage (*52*). Our observation of a single major ICL-forming species by MS and NMR calls this into question and suggests that there may be additional factors influencing colibactin production and/or stabilization that are still uncharacterized.

The elucidation of the colibactin-DNA ICL structure reveals how the structural features of colibactin contribute to its DNA-damaging activity (fig. S15). Colibactin can adopt an extended, concave conformation that complements the shape of the narrow, AT-rich minor groove. The heteroaromatic rings flanking the central α-ketoiminium and at the termini of colibactin lie parallel to the walls of the minor groove, enabling hydrophobic and van der Waals interactions. Notably, the orientation of the two thiazole rings reinforces colibactin's concave shape. The regions that link the central scaffold to the two electrophilic warheads contain multiple methylene groups, derived biosynthetically from glycine and malonyl–coenzyme A (*53*), making them

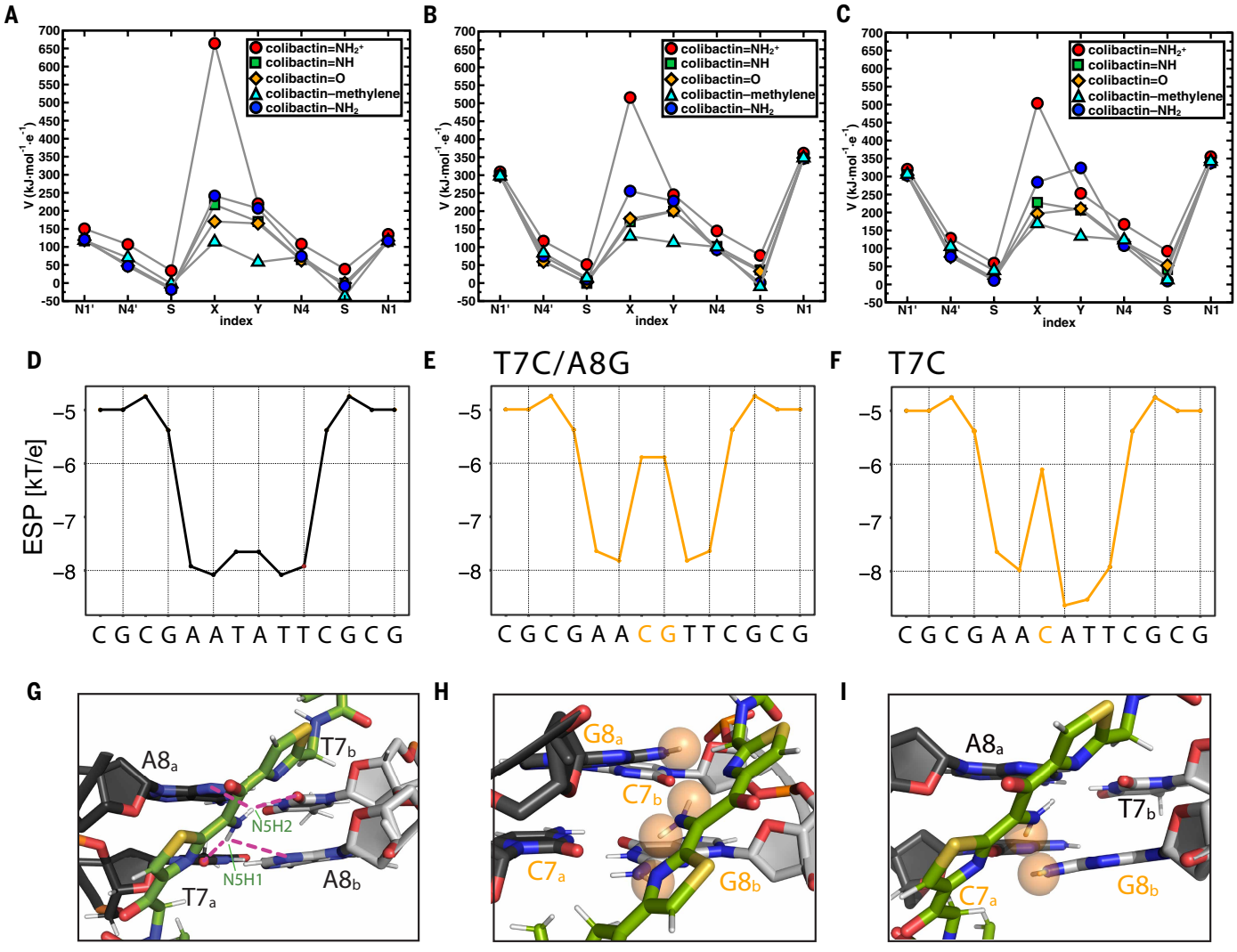


Fig. 6. Electrostatic and steric interactions involving colibactin's unstable central motif likely drive sequence specificity for DNA alkylation. (**A** to **C**) ESP values of proposed colibactin structures [volts (V) in kJ·mol^{-1}·e^{-1}] obtained from DFT optimization calculations of proposed colibactin structures at the B3LYP-D3/6-31G* level of theory. ESP values were calculated for (A) proposed structures of free colibactin, (B) proposed colibactin structures cross-linked to the doubly charged DNA sequence GAATATTC, and (C) proposed colibactin structures cross-linked to the doubly charged double-mutant DNA sequence 5′-GAACGTTC-3′. Proposed colibactin structures included the α-ketoiminium (colibactin=NH$_2^+$), α-ketoimine (colibactin=NH), diketone (colibactin=O), CH$_2$–CH$_2$ (colibactin–methylene), and enolamine (colibactin–NH$_2$) central functional groups. The indices X and Y correspond to the heavy element indices of the central functional group. X=N and Y=O for colibactin=NH$_2^+$, colibactin=NH, and colibactin–NH$_2$; X=O and Y=O for colibactin=O; and X=C and Y=C for colibactin–methylene. (**D** to **F**) Electrostatic potential calculations using DNAPhi predicting the highly electronegative environment in the sequence containing the preferred AATATT motif (black) and the decrease in electronegativity upon sequence substitution in the central motif (orange). (**G** to **I**) Comparison of the NMR structure with structural modeling of the substituted sequences shown in (E) and (F), showing potential steric clashes between colibactin and DNA (orange spheres).

conformationally flexible and likely facilitating curvature. Colibactin's positively charged functional groups should also enhance its affinity for DNA. Multiple nitrogen atoms within colibactin form hydrogen bonds and/or electrostatic interactions that influence binding specificity, including the central α-ketoiminium, the amide nitrogens of the two linker regions, and the two pyrrolidinone rings of the electrophilic warheads. Together, these interactions likely position the reactive cyclopropane rings close to N3 of the attacking adenines. This conformation may also be favored by the steric bulk of the pyrroline rings, which are too large to point into the minor groove. Lastly, an internal hydrogen bond within the electrophilic warhead between the carbonyl of the pyrrolidinone and a protonated pyrrolinium may be critical for enabling alkylation by enhancing electrophilicity and enabling proper stereochemical alignment of the breaking C–C bond of the spirocyclopropane with the extended π-system of the warhead.

Many of the chemical principles likely involved in mediating the specificity of colibactin binding and alkylation (convex shape, aromatic rings, positive charge, hydrogen bond donors) are well established to be important for AT-rich minor groove recognition by other natural products (*54*–*58*), including the polyamide DNA-binding compounds netropsin and distamycin (*38*) as well as the DNA-alkylating duocarmycins, which form monoadducts (*59*, *60*). However, colibactin distinctively combines features of AT-rich minor groove binding specificity with the capacity for interstrand cross-link formation, distinguishing it from other natural products that target DNA. Like other DNA-damaging natural products, the molecular understanding gained

in our studies may inspire the design of synthetic agents of therapeutic potential.

Notably, the interaction between the positively charged, central α-ketoiminium of colibactin and the minor groove of DNA is particularly distinctive among AT-rich minor groove–binding and –alkylating compounds, which typically contain positively charged functional groups at their termini. The decreased electrostatic potential of the central AATATT motif within the minor groove likely favors binding of the protonated α-ketoiminium of colibactin, contributing to specificity (*61, 62*). This interaction structurally resembles those of DNA binding proteins and small molecules, including the positively charged nitrogen atoms of the guanidinium side chain of arginine, which is enriched in proteins that bind AT-rich minor grooves (*57*) as well as the terminal amidine of the minor groove–binding natural product netropsin (*38*) (fig. S23).

The ICL structure, experiments with a synthetic colibactin analog, and calculations highlight the importance of colibactin's central structural motif for its sequence specificity. Notably, the pseudosymmetric structure of colibactin places the positively charged iminium equidistant from the two electrophilic warheads, suggesting that it is important for positioning these functional groups. Although the instability of the central portion of colibactin motivated the development of a "stable" synthetic colibactin analog lacking this functional group, we found that this analog has an altered sequence specificity compared with that of colibactin produced by *pks*$^+$ *E. coli*. This difference suggests that the synthetic analog will not phenocopy the effects of colibactin and calls into question its use as a biologically appropriate surrogate for the natural genotoxin. Our findings raise the questions of why such a chemically unstable structural motif would have evolved to play a notable role in colibactin's biological activity and how it is stabilized and/or protected within the cellular environment. Lastly, our work highlights innovative experimental strategies that may be used to discover and characterize additional unstable DNA-damaging microbial natural products that may be recalcitrant to traditional isolation approaches.

Materials and methods

Cell lines and cultures

The *pks*$^-$ strain used in this study was *E. coli* BW25113 possessing empty vector (pBeloBAC11) obtained from New England Biolabs. The *pks*$^+$ strain was *E. coli* BW25113 possessing pBeloBAC11-*pks* and was a generous gift from the Bonnet Laboratory. Starter cultures of all bacteria were grown overnight aerobically with shaking at 37°C in 5 mL of Luria-Bertani (LB-Lennox, RPI) broth containing 34 μg/mL chloramphenicol. Cultures were inoculated from the desired frozen glycerol stock. For all DNA incubation experiments unless stated otherwise, aliquots of each overnight culture were diluted 1:10 in room-temperature Dulbecco's modified Eagle's Medium (DMEM) supplemented with 25 mM HEPES (pH 7.4) and incubated at 37°C with constant shaking (190 RPM) until the optical density at 600 nm (OD_{600}) reached ~0.4 – 0.5.

Annealing of synthetic oligonucleotides

Oligonucleotides used in this study were purchased dry from Sigma Aldrich, Integrated DNA Technologies, or Genewiz and reconstituted upon arrival in TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0) to make 100 μM stock solutions. Aliquots of complementary oligonucleotide solutions were combined in a 1:1 equimolar ratio in annealing buffer (10 mM Tris-HCl, 50 mM NaCl, pH 8.0) to make a 10 μM solution. For example, 10 μL of each oligo stock solution were mixed with 80 μL of annealing buffer. The resulting solutions were incubated at 95°C for 5 min in a heat block, then slowly cooled to room temperature over ~2 hours. To verify proper annealing, samples were verified by size using agarose gel electrophoresis. 4% Tris-Acetate-EDTA (TAE) gels were made using NuSieve 3:1 agarose (Lonza) and run at 80 V for 60 min. Gels were pre-stained with Sybr Safe (Thermo Fisher) and visualized with an Azure Biosystems 400 Imager. A table of annealed oligonucleotide substrates is provided in the Supporting Information.

DNA cross-linking assay – bacterial incubations

The protocol was adapted from previously reported assays (*20, 24, 35*). 2 μg of the desired oligo nucleotide substrate was diluted to 140 μL of DMEM-HEPES medium in either 1.5 mL Eppendorf tubes or a 96-well culture plate (VWR). DNA solutions were then inoculated with 60 μL of freshly grown bacteria at an OD_{600} = 0.4 – 0.5 (*pks*$^-$ or *pks*$^+$ *E. coli*; ~20 × 10^6 cells) and incubated at 37°C aerobically for 5 hours without shaking. If cultured in a 96-well plate, OD_{600} measurements were taken every 5 min to assess growth with brief shaking before each measurement. After this time, the bacteria were pelleted by centrifugation at 10,000 × g for 3 min (4°C) and the DNA-containing supernatants were transferred to fresh Eppendorf tubes containing 20 μL of 3 M NaOAc, pH 5. Cell pellets were either saved for *N*-myristoyl-D-asparagine quantitation or discarded. DNA was precipitated by adding 660 μL of cold, 95% EtOH (aqueous, v/v) to the acidified supernatants and storing samples at –20°C for ~16 hours. The precipitated DNA was isolated by centrifugation at 16,100 × g for 20 min (4°C) and removal of the supernatant. The resulting DNA pellet was briefly washed with 200 μL of 70% EtOH (aqueous, v/v) by inversion and re-pelleted by centrifugation at 16,100 × g for 20 min. The supernatant was removed, and the DNA pellet was air dried for ~5 – 10 min before reconstitution in 51 μL of TE buffer. The concentrations of all DNA samples were determined by analyzing a 1 μL aliquot on a Nanodrop 2000 spectrophotometer (Thermo Fisher). From this point on, all DNA samples were kept on ice while in use or stored at –20°C to reduce degradation of colibactin-ICLs.

DNA cross-linking assay – synthetic colibactin analog

The protocol was adapted from previously reported methods and the procedure described above (*24*). Briefly, 200 ng of the desired oligo nucleotide substrate was diluted in a citrate buffer (10 mM, pH 5.0) and treated with a mixture of synthetic colibactin analogs (100 μM of a 1:8 mixture of uncyclized and cyclized analogs in DMSO) to a final volume of 20 μL (5% DMSO v/v in water). Assay mixtures were incubated at 37°C for 1 hour without shaking and were used immediately for electrophoresis.

Interstrand cross-link analysis by denaturing gel electrophoresis

The protocol was adapted from previously reported methods (*35*). Briefly, 10 μL aliquots of all DNA samples to be analyzed were diluted to 10 ng/μL in TE buffer. DNA concentrations were verified by Nanodrop and adjusted if needed. While on ice, 5 μL of '1% Denaturing Buffer' (1% m/v NaOH, 6% m/v sucrose, 0.01% m/v Orange G) were added to 10 μL of each diluted sample (~100 ng DNA). The denatured samples were then loaded into a 4% Tris-Acetate-EDTA (TAE) gel pre-stained with Sybr Gold (Thermo Fisher). The gel was run at 80 V for 60 min and visualized on an Azure Biosystems 400 Imager. Gel bands were further quantified using ImageJ. Percent cross-linking was calculated by dividing the ICL band intensity by the sum of both ICL and native DNA bands and multiplying by a factor of 100.

DNA strand cleavage assay

Individual oligo substrates were subjected to the DNA cross-linking protocol described above except that the resulting DNA pellets were redissolved in 45 μL of TE buffer. Samples were then heated to 90°C for 1 hour using a heat block to induce depurination of colibactin ICLs. After samples cooled to room temperature, 5 μL of 3 M NaOAc, pH 5 and 150 μL of cold, 95% EtOH were added in sequence. Samples were stored at –20°C overnight to induce DNA precipitation. Precipitated DNA was pelleted by centrifugation (20 min at 16,100 × g) and the supernatant was removed by pipetting. DNA pellets were briefly air dried (~5 min) and then dissolved in 50 μL of 1 M aqueous piperidine. Samples were heated to 90°C for 30 min, cooled to room temperature,

and evaporated to dryness using a Labconco Centrivap (45°C for ~1 hour). The resulting DNA pellets were reconstituted in 51 μL of TE buffer. The sample concentration was determined by Nanodrop analysis. All samples were stored at −20°C until further analysis. 5′-FAM-labeled DNA samples were analyzed by sequencing polyacrylamide gel electrophoresis while unlabeled samples were subjected to LC-HRAM-MS.

Analysis of strand cleavage products by sequencing polyacrylamide gel electrophoresis

150 mL of 15% UreaGel solution was prepared by combining SequaGel UreaGel 19:1 denaturing gel reagents (National Diagnostics) according to the manufacturer's protocol. Polymerization was initiated by adding 60 μL of tetramethylethylenediamine (TEMED; Sigma Aldrich) and 1.2 mL of freshly prepared 10% ammonium persulfate (VWR). This solution was used to cast a 12 in × 16 in × 0.75 mm gel which solidified over 1 hour. Prior to sample loading, the gel was pre-run for 30 min at 50 W in Tris-Borate-EDTA (TBE) buffer and all wells were flushed to remove residual urea. Strand cleavage samples were prepared for loading by mixing a 12 μL aliquot (35 ng/μL) with 12 μL of 2X TBE-Urea Sample Buffer (Novex) and heating the mixture to 95°C for 5 min. The corresponding A+G Maxam-Gilbert ladders were generated by a previously reported method (*63*). Ladders were diluted to 125 ng/μL and prepared for loading analogously to the strand cleavage samples. After ladder and sample loading, gels were run at a constant 50 W for 4 hours. Gel imaging was performed using an Azure Sapphire Imager set for fluorescence detection (ex. 488/ em. 518). For the experiments in Fig. 1C, 2, 5B, S7C, and S11B the "auto-exposure to region" setting was used during "Live Mode", which focuses the exposure on a selected region to avoid under- or overexposure. This was used in conjunction with the "capture selected region" setting which images a selected portion of a gel. The exposure setting was "Wide Dynamic Range" for all gels. For the experiments in Fig. 5C and S17, full gel images were processed in Adobe Photoshop to add an invert layer, and the brightness, contrast and exposure were globally modified to enhance visibility of bands against the background. See Appendix S1 for gel images.

Analysis of intact colibactin-DNA interstrand cross-links by LC-MS

Sample preparation: 2-3 μg of each of the oligonucleotides was dissolved in 50 μL of 10 mM Tris Buffer containing 1 mM EDTA and transferred to 1.2 mL silanized vials for injection on the LC-MS.

Chromatography for LC-MS: For each sample, 5 μL were injected onto an UltiMate 3000 RSLCnano UPLC (Thermo Scientific, Waltham, MA) system. Separation was performed using a Shodex HILICpak VN-50 2.0 × 150 mm column (Showa Denko K.K., Tokyo, Japan) maintained at 15°C using (A) water with 10 mM ammonium acetate and (B) 90:10 acetonitrile:water with 10 mM ammonium acetate. Initially 100% B at a flow rate of 200 μL/min was maintained for 6 min followed by linear gradient to 15% B and flow rate to 150 μL/min with a 3 min re-equilibration between injections at the initial conditions.

Mass spectrometry: All mass spectrometric data was acquired with an Orbitrap Lumos mass spectrometer (Thermo Scientific, Waltham, MA). Negative mode electrospray ionization was used with a source voltage of 2.5 kV, a sheath gas setting of 15, and a capillary temperature of 400°C. Data was collected in profile mode at a resolution setting of 120,000 with a Mass Range setting at High. The S-Lens RF level setting was 60% with full scan detection of *m/z* 1500-2500 using a normalized AGC Target of 250%, a Maximum Injection Time of 100 ms, and 10 microscans.

Analysis of colibactin-induced strand cleavage products by LC-MS

Sample preparation: A total amount of 2 μg of each of the cleavage product samples was dissolved in 50 μL of 10 mM Tris Buffer containing 1 mM EDTA and transferred to 1.2 mL silanized vials for injection on the LC-MS.

Chromatography for LC-MS: From each sample 5 μL were injected onto an UltiMate 3000 RSLCnano UPLC (Thermo Scientific, Waltham, MA) system. Separation was performed using a Shodex HILICpak VN-50 2.0 × 150 mm column (Showa Denko K.K., Tokyo, Japan) maintained at 15°C using (A) water with 10 mM ammonium acetate and (B) 90:10 acetonitrile:water with 10 mM ammonium acetate. Initially, 100% B at a flow rate of 200 μL/min was maintained for 6 min followed by linear gradient to 40% B over 36 min followed by 0% B over 7 min, with a 4 min hold at 0% B and a 5 min re-equilibration between injections at the initial conditions.

Mass spectrometry: All mass spectrometric data was acquired with an Orbitrap Lumos mass spectrometer (Thermo Scientific, Waltham, MA). Negative mode electrospray ionization was used with a source voltage of 2.5 kV, a sheath gas setting of 15, and a capillary temperature of 400°C. Data was collected in profile mode at a resolution setting of 120,000 with a Mass Range setting at High. The S-Lens RF level setting was 100% with full scan detection of *m/z* 700-2500 using a normalized AGC Target of 250%, a Maximum Injection Time of 200 ms, and 5 microscans.

Oligonucleotide data analysis: Data analysis was performed using Thermo Scientific's Protein Deconvolution and FreeStyle software packages and the online Mongo Oligo Mass Calculator tool.

Cleavage assay MS data analysis: Identifications and relative abundance measurements of the base treatment-induced cleavage sites of the colibactin-exposed double strand oligonucleotides were made using isotopically-resolved charge-state mass deconvolution of their LC-HRAM-MS spectra. The molecular formulas of the base-treatment cleavage products are the same as the "w" and "d" product ions formed upon MS^2 collisional induced dissociation (CID) of negatively charged unmodified single strand oligonucleotides (*64*). The "w" and "d" product ions for each single strand oligonucleotide were calculated online using the Mongo Oligo Mass Calculator v2.06 (http://mass.rega.kuleuven.be/mass/mongo.htm) with the "CID fragments" feature and the "monoisotopic mass", "negative mode", "DNA", and the "5'-OH" and "3'-OH" terminals selected. Mass identities were verified by comparing to calculated masses of proposed structures in ChemDraw. The measured masses of the cleavage products were determined by deconvoluting the multiple charge states seen in full scan spectra acquired during the cleavage product retention period using Freestyle software (Thermo Scientific, Waltham, MA). The deconvoluted experimental masses from FreeStyle were compared to cleavage products masses calculated using the Mongo software (accounting for the charge state) to assign the cleavage products. The assigned cleavage products were then used to identify the location of the colibactin adduct of the intact oligonucleotide. Signal intensities from both the [M+H] and [M + Na] ions from each cleavage product were summed and normalized to the amount of DNA injected. Values plotted are the difference between the average signal intensities observed in assays with pks^+ and pks^- *E. coli*.

DNA cross-linking assay in the presence of groove binders

AAATTAATA 50-nucleotide oligomer (2 μg) was subjected to a modified version of the DNA cross-linking assay conditions described above in which individual DNA groove binding small molecules were added to the assay mixture just before addition of bacteria. Groove binders tested were: netropsin (Enzo Chemicals), DAPI (4′,6-diamidino-2-phenylindole, Sigma Aldrich), methyl green (Chem Impex), and actinomycin D (Acros Organics). Groove binders were added as solutions in DMSO to final concentrations of 0, 5, 10, 50, 100, 500, and 1000 nM while keeping the DMSO concentration ≤ 2% (v/v). The extent of ICL formation was determined by the denaturing gel electrophoresis protocol described above.

Quantitation of *N*-myristoyl-D-asparagine ("prodrug motif")

This assay was adapted from a previous report and was used to confirm production of colibactin in assay mixtures (*65*). Cell pellets obtained from the DNA cross-linking assay were resuspended in 200 μL of LC-MS grade methanol (Honeywell) containing 100 nM d_{27}-*N*-myristoyl-D-asparagine, which served as an internal standard and was prepared as described previously (*33*). Cell suspensions were sonicated for 2 min in a bath sonicator and vigorously vortexed. This was repeated once more before centrifuging the samples at 16,100 × g for 10 min to pellet all cell debris. Sample supernatants were passed through a centrifugal, AcroPrep Advance 96-well 0.2 μM PTFE filter plate (4000 RPM for 10 min, Pall Corp.) and collected in a 96-well clear bottom plate. Samples were either frozen at −20°C to prevent evaporation or immediately analyzed by a previously reported liquid chromatography negative electrospray–ionization tandem mass spectrometry (UPLC-ESI$^-$-MS/MS) method.

Liquid chromatography was performed using a Waters Acquity UPLC H-Class System (Waters Corporation) equipped with an Agilent Poroshell 120 EC-C18, 2.7 μm, 4.7 mm × 50 mm column using a multistep gradient. Conditions started at 10% solvent B at 650 μL/min for 0.5 min, followed by a linear gradient to 95% B over 0.5 min, a hold at 95% B for 1 min, and a linear gradient back to 10% B over 0.5 min where the column re-equilibrated for 1 min (solvent A, 95:5 water/methanol + 0.03% NH_4OH; solvent B, 80:15:5 isopropanol/methanol/water; injection volume = 5 μL). Mass spectrometry was performed using a Waters Xevo TQ-S UPLC-triple quadrupole mass spectrometer. The multireaction monitoring (MRM) transitions were *m/z* 341.3 → *m/z* 226.3 (collision energy (CE), 24 V; cone voltage, 50) for unlabeled prodrug motif and *m/z* 368.5 → *m/z* 253.3 (CE, 28 V; cone voltage, 58 V) for D_{27}-prodrug motif. Data analysis was conducted using TargetLynx software. Unlabeled prodrug concentrations were calculated by converting the peak area ratios (unlabeled prodrug/d_{27}-prodrug) to concentration ratios using a freshly run calibration curve of varying unlabeled prodrug containing 100 nM d_{27}-prodrug and multiplying by internal standard concentration (100 nM).

Large-scale production of a colibactin-DNA ICL for NMR spectroscopy

~192 μg of double-stranded 2′-fluoro-14mer (Genewiz) were diluted in 13.5 mL of DMEM-HEPES. The resulting solution was dispensed into a 96-well plate in 140 μL aliquots and subjected to the DNA cross-linking assay described above except incubations were performed for 16 hours at 30°C. Two columns of the 96-well plate were used as negative controls (media blank and *pks*$^-$ *E. coli*); all other columns contained *pks*$^+$ *E. coli*. After the incubation, half of the wells in each column were combined to give two samples per column (24 samples total, ~800 μL) which were centrifuged at 10,000 × g for 5 min to pellet cells. The supernatants were then divided into two 400 μL aliquots and each aliquot was treated with 40 μL of 3M NaOAc, pH 5 and added to 880 μL of 95% EtOH (aqueous, v/v) to precipitate DNA. Samples were stored at −20°C overnight.

DNA-precipitated samples were centrifuged at 16,100 × g for 30 min (4°C) to pellet DNA. The supernatants were removed, and each DNA pellet was washed with 200 μL of 70% EtOH (aqueous, v/v). The DNA was re-pelleted by centrifugation (16,100 × g for 15 min) and the supernatant was removed. The DNA pellets were air-dried for ~5 min. Samples originating from the same column on the 96-well plate were reconstituted in 125 μL of TE buffer and pooled to make 12 samples (500 μL total volume). The extent of ICL formation in each sample was checked using the previously denaturing gel electrophoresis method. Finally, *pks*$^+$ samples were pooled, desalted with water and concentrated to ~50 μL using Amicon Ultra 3K – 0.5 mL spin filters (16,100 × g for 30 min, Millipore) and pooled.

This entire protocol was repeated twice to generate a total of ~410 μg of partially cross-linked DNA from 3 96-well plates. All DNA from *pks*$^+$ samples was combined to generate the final sample, which was stored at −20°C until further processing and analysis by NMR spectroscopy.

Production of a [^{15}N,^{13}C]-colibactin-DNA ICL for NMR spectroscopy

DNA containing an ICL with isotopically labeled colibactin was produced analogously to the unlabeled sample except all incubations were performed in M9 minimal medium containing [^{15}N]-ammonium chloride (99%, Cambridge Isotope Laboratories, Inc.) and [$^{13}C_6$]-D-glucose (99%, Cambridge Isotope Laboratories, Inc.). Additionally, overnight starter cultures were sub-cultured in this isotopically labeled medium instead of DMEM-HEPES. Isotope incorporation was assessed by mass spectrometry using the protocol described above for intact colibactin-DNA interstrand cross-links and revealed >90% isotope incorporation.

NMR spectroscopy

CYANA library residue design: As no CYANA library residue for a colibactin modified base currently exists, it had to be constructed anew. The colibactin molecule was drawn using ChemDraw 21.0.0 according to the structure inferred from the analysis of intact colibactin-DNA interstrand cross-links described above and knowledge from prior studies (*21*, *22*). An adenosine monophosphate residue was N3-linked onto the opened warhead to mimic a monoalkylated residue on the imino-facing side of colibactin. The opposite end of the colibactin molecule had a single methyl group of the opened warhead removed. This was necessary, as CYANA is unable to model the interstrand cross-linked adenosines as a single molecule in the context of any DNA or RNA chain. Thus, in order to accurately represent the link, an N3-methyladenosine residue was also constructed using ChemDraw, representing the colibactin modified base on the ketone-facing side. The ChemDraw structures were transformed to PDB files using OPENBABEL's cdxml to pdb function (*66*).

The exported file was then manually adjusted to fit the CYANA library format. A custom R script was then used to center the coordinates according to CYANA specifications and rearrange the atom order to the default of adenosines in the CYANA library such that all bonding and dihedral parameters specify the correct atoms on the new base. Dihedral parameters for the adenosine atoms were copied from standard adenosine residues, all colibactin dihedral angles were manually specified by their degrees of freedom according to the known stereochemistry, sp-hybridization, and aromaticity. Finally, for CYANA calculations, a bond between the warhead-methyl end of colibactin and the N3-methyl of the opposite strand A+3 residue of length 1.48-1.52 Å was included, thereby mimicking the complete colibactin ICL.

Xplor library residue design: For Xplor, the topology and parameter file templates were generated using the ChemDraw structures and PRODRG version AA100323.0717 (*67*). The atom parameters and topology values in the exported templates were manually adjusted according to the stereochemistry of proposed colibactin structures as determined by prior biosynthetic and synthetic studies. The pyrrolidinone, pyrroline, and thiazole ring parameters were adjusted according to the following references, respectively (*68*–*70*). Planarity, angles, and dihedrals were adjusted according to the stereochemistry of proposed colibactin structures as determined by prior biosynthetic and synthetic studies. Even calculations where these angles were freely rotated converged with prior stereochemistry. The end connectors of colibactin and the N3-methyladenosine were also added manually and their repulsion lowered to comply with the features of the covalent C–C bond. The N3-methyl residue on the connecting adenosine was written and adjusted manually. Similarly, for cases in which the structural features were not known, they were modeled manually, such as the axial vs equatorial positions of the terminal pyrrolidines, the single or double protonation of the α-iminoketone and the *cis*- vs *trans*-orientation of the thiazole rings, the parameter and topology files were adjusted manually.

NMR data acquisition and resonance assignment: DNA samples were suspended in buffer (10 mM Tris-HCl pH 8.2, 10mM NaCl) by washing them five times using Amicon centrifugal filters. All NMR experiments

were acquired in 5 mm Shigemi tubes with a Bruker 800 MHz instrument containing a cryogenic probe. Spectra for observing non-exchangeable protons were collected in 100% D_2O at 298 K and for exchangeable protons in 90% H_2O at 278 K. ^1H-^1H NOE spectroscopy and total correlation spectroscopy were recorded with unlabeled samples and ^{15}N-HSQC spectra were collected by using $^{13}C^{15}N$-labeled samples. All data was analyzed using NMRDraw v11.1, NMRviewJ 8.0.3, and NMRFx Analyst v11.2.4-c (*71*).

Structural modeling: Initial structural models were generated using manually assigned restraints in CYANA, where upper-limit distance restraints of 2.7, 3.3, and 5.0 Å were employed for direct NOE cross-peaks of strong, medium, and weak intensities, respectively (*72*). To prevent the generation of structures with collapsed major grooves, cross-helix P–P distance restraints (with 20% weighting coefficient) were employed for B-form helical segments. Standard torsion angle restraints were used for the B-helical geometry, allowing for ±50° deviations from ideality ($\alpha = -68°$, $\beta = -147°$, $\gamma = 46°$, $\delta = 135°$, $\varepsilon = -150°$, $\zeta = -100°$) (*73*). Standard hydrogen-bonding restraints with approximately linear NH–N and NH–O bond distances of 1.9 ± 0.1 Å and N–N and N–O bond distances of 2.9 ± 0.01 Å.

The CYANA structure with the lowest target function was used as the initial model for structure calculations Xplor-NIH to incorporate electrostatic constraints. First, structures were calculated using annealing from 2000 °C to 25 °C in steps of 12.5 °C. Standard energy potential terms for bonds, angles, torsion angles, van der Waals interactions, and interatomic repulsions were included. Energy potentials for NOEs, hydrogen bonds, and planarity were incorporated with restraints derived from NMR data. All restraints used in CYANA were included except for phosphate-phosphate distances. The structures were sorted by energy using bond, angle, dihedral, and NOE energy potential terms and the ten percent of the structures with the lowest sort energy. The lowest ten percent of these were deposited in the RCSB data bank.

Structure deposition: NMRFx Analyst was used to confirm distance restraints used for the structure calculations and generate NMR-STAR format files for uploading to the BMRB. To do this, an N3-methyladenosine residue with a fluorine at the F2' position was constructed for the NMRFx residue library. This allowed loading in the DNA sequence with the modified residue in each of the two strands. A PDB file, containing the atoms specific to the colibactin molecule was extracted from the file generated by XPLOR. This PDB file was read to load in the colibactin molecule to form the complete complex. Next, a peak list file (in NMRFX .xpk2 format) was loaded containing NOE cross peaks. This peak list was used to populate a table of distance restraints and violations (given the loaded 3D structure). The molecular viewer in NMRFx was also used to visualize constraints in the context of the 3D structure. NMRFx was then used to export the NMR-STAR file containing the molecular assembly and chemical shift assignments.

Computational modeling

Calculations of colibactin electrostatic potential: We computed the ESP of specific atoms on proposed colibactin structures and doubly charged DNA with sequences 5′-GAATATTC-3′ and 5′-GAACGTTC-3′ following previously reported protocols (*74*–*76*). We computed the partial charges on atoms of interest using iterative Hirshfeld (Hirshfeld-I) charges (*77*, *78*) as implemented in Multiwfn version 3.7 (*79*) and applied these charges to compute the ESP at these atoms (*76*). All ESP values were obtained from geometry optimization calculations carried out with density functional theory (DFT) at the B3LYP (*80*–*82*)–D3 (*83*)/6-31G* (*84*–*87*) level of theory.

Calculation of DNA electrostatic potential: We used the online tool DNAphi provided by the Rohs lab (https://rohslab.usc.edu/DNAphi/index.html) (*63*) which predicts ESP of minor grooves by solving the non-linear Poisson-Boltzmann equation of DNA fragments.

Modeling of colibactin ICLs containing alternative central base pairs: We used the restraints generated through NMR (see above) and substituted the central or flanking base pairs for generating initial models of non-ideal motifs cross-linked to colibactin by CYANA followed by the introduction of electrostatics by Xplor as described above.

REFERENCES AND NOTES

1. G. El Tekle, N. Andreeva, W. S. Garrett, The role of the microbiome in the etiopathogenesis of colon cancer. *Annu. Rev. Physiol.* **86**, 453–478 (2024). doi: 10.1146/annurev-physiol-042022-025619; pmid: 38345904
2. M. T. White, C. L. Sears, The microbial landscape of colorectal cancer. *Nat. Rev. Microbiol.* **22**, 240–254 (2024). doi: 10.1038/s41579-023-00973-4; pmid: 37794172
3. M. W. Dougherty, C. Jobin, Shining a light on colibactin biology. *Toxins (Basel)* **13**, 346 (2021). doi: 10.3390/toxins13050346; pmid: 34065799
4. K. M. Wernke *et al.*, Structure and bioactivity of colibactin. *Bioorg. Med. Chem. Lett.* **30**, 127280 (2020). doi: 10.1016/j.bmcl.2020.127280; pmid: 32527463
5. J.-P. Nougayrède *et al.*, *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848–851 (2006). doi: 10.1126/science.1127059; pmid: 16902142
6. G. Cuevas-Ramos *et al.*, *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11537–11542 (2010). doi: 10.1073/pnas.1001261107; pmid: 20534522
7. A. Cougnoux *et al.*, Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**, 1932–1942 (2014). doi: 10.1136/gutjnl-2013-305257; pmid: 24658599
8. J. C. Arthur *et al.*, Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012). doi: 10.1126/science.1224820; pmid: 22903521
9. M. Bonnet *et al.*, Colonization of the human gut by *E. coli* and colorectal cancer risk. *Clin. Cancer Res.* **20**, 859–867 (2014). doi: 10.1158/1078-0432.CCR-13-1343; pmid: 24334760
10. S. Tomkovich *et al.*, Locoregional effects of microbiota in a preclinical model of colon carcinogenesis. *Cancer Res.* **77**, 2620–2632 (2017). doi: 10.1158/0008-5472.CAN-16-3472; pmid: 28416491
11. C. M. Dejea *et al.*, Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592–597 (2018). doi: 10.1126/science.aah3648; pmid: 29420293
12. T. Miyasaka *et al.*, *Pks*-positive *Escherichia coli* in tumor tissue and surrounding normal mucosal tissue of colorectal cancer patients. *Cancer Sci.* **115**, 1184–1195 (2024). doi: 10.1111/cas.16088; pmid: 38297479
13. E. Buc *et al.*, High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. *PLOS ONE* **8**, e56964 (2013). doi: 10.1371/journal.pone.0056964; pmid: 23457644
14. V. Eklöf *et al.*, Cancer-associated fecal microbial markers in colorectal cancer detection. *Int. J. Cancer* **141**, 2528–2536 (2017). doi: 10.1002/ijc.31011; pmid: 28833079
15. C. A. Brotherton, M. Wilson, G. Byrd, E. P. Balskus, Isolation of a metabolite from the *pks* island provides insights into colibactin biosynthesis and activity. *Org. Lett.* **17**, 1545–1548 (2015). doi: 10.1021/acs.orglett.5b00432; pmid: 25753745
16. X. Bian, A. Plaza, Y. Zhang, R. Müller, Two more pieces of the colibactin genotoxin puzzle from *Escherichia coli* show incorporation of an unusual 1-aminocyclopropanecarboxylic acid moiety. *Chem. Sci.* **6**, 3154–3160 (2015). doi: 10.1039/C5SC00101C; pmid: 28706687
17. M. I. Vizcaino, J. M. Crawford, The colibactin warhead cross-links DNA. *Nat. Chem.* **7**, 411–417 (2015). doi: 10.1038/nchem.2221; pmid: 25901819
18. M. R. Wilson *et al.*, The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019). doi: 10.1126/science.aar7785; pmid: 30765538
19. M. Xue, E. Shine, W. Wang, J. M. Crawford, S. B. Herzon, Characterization of natural colibactin-nucleobase adducts by tandem mass spectrometry and isotopic labeling. Support for DNA alkylation by cyclopropane ring opening. *Biochemistry* **57**, 6391–6394 (2018). doi: 10.1021/acs.biochem.8b01023; pmid: 30365310
20. N. Bossuet-Greif *et al.*, The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *mBio* **9**, e02393-17 (2018). doi: 10.1128/mBio.02393-17; pmid: 29559578
21. Y. Jiang *et al.*, Reactivity of an unusual amidase may explain colibactin's DNA cross-linking activity. *J. Am. Chem. Soc.* **141**, 11489–11496 (2019). doi: 10.1021/jacs.9b02453; pmid: 31251062
22. M. Xue *et al.*, Structure elucidation of colibactin and its DNA cross-links. *Science* **365**, eaax2685 (2019). doi: 10.1126/science.aax2685; pmid: 31395743
23. A. R. Healy *et al.*, Synthesis and reactivity of precolibactin 886. *Nat. Chem.* **11**, 890–898 (2019). doi: 10.1038/s41557-019-0338-2; pmid: 31548676
24. K. M. Wernke *et al.*, Probing microbiome genotoxicity: A stable colibactin provides insight into structure–activity relationships and facilitates mechanism of action studies. *J. Am. Chem. Soc.* **143**, 15824–15833 (2021). doi: 10.1021/jacs.1c07559; pmid: 34524796

25. C. Pleguezuelos-Manzano *et al.*, Mutational signature in colorectal cancer caused by genotoxic *pks*$^{+}$ *E. coli*. *Nature* **580**, 269–273 (2020). doi: 10.1038/s41586-020-2080-8; pmid: 32106218
26. P. J. Dziubańska-Kusibab *et al.*, Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat. Med.* **26**, 1063–1069 (2020). doi: 10.1038/s41591-020-0908-2; pmid: 32483361
27. A. Boot *et al.*, Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Res.* **30**, 803–813 (2020). doi: 10.1101/gr.255620.119; pmid: 32661091
28. D. Terlouw *et al.*, Enrichment of colibactin-associated mutational signatures in unexplained colorectal polyposis patients. *BMC Cancer* **24**, 104 (2024). doi: 10.1186/s12885-024-11849-y; pmid: 38238650
29. A. Rosendahl Huber *et al.*, Improved detection of colibactin-induced mutations by genotoxic *E. coli* in organoids and colorectal cancer. *Cancer Cell* **42**, 487–496.e6 (2024). doi: 10.1016/j.ccell.2024.02.009; pmid: 38471458
30. D. Terlouw *et al.*, Recurrent APC splice variant c.835-8A>G in patients with unexplained colorectal polyposis fulfilling the colibactin mutational signature. *Gastroenterology* **159**, 1612–1614.e5 (2020). doi: 10.1053/j.gastro.2020.06.055; pmid: 32603656
31. B. Chen *et al.*, Contribution of *pks*$^{+}$ *E. coli* mutations to colorectal carcinogenesis. *Nat. Commun.* **14**, 7827 (2023). doi: 10.1038/s41467-023-43329-5; pmid: 38030613
32. M. Díaz-Gay *et al.*, Geographic and age variations in mutational processes in colorectal cancer. *Nature* **643**, 230–240 (2025). doi: 10.1038/s41586-025-09025-8; pmid: 40267983
33. T. Mäklin *et al.*, Geographical variation in the incidence of colorectal cancer and urinary tract cancer is associated with population exposure to colibactin-producing *Escherichia coli*. *Lancet Microbe* **6**, 101015 (2025). doi: 10.1016/j.lanmic.2024.101015; pmid: 39644909
34. A. Mandarino Alves *et al.*, Dysfunctional mucus structure in cystic fibrosis increases vulnerability to colibactin-mediated DNA adducts in the colon mucosa. *Gut Microbes* **16**, 2387877 (2024). doi: 10.1080/19490976.2024.2387877; pmid: 39133871
35. M. Xue, K. M. Wernke, S. B. Herzon, Depurination of colibactin-derived interstrand cross-links. *Biochemistry* **59**, 892–900 (2020). doi: 10.1021/acs.biochem.9b01070; pmid: 31977191
36. M. R. Volpe *et al.*, In vitro characterization of the colibactin-activating peptidase ClbP enables development of a fluorogenic activity probe. *ACS Chem. Biol.* **14**, 1097–1101 (2019). doi: 10.1021/acschembio.9b00069; pmid: 31059217
37. A. M. Maxam, W. Gilbert, A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560–564 (1977). doi: 10.1073/pnas.74.2.560; pmid: 265521
38. M. L. Kopka, C. Yoon, D. Goodsell, P. Pjura, R. E. Dickerson, The molecular origin of DNA-drug specificity in netropsin and distamycin. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1376–1380 (1985). doi: 10.1073/pnas.82.5.1376; pmid: 2983343
39. T. A. Larsen, D. S. Goodsell, D. Cascio, K. Grzeskowiak, R. E. Dickerson, The structure of DAPI bound to DNA. *J. Biomol. Struct. Dyn.* **7**, 477–491 (1989). doi: 10.1080/07391102.1989.10508505; pmid: 2627296
40. S. K. Kim, B. Nordén, Methyl green. A DNA major-groove binding drug. *FEBS Lett.* **315**, 61–64 (1993). doi: 10.1016/0014-5793(93)81133-K; pmid: 8416812
41. S. Kamitori, F. Takusagawa, Crystal structure of the 2:1 complex between d(GAAGCTTC) and the anticancer drug actinomycin D. *J. Mol. Biol.* **225**, 445–456 (1992). doi: 10.1016/0022-2836(92)90931-9; pmid: 1593629
42. R. Cosstick *et al.*, Molecular recognition in the minor groove of the DNA helix. Studies on the synthesis of oligonucleotides and polynucleotides containing 3-deaza-2′-deoxyadenosine. Interaction of the oligonucleotides with the restriction endonuclease EcoRV. *Nucleic Acids Res.* **18**, 4771–4778 (1990). pmid: 2395641
43. T. Zhou *et al.*, Isolation of new colibactin metabolites from wild-type *Escherichia coli* and in situ trapping of a mature colibactin derivative. *J. Am. Chem. Soc.* **143**, 5526–5533 (2021). doi: 10.1021/jacs.1c01495; pmid: 33787233
44. M. Ikehara, 2′-Substituted 2′-deoxypurine nucleotides: Their conformation and properties. *Heterocycles* **21**, 75–90 (1984). doi: 10.3987/S-1984-01-0075
45. C. A. Fitch, G. Platzer, M. Okon, B. E. Garcia-Moreno, L. P. McIntosh, Arginine: Its pK_a value revisited. *Protein Sci.* **24**, 752–761 (2015). doi: 10.1002/pro.2647; pmid: 25808204
46. L. B. Alexandrov *et al.*, ; The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020). doi: 10.1038/s41586-020-1943-3; pmid: 32025018
47. G. Koh, A. Degasperi, X. Zou, S. Momen, S. Nik-Zainal, Mutational signatures: Emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer* **21**, 619–637 (2021). doi: 10.1038/s41568-021-00377-7; pmid: 34316057
48. W. O. Foye, *Review of DNA-binding/damaging small molecules in cancer chemotherapy. Cancer Chemotherapeutic Agents* (American Chemical Society, 1995).
49. P. B. Dervan, Molecular recognition of DNA by small molecules. *Bioorg. Med. Chem.* **9**, 2215–2235 (2001). doi: 10.1016/S0968-0896(01)00262-0; pmid: 11553460
50. W. C. Tse, D. L. Boger, Sequence-selective DNA recognition: Natural products and nature's lessons. *Chem. Biol.* **11**, 1607–1617 (2004). doi: 10.1016/j.chembiol.2003.08.012; pmid: 15610844
51. K. M. Wernke, "Synthetic studies of colibactin, a carcinogenic gut microbiome metabolite," thesis, Yale University (2022).
52. Z. R. Li *et al.*, Macrocyclic colibactin induces DNA double-strand breaks via copper-mediated oxidative cleavage. *Nat. Chem.* **11**, 880–889 (2019). doi: 10.1038/s41557-019-0317-7; pmid: 31527851
53. L. Zha, M. R. Wilson, C. A. Brotherton, E. P. Balskus, Characterization of polyketide synthase machinery from the *pks* island facilitates isolation of a candidate precolibactin. *ACS Chem. Biol.* **11**, 1287–1295 (2016). doi: 10.1021/acschembio.6b00014; pmid: 26890481
54. D. E. Wemmer, P. B. Dervan, Targeting the minor groove of DNA. *Curr. Opin. Struct. Biol.* **7**, 355–361 (1997). doi: 10.1016/S0959-440X(97)80051-6; pmid: 9204277
55. W. A. Denny, DNA minor groove alkylating agents. *Curr. Med. Chem.* **8**, 533–544 (2001). doi: 10.2174/0929867003373283; pmid: 11281840
56. S. Neidle, DNA minor-groove recognition by small molecules. *Nat. Prod. Rep.* **18**, 291–309 (2001). doi: 10.1039/a705982e; pmid: 11476483
57. Z. Morávek, S. Neidle, B. Schneider, Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Res.* **30**, 1182–1191 (2002). doi: 10.1093/nar/30.5.1182; pmid: 11861910
58. A. Rahman, P. O'Sullivan, I. Rozas, Recent developments in compounds acting in the DNA minor groove. *MedChemComm* **10**, 26–40 (2018). doi: 10.1039/C8MD00425K; pmid: 30774852
59. L. H. Hurley, V. L. Reynolds, D. H. Swenson, G. L. Petzold, T. A. Scahill, Reaction of the antitumor antibiotic CC-1065 with DNA: Structure of a DNA adduct with DNA sequence specificity. *Science* **226**, 843–844 (1984). doi: 10.1126/science.6494915; pmid: 6494915
60. D. L. Boger, D. S. Johnson, CC-1065 and the duocarmycins: Unraveling the keys to a new class of naturally derived DNA alkylating agents. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3642–3649 (1995). doi: 10.1073/pnas.92.9.3642; pmid: 7731958
61. R. Rohs *et al.*, The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248–1253 (2009). doi: 10.1038/nature08473; pmid: 19865164
62. T.-P. Chiu, S. Rao, R. S. Mann, B. Honig, R. Rohs, Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. *Nucleic Acids Res.* **45**, 12565–12576 (2017). doi: 10.1093/nar/gkx915; pmid: 29040720
63. A. M. Maxam, W. Gilbert, Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**, 499–560 (1980). doi: 10.1016/S0076-6879(80)65059-9; pmid: 6246368
64. S. A. McLuckey, G. J. Van Berkel, G. L. Glish, Tandem mass spectrometry of small, multiply charged oligonucleotides. *J. Am. Soc. Mass Spectrom.* **3**, 60–70 (1992). doi: 10.1016/1044-0305(92)85019-G; pmid: 24242838
65. M. R. Volpe *et al.*, A small molecule inhibitor prevents gut bacterial genotoxin production. *Nat. Chem. Biol.* **19**, 159–167 (2023). doi: 10.1038/s41589-022-01147-8; pmid: 36253549
66. N. M. O'Boyle *et al.*, Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011). doi: 10.1186/1758-2946-3-33; pmid: 21982300
67. A. W. Schüttelkopf, D. M. van Aalten, PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 1355–1363 (2004). doi: 10.1107/S0907444904011679; pmid: 15272157
68. M. S. Yuan, Q. Fang, (*E*)-1,1′-Dibutyl-3,3′-biindolinyl-idene-2,2′-dione. *Acta Crystallogr. Sect. E Struct. Rep. Online* **67**, o52 (2010). doi: 10.1107/S160053681005066X; pmid: 21522762
69. B. A. Stenfors, R. J. Staples, S. M. Biros, F. N. Ngassa, Crystal structure of 1-[(4-methylbenzene)sulfonyl]pyrrolidine. *Acta Crystallogr. E Crystallogr. Commun.* **76**, 452–455 (2020). doi: 10.1107/S205698902000208X; pmid: 32148893
70. B. J. Esselman *et al.*, Precise equilibrium structure of thiazole (c-C_3H_3NS) from twenty-four isotopologues. *J. Chem. Phys.* **155**, 054302 (2021). doi: 10.1063/5.0057221; pmid: 34364360
71. E. Koag, S. G. Hulse, G. L. Helms, K. M. Call, M. F. Summers, J. Marchant, B. A. Johnson, NMRFx: Integrated Software for NMR Data Processing, Visualization, Analysis and Structure Calculation. bioRxiv 2025.08.26 [Preprint] (2025); https://doi.org/10.1101/2025.08.26.672401.
72. P. Güntert, C. Mumenthaler, K. Wüthrich, Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298 (1997). doi: 10.1006/jmbi.1997.1284; pmid: 9367762
73. B. S. Tolbert *et al.*, Major groove width variations in RNA structures determined by NMR and impact of 13C residual chemical shift anisotropy and 1H-13C residual dipolar coupling on refinement. *J. Biomol. NMR* **47**, 205–219 (2010). doi: 10.1007/s10858-010-9424-x; pmid: 20549304
74. A. Nazemi, A. H. Steeves, D. W. Kastner, H. J. Kulik, Influence of the Greater Protein Environment on the Electrostatic Potential in Metalloenzyme Active Sites: The Case of Formate Dehydrogenase. *J. Phys. Chem. B* **126**, 4069–4079 (2022). doi: 10.1021/acs.jpcb.2c02260; pmid: 35609244
75. R. Mehmood, H. J. Kulik, Both Configuration and QM Region Size Matter: Zinc Stability in QM/MM Models of DNA Methyltransferase. *J. Chem. Theory Comput.* **16**, 3121–3134 (2020). doi: 10.1021/acs.jctc.0c00153; pmid: 32243149
76. C. R. Reinhardt *et al.*, Computational Screening of Putative Catalyst Transition Metal Complexes as Guests in a $Ga_4L_6^{12-}$ Nanocage. *Inorg. Chem.* **63**, 14609–14622 (2024). doi: 10.1021/acs.inorgchem.4c02113; pmid: 39049593

77. P. Bultinck, C. Van Alsenoy, P. W. Ayers, R. Carbó-Dorca, Critical analysis and extension of the Hirshfeld atoms in molecules. *J. Chem. Phys.* **126**, 144111 (2007). doi: 10.1063/1.2715563; pmid: 17444705
78. S. Van Damme, P. Bultinck, S. Fias, Electrostatic Potentials from Self-Consistent Hirshfeld Atomic Charges. *J. Chem. Theory Comput.* **5**, 334–340 (2009). doi: 10.1021/ct800394q; pmid: 26610109
79. T. Lu, F. Chen, Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **33**, 580–592 (2012). doi: 10.1002/jcc.22885; pmid: 22162017
80. C. Lee, W. Yang, R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B Condens. Matter* **37**, 785–789 (1988). doi: 10.1103/PhysRevB.37.785; pmid: 9944570
81. A. D. Becke, Density-Functional Thermochemistry. 3. The Role of Exact Exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993). doi: 10.1063/1.464913
82. P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **98**, 11623–11627 (1994). doi: 10.1021/j100096a001
83. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010). doi: 10.1063/1.3382344; pmid: 20423165
84. R. Ditchfield, W. J. Hehre, J. A. Pople, Self-Consistent Molecular-Orbital Methods. 9. Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, 724–728 (1971). doi: 10.1063/1.1674902
85. P. C. Hariharan, J. A. Pople, The Influence of Polarization Functions on Molecular-Orbital Hydrogenation Energies. *Theor. Chim. Acta* **28**, 213–222 (1973). doi: 10.1007/BF00533485
86. W. J. Hehre, R. Ditchfield, J. A. Pople, Self-Consistent Molecular-Orbital Methods. 12. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular-Orbital Studies of Organic-Molecules. *J. Chem. Phys.* **1972**, 2257–2261 (1972). doi: 10.1063/1.1677527
87. M. M. Francl *et al.*, Self-Consistent Molecular-Orbital Methods. 23. A Polarization-Type Basis Set for 2nd-Row Elements. *J. Chem. Phys.* **77**, 3654–3665 (1982). doi: 10.1063/1.444267
88. P. Villalta *et al.*, Data from: "The specificity and structure of DNA cross-linking by the gut bacterial genotoxin colibactin," Dryad (2025); https://doi.org/10.5061/dryad.vmcvdnd5g.
89. R. Haslecker, CYANA and Xplor run files for: "The specificity and structure of DNA cross-linking by the gut bacterial genotoxin colibactin," Zenodo (2025); https://doi.org/10.5281/zenodo.17114378.
90. E. Carlson *et al.*, Computational Data Information for "The specificity and structure of DNA cross-linking by the gut bacterial genotoxin colibactin," Zenodo (2025); https://doi.org/10.5281/zenodo.17075479.

ACKNOWLEDGMENTS

We acknowledge P. Boudreau and J. Wong for performing preliminary experiments. MS was performed in the Analytical Biochemistry Shared Resource (ABSR) of the University of Minnesota Masonic Cancer Center (MCC). We thank B. Carlson at the MCC for editorial assistance. We acknowledge M. Manetsch for providing the script to compute ESP values. E.P.B. is a Howard Hughes Medical Institute (HHMI) investigator. **Funding:** National Institutes of Health grant R01CA208834 (E.P.B.); National Institutes of Health fellowship F32CA254165 (E.S.C.); National Institutes of Health grant R35GM152027 (H.J.K.); National Institutes of Health grant R50CA211256 (P.W.V.); National Institutes of Health grant P30CA077598 ABSR; National Science Foundation grant CBET-1846426 (H.J.K.); HHMI grant 55108516 (V.M.D.); National Institutes of Health grant R01GM123012 (B.A.J.). **Author contributions:** Conceptualization: E.P.B., E.S.C.; Methodology: P.W.V., L.H., R.H., E.S.C., B.A.J.; Investigation: P.W.V., S.B., L.H., R.H., V.V., E.S.C., C.L., A.S., M.A.A.R., E.S.M.; Visualization: E.S.C.; Funding acquisition: E.P.B., V.M.D., E.S.C.; Project administration: E.P.B., V.M.D.; Supervision: E.P.B., S.B., V.M.D., H.J.K.; Writing – original draft: E.P.B., R.H., V.M.D., E.S.C.; Writing – review & editing: E.P.B., P.W.V., S.B., R.H., V.M.D., H.J.K., V.V., E.S.C., C.L., M.A.A.R. **Competing interests:** E.P.B. is an inventor on patent applications related to detection and inhibition of colibactin biosynthesis (US patents 11,617,759; 12,115,176; and 11,040,951). All other authors declare that they have no competing interests. **Data and materials availability:** Atomic coordinates have been deposited in the Protein Data Bank under accession codes pdb_00009oso. Chemical shifts have been deposited in the Biological Magnetic Resonance Data Bank under accession code 31249. MS data has been submitted to Dryad (*88*). Scripts and calculations related to the NMR structure have been uploaded to Zenodo (*89*). Optimized colibactin geometries, input files, and ESP calculation scripts have been uploaded to Zenodo (*90*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ady3571
Supplementary Text; Figs. S1 to S23; Tables S1 to S5; Appendix S1; References (*91*–*94*); MDAR Reproducibility Checklist

Submitted 3 June 2025; accepted 3 October 2025

10.1126/science.ady3571