Title: The specificity and structure of DNA crosslinking by the gut bacterial genotoxin colibactin

Authors: 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 D'Souza[2]*, Emily P. Balskus[1,8,9]*

Affiliations:

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

[2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

[3]Masonic Cancer Center University of Minnesota, Minneapolis, MN 55455, USA

[4]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[5]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[6]Structural Biology Initiative, City University of New York (CUNY) Advanced Science Research Center, New York, NY 10031, USA

[7]Department of Medicinal Chemistry, University of Minnesota, Minneapolis, MN 55455, USA

[8]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA

[9]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

*Corresponding authors. Email: balskus@chemistry.harvard.edu, dsouza@g.harvard.edu, villa001@umn.edu, balbo006@umn.edu

†Contributed equally

Current affiliation: Department of Chemistry, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

Abstract: 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 crosslinks (ICLs). However, we currently lack direct information regarding the structure of this lesion. Here, we combine mass spectrometry and nuclear magnetic resonance spectroscopy to elucidate the specificity and structure of the colibactin-DNA ICL. We find that colibactin alkylates within the minor groove of AT-rich DNA, explaining the origins of mutational signatures. Unexpectedly, we discover that the chemically unstable central motif of colibactin mediates the sequence specificity of crosslinking. 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.

Main Text:

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. 1 and Fig. S1A) produced by commensal Enterobacteriaceae, including strains of *E. coli*, that harbor the *pks* (or *clb*) gene cluster (*5*). This gene cluster encodes a biosynthetic pathway that employs 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*). Importantly, *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 crosslinks (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), important 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 T>C single-base substitutions and indels that occur within AT-rich sequence motifs (i.e. 5'-AAWWTT-3' where W = 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 the colibactin-DNA ICL spans 4-5.5 Å (or 3-4 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 5-20% of CRC genomes, occur in driver genes such as *APC*, and are correlated with early-onset CRC (*25-34*). We also recently identified colibactin-DNA adducts in human colonoscopy samples (*35*). Together, these

data indicate that humans are exposed to colibactin, strengthening its connection to cancer development. They also provide indirect evidence that colibactin crosslinks 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 crosslinked 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 crosslinked efficiently. Systematically truncating this 500mer identified a 50mer that is crosslinked (71% ICL formation) after a 5 h incubation at 37 ºC with *pks*$^+$ *E. coli* in DMEM-HEPES medium, pH 7.4 (Fig. 1B-C). No ICL formation was observed when incubating the 50mer with an isogenic *pks*$^-$ *E. coli* strain. We noted that this 50mer contained an AT-rich motif (5′-AAATTAATA-3′) that is consistent with previously reported colibactin mutational signatures (*25, 26*), leading us to hypothesize 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, Fig. S2). Colibactin-ICLs are thermally unstable and spontaneously depurinate to form abasic sites when heated to 90 ºC (*36, 37*). 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. 1D-E). We applied this assay to a *pks*$^+$ *E. coli*-exposed 25mer containing the AT-rich motif identified from the 50mer. 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**A<u>TT</u>AATA-3′ (where alkylated nucleotides are bolded or opposite to underlined nucleotides) (Fig. 1D-E, Fig. S2,3). The accuracy of our LC-MS approach was confirmed by subjecting a 5′-6-carboxyfluorescein-labeled 50mer to our crosslinking and strand cleavage protocol and determining alkylation site locations using traditional Maxam–Gilbert gel sequencing (Fig. S4-6) (*38*).

This result suggests the colibactin-DNA ICL spans 3-5 bp. To determine the precise length of the ICL, we repeated our crosslinking and strand cleavage assays with a 50mer and 25mer containing the sequence 5′-AATATTATA-3′ which has the potential to form ICLs between pairs of adenines 2 – 8 bp apart. The results of this assay indicated colibactin forms a single ICL between two adenines spanning 4 bp (5′-A**A**TA<u>T</u>TATA-3′). This is consistent with previous studies showing colibactin-induced DSBs have 2 bp overhangs (*26*) and suggests that the four sites of alkylation within the 5′-AAATTAATA-3′ motif we initially tested likely correspond to a mixture of two colibactin-DNA ICLs (Fig. 1D-E, Fig. S4). The extent of ICL formation was reduced compared to the initial sequence perhaps due to this sequence possessing only one recognition site (Fig. 1C-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′-A<u>T</u>TA**A**TATA-3′); if not, we would expect the ICL to shift to a new location (5′-ATTA**A**TA<u>T</u>A-3′). ICL formation and strand cleavage analysis revealed the ICL in the new location, suggesting colibactin preferentially crosslinks at a TA base pair three positions downstream from the initial A (Fig. 1C-E). Finally, we detected little to no alkylation of dsODNs containing 5′-ATTAATAAA-3′, which has no 5′-AWWT-3′ motifs (Fig. 1C-E). Taken together, this data suggests that colibactin-DNA ICLs form at 5′-W**A**WW<u>T</u>W-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 crosslinked at comparable levels to 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. S7C-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. Based on the structural characterization of colibactin-DNA adducts, colibactin is thought to alkylate *N3* of adenine, which is accessible via 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 50mer 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 these molecules had no effect on bacterial growth or colibactin production (Fig. S9-10). AT-rich minor groove binders (netropsin and DAPI) (*39, 40*) inhibited ICL formation in a dose-dependent manner (Fig. 2). By contrast, a major groove binder (methyl green) (*41*) and a GC-rich minor groove binder (actinomycin D) (*42*) 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 *N*3-adenine colibactin-DNA adduct (*18*). Additional adenine adducts containing the other 'half' of colibactin were identified using mass spectrometry but were not characterized by NMR spectroscopy, leaving their connectivity unknown (*21, 22*). Hypothesizing that both colibactin cyclopropanes alkylate the *N*3 position of adenine, which would be consistent with selective minor groove binding, we incubated *pks*$^{+}$ *E. coli* with 50mers containing a 5′-A**A**TATT-3′ sequence motif in which the bolded residue was replaced with *N*3-deaza-deoxyadenosine (dAdo) (*43*), eliminating the possibility for *N*3-dAdo adducts to form. ICL formation was abolished when *N*3-deaza-dAdo was incorporated into the forward, reverse, or both strands (Fig. S11). Strand cleavage analysis of assays with 25mers containing these sequence variants also showed no alkylation of *N*3-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. S11C-D). Taken together, these data indicate colibactin exclusively forms bis-*N*3-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 bp 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 crosslinked, 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 *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, 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, 44*). 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 via MS helps to resolve important aspects of colibactin's structure.

To understand the molecular details of how colibactin binds and crosslinks DNA, we used solution-state NMR. This required generating the colibactin-DNA ICL on a large scale using *pks*$^+$ *E. coli*. We chose a 14mer dsODN with a palindromic sequence (5′-CGCGA**A**TA<u>T</u>TCGCG-3′) and substituted A6 with 2′-fluoro-deoxyadenosine to increase the glycolytic bond strength and minimize depurination (*45*). 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 crosslinked DNA (~55-65%) and free DNA. Thus, we first assigned the spectra of free DNA (with and without 2′-fluoro-A6 modification), which allowed for comparative, unambiguous assignments of DNA chemical shifts that are perturbed upon crosslinking to colibactin. Overall, we do not observe any substantial changes in base-to-ribose NOE walks throughout the DNA molecule, indicating that the overall B-form groove parameters are maintained upon colibactin crosslinking. 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. In comparison to 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 ppm, and the H8 by ~0.83 and ~0.78 ppm, respectively (Fig. S14A). In contrast, the average change for the other 2′-deoxyadenosines in the DNA was less than 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 $T10_a$-A5b, 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 A6a and A6b H8 differ by 0.05 ppm, indicating that they are in slightly different environments (Fig. S14A). Second, while 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 observe 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: 1) no connections from the terminal pyrrolidine rings; 2) the thiazole hydrogens C34H and C39H give connections to the outer deoxyribose hydrogens 9a and 10b (H4'/H5'), respectively; 3) 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. 4B,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 outwards. 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. 4B,D and Fig. S14C). The curved shape and close contacts between its heterocyclic aromatic rings and the hydrophobic surfaces of the minor groove walls suggests shape complementarity as well as van der Waals and hydrophobic interactions all contribute to colibactin-DNA binding. As discussed below, similar features are 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 confirm the presence of a nitrogen in the central scaffold of colibactin, surprisingly we find this nitrogen within an iminium functional group. Two distinct proton peaks are correlated with N37 in the ^{15}N-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 A8a and electrostatic interactions with O2 of T7b, O2 of T7, and N3 of A8b. Second, the equivalent colibactin H13 and H13' amide protons show connectivities to the flanking T9a and T10b 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 A5bH1'/C11aH2' and A5aH1'/ T10bH2' 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 S15). This orientation would position the cyclopropane rings of the colibactin warheads in close proximity 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. Importantly, we see no evidence for intramolecular cyclization of colibactin (*24*), indicating 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 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. Finally, 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 spiro-cyclopropanes 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, exchanging the two central TA base pairs for CG base pairs should diminish alkylation due 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. Strikingly, 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. Finally, we compared the reactivity of colibactin and the synthetic analog toward additional oligos 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 crosslinked to DNA (Fig. 6A and Fig. S18,19). We observe the highest ESP on the iminium nitrogen atom of colibactin bearing a central α-ketoiminium (ca. 664 $kJ \cdot mol^{-1} \cdot e^{-1}$), consistent with the α-ketoiminium having a positive charge. We calculate a lower ESP on the iminium nitrogen of the α-ketoiminium when this proposed colibactin is crosslinked to DNA compared to that of the corresponding free colibactin (Fig. 6B-C and Fig. S20,21). This decrease could indicate stabilization by electrostatic or other non-covalent (e.g. hydrogen bonding) interactions with the DNA. However, we do not observe a substantial decrease in ESP of the heavy atoms of the

central functional groups in the other proposed colibactin structures when they are crosslinked to DNA, suggesting 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 find 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 of N37 of colibactin and its protonation (Fig. 6D) (*46*). Substitution of a single central base pair with a CG base pair increases ESP at this position while substituting both central base pairs, which abolished ICL formation by colibactin, results in a loss of this ESP differential (Fig. 6E,F). Modeling interactions between colibactin and these different sequences shows distortions when guanines are introduced due to repulsions from the large amino moieties of these central base pairs (Fig. 6G-I). The outer base pairs flanking the alkylation site ($A5_a$ - $T10_b$ and $T10_a$ - $A5_b$) also contribute to the negative ESP, and modeling GC base pair substitutions at these positions also results 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). In summary, this structure, experimental data, and computations reveal and explain colibactin's specificity for DNA alkylation, highlighting an especially important 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 (*47, 48*), 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 (*49*). 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 (*50, 51*). The high prevalence of *pks*$^+$ *E. coli* in many geographic locations (*34*), the accumulating evidence linking colibactin to CRC development (*5-14*, *25-34*), 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, 52*), though the timing of this oxidation relative to DNA alkylation is unclear. The presence of this central functional group was unexpected due 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, 44*), 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 based on the identification of shunt products that the colibactin biosynthetic pathway produces multiple metabolites with the potential to cause DNA damage (*53*). Our observation of a single major ICL-forming species by MS and NMR calls this into question and suggests 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 which 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-CoA (*54*), making them 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. Finally, 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 spiro-cyclopropane 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 (*55-59*), including the polyamide DNA-binding compounds netropsin and distamycin (*39*), as well as the DNA alkylating duocarmycins, which form monoadducts (*60, 61*). However, colibactin uniquely combines features of AT-rich minor-groove binding specificity with the capacity for interstrand crosslink 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 (*62, 63*). Interestingly, 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 (*58*) as well as the terminal amidine of the minor groove-binding natural product netropsin (*39*) (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. Strikingly, the pseudosymmetric structure of colibactin places the positively charged iminium equidistant from

the two electrophilic warheads, suggesting 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 find this analog has an altered sequence specificity compared to colibactin produced by *pks*$^+$ *E. coli*. This dramatic difference suggests 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 fascinating questions of why such a chemically unstable structural motif would have evolved to play an important role in colibactin's biological activity and how it is stabilized and/or protected within the cellular environment. Finally, 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 ºC 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 ºC 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 ºC for 5 minutes in a heat block, then slowly cooled to room temperature over ~2 h. 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 crosslinking assay – bacterial incubations

The protocol was adapted from previously reported assays (*20,24,36*). 2 µg of the desired oligo 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 x 10^6 cells) and incubated at 37 ºC

aerobically for 5 h 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 x g for 3 min (4 ºC) 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 ºC for ~16 h. The precipitated DNA was isolated by centrifugation at 16,100 x g for 20 min (4 ºC) 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 x 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 ºC to reduce degradation of colibactin-ICLs.

DNA crosslinking 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 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 ºC for 1 hour without shaking and were used immediately for electrophoresis.

Interstrand crosslink analysis by denaturing gel electrophoresis

The protocol was adapted from previously reported methods (*36*). 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 crosslinking 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 crosslinking protocol described above except that the resulting DNA pellets were redissolved in 45 µL of TE buffer. Samples were then heated to 90 ºC for 1 h 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 ºC overnight to induce DNA precipitation. Precipitated DNA was pelleted by centrifugation (20 min @ 16,100 x 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 ºC for 30 min, cooled to room temperature, and evaporated to dryness using a Labconco Centrivap (45 ºC for ~1 h). 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 ºC until further analysis. 5′-FAM-labeled DNA samples were analyzed by sequencing polyacrylamide gel electrophoresis while unlabeled samples were subjected to liquid-chromatography high-resolution accurate-mass mass spectrometry (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 x 16 in x 0.75 mm gel which solidified over 1 h. 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 ºC for 5 min. The corresponding A+G Maxam-Gilbert ladders were generated by a previously reported method (*64*). 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 h. 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.

Analysis of intact colibactin-DNA interstrand crosslinks 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 x 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. For intact sample analysis, 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. For cleavage samples, 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 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.

<u>Analysis of colibactin-induced strand cleavage products by LC-MS</u>

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 x 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. For intact sample analysis, 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. For cleavage samples, 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 (*65*). 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 crosslinking assay in the presence of groove binders

AAATTAATA-50mer (2 µg) was subjected to a modified version of the DNA crosslinking 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 (*66*). Cell pellets obtained from the DNA crosslinking 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 (*34*). 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 x 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 ºC 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 x 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 multi-reaction 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 crosslinking assay described above except incubations were performed for 16 h at 30 ºC. 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 x 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 ºC overnight.

DNA-precipitated samples were centrifuged at 16,100 x g for 30 min (4 ºC) 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 x 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 x g for 30 min, Millipore) and pooled.

This entire protocol was repeated twice to generate a total of ~410 µg of partially crosslinked DNA from 3 96-well plates. All DNA from *pks*$^+$ samples was combined to generate the final sample, which was stored at –20 ºC 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 crosslinks 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 crosslinks 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 crosslinked 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 (*67*).

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 (*68*). 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 (*69-71*). 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 ºK and for exchangeable protons in 90% H_2O an at 278 ºK. ^1H-^1H NOESY and TOCSY 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 (*72*).

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 a strong, medium, and weak intensities, respectively (*73*). 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 ($\alpha = -68°$, $\beta = -147°$, $\gamma = 46°$, $\delta = 135°$, $\varepsilon = -150°$, $\zeta = -100°$) (*74*). 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.

<u>Computational Modeling</u>

Calculations of colibactin electrostatic potential

We computed the electrostatic potential (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 (*75-77*). We computed the partial charges on atoms of interest using iterative Hirshfield (Hirshfield-I) charges (*78, 79*) as implemented in Multiwfn version 3.7 (*80*) and applied these charges to compute the ESP at these atoms (*77*). All ESP values were obtained from geometry optimization calculations carried out with density functional theory (DFT) at the B3LYP (*81-83*) -D3 (*84*) /6-31G* (*85–88*) 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) (*89*) 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 via NMR (see above) and substituted the central or flanking base pairs for generating initial models of non-ideal motifs crosslinked 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).
2. M. T. White, C. L. Sears, The microbial landscape of colorectal cancer. *Nat. Rev. Microbiol.* **22**, 240–254 (2024).
3. M. W. Dougherty, C. Jobin, Shining a light on colibactin biology. *Toxins (Basel)* **13**, 346 (2021).
4. K. M. Wernke, M. Xue, A. Tirla, C. S. Kim, J. M. Crawford, S. B. Herzon, Structure and bioactivity of colibactin. *Bioorg. Med. Chem. Lett.* **30**, 127280 (2020).
5. J.-P. Nougayrède, S. Homburg, F. Taieb, M. Boury, E. Brzuszkiewicz, G. Gottschalk, C. Buchrieser, J. Hacker, U. Dobrindt, E. Oswald, *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848–851 (2006).
6. G. Cuevas-Ramos, C. R. Petit, I. Marcq, M. Boury, E. Oswald, J.-P. Nougayrède, *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. USA* **107**, 11537–11542 (2010).
7. A. Cougnoux, G. Dalmasso, R. Martinez, E. Buc, J. Delmas, L. Gibold, P. Sauvanet, C. Darcha, P. Déchelotte, M. Bonnet, D. Pezet, H. Wodrich, A. Darfeuille-Michaud, R. Bonnet, Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**, 1932–1942 (2014).
8. J. C. Arthur, E. Perez-Chanona, M. Mühlbauer, S. Tomkovich, J. M. Uronis, T. J. Fan, B. J. Campbell, T. Abujamel, B. Dogan, A. B. Rogers, J. M. Rhodes, A. Stintzi, K. W. Simpson, J. J. Hansen, T. O. Keku, A. A. Fodor, C. Jobin, Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
9. M. Bonnet, E. Buc, P. Sauvanet, C. Darcha, D. Dubois, B. Pereira, P. Déchelotte, R. Bonnet, D. Pezet, A. Darfeuille-Michaud, Colonization of the human gut by *E. coli* and colorectal cancer risk. *Clin. Cancer Res.* **20**, 859–867 (2014).
10. S. Tomkovich, Y. Yang, K. Winglee, J. Gauthier, M. Mühlbauer, X. Sun, M. Mohamadzadeh, X. Liu, P. Martin, G. P. Wang, E. Oswald, A. A. Fodor, C. Jobin, Locoregional effects of microbiota in a preclinical model of colon carcinogenesis. *Cancer Res.* **77**, 2620–2632 (2017).
11. C. M. Dejea, P. Fathi, J. M. Craig, A. Boleij, R. Taddese, A. L. Geis, X. Wu, C. E. DeStefano Shields, E. M. Hechenbleikner, D. L. Huso, R. A. Anders, F. M. Giardiello, E. C. Wick, H. Wang, S. Wu, D. M. Pardoll, F. Housseau, C. L. Sears, Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592–597 (2018).
12. T. Miyasaka, T. Yamada, K. Uehara, H. Sonoda, A. Matsuda, S. Shinji, R. Ohta, S. Kuriyama, Y. Yokoyama, G. Takahashi, T. Iwai, K. Takeda, K. Ueda, S. Kanaka, R. Ohashi, H. Yoshida, *Pks*-positive *Escherichia coli* in tumor tissue and surrounding normal mucosal tissue of colorectal cancer patients. *Cancer Sci.* **115**, 1184–1195 (2024).
13. E. Buc, D. Dubois, P. Sauvanet, J. Raisch, J. Delmas, A. Darfeuille-Michaud, D. Pezet, R. Bonnet, High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. *PLoS One* **8**, e56964 (2013).
14. V. Eklöf, A. Löfgren-Burström, C. Zingmark, S. Edin, P. Larsson, P. Karling, O. Alexeyev, J. Rutegård, M. L. Wikberg, R. Palmqvist, Cancer-associated fecal microbial markers in colorectal cancer detection. *Int. J. Cancer* **141**, 2528–2536 (2017).

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).
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 aminocyclopropanecarboxylic acid moiety. *Chem. Sci.* **6**, 3154–3160 (2015).
17. M. I. Vizcaino, J. M. Crawford, The colibactin warhead crosslinks DNA. *Nat. Chem.* **7**, 411–417 (2015).
18. M. Wilson, Y. Jiang, P. W. Villalta, A. Stornetta, P. D. Boudreau, A. Carrà, C. A. Brennan, E. Chun, L. Ngo, L. D. Samson, B. P. Engelward, W. S. Garrett, S. Balbo, E. P. Balskus, The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019).
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).
20. N. Bossuet-Greif, J. Vignard, F. Taieb, G. Mirey, D. Dubois, C. Petit, E. Oswald, J.-P. Nougayrède, The colibactin genotoxin generates DNA interstrand crosslinks in infected cells. *mBio* **9**, e02393-17 (2018).
21. Y. Jiang, A. Stornetta, P. W. Villalta, M. R. Wilson, P. D. Boudreau, L. Zha, S. Balbo, E. P. Balskus, Reactivity of an unusual amidase may explain colibactin's DNA cross-linking activity. *J. Am. Chem. Soc.* **141**, 11489–11496 (2019).
22. M. Xue, C. S. Kim, A. R. Healy, K. M. Wernke, Z. Wang, M. C. Frischling, E. E. Shine, W. Wang, S. B. Herzon, J. M. Crawford, Structure elucidation of colibactin and its DNA cross-links. *Science* **365**, eaax2685 (2019).
23. A. R. Healy, K. M. Wernke, C. S. Kim, N. R. Lees, J. M. Crawford, S. B. Herzon, Synthesis and reactivity of precolibactin 886. *Nat. Chem.* **11**, 890–898 (2019).
24. K. M. Wernke, A. Tirla, M. Xue, Y. V. Surovtseva, F. S. Menges, S. B. Herzon, 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).
25. C. Pleguezuelos-Manzano, J. Puschhof, A. Rosendahl Huber, A. van Hoeck, H. M. Wood, J. Nomburg, C. Gurjao, F. Manders, G. Dalmasso, P. B. Stege, F. L. Paganelli, M. H. Geurts, J. Beumer, T. Mizutani, Y. Miao, R. van der Linden, S. van der Elst, Genomics England Research Consortium, K. C. Garcia, J. Top, R. J. L. Willems, M. Giannakis, R. Bonnet, P. Quirke, M. Meyerson, E. Cuppen, R. van Boxtel, H. Clevers, Mutational signature in colorectal cancer caused by genotoxic *pks*+ *E. coli*. *Nature* **580**, 269–273 (2020).
26. P. J. Dziubańska-Kusibab, H. Berger, F. Battistini, B. A. M. Bouwman, A. Iftekhar, R. Katainen, T. Cajuso, N. Crosetto, M. Orozco, L. A. Aaltonen, T. F. Meyer, Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat. Med.* **26**, 1063–1069 (2020).
27. A. Boot, A. W. T. Ng, F. T. Chong, S. C. Ho, W. Yu, D. S. W. Tan, N. G. Iyer, S. G. Rozen, 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).
28. D. Terlouw, A. Boot, Q. R. Ducarmon, S. Nooij, M. Suerink, M. E. van Leerdam, D. van Egmond, C. M. Tops, R. D. Zwittink, D. Ruano, A. M. J. Langers, M. Nielsen, T. van

Wezel, H. Morreau, Enrichment of colibactin-associated mutational signatures in unexplained colorectal polyposis patients. *BMC Cancer* **24**, 104 (2024).
29. A. R. Huber, C. Pleguezuelos-Manzano, J. Puschhof, J. Ubels, C. Boot, A. Saftien, M. Verheul, L. T. Trabut, N. Groenen, M. van Roosmalen, K. S. Ouyang, H. Wood, P. Quirke, G. Meijer, E. Cuppen, H. Clevers, R. van Boxtel, Improved detection of colibactin-induced mutations by genotoxic *E. coli* in organoids and colorectal cancer. *Cancer Cell* **42**, 487–496.e6 (2024).
30. D. Terlouw, M. Suerink, A. Boot, T. van Wezel, M. Nielsen, H. Morreau, 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).
31. B. Chen, D. Ramazzotti, T. Heide, I. Spiteri, J. Fernandez-Mateos, C. James, L. Magnani, T. A. Graham, A. Sottoriva, Contribution of *pks*+ *E. coli* mutations to colorectal carcinogenesis. *Nat. Commun.* **14**, 7827 (2023).
32. 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 (2024).
33. M. Díaz-Gay et al. Geographic and age variations in mutational processes in colorectal cancer. *Nature* **643**, 230 (2025).
34. 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).
35. A. Mandarino Alves, C. Lecchi, S. Lopez, A. Stornetta, P. P. Mathai, P. W. Villalta, S. Ishii, E. P. Balskus, S. Balbo, A. Khoruts, Dysfunctional mucus structure in cystic fibrosis increases vulnerability to colibactin-mediated DNA adducts in the colon mucosa. *Gut Microbes* **16**, 2387877 (2024).
36. M. Xue, K. M. Wernke, S. B. Herzon, Depurination of colibactin-derived interstrand cross-links. *Biochemistry* **59**, 892–900 (2020).
37. M. R. Volpe, M. R. Wilson, C. A. Brotherton, E. S. Winter, S. E. Johnson, E. P. Balskus, In vitro characterization of the colibactin-activating peptidase ClbP enables development of a fluorogenic activity probe. *ACS Chem. Biol.* **14**, 1097–1101 (2019).
38. A. M. Maxam, W. Gilbert, A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**, 560–564 (1977).
39. 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. USA* **82**, 1376–1380 (1985).
40. 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).
41. S. K. Kim, B. Nordén, Methyl green. A DNA major-groove binding drug. *FEBS Lett.* **315**, 61–64 (1993).
42. 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).
43. R. Cosstick, X. Li, D. K. Tuli, D. M. Williams, B. A. Connolly, P. C. Newman, Molecular recognition in the minor groove of the DNA helix. Studies on the synthesis of oligonucleotides and polynucleotides containing 3-deaza-2′-deoxyadenosine. *Nucleic Acids Res.* **18**, 4771–4778 (1990).
44. T. Zhou, Y. Hirayama, Y. Tsunematsu, N. Suzuki, S. Tanaka, N. Uchiyama, Y. Goda, Y. Yoshikawa, Y. Iwashita, M. Sato, N. Miyoshi, M. Mutoh, H. Ishikawa, H. Sugimura, K. Wakabayashi, K. Watanabe, 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).
45. M. Ikehara, 2'-Substituted 2'-deoxypurine nucleotides: Their conformation and properties. *Heterocycles* **21**, 75–90 (1984).
46. C. A. Fitch, G. Platzer, M. Okon, B. E. Garcia-Moreno, L. P. McIntosh Arginine: Its pKa value revisited. *Protein Sci.* **24**, 752-761 (2015).
47. L. B. Alexandrov, J. Kim, N. J. Haradhvala, et al., The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
48. 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).
49. W. O. Foye, Review of DNA-binding/damaging small molecules in cancer chemotherapy. Cancer Chemotherapeutic Agents. American Chemical Society (1995).
50. P. B. Dervan, Molecular recognition of DNA by small molecules. *Bioorg. Med. Chem.* **9**, 2215–2235 (2001).
51. W. C. Tse, D. L. Boger, Sequence-selective DNA recognition: natural products and nature's lessons. *Chem. Biol.* **11**, 1607–1617 (2004).
52. K. M. Wernke, Synthetic studies of colibactin, a carcinogenic gut microbiome metabolite. Thesis, Yale University (2022).
53. Z. R. Li, J. Li, W. Cai, J. Y. H. Lai, S. M. K. McKinnie, W. P. Zhang, B. S. Moore, W. Zhang, P. Y. Qian, Macrocyclic colibactin induces DNA double-strand breaks via copper-mediated oxidative cleavage. *Nat. Chem.* **11**, 880–889 (2019).
54. 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).
55. D. E. Wemmer, P. B. Dervan, Targeting the minor groove of DNA. *Curr. Opin. Struct. Biol.* **7**, 355–361 (1997).
56. W. A. Denny, DNA minor groove alkylating agents. *Curr. Med. Chem.* **8**, 533–544 (2001).
57. S. Neidle, DNA minor-groove recognition by small molecules. *Nat. Prod. Rep.* **18**, 291–309 (2001).
58. Z. Morávek, S. Neidle, B. Schneider, Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Res.* **30**, 1182–1191 (2002).
59. A. Rahman, P. O'Sullivan, I. Rozas, Recent developments in compounds acting in the DNA minor groove. *Medchemcomm* **10**, 26–40 (2018).
60. 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).
61. 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. USA* **92**, 3642–3649 (1995).
62. R. Rohs, S. West, A. Sosinsky, P. Liu, R. S. Mann, B. Honig, The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
63. 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).
64. A. M. Maxam, W. Gilbert, Sequencing end-labeled DNA with base-specific chemical cleavages. *Meth. Enzymol*. **65**, 499–560 (1980).

65. 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).
66. M. R. Volpe, J. A. Velilla, M. Daniel-Ivad, J. J. Yao, A. Stornetta, P. W. Villalta, H. C. Huang, D. A. Bachovchin, S. Balbo, R. Gaudet, E. P. Balskus, A small molecule inhibitor prevents gut bacterial genotoxin production. *Nat. Chem. Biol.* **19**,159–167 (2023).
67. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
68. 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).
69. 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).
70. 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).
71. B. J. Esselman, M. A. Zdanovskaia, A. N. Owen, J. F. Stanton, R. C. Woods, R. J. McMahon, Precise equilibrium structure of thiazole (c-C3H3NS) from twenty-four isotopologues. *J. Chem. Phys.* **155**, 054302 (2021).
72. 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 2025.08.26.672401.
73. P. Güntert, C. Mumenthaler, K. Wüthrich, Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–98 (1997).
74. B. S. Tolbert, Y. Miyazaki, S. Barton, B. Kinde, P. Starck, R. Singh, A. Bax, D. A. Case, M. F. Summers, 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).
75. 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).
76. 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).
77. C. R. Reinhardt, M. T. Manetsch, W.-L. Li, Y. Román-Leshkov, T. Head-Gordon, H. J. Kulik, Computational Screening of Putative Catalyst Transition Metal Complexes as Guests in a Ga4L612– Nanocage. *Inorg. Chem.* **63**, 14609–14622 (2024).
78. P. Bultinck, C. V. Alsenoy, P. W. Ayers, R. Carbó-Dorca, Critical Analysis and Extension of the Hirshfeld Atoms in Molecules. *J. Chem. Phys.* **126,** 144111 (2007).
79. S. V. Damme, P. Bultinck, S. Fias, Electrostatic Potentials from Self-Consistent Hirshfeld Atomic Charges. *J. Chem. Theory Comput.* **5**, 334-340 (2009).
80. T. Lu, F. Chen, Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **33**, 580–592 (2011).
81. C. Lee, W. Yang, R. G. Parr, Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Physical Review B* **37**, 785–789 (1988).
82. A. D. Becke, Density-Functional Thermochemistry .3. The Role of Exact Exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).

83. 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).
84. 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).
85. 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**, **54**, 724–728 (1971).
86. P. C. Hariharan, J. A. Pople, The Influence of Polarization Functions on Molecular-Orbital Hydrogenation Energies. *Theor. Chim. Acta.* **28**, 213–222 (1973).
87. 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**, **56**, 2257–2261 (1972).
88. M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. Defrees, J. A. Pople, Self-Consistent Molecular-Orbital Methods .23. A Polarization-Type Basis Set for 2nd-Row Elements. *J. Chem. Phys.* **77**, 3654–3665 (1982).
89. 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).
90. M. Evich, A. M. Spring-Connell, M. W. Germann, Impact of modified ribose sugars on nucleic acid conformation and function. *Heterocyclic Commun.* **23**, 155–165 (2017).
91. R. F. W. Bader, A Quantum-Theory of Molecular-Structure and Its Applications. *Chem. Rev.* **91**, 893–928 (1991).
92. R. F. W. Bader, Atoms in molecules. *Acc. Chem. Res.* **18**, 9–15 (1985).
93. E. Espinosa, E. Molins, C. Lecomte, Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities. *Chem. Phys. Lett.* **285**, 170–173 (1998).

Acknowledgments: We acknowledge Paul Boudreau and Joel Wong for performing preliminary experiments. Mass spectrometry was performed in the Analytical Biochemistry Shared Resource (ABSR) of the University of Minnesota Masonic Cancer Center (MCC). We thank Bob Carlson at the MCC for editorial assistance. We acknowledge Melissa Manetsch for providing the script to compute electrostatic potential (ESP) values. EPB is a Howard Hughes Medical Institute Investigator.

Funding:

National Institutes of Health grant R01CA208834 (EPB)

National Institutes of Health fellowship F32CA254165 (ESC)

National Institutes of Health grant R35GM152027 (HJK)

National Institutes of Health grant R50CA211256 (PWV)

National Institutes of Health grant P30CA077598 (ABSR)

National Science Foundation grant CBET-1846426 (HJK)

HHMI grant 55108516 (VMD)

National Institutes of Health grant R01GM123012 (BAJ)

Author contributions:

Conceptualization: EPB, ESC

Methodology: PWV, LH, RH, ESC, BAJ

Investigation: PWV, SB, LH, RH, VV, ESC, CL, AS, MAR, ESM

Visualization: ESC

Funding acquisition: EPB, VMD, ESC

Project administration: EPB, VMD

Supervision: EPB, SB, VMD, HJK

Writing – original draft: EPB, RH, VMD, ESC

Writing – review & editing: EPB, PWV, SB, RH, VMD, HJK, VV, ESC, CL, MAR

Competing interests:

EPB is an inventor on patent applications related to detection and inhibition of colibactin biosynthesis (U.S. Patent 11,617,759, U.S. Patent 12,115,176, U.S. Patent 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 codes 31249. Mass spectrometry data has been submitted to Dryad: https://doi.org/10.5061/dryad.vmcvdnd5g. Scripts and calculations related to the NMR structure have been uploaded to: https://doi.org/10.5281/zenodo.17114378. Optimized colibactin geometries, input files, and ESP calculation scripts have been uploaded to: https://doi.org/10.5281/zenodo.17075479.

Supplementary Materials

Supplementary Text

Figs. S1 to S23

Tables S1 to S5

References (*90–93*)

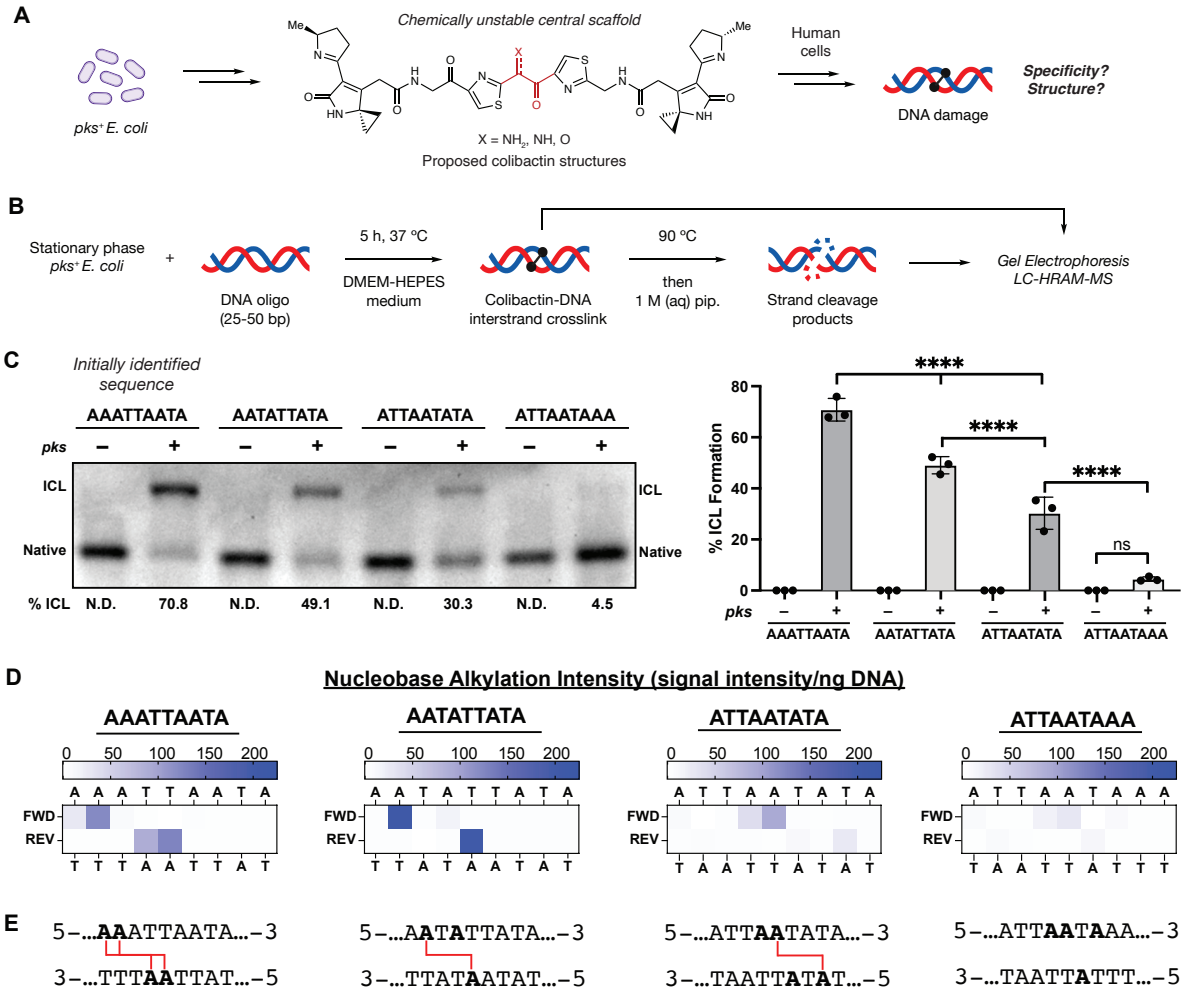

Fig. 1. The chemically unstable gut bacterial genotoxin colibactin forms DNA interstrand crosslinks (ICLs) in a sequence specific manner. (A) The chemical structure of colibactin and the structure and specificity of the colibactin-DNA ICL are unresolved. (B) Experimental workflow for characterizing ICL formation on 25-50 bp double-stranded oligonucleotides using colibactin produced in situ by *pks*$^+$ *E. coli*. (C) Extent of ICL formation within 50mers possessing the indicated sequence motifs. Crosslink formation was identified by denaturing gel electrophoresis and quantified by densitometry. Quantified values are displayed as a column plot. (D) Residue-specific alkylation of 25mers containing the indicated sequence motifs. Alkylation intensities were determined through liquid chromatography-high resolution accurate mass-mass spectrometry (LC-HRAM-MS) analysis and normalized to total DNA injected. Intensity reported is the difference between the average detected in assays with *pks*$^+$ *E. coli* compared to the average detected in assays with *pks*$^-$ *E. coli*. (E) Inferred ICL locations within tested sequence motifs. All alkylated residues are bolded and those crosslinked are connected by red lines. All data are mean +/- s.d. with n=3 biological replicates. Full 50 bp and 25 bp sequences are available in SI. N.D.: Not detected.

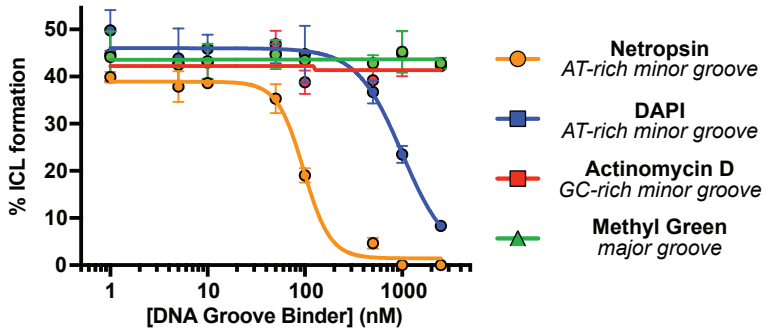

Fig. 2. Colibactin binds AT-rich regions of the minor groove. ICL formation by *pks*$^{+}$ *E. coli* in the presence of select DNA groove binding small molecules. Dose-dependent inhibition of ICL formation by AT-rich minor groove binders was observed by denaturing gel electrophoresis and quantified by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

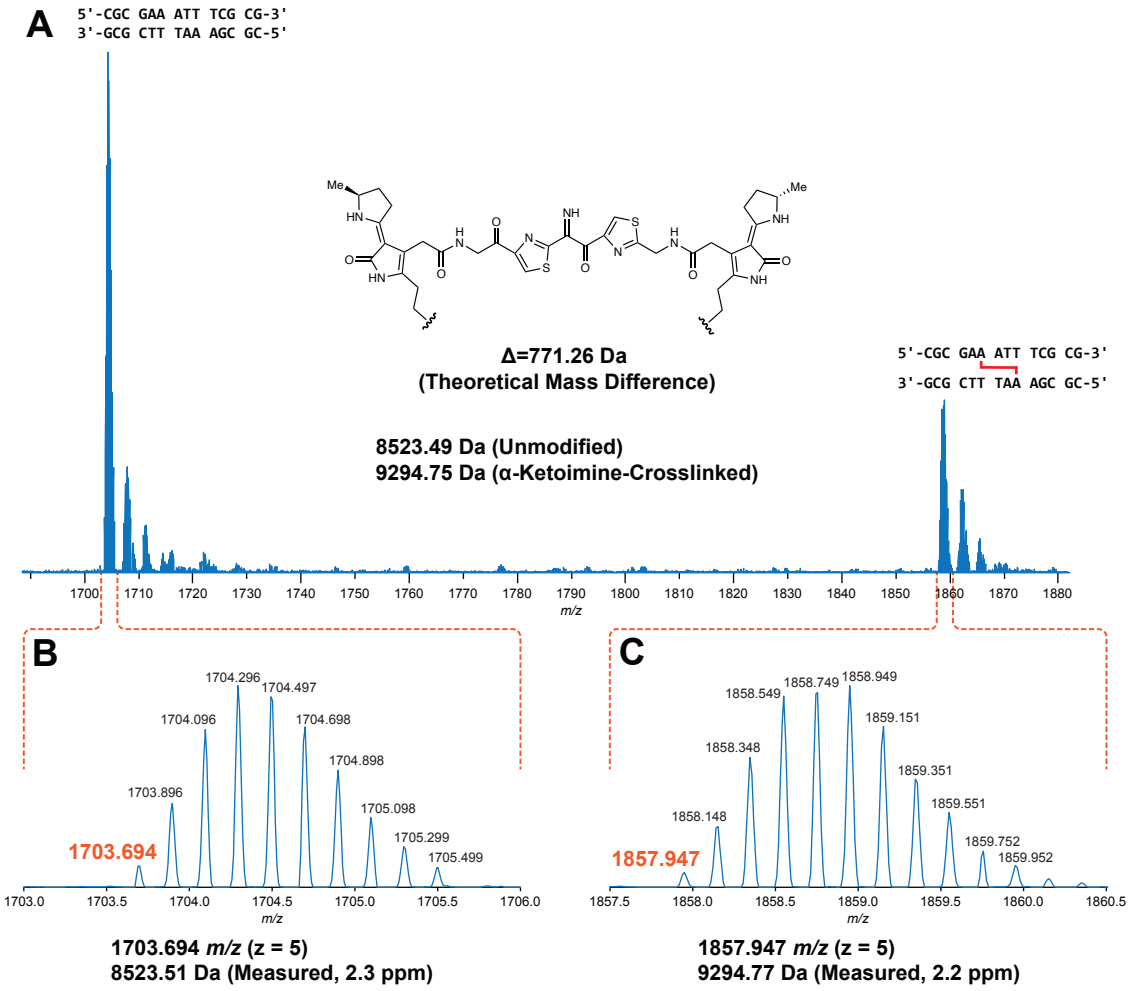

Fig. 3. Mass spectrometry analysis of an intact ICL reveals a central α-ketoimine in colibactin. HRAM LC analysis of crosslinked 14-mer DNA double strand oligonucleotide exposed to *pks*$^{+}$ *E. coli*. (A) Full scan spectrum containing –5 charge state signal of unmodified and modified 5'-CGCGAAATTTCGCG-3' double strand oligonucleotides. (B) Expanded region of spectrum corresponding to the unmodified oligonucleotide. (C) Expanded region of the spectrum corresponding to the modified oligonucleotide.

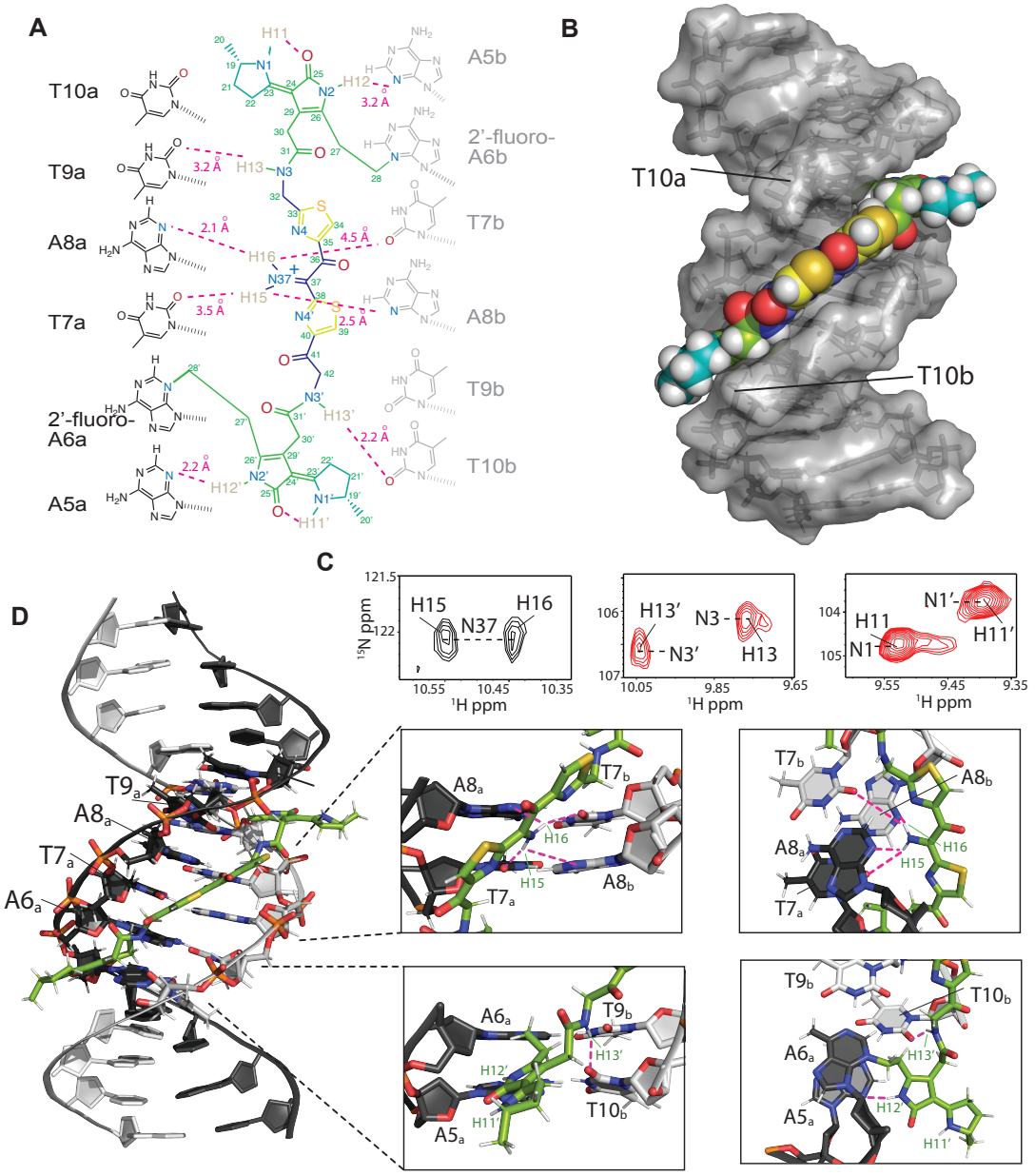

Fig. 4. Structure of the colibactin-DNA ICL. (A) Schematic and atom numbering of colibactin, and its hydrogen bonding and electrostatic interaction with nucleotide bases as denoted by magenta dashed lines. (B) Surface representation of the structure showing the pyrrolidine, thiazole and pyrrole rings stacked in between the minor groove edges of the DNA parallel to the long axis. (C) Portions of a 2D 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 pseudo-symmetric nature of the interaction. (D) Cartoon representation of the interaction and zoom-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 T10b and A5a nucleobase (bottom). For clarity, only the inner motif residues (A6-T9) on the A-strand have been labeled.

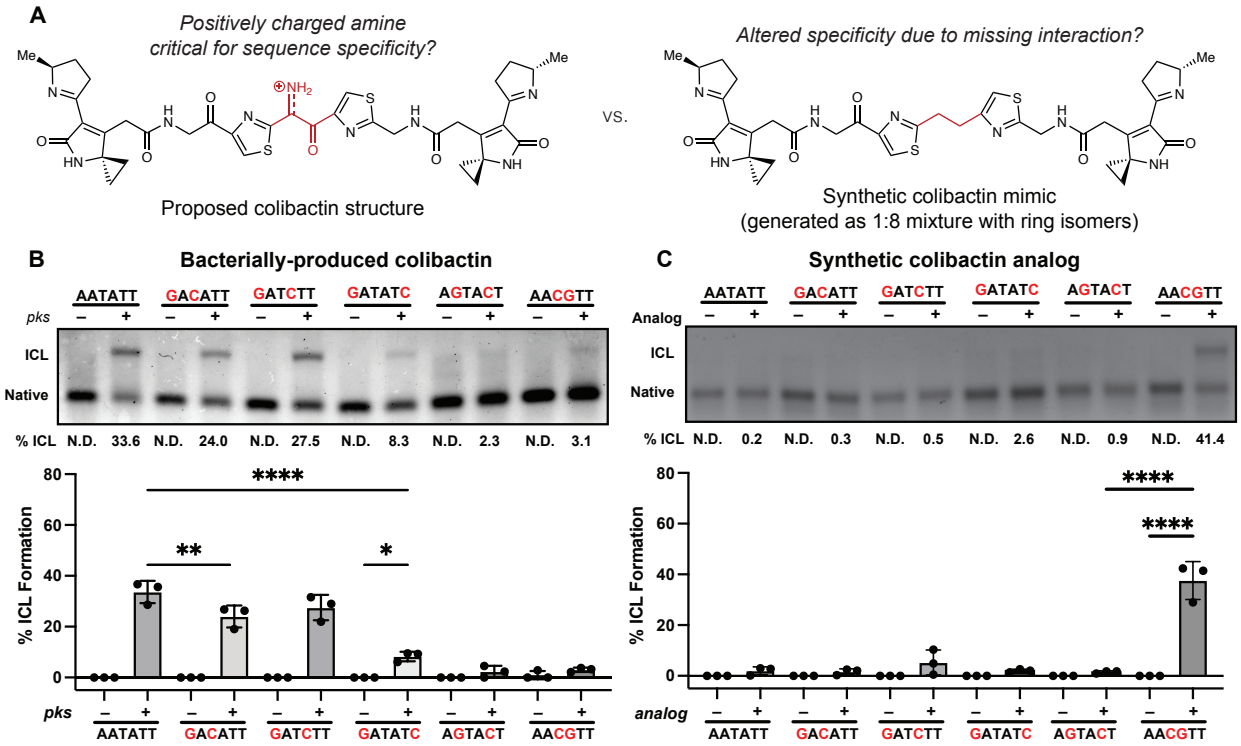

Fig. 5. The unstable central motif of colibactin influences its sequence specificity for ICL formation. (A) Structural comparison of bacterially produced colibactin and a synthetic 'stable' colibactin analog suggests the analog may have altered specificity for ICL formation. (B) Quantification of ICL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

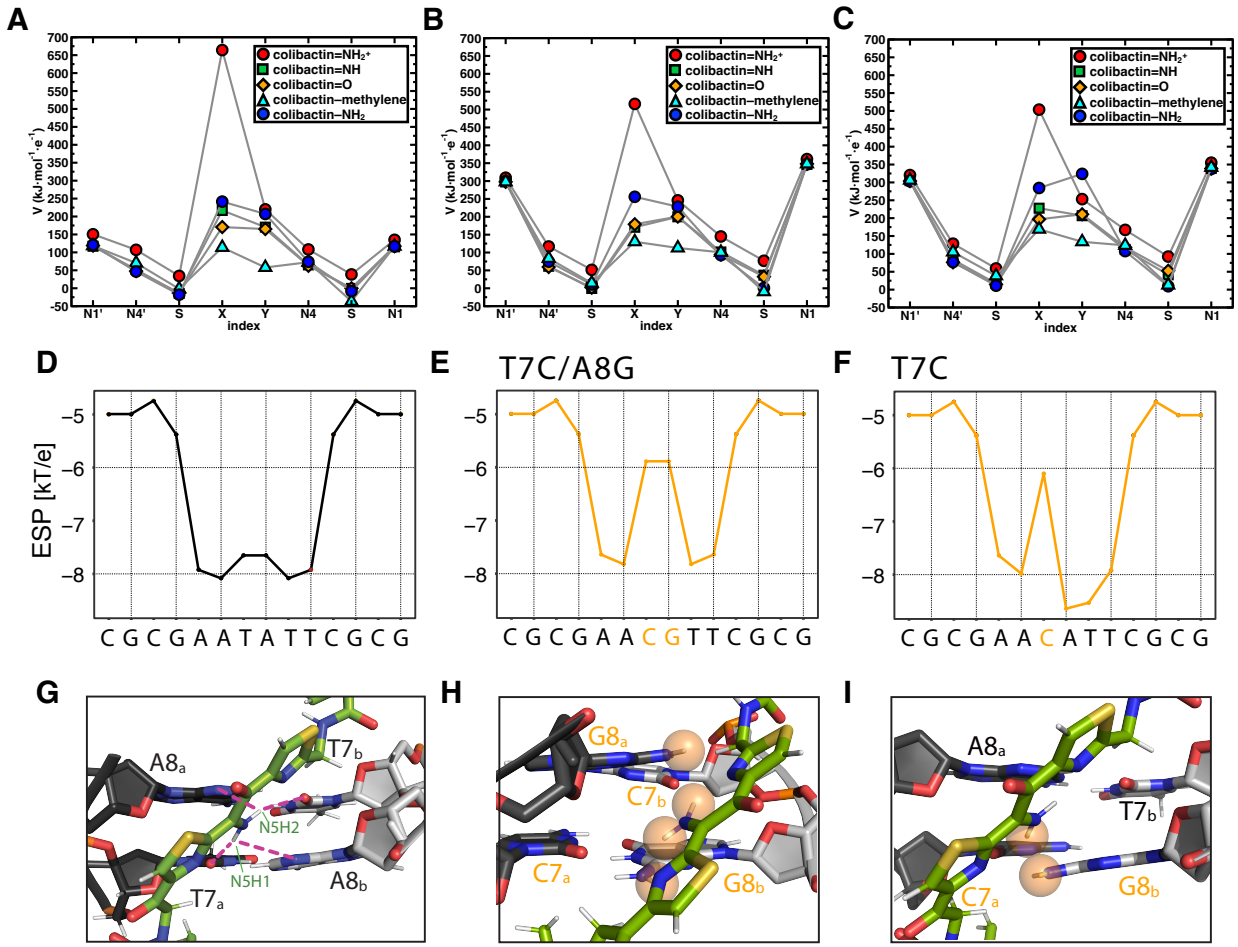

Fig. 6. Electrostatic and steric interactions involving colibactin's unstable central motif likely drive sequence specificity for DNA alkylation. Electrostatic potential (ESP) values of proposed colibactin structures (V in $kJ \cdot mol^{-1} \cdot e^{-1}$) obtained from density functional theory (DFT) 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 crosslinked to the doubly charged DNA sequence GAATATTC (C) proposed colibactin structures crosslinked 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. X=C and Y=C for colibactin–methylene. (D–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-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).