An SE(3)-equivariant Crystal Structure Prediction Framework for Prospective Screening and Discovery of Bioactive Nucleoside Self-Assembling Materials

Zheng Wang[1†], Tianyi Wang[1†], Haixian Zhang[2,*], Zhenyuan Huang[1], Chengxi Zhao[3], Changfu Li[4], Tiannan Liu[1], Ding Bai[1], Xianglong Han[1,*], Hang Zhao[1,*], Han Wang[5,*]

[1]State Key Laboratory of Oral Diseases & National Center for Stomatology & National Clinical Research Center for Oral Diseases & Research Unit of Oral Carcinogenesis and Management & Chinese Academy of Medical Sciences, West China Hospital of Stomatology, Sichuan University, Chengdu, China

[2]School of Artificial Intelligence, Sichuan University, Chengdu, China

[3]State Key Laboratory of Precision and Intelligent Chemistry, Hefei National Research Center for Physical Sciences at the Microscale, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, China

[4]Key Laboratory of Green Chemistry & Technology of Ministry of Education, College of Chemistry, Sichuan University, Chengdu, China

[5]State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing, China

[†]These authors contributed equally to this work.

*Correspondence:

Email: zhanghaixian@scu.edu.cn; xhan@scu.edu.cn; zhaohangahy@scu.edu.cn; wanghan0501@bjmu.edu.cn

Contents

1. Experiment Section

1.1. Supplementary Introduction to GATr Architecture

1.1.1. Clifford Algebra and Multivector Representations

The GATr architecture operates on multivector representations grounded in Clifford algebra $\mathrm{Cl}(3,0,1)$. This algebra serves as the projective geometric algebra (PGA) of three-dimensional space. This algebraic framework provides a unified language for encoding geometric primitives, including points, lines, planes, rotations, and translations, as well as the relationships between them, all within a single algebraic structure.

In the $\mathrm{Cl}(3,0,1)$ algebra, a multivector is a 16D object comprising components across four grades: one scalar (grade 0), four vector components (grade 1), six bivector components (grade 2), four trivector components (grade 3), and one pseudoscalar (grade 4). The four vector components include the homogeneous basis element e_0, which plays a qualitatively different role from the three spatial vectors e_1, e_2, and e_3. Specifically, e_0 encodes the projective or homogeneous direction, enabling the algebra to represent translations and distances within a single unified framework, whereas the spatial vectors encode directional information in Euclidean 3D space. This unified representation naturally encapsulates both magnitude and orientation information in a single data structure, enabling the network to reason about geometric relationships, such as distances, angles, and relative orientations, through purely algebraic operations on multivector components. An input 3D point $r = (x, y, z)$ is embedded as a trivector in the multivector representation via a projective embedding:

$$r = -x\, e_{023} + y\, e_{013} - z\, e_{012} + e_{123},$$

where the component e_{123} serves as a homogeneous coordinate set to 1. This trivector embedding encodes the 3D point in a form that naturally supports the geometric product operations central to the GATr architecture, while higher-grade components of the multivector remain initially zero and are populated through learned transformations as the data propagates through the network.

The multivector representation is crucial because it enables SE(3)-equivariant operations through the geometric product, which is the fundamental bilinear operation of Clifford algebra that combines the inner product and outer product into a single

unified operation. The geometric product between two multivectors yields a third multivector that encodes geometric relationships such as alignment, orthogonality, and relative orientation, all in a manner that is intrinsically equivariant to rotations and translations. Under a rotation $R \in \mathrm{SO}(3)$, multivectors transform via the sandwich product $M' = RMR^{\dagger}$, where $R^{\dagger}$ is the reverse of the rotor R. This transformation rule guarantees that all network operations preserve equivariance by construction, eliminating the need for explicit equivariance constraints or data augmentation. Furthermore, invariant scalar features, including interatomic distances and energy-related quantities, can be extracted from equivariant multivector representations through the geometric product with a reference multivector. This provides a principled mechanism for combining equivariant and invariant information within a single architecture.

1.1.2. GATr Architecture

The GATr backbone in SE3CSP processes molecular representations through a sequence of equivariant transformer blocks, each combining geometric self-attention with equivariant feed-forward processing[1]. The architecture operates on two parallel streams throughout. The first is a multivector stream that preserves full geometric information through the 16-dimensional multivector representation, and the second is a scalar stream that accumulates invariant features for tasks such as token classification.

The backbone begins with an initial equivariant linear layer (EquiLinear) that projects the input multivector and scalar features into the hidden representation space, and the core of the backbone comprises a sequence of stacked GATr transformer blocks. Each block follows a pre-norm architecture consisting of two sequential sublayers, with residual connections around each sublayer:

Equivariant Layer Normalization: Before each sublayer, an equivariant layer normalization (EquiLayerNorm) is applied independently to both the multivector and scalar streams. Unlike standard layer normalization, which normalizes across all features regardless of their geometric meaning, EquiLayerNorm preserves the algebraic structure of the multivector by normalizing each grade component independently. For the multivector stream, normalization is applied separately to the scalar, vector, bivector, and trivector components, ensuring that the relative magnitudes of geometrically meaningful quantities are maintained. For the scalar stream, standard layer normalization is applied. This grade-aware normalization is critical for training

stability, as the different grades of a multivector can have vastly different magnitude scales during the early phases of training.

Geometric Self-Attention: Multi-head geometric self-attention operates on both the multivector and scalar streams. Queries, keys, and values are computed via equivariant linear projections. Rotary positional encoding is applied to the scalar channels of queries and keys to inject sequential positional information. The attention mechanism computes geometric attention weights that respect the algebraic structure of multivectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\langle Q, K\rangle_{\text{MV}}}{\sqrt{d_k}}\right)V$$

where $\langle\cdot,\cdot\rangle_{\text{MV}}$ denotes the geometric inner product between multivectors.

Geometric Multilayer Perceptron: Following attention, a two-layer equivariant geometric multilayer perceptron (GeoMLP) processes the features, providing the network with nonlinear expressiveness while maintaining equivariance. The GeoMLP architecture is carefully designed to respect the algebraic constraints of geometric algebra. The GeoMLP uses equivariant bilinear layers to process multivector features, ensuring that geometric transformations applied to the input are preserved through the nonlinear transformation. A reference multivector constructed from the mean of input multivectors is used to define the equivariant join operation within the bilinear layers.

Reference Multivector and Equivariant Join: A critical component of the GATr architecture is the reference multivector, which serves as a global geometric anchor for the network. The reference multivector is constructed from the mean value of the input multivector representations, effectively encoding the centroid of the molecular system as a trivector in projective geometric algebra. This centroid encoding provides the network with a fixed geometric reference point against which all local geometric features can be measured.

The reference multivector participates in two key operations within the network. First, it enables the equivariant join operation within the GeoMLP. The join combines each atom's multivector representation with the reference multivector via the geometric product, producing an invariant scalar feature that measures the atom's geometric relationship to the molecular centroid. This effectively encodes the atom's distance and orientation relative to the center of mass. This invariant feature is then used to modulate the equivariant multivector stream through the scalar gating mechanism, allowing the

network to produce position-dependent nonlinear transformations that remain equivariant by construction. Second, the reference multivector provides a stable geometric anchor during coordinate extraction at the prediction heads, where the pseudoscalar component of the reference is used to normalize predicted coordinates and ensure numerical stability.

1.2. Supplementary Overview of Loss Functions

1.2.1. Conformational Loss Functions

In the conformation learning stage, the following losses provide supervision for the three prediction heads, offering guidance across the hierarchical levels of molecular structure. Specifically, these losses comprise the token loss, pointer loss, position loss, and four derivative losses derived from atomic positions: bond length loss, bond angle loss, dihedral loss, and self-distance loss.

Token Loss: A cross-entropy loss supervises DSMILES token prediction, ensuring that the model correctly identifies the chemical identity of each atom in the sequence:

$$\mathcal{L}_{\text{Token}} = \text{CrossEntropy}(\hat{t}_i, t_i),$$

This loss function serves as an auxiliary objective, ensuring that the model learns to recognize the molecular constituents it predicts before advancing to the accurate spatial position of atoms.

Pointer Loss: A cross-entropy loss supervises positional pointer prediction, and serves as an auxiliary loss:

$$\mathcal{L}_{\text{Ptr}} = \text{CrossEntropy}(\hat{p}_i, p_i),$$

This loss ensures that the model correctly reconstructs the molecular graph by predicting the inter-fragment connectivity encoded in the positional pointer sequence.

Position Loss: A smooth L1 loss supervises 3D coordinate prediction, applied only to non-padding positions:

$$\mathcal{L}_{\text{Pos}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \text{SmoothL1}(\hat{r}_i, r_i),$$

where $\mathcal{M}$ denotes the set of non-padding indices. The smooth L1 loss is preferred to the standard L2 loss due to its greater robustness to outliers. This property prevents a

small number of poorly predicted atoms from dominating the gradient updates, thereby enhancing training stability.

Bond Length Loss: A smooth L1 loss is employed to supervise 3D coordinate prediction by minimizing the difference between the calculated and ground truth bond lengths. These bond lengths are derived from the distances between each atom and its corresponding parent atom, as determined by the finding parent indices algorithm (Supplementary Algorithm S1):

$$\mathcal{L}_{\text{BL}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \text{SmoothL1}\left(\hat{d}_i, d_i\right),$$

where d_i denotes the bond length between the parent atom $R_1(i)$ of atom i and atom i itself, and $\mathcal{B}$ is the set of atoms with valid bond length targets.

Bond Angle Loss: A smooth L1 loss supervises 3D coordinate prediction by calculating the difference between the calculated and ground truth bond angles:

$$\mathcal{L}_{\text{BA}} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \text{SmoothL1}\left(\hat{\theta}_i, \theta_i\right),$$

where $\mathcal{A}$ denotes the set of atoms with valid bond angle targets, and θ_i denotes the bond angle are computed from the triplets of atoms $(R_2(i), R_1(i), i)$ identified by the find parent indices algorithm (Supplementary Algorithm S1), where $R_2(i)$ is the grandparent of atom i.

Dihedral Angle Loss. A smooth L1 loss is employed to measure the difference between the calculated dihedral angles and the corresponding ground-truth values:

$$\mathcal{L}_{\text{DA}} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \text{SmoothL1}\left(\hat{\phi}_i, \phi_i\right),$$

where $\mathcal{D}$ denotes the set of atoms with valid dihedral angle targets, and ϕ_i denotes the dihedral angle are computed from the quadruples of atoms $(R_3(i), R_2(i), R_1(i), i)$ identified by the find parent indices algorithm (Supplementary Algorithm S1), where $R_3(i)$ is the parent of grandparent of atom i.

Self-Distance Loss. To enforce global structural consistency beyond local bond geometry, a pairwise distance matrix loss is computed between all atom pairs:

$$\mathcal{L}_{\text{SD}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \text{SmoothL1}\left(\left\|\hat{r}_i - \hat{r}_j\right\|, \left\|r_i - r_j\right\|\right),$$

where $\mathcal{P}$ includes all pairs of non-padding atoms. This loss captures long-range geometric relationships that are not directly supervised by the bond length, angle, and dihedral losses. For example, the distance between two atoms on opposite sides of a ring system or the overall molecular diameter.

1.2.2. Unit-cell Parameter Loss Functions

During the crystal parameter learning stage, these losses supervise the lattice prediction module. Specifically, the loss set comprises six unit-cell parameter losses, lattice basis vector loss, space group loss, and density loss.

Unit-cell Parameter Losses: The Smooth L1 losses are computed on the six unit-cell parameters: lattice lengths (a, b, c) and lattice angles (α, β, γ):

$$\mathcal{L}_{\text{Lattice}} = \sum_{l\in\{a,b,c\}} \text{SmoothL1}\left(\hat{l}, l\right) + \sum_{\kappa\in\{\alpha,\beta,\gamma\}} \text{SmoothL1}(\hat{\kappa}, \kappa).$$

The unit-cell parameter losses supervise the unit cell in the physically interpretable parameterization that crystallographers use directly, ensuring that the predicted values can be compared directly with experimental crystallographic data.

Unit-cell Basis Vector Losses: To provide enhanced geometric supervision for the lattice, the predicted unit-cell parameters are transformed into Cartesian basis vectors. This conversion follows standard crystallographic conventions, whereby the six scalar lattice parameters $(a, b, c, \alpha, \beta, \gamma)$ are mapped to three 3D vectors $(\mathbb{A}, \mathbb{B}, \mathbb{C})$ that span the lattice. In this convention, the $\mathbb{A}$ vector is aligned with the x-axis, and the $\mathbb{B}$ vector lies within the xy-plane:

$$\begin{cases} \mathbb{A} = (a, 0, 0) \\ \mathbb{B} = (b\cos\gamma, b\sin\gamma, 0) \\ \mathbb{C} = \left(c\cos\beta, c\dfrac{\cos\alpha - \cos\beta\cos\gamma}{\sin\gamma}, \sqrt{c^2 - c_x^2 - c_y^2}\right). \end{cases}$$

Here, the smooth L1 losses are then applied to each Cartesian component of the predicted basis vectors, offering direct geometric supervision over the shape of the lattice:

$$\mathcal{L}_{\text{Basis}} = \sum_{\mathbb{v}\in\{\mathbb{A},\mathbb{B},\mathbb{C}\}} \frac{1}{3}\sum_{k=1}^{3} \text{SmoothL1}(\hat{\mathbb{v}}_k, \mathbb{v}_k).$$

Space Group Loss: A cross-entropy loss is used on the space group classification:

$$\mathcal{L}_{\text{SG}} = \text{CrossEntropy}\left(\hat{G}, G\right).$$

Density Loss. A smooth L1 loss is employed on crystal density prediction:

$$\mathcal{L}_{\text{Density}} = \text{SmoothL1}(\hat{\rho}, \rho).$$

1.3. Computational Environment

Computational tasks were conducted on a high-performance workstation equipped with an Intel Core Ultra 9 285K processor (featuring 24 cores and 24 threads) and 64 GB of system memory. Hardware acceleration was provided by an NVIDIA GeForce RTX 4090 graphics processing unit with 24 GB of dedicated VRAM, delivering the parallel processing capabilities required for intensive computing workflows.

1.4. Single-Crystal X-ray Diffraction (SCXRD)

Single crystals suitable for X-ray diffraction were obtained by slow solvent evaporation. Diffraction data were collected and integrated using the SAINT software package, and absorption corrections were applied using SADABS. Structures were solved by direct methods with SHELXS97 and refined by full-matrix least-squares refinement on F^2 using SHELXL-2016. Non-hydrogen atoms were refined anisotropically. Hydrogen atoms bonded to N and O atoms were located from difference Fourier maps and refined isotropically, whereas hydrogen atoms bonded to carbon were placed in geometrically calculated positions and refined using a riding model. Crystallographic parameters and data-collection details are summarized in Supplementary Table 12-16.

1.5. 8BA Morphology Characterization

The crystal morphology of 8BA was examined using light microscopy. Briefly, the as-synthesized 8BA crystals were dispersed in PBS and double-distilled water (ddH_2O) by gentle vortexing to obtain a homogeneous suspension. A drop of the suspension (approximately 20 μL) was placed onto a clean glass slide and covered with a coverslip. The specimens were observed under a Zeiss inverted fluorescence microscope (Axio Observer Z1, Carl Zeiss, Germany) equipped with a digital camera. Images were captured at room temperature under bright-field illumination with appropriate magnification (10×, 20×, and 40× objectives). The morphological characteristics of the crystals, including size, shape, and surface morphology, were recorded and analyzed.

1.5.1. Zeta Potential Measurement

The surface charge properties of 8BA, PAMAM-G3, 2FA, and 2-Amino-8-phenyl-2'-deoxy-2'-fluoro-D-adenosine (8PA) were characterized by nanoparticle tracking

analysis (NTA) using a Particle Metrix Zetaview instrument (Particle Metrix GmbH, Germany). Prior to measurement, each sample was diluted to an appropriate concentration with double-distilled water (ddH_2O) to ensure optimal particle density for analysis. The movement of individual particles was tracked using a camera operating at 30 frames per second, and the diffusion coefficient was calculated from the Brownian motion of particles. The zeta potential was derived from the electrophoretic mobility using the Smoluchowski equation. Each sample was measured in triplicate at 25°C, and the results were expressed as mean ± standard deviation (SD).

1.5.2. Scanning Electron Microscopy (SEM)

The surface morphology and microstructure of 8BA crystals were characterized by scanning electron microscopy (SEM). Briefly, 8BA crystals were dispersed in phosphate-buffered saline (PBS) or double-distilled water (ddH_2O), rapidly frozen in liquid nitrogen, and freeze-dried (FreeZone 2.5, Labconco, USA) for 12 h. The dried samples were mounted on silica wafers using conductive carbon tape and sputter-coated with a thin gold layer (~10 nm) prior to imaging. SEM observations were performed using a JEOL JSM-6390LV microscope at an accelerating voltage of 15 kV. High-resolution morphological analysis was further conducted using an SPI4000 scanning probe microscope (Seiko Instruments Inc., Japan).

1.5.3. High-Angle Annular Dark-Field Scanning Transmission Electron Microscopy (HAADF-STEM)

The internal structure of 8BA crystals were characterized by high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM). Briefly, 8BA crystals dispersed in ddH_2O were embedded in epoxy resin, sectioned into ultrathin slices (~70 nm) using an ultramicrotome (Leica EM UC7), and mounted onto copper grids. HAADF-STEM imaging was performed on a JEOL JEM-2100F microscope operated at 200 kV. This technique enabled visualization of the internal pore structure, crystal defects and density variations within the 8BA crystals.

1.5.4. Thermogravimetric Analysis (TGA)

The thermal stability and decomposition behavior of 8BA crystals were investigated using thermogravimetric analysis (TGA). Approximately 5-10 mg of the dried 8BA sample was accurately weighed and placed into a platinum crucible. TGA measurements were performed on a PerkinElmer Pyris 1 thermogravimetric analyzer (PerkinElmer Inc., Waltham, MA, USA). The sample was heated from 25°C to 400°C

at a constant heating rate of 10 °C/min under a nitrogen atmosphere with a flow rate of 50 mL/min to prevent oxidation during heating. The mass of the sample was continuously recorded as a function of temperature, and the resulting thermogravimetric curve was used to determine the thermal decomposition temperature, weight loss percentage, and stability profile of 8BA. The onset decomposition temperature was identified from the derivative thermogravimetric (DTG) curve.

1.5.5. Powder X-ray Diffraction (PXRD)

PXRD measurements were carried out using a Panalytical Empyrean X-ray diffractometer (Panalytical, Netherlands). The PXRD experiments followed the procedure outlined previously. Crystals of 8BA were prepared as samples. The experiments were conducted under the following conditions: the applied voltage and anode current were 40 kV and 40 mA, respectively, with a radius of 240.00 mm. The active length was set to 3.3482°, and the diffractometer settings were as follows: 2Theta = 9.9972°, Omega offset = 0.0000°, Z = 7.550 mm. The Cu Kα radiation with a wavelength of 1.540598 Å was used, and the scanning mode employed a 0.01° interval with a 0.05 s counting time. The XRD patterns were collected in the range from 5° to 60° (2θ).

1.6. Molecular Dynamics (MD) Simulations

MD simulations were performed to investigate the interaction mechanism between cell-free DNA (cfDNA) and two nanomaterial systems (8BA and 8PA). The three-dimensional structure of cfDNA was first constructed using the Builder module in Chimera software, followed by energy minimization and optimization to obtain a realistic helical structure.

The cfDNA structure was then placed approximately 3 nm above the surface of each material (8BA and 8PA) to simulate the initial adsorption configuration. All MD simulations were carried out using GROMACS 2024.4 software under periodic boundary conditions in the NPT ensemble (constant number of particles, pressure, and temperature). The Amber14SB all-atom force field was employed to describe the interatomic interactions, and the TIP3P water model was used for solvation. To maintain the integrity of hydrogen-containing bonds during the simulation, the LINCS algorithm was applied with an integration time step of 2 fs. Electrostatic interactions were calculated using the Particle-Mesh Ewald (PME) method with a cutoff of 10 Å for van der Waals interactions, and the non-bonded pair list was updated every 10 steps.

The simulation temperature was maintained at 300 K using the V-rescale temperature coupling method, and the pressure was controlled at 1 bar using the Parrinello-Rahman method.

For each system, the simulation protocol was as follows: (1) Energy minimization using the steepest descent algorithm to eliminate close contacts between atoms; (2) 100 ps NVT equilibration at 300 K; (3) 200 ns production MD simulation for data collection. Trajectories were saved every 40 ps, resulting in 5000 frames for analysis. Visualization and analysis of the simulation results were performed using GROMACS built-in tools and VMD software. The final configurations and trajectory files were used to analyze the adsorption behavior, binding energy, and interaction mechanisms between cfDNA and the 8BA/8PA materials.

1.7. Biocompatibility and Biosafety Assessment of 8BA

1.7.1. Hemolytic Activity Assay

Erythrocytes were obtained by centrifuging mouse blood at 100× g for 5 minutes. After washing the blood three times with PBS, the purified erythrocytes were further diluted to a final concentration of 5% (v/v). Then, 100 μL of 8BA and PAMAM-G3 solutions were added to 400 μL of the 5% erythrocyte suspension to achieve final concentrations of 8BA and PAMAM-G3 of 5 mM, 2 mM, 1 mM, 0.5 mM, 0.1 mM, and 0.05 mM. After incubation for 1 hour at 37 °C, all samples were centrifuged at 100× g for 10 minutes. The tubes of all the samples were then photographed and recorded. The supernatants were transferred into a new 96-well clear plate, and the absorbance of the solutions at 540 nm was measured using a microplate reader (Thermo Fisher, USA). PBS was used as the negative control, and 0.1% Triton X-100 served as the positive control. Hemolysis was calculated using the equation:

$$\text{Hemolysis}(\%) = \frac{(A_{\text{materials}} - A_{\text{PBS}})}{(A_{\text{Triton}} - A_{\text{PBS}})} \times 100\%$$

where $A_{\text{materials}}$ is the absorbance value of the different concentrations of 8BA and PAMAM-G3, A_{PBS} is the absorbance value of the negative control (PBS), and A_{Triton} is the absorbance value of the positive control.

1.7.2. *In Vivo* Pharmacokinetic Study of 8BA

The *in vivo* pharmacokinetic profile of 8BA was evaluated in male BALB/c mice following intravenous or oral administration. Briefly, 8BA stock solution (50 mg mL^{-1} in DMSO) was diluted with PEG300, 30% HS15 and saline to prepare working

solutions for oral (1.0 mg mL^{-1}) or intravenous (0.5 mg mL^{-1}) administration. Male BALB/c mice (6–8 weeks old, n = 3 per group) received 8BA intravenously (5 mg kg^{-1}) or orally (10 mg kg^{-1}). Blood samples were collected at predetermined time points (5 min to 24 h), and plasma was isolated by centrifugation for LC–MS/MS analysis. Quantification of 8BA was performed using a Shimadzu UFLC system coupled to a QTRAP 5500 mass spectrometer equipped with a Waters Acquity UPLC® BEH C18 column. Pharmacokinetic parameters, including AUC, C_{max}, T_{max}, $t_{1/2}$ and clearance (CL), were calculated by non-compartmental analysis. Detailed data can be found in Supplementary Table 5-6.

1.7.3. Live/dead Staining and Cell Viability

Cells were plated in 24-well plates at a density of 5×10^4 cells per well. At 1, 2, and 3 days post-seeding, cell viability was assessed using a live/dead cell staining kit (KeyGEN BioTECH) for 30 minutes. Images were captured using a confocal laser scanning microscope (CLSM). Additionally, cell proliferation after 1, 2, and 3 days of co-culture with various concentrations of 8BA was evaluated using the CCK8 assay.

1.8. Transcriptomics Analysis

After administering different concentrations of 8BA (1 mM and 2 mM) and LPS for 24 hours, total RNA was extracted from RAW264.7 cells using TRIzol® Reagent according to the manufacturer's instructions (Invivogen, Carlsbad, CA, USA), and genomic DNA was removed using DNase I (TaKara Bio Inc., Shiga, Japan). RNA quality was assessed using the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and quantified with the ND-2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Only RNA samples meeting the quality criteria (OD260/280 = 1.8-2.0, OD260/230 ≥ 2.0, RIN ≥ 6.5, 28S:18S ≥ 1.0, ≥ 100 ng/μL, ≥ 2 μg) were used for library construction. The experimental groups were designed as follows: (1) Control (Ctrl), (2) 1 mM 8BA treatment, (3) 2 mM 8BA treatment, and (4) LPS treatment. Each group was treated for 24 hours before RNA extraction. RNA-seq transcriptome libraries were prepared using 2 μg of total RNA following the TruSeqTM RNA sample preparation Kit from Illumina (San Diego, CA, USA). The constructed libraries were sequenced on the Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA) to generate 150 bp paired-end reads. All bioinformatics analyses were performed using the Majorbio Cloud Platform from Fujian Berry Genomics Technology Co., Ltd. (Fujian, China). High-quality reads from each sample were

aligned to the mouse reference genome (GCF_000743255.1_ASM74325v1, https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/743/255/GCF_000743255.1_ASM74325v1/) using STAR software. To assess rRNA contamination, 10,000 raw reads from each sample were randomly selected and aligned to the Rfam database (http://rfam.xfam.org/) using BLAST, and the percentage of rRNA was calculated to evaluate the contamination level. Gene expression quantification was performed using Salmon software. Differentially expressed genes (DEGs) were identified using the DESeq2 package (http://bioconductor.org/packages/release/bioc/html/DESeq2.html) with a cutoff of $|\log2FoldChange| > 1$ and p-value < 0.05. The DEGs were then subjected to Gene Ontology (GO) enrichment analysis (http://www.geneontology.org/) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using the clusterProfiler package. Additionally, Gene Set Enrichment Analysis (GSEA) was performed to identify enriched gene sets and pathways. GSEA was conducted using the GSEA software (Broad Institute, MIT, USA) with the Molecular Signatures Database (MSigDB) as the reference gene set database, including both hallmark gene sets and KEGG pathways. Enrichment results were considered significant when the false discovery rate (FDR) was less than 0.25.

1.9. *In vitro* Assessment of TLR Inhibition by 8BA

1.9.1. qRT-PCR

Total RNA from cells or tissues was extracted using SteadyPure RNA Extraction kit (Accurate Biology, AG21024) according to the manufacturer's protocol. RNA concentration and purity were determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific). Reverse transcription experiments were performed with Evo M-MLV RT Kit with gDNA Clean for qPCR (Accurate Biology, AG11705). For the real-time qPCR experiments, the SYBR Green Premix Pro Tag HS qPCR Kit (Accurate Biology, AG11701) was used. The primer sequences are as follows: Gapdh-F, TCTGAGCCTCCTCCAATTCAAC, Gapdh-R, AAATCCGTTCACACCGACCT; Tlr3-F, GTGAGATACAACGTAGCTGACTG, Tlr3-R, TCCTGCATCCAAGATAGCAAGT; Tlr4-F, ATGGCATGGCTTACACCACC, Tlr4-R, GAGGCCAATTTTGTCTCCACA; Tlr7-F, ATGTGGACACGGAAGAGACAA; Tlr7-R, GGTAAGGGTAAGATTGGTGGTG; Tlr9-F, ATGGTTCTCCGTCGAAGGACT; Tlr9-R, GAGGCTTCAGCTCACAGGG; Tlr13-F, CCAACTTGACAGAGGCCATTAG, Tlr13-R, GCACCTTCGTCGATCTTCCAA.

1.9.2. Western Blotting (WB)

Cells were seeded in 6-well plates at 2×10^5 cells per well and administrated with 100ng/ml LPS and 1mM 8BA. Proteins were collected by lysing in RIPA buffer (Beyotime) and centrifuging. WB was then conducted under manufactures' instructions (Bio-Rad). First antibodies include TLR4 (affbiotech, AF7017), TLR9 (affbiotech, DF2970), pJNK (Abclonal, AP0631), JNK (Abcam, ab179461), ERK (SAB,40903), pERK (Abclonal, AP0974), PI3K (CST,4292), IkB-α (CST,4812), vinculin (Abclonoal, A1758), b-tubulin (Abclonoal, AC015) and GAPDH (Proteintech, 60004-1-Ig).

1.9.3. Immunofluorescence Staining and Confocal Imaging

Cells were fixed and permeabilized administrated with 1 mM and 2 mM 8BA, followed by incubation with primary antibodies against TLR4 and TLR9. After washing, the cells were incubated with a secondary antibody conjugated to FITC (488 nm). Nuclei were counterstained with DAPI. The stained cells were then examined using a confocal laser scanning microscope (CLSM) to capture fluorescent images.

1.10. Characterization of Osteoblastic Phenotypes *In Vitro*

Osteogenic differentiation of primary BMSCs was induced using osteogenic medium composed of α-MEM with 10% FBS (Gibco), 100 μM ascorbic acid, 10 mM β-glycerophosphate, and 10 nM dexamethasone (Sigma-Aldrich, St. Louis, MO, USA). After 7 days of induction, cells were fixed and stained using an alkaline phosphatase (ALP) staining kit (Beyotime Biotechnology, Shanghai, China) according to the manufacturer's instructions. The stained samples were observed under a light microscope (Olympus CX23, Olympus Corporation, Tokyo, Japan), and images were captured for qualitative and quantitative analysis.

Primary BMMs were seeded in 6-well plates at a density of 2×10^5 cells per well and cultured in α-MEM supplemented with 10% FBS, 1% penicillin-streptomycin, and 30 ng/mL recombinant murine M-CSF (PeproTech, Rocky Hill, NJ, USA) for 3 days to maintain cell survival. The medium was then replaced with osteoclastogenic medium consisting of α-MEM with 10% FBS, 30 ng/mL M-CSF, and 50 ng/mL RANKL (PeproTech). Cells were cultured for 7-10 days with medium refreshed every 2-3 days. Osteoclast formation was assessed by TRAP (tartrate-resistant acid phosphatase) staining using a TRAP staining kit (Cosmo Bio Co., Ltd., Tokyo, Japan) according to the manufacturer's instructions. Briefly, cells were fixed with 4% paraformaldehyde for 15 min, then incubated with the staining solution at 37°C for 30 min in the dark. TRAP-

positive cells with more than three nuclei were considered mature osteoclasts. Images were captured under a light microscope (Olympus CX23, Olympus Corporation, Tokyo, Japan) for qualitative analysis, and the number of TRAP-positive multinucleated cells was counted to evaluate osteoclastogenesis.

1.11. Histological and Histochemical Staining Analysis

1.11.1. Tartrate-Resistant Acid Phosphatase (TRAP) Staining

For histochemical detection of osteoclasts in bone tissue sections, tartrate-resistant acid phosphatase (TRAP) staining was performed using a TRAP staining kit (Sigma-Aldrich, St. Louis, MO, USA) according to the manufacturer's instructions. Briefly, the bone tissue sections were fixed in 4% paraformaldehyde for 15 min at room temperature, then washed with distilled water three times. The sections were incubated with the TRAP staining solution at 37°C for 30 min in the dark, protected from light. After incubation, the sections were washed with distilled water and counterstained with hematoxylin (Vector Laboratories, Burlingame, CA, USA) for 1 min. The stained sections were then dehydrated through a graded ethanol series (70%, 95%, 100%) and mounted with Permount mounting medium (Fisher Scientific, Waltham, MA, USA).

TRAP-positive cells (osteoclasts) appeared as red-purple multinucleated cells. The number of osteoclasts per bone surface (OcN/BS) was quantified using ImageJ software (National Institutes of Health, Bethesda, MD, USA). For each section, at least three non-overlapping fields of view were randomly selected and analyzed under a light microscope (Olympus CX23, Olympus Corporation, Tokyo, Japan) at 200× magnification.

1.11.2. Hematoxylin and Eosin (H&E) Staining

For histopathological evaluation of periodontal tissues and skin lesions, H&E staining was performed. Briefly, tissue samples were fixed in 4% paraformaldehyde for 24-48 h at room temperature, then dehydrated through a graded ethanol series (70%, 80%, 95%, 100%) and embedded in paraffin. Paraffin-embedded tissues were sectioned at 5 μm thickness using a microtome (Leica RM2235, Leica Microsystems, Wetzlar, Germany). The sections were deparaffinized in xylene (two changes, 10 min each) and rehydrated through a graded ethanol series (100%, 95%, 80%, 70%) to distilled water.

For H&E staining, the sections were first stained with hematoxylin (Sigma-Aldrich) for 5 min, then rinsed in running tap water for 5 min to remove excess stain. The sections were then differentiated in 1% acid alcohol (1% HCl in 70% ethanol) for

a few seconds, followed by rinsing in running tap water for 5 min. After blueing in 0.2% ammonia water or lithium carbonate solution for 1 min, the sections were counterstained with eosin (Sigma-Aldrich) for 2-3 min. Finally, the sections were dehydrated through a graded ethanol series (70%, 95%, 100%), cleared in xylene (two changes), and mounted with Permount mounting medium (Fisher Scientific). The stained sections were observed under a light microscope (Olympus CX23, Olympus Corporation, Tokyo, Japan), and images were captured for histopathological analysis.

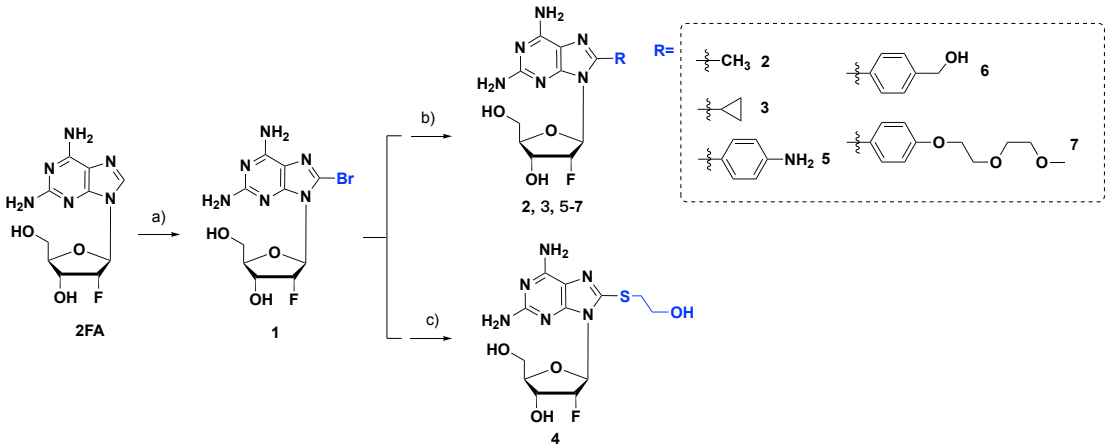
1.11.3. Masson's Trichrome Staining

For detection of collagen fibers in connective tissues, Masson's trichrome staining was performed. Tissue sections were prepared as described above for H&E staining. After deparaffinization and rehydration, the sections were fixed in Bouin's solution (Sigma-Aldrich) at 56°C for 1 h, then rinsed in running tap water for 10 min to remove the yellow coloration. The sections were then stained with Weigert's hematoxylin (Sigma-Aldrich) for 5 min, followed by rinsing in running tap water for 5 min. After differentiation in 1% acid alcohol for a few seconds and rinsing, the sections were stained with ponceau acid fuchsin solution (Sigma-Aldrich) for 5-10 min. The sections were then differentiated in 1% phosphotungstic acid solution (Sigma-Aldrich) for 5-10 min, followed by direct transfer to methyl blue solution (Sigma-Aldrich) for 5 min. After brief rinsing in 1% acetic acid solution, the sections were dehydrated through a graded ethanol series, cleared in xylene, and mounted. In Masson's trichrome staining, collagen fibers appeared blue, muscle fibers appeared red, and nuclei appeared dark blue to black. Sections were imaged under a light microscope (Olympus CX23) for qualitative and semi-quantitative collagen analysis.

1.11.4. Immunohistochemistry (IHC)

Immunohistochemistry was performed under manufactures' instructions. First antibodies include TLR4 (Abclonal), TLR9 (Abclonal), TNF-a (Abcam) and IL-6 (Abcam). Three representative slides from different mice in each group were selected for quantification. Quantification of TLR4, TLR9, TNF-a and IL-6 positive cells were conducted manually at a magnification of ×100 in a blinded manner using Image J.

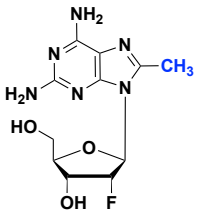
1.12. Chemical synthesis of Compounds 1-7



Supplementary Fig. 1. Synthesis of compounds **1**–**7**. Reagents and conditions: a) NBS, MeCN/H_2O, r.t., 1 h; b) corresponding boric acid/boric acid ester, $Pd(PPh_3)_4$, K_2CO_3, 1,4-Dioxane/H_2O, 100 °C, 4-10 h; c) 2-mercaptoethanol, Et_3N, H_2O, 100 °C, 3 h.

Compounds **1**, **5**, and **7** were synthesized according to our previously reported procedures[2]. Compounds **2**, **3** and **6** were prepared via Suzuki coupling reactions using the corresponding boronic acids or boronic acid esters as coupling partners[3]. Detailed synthetic procedures and isolated yields are provided below.

2-amino-8-methyl-2'-deoxy-2'-fluoro-D-adenosine (2).



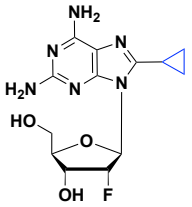
Compound **2** was obtained as a white solid with 78% yield starting from compound **1** and 2,4,4,5,5-pentamethyl-1,3,2-dioxaborolane following the general procedure.

^{1}H NMR (400 MHz, DMSO-*d*$_6$): δ 6.71 (s, 2H), 5.97 (dd, *J* = 19.7, 3.9 Hz, 1H), 5.78 – 5.55 (m, 4H), 5.42 – 5.31 (m, 1H), 4.54 – 4.43 (m, 1H), 3.95 – 3.86 (m, 1H), 3.77 – 3.64 (m, 1H), 3.57 – 3.48 (m, 1H), 2.44 (s, 3H).

^{13}C NMR (101 MHz, DMSO-*d*$_6$): 159.54, 155.52, 152.10, 144.59, 111.75, 91.75 (d, *J*C-F= 186.2 Hz), 85.71 (d, *J*C-F = 33.4 Hz), 84.22, 68.71 (d, *J*C-F = 15.4 Hz), 61.20, 14.16.

HRMS (ESI+) $[M+H]^+$ calcd m/z for $[C_{11}H_{16}FN_6O_3]^+$: 299.1268, found: 299.1265.

2-amino-8-cyclopropane-2'-deoxy-2'-fluoro-D-adenosine (3).



Compound **3** was obtained as a white solid with 68% yield starting from compound **1** and cyclopropylboronic acid following the general procedure.

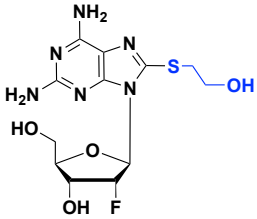
^{1}H NMR (400 MHz, DMSO-*d*$_6$): δ 6.65 (s, 2H), 6.19 (dd, *J* = 19.3, 4.2 Hz, 1H), 5.88 – 5.54 (m, 4H), 5.40 (dd, *J* = 6.9, 4.7 Hz, 1H), 4.57 – 4.45 (m, 1H), 4.08 – 3.82 (m, 1H), 3.76 – 3.62 (m, 1H), 3.63 – 3.46 (m, 1H), 2.27 – 2.06 (m, 1H), 1.08 – 0.80 (m, 4H).

^{13}C NMR (101 MHz, DMSO-*d*$_6$): δ 159.49, 155.52, 152.16, 149.14, 111.58, 91.75 (d, *J*C-F = 186.9 Hz), 85.37 (d, *JC-F* = 33.4 Hz), 84.41, 68.71 (d, *JC-F* = 15.6 Hz), 61.30, 7.60, 7.56, 6.89.

HRMS (ESI+) $[M+H]^+$ calcd m/z for $[C_{13}H_{18}FN_6O_3]^+$: 325.1419, found: 325.1420.

2-amino-8-(2-hydroxyethylthio)-2'-deoxy-2'-fluoro-D-adenosine (4).

A mixture of compound **1** (72.2 mg, 0.2 mM), 2-mercaptoethanol (47 mg, 0.6 mmol), Et_3N (102 mg, 2.0 mmol) and H_2O (1 ml) in a sealed tube was stirred at 100 °C for 3 h, then evaporated the solvent and purified by column chromatography (DCM/MeOH 10:1) to obtained a white solid yield (39.5 mg, 55% yield).

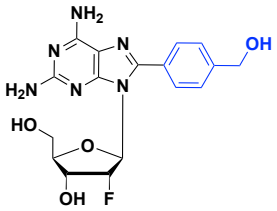


^{1}H NMR (800 MHz, DMSO-*d*$_6$): δ 6.81 (s, 2H), 6.01 (dd, *J* = 19.7, 4.0 Hz, 1H), 5.78 (s, 2H), 5.75 – 5.65 (m, 1H), 5.62 (d, *J* = 6.2 Hz, 1H), 5.33 – 5.30 (m, 1H), 4.99 (t, *J* = 5.4 Hz, 1H), 4.56 – 4.50 (m, 1H), 3.92 (q, *J* = 4.5 Hz, 1H), 3.72 – 3.67 (m, 1H), 3.64 (q, *J* = 6.1 Hz, 1H), 3.56 – 3.51 (m, 1H), 3.23 (qt, *J* = 13.1, 6.5 Hz, 2H).

^{13}C NMR (201 MHz, DMSO-*d*$_6$): δ 159.60, 155.01, 152.47, 142.35, 113.60, 91.52 (d, *J*C-F = 186.4 Hz), 86.14 (d, *J*C-F = 33.9 Hz), 84.71, 68.99 (d, *J*C-F = 15.7 Hz), 61.65, 59.87, 35.99.

HRMS (ESI+) $[M+H]^+$ calcd m/z for $[C_{12}H_{18}FN_6O_4S]^+$: 361.1089, found: 361.1075.

2-amino-8-(4-(hydroxymethyl)phenyl)-2'-deoxy-2'-fluoro-D-adenosine (8BA) (6).



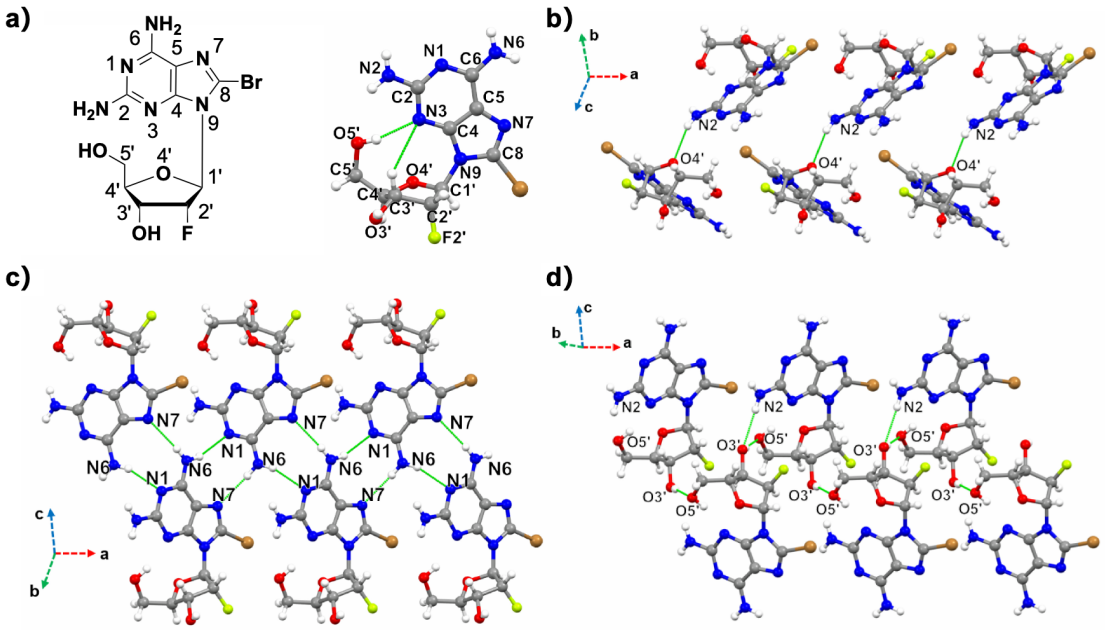
Compound **6** was obtained as a white solid with 75% yield starting from compound **1** and 4-(hydroxymethyl)phenylboronic acid following the general procedure.

^{1}H NMR (400 MHz, DMSO-*d*$_6$): δ 7.60 (d, *J* = 8.1 Hz, 2H), 7.50 (d, *J* = 8.0 Hz, 2H), 6.95 (s, 2H), 5.96 – 5.72 (m, 4H), 5.55 (d, *J* = 6.1 Hz, 1H), 5.44 (dd, *J* = 7.4, 4.6 Hz, 1H), 5.33 (t, *J* = 5.7 Hz, 1H), 4.62 – 4.50 (m, 3H), 3.91 (d, *J* = 5.0 Hz, 1H), 3.77 – 3.67 (m, 1H), 3.64 – 3.51 (m, 1H).

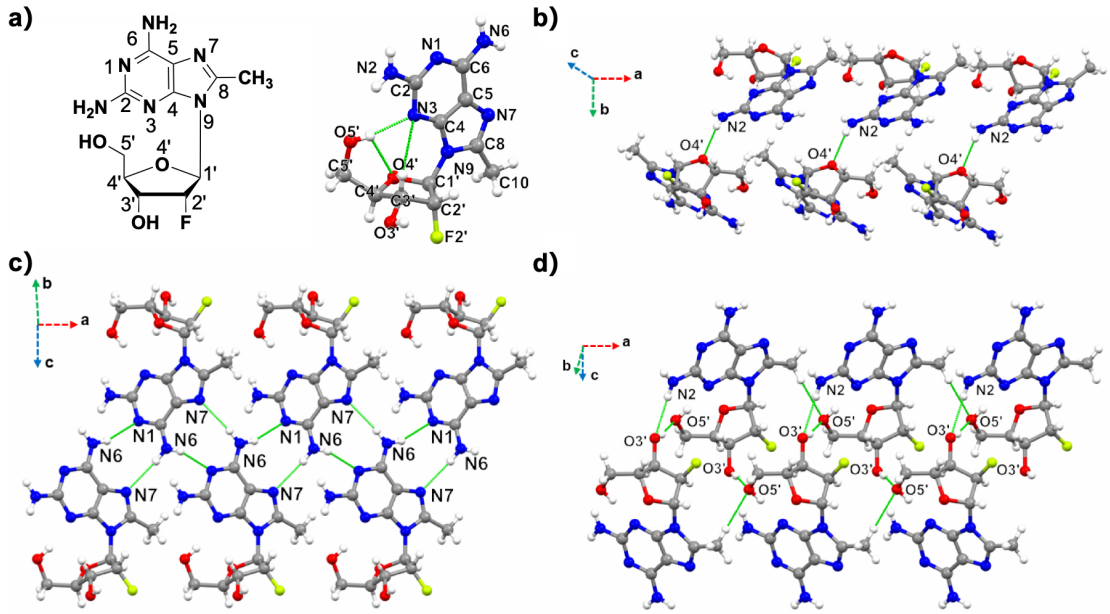
^{13}C NMR (101 MHz, Chloroform-*d*$_6$): δ 159.79, 156.27, 152.15, 146.85, 144.27, 128.91, 127.99, 126.67, 113.14, 91.30 (d, *J*C-F = 185.3 Hz), 86.82 (d, *J*C-F= 53.0 Hz), 84.93, 69.05 (d, *J*C-F = 15.4 Hz), 62.52, 61.77

HRMS (ESI+) $[M+H]^+$ calcd m/z for $[C_{13}H_{18}FN_6O_3]^+$: 391.1530, found: 391.1531.

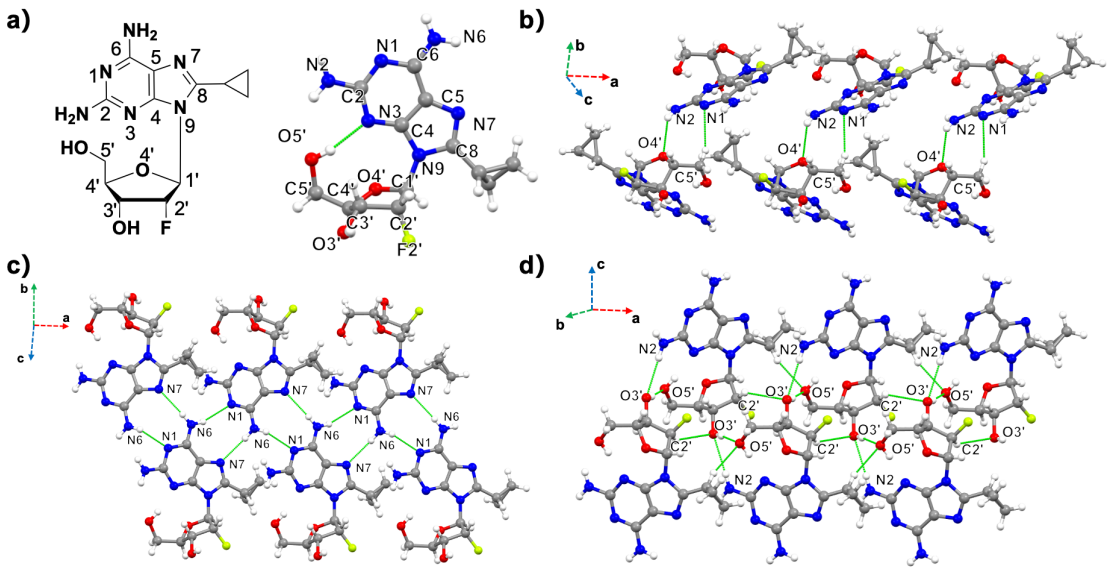
2. Supplementary Figures



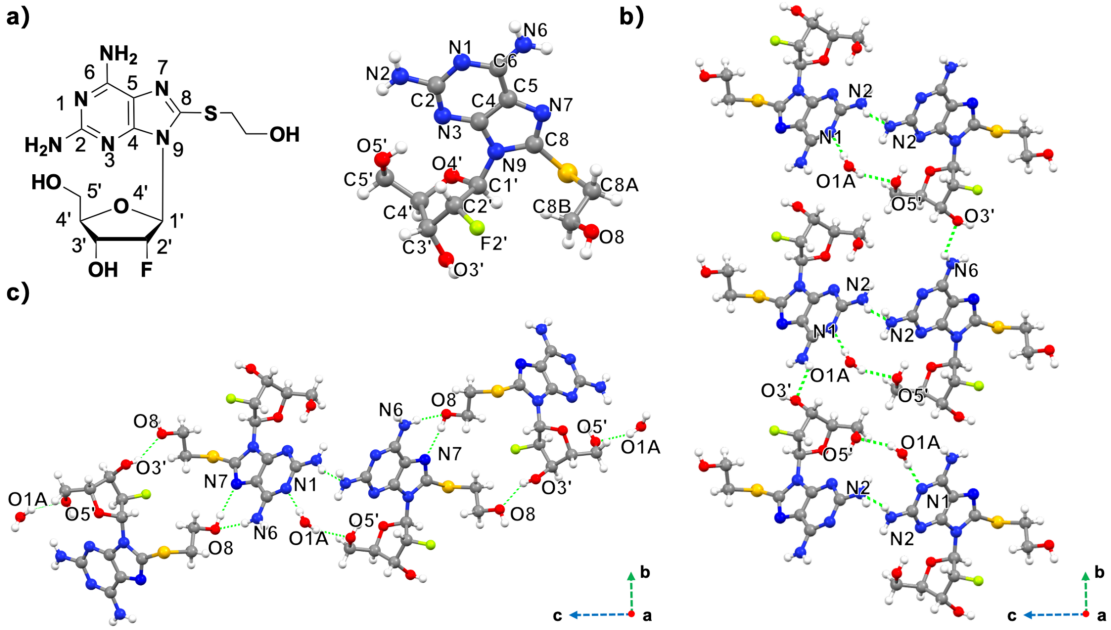
Supplementary Fig. 2. a) The chemical structure and crystal structure of **1**. The H-bonds for the base-pair (b) and sugar-sugar interactions (c, d) in **1**. Atoms are coded as follows: red, oxygen; blue, nitrogen; gray, carbon; white, hydrogen; green, fluorine; brown, bromine.



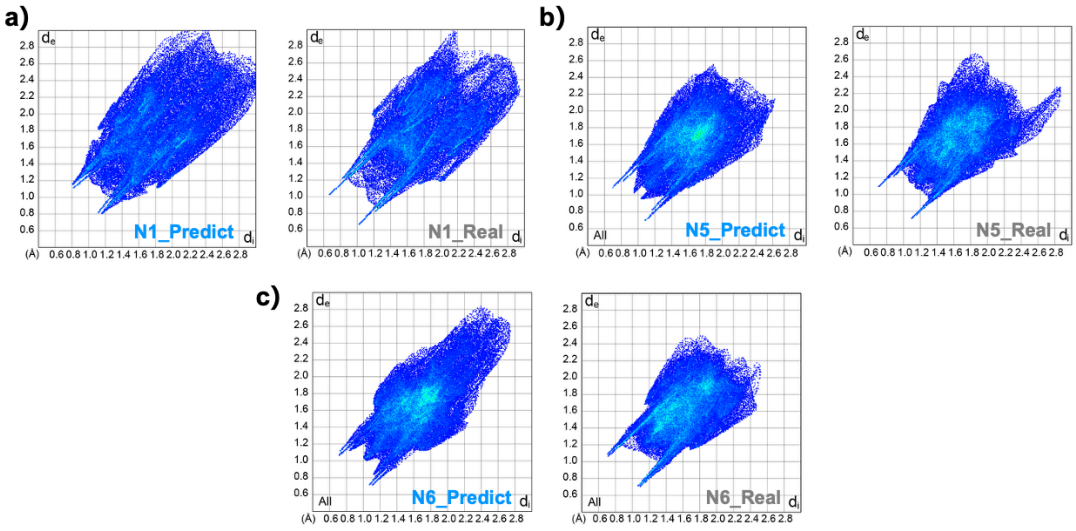
Supplementary Fig. 3. a) The chemical structure and crystal structure of **2**. The H-bonds for the base-pair (b) and sugar-sugar interactions (c, d) in **2**. Atoms are coded as follows: red, oxygen; blue, nitrogen; gray, carbon; white, hydrogen; green, fluorine.



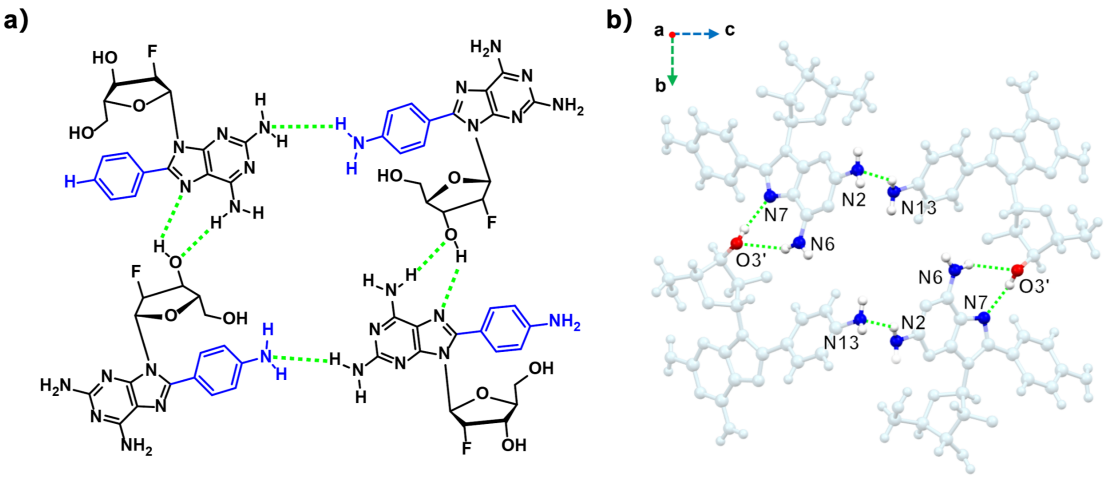
Supplementary Fig. 4. a) The chemical structure and crystal structure of **3**. The H-bonds for the base-pair (b) and sugar-sugar interactions (c, d) in **3**. Atoms are coded as follows: red, oxygen; blue, nitrogen; gray, carbon; white, hydrogen; green, fluorine.



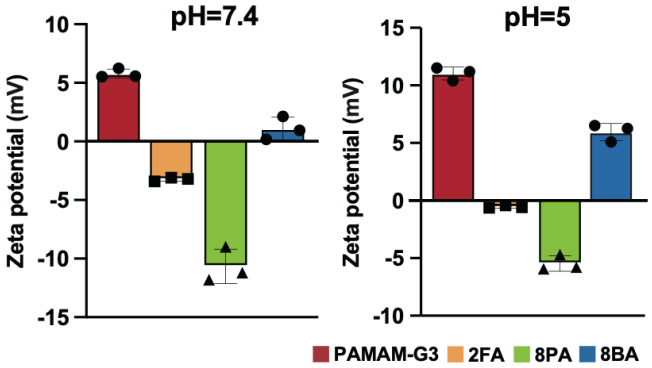
Supplementary Fig. 5. a) The chemical structure and crystal structure of **4**. b, c) The H-bonds for the base-pair and sugar-sugar interactions in **4**. Atoms are coded as follows: red, oxygen; blue, nitrogen; gray, carbon; white, hydrogen; green, fluorine.



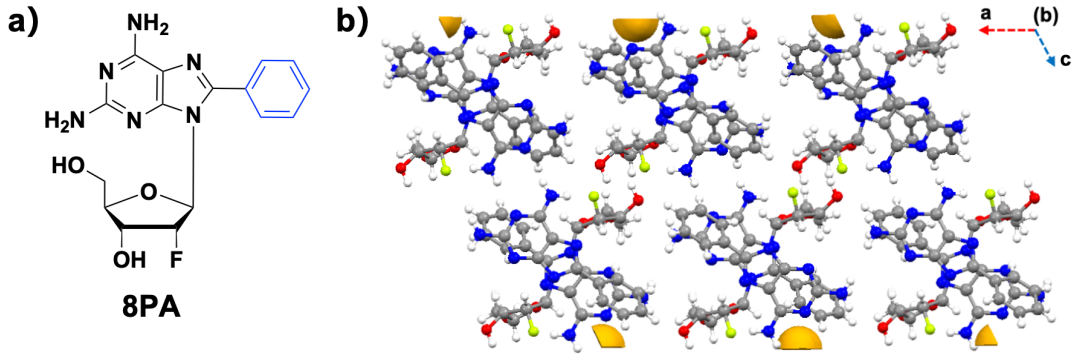
Supplementary Fig. 6. Hirshfeld fingerprint plots of SE3CSP-predicted and experimental structures for compounds **1**, **5** and **6**.



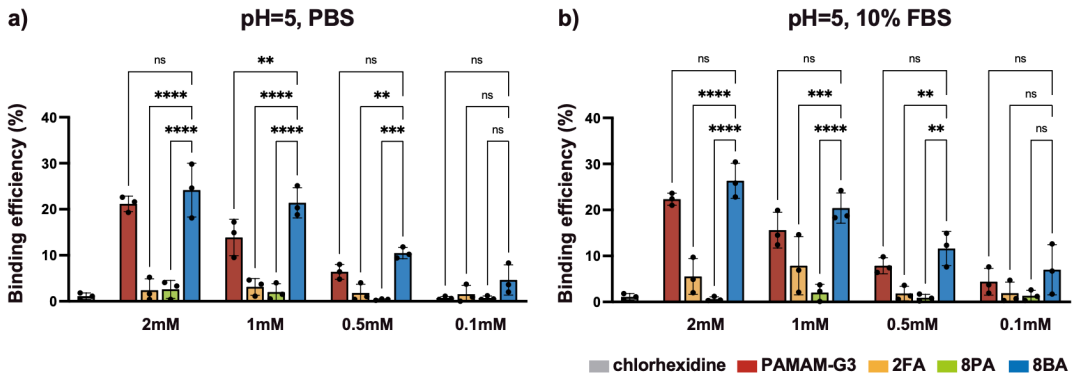
Supplementary Fig. 7. Tetrameric chemical structure (a) and single-crystal structure (b) of **5**. Interactions between the nucleobases and sugar rings lead to the formation of a water-channel-like structure at the center.



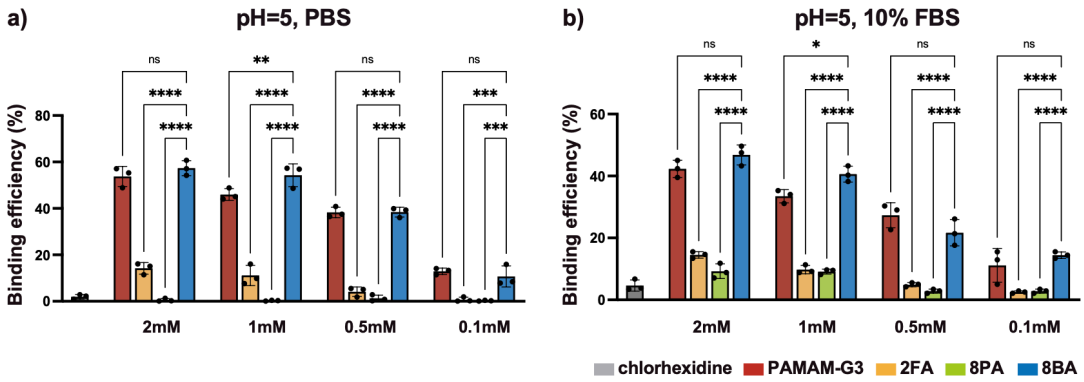
Supplementary Fig. 8. Zeta potential of PAMAM-G3, 2FA, 8PA and 8BA at pH 7.4 and pH 5. Bar plot showing zeta potential (mV) of PAMAM-G3, 2FA, 8PA and 8BA measured at physiological pH (7.4, blue bars) and acidic pH (5, red bars). Data are presented as mean ± SD, n = 3 independent measurements. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



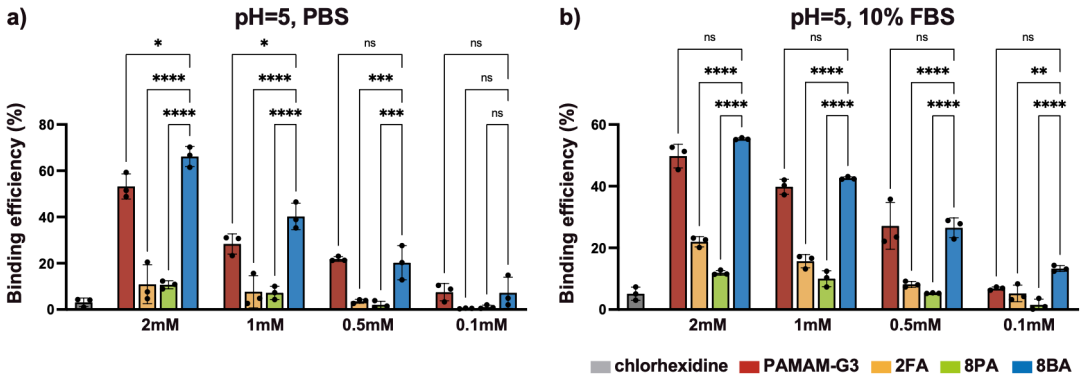
Supplementary Fig. 9. Chemical structure of the 8PA (a) and pore analysis performed using Mercury software (b). The orange regions in (b) represent the visualized void space, showing that no channel-like porous structure is present in 8PA. Atoms are coded as follows: red, oxygen; blue, nitrogen; gray, carbon; white, hydrogen; green, fluorine.



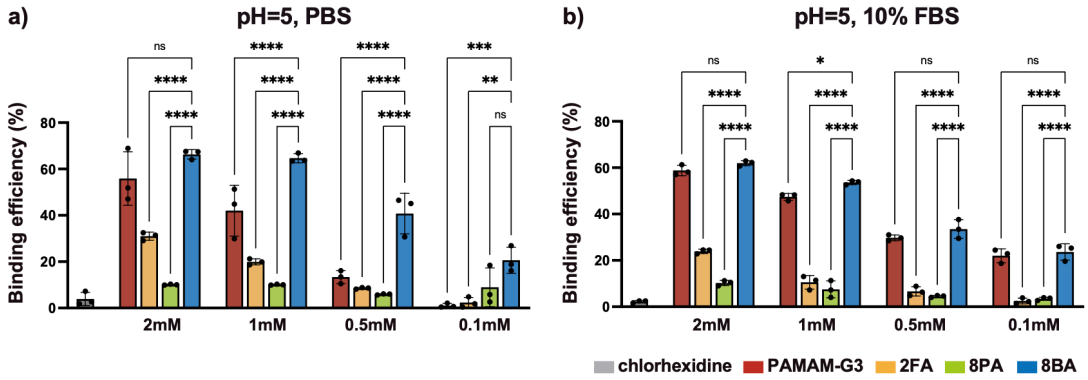
Supplementary Fig. 10. Acidic pH-dependent cfDNA adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward cell-free DNA (cfDNA) at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM under acidic conditions (pH 5.0). b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward cfDNA at the same concentrations under acidic conditions (pH 5.0) with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



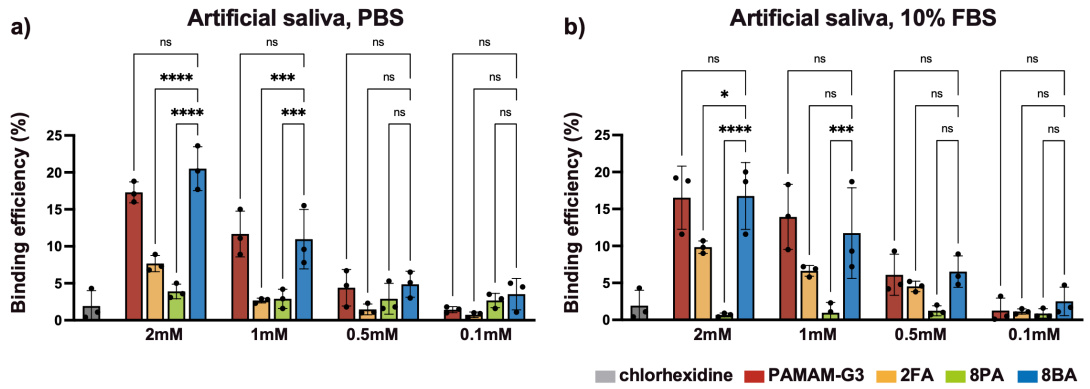
Supplementary Fig. 11. Acidic pH-dependent LPS adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward LPS at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM under acidic conditions (pH 5.0). b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward LPS at the same concentrations under acidic conditions (pH 5.0) with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



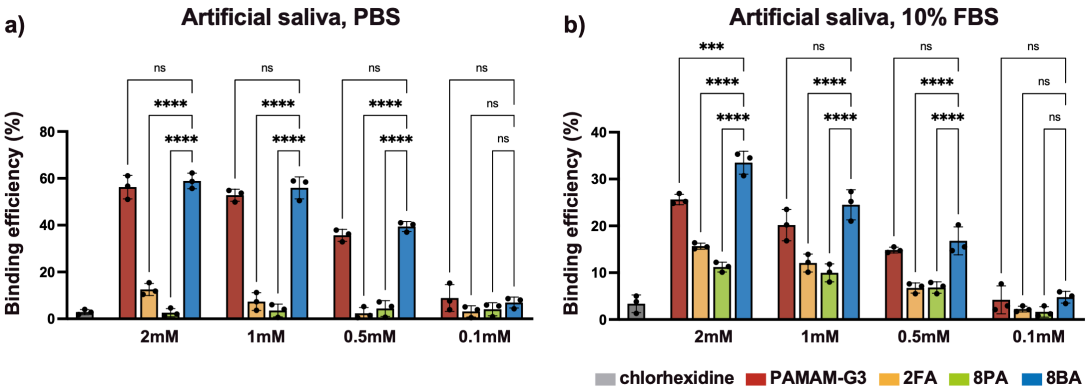
Supplementary Fig. 12. Acidic pH-dependent TNF-α adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward TNF-α at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM under acidic conditions (pH 5.0). b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward TNF-α at the same concentrations under acidic conditions (pH 5.0) with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



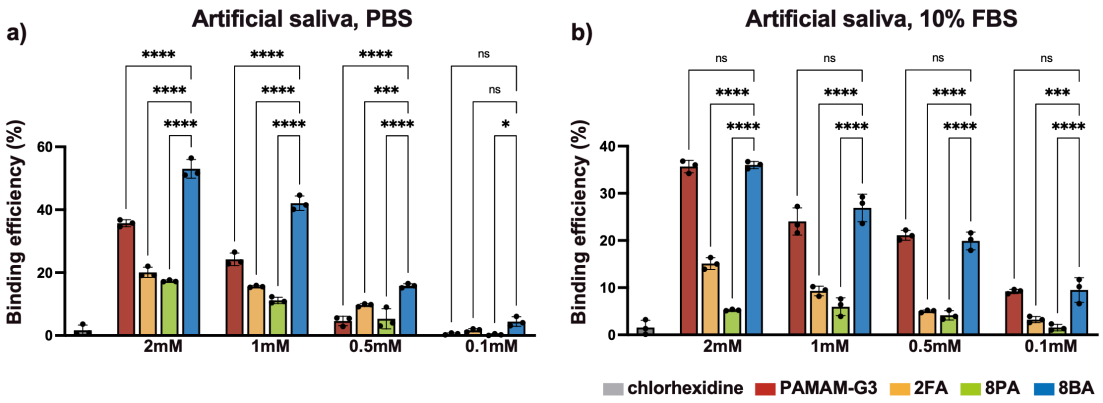
Supplementary Fig. 13. Acidic pH-dependent IL-6 adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward IL-6 at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM under acidic conditions (pH 5.0). b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward IL-6 at the same concentrations under acidic conditions (pH 5.0) with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



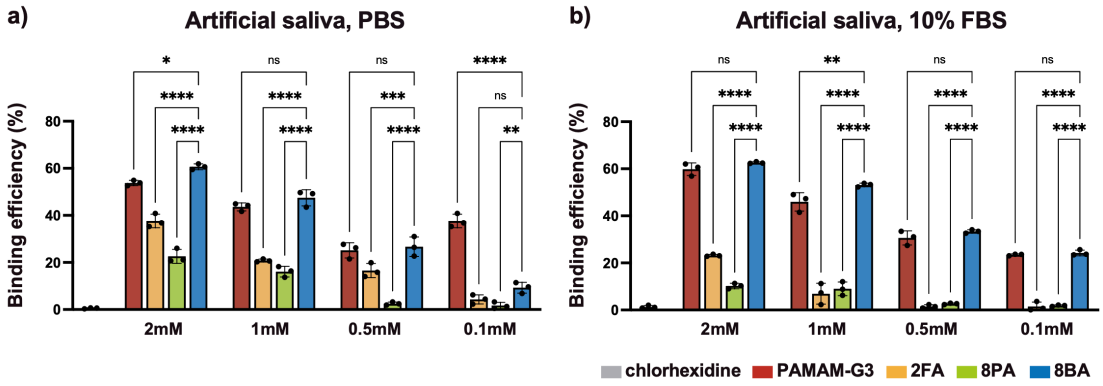
Supplementary Fig. 14. Artificial saliva-dependent cfDNA adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward cfDNA at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM in artificial saliva environment. b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward cfDNA at the same concentrations in artificial saliva environment with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



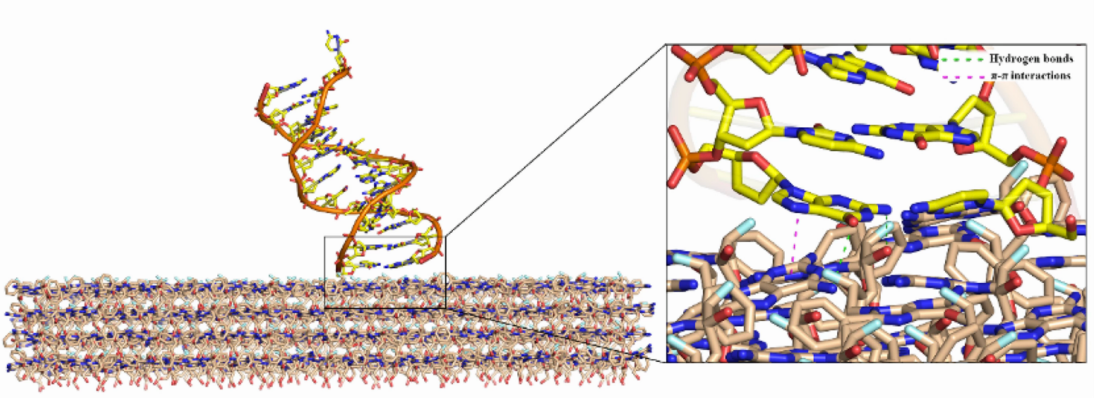
Supplementary Fig. 15. Artificial saliva-dependent LPS adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward LPS at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM in artificial saliva environment. b, Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward LPS at the same concentrations in artificial saliva environment with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



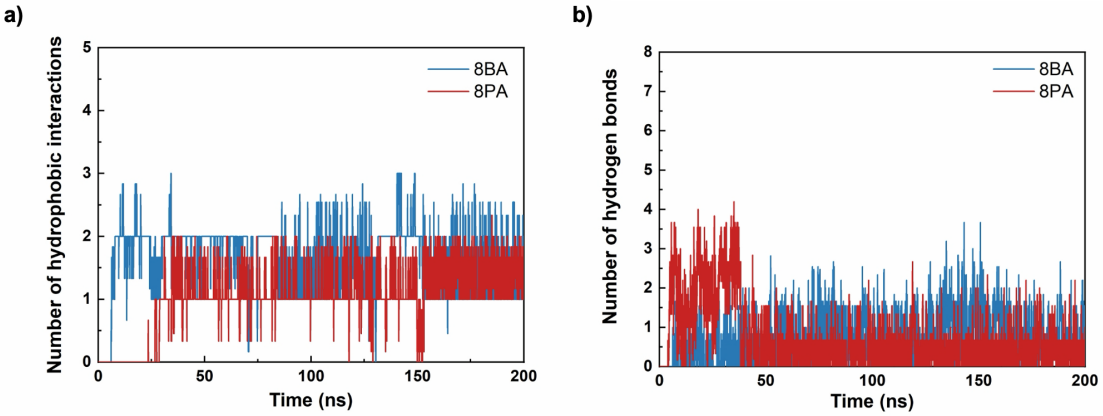
Supplementary Fig. 16. Artificial saliva-dependent TNF-α adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward TNF-α at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM in artificial saliva environment. b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward TNF-α at the same concentrations in artificial saliva environment with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



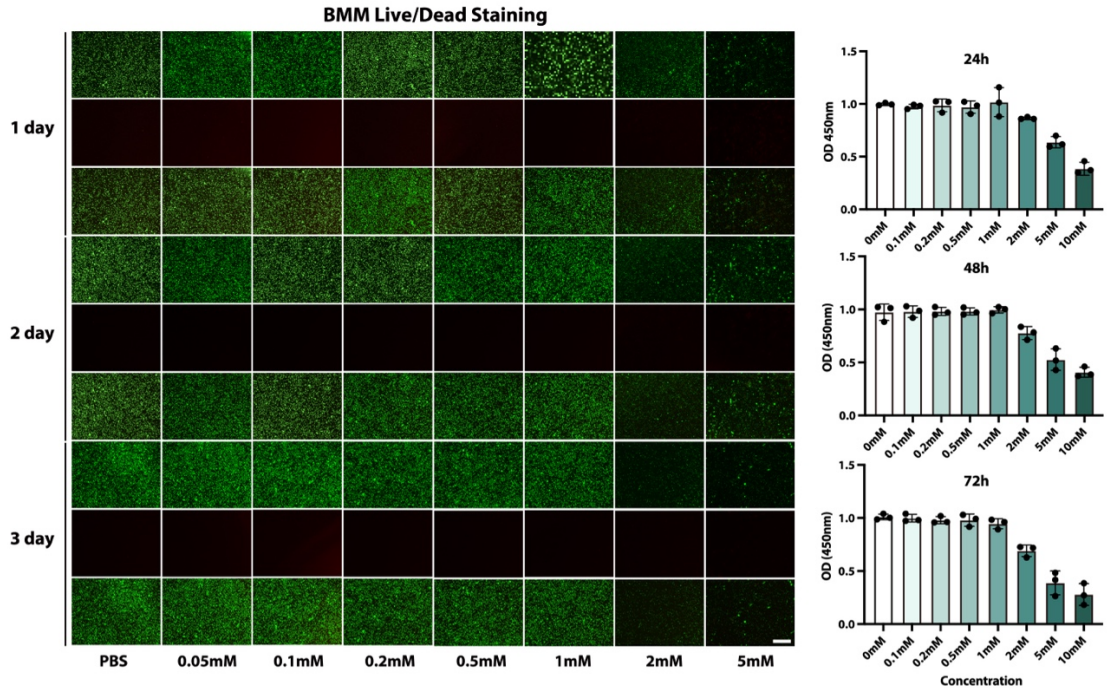
Supplementary Fig. 17. Artificial saliva-dependent IL-6 adsorption efficiency of various materials in the absence or presence of serum proteins. a) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward IL-6 at concentrations of 2 mM, 1 mM, 0.5 mM and 0.1 mM in artificial saliva environment. b) Bar plot showing adsorption efficiency of chlorhexidine, PAMAM-G3, 2FA, 8PA and 8BA toward IL-6 at the same concentrations in artificial saliva environment with 10% FBS. Data are presented as mean ± SD, n = 3 independent experiments. Statistical analyses were performed using two-way ANOVA with Tukey's multiple comparisons test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.



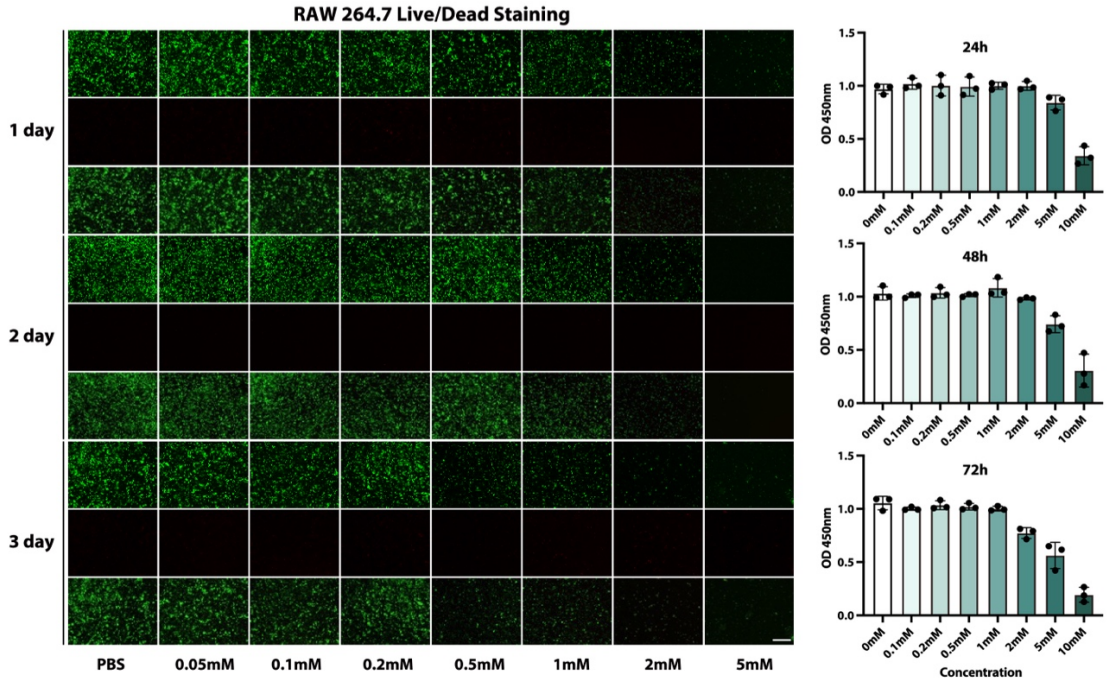
Supplementary Fig. 18. MD simulation of cfDNA binding on 8PA material surface, showing hydrogen bonds and π-π interactions. MD simulation of cfDNA binding on 8PA material surface. Green dashed lines indicate hydrogen bonds, red dashed lines indicate π-π interactions.



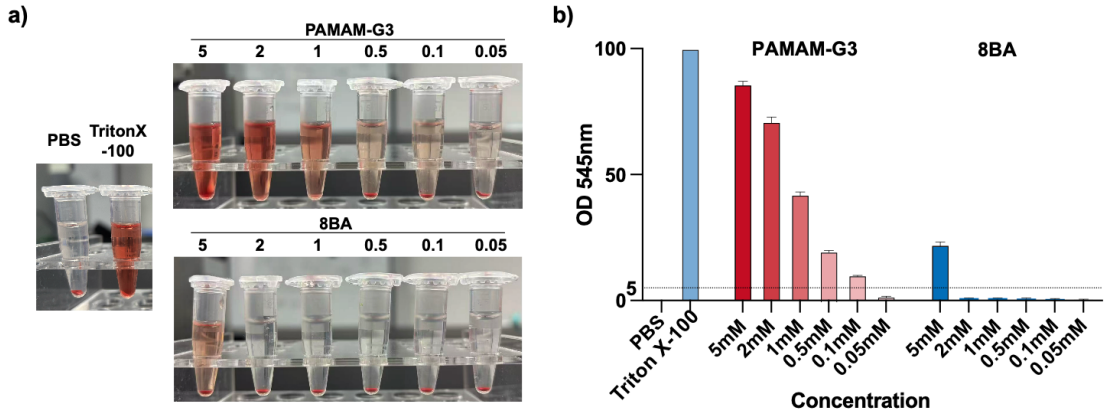
Supplementary Fig. 19. Number of hydrophobic interactions or hydrogen bonds in 8BA (blue) and 8PA (red) systems.



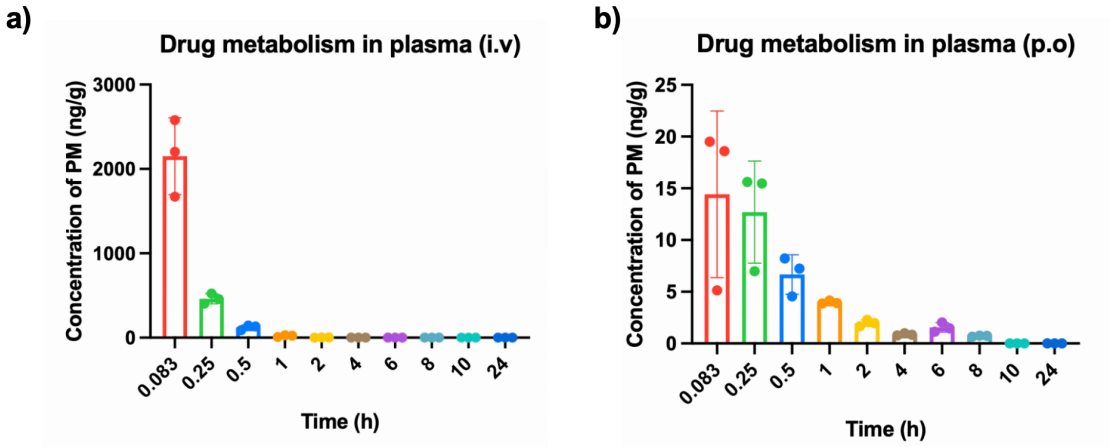
Supplementary Fig. 20. Viability and cytotoxicity of BMMs cultured with various concentrations of 8BA for three days. a) Representative live/dead staining images of BMMs after co-culture with indicated concentrations of 8BA for three days. Live cells are shown in green and dead cells in red. Scale bar: 500 μm. b) Bar plot showing cell viability measured by CCK-8 assay under the same conditions, n = 3 independent experiments. Data are presented as mean ± SD.



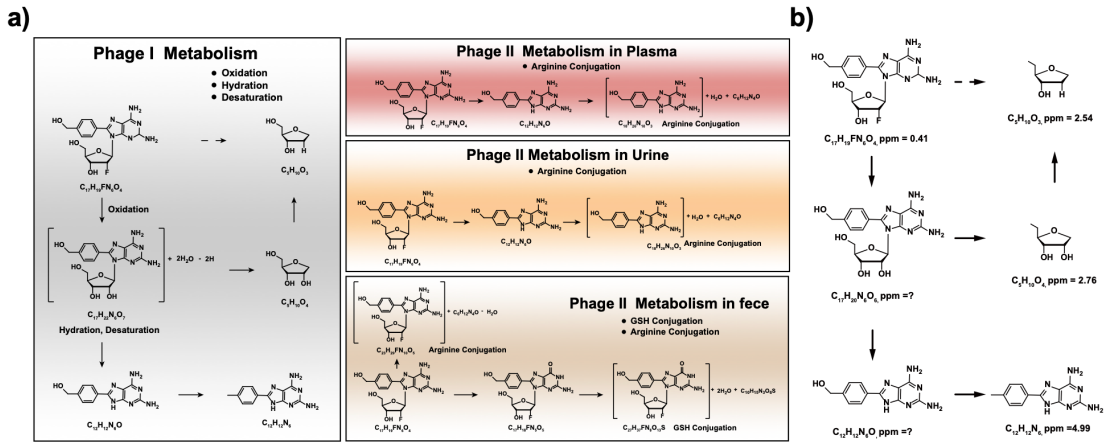
Supplementary Fig. 21. Viability and cytotoxicity of RAW 264.7 cells cultured with various concentrations of 8BA for three days. a) Representative live/dead staining images of RAW 264.7 cells after co-culture with indicated concentrations of 8BA for three days. Live cells are shown in green and dead cells in red. Scale bar: 500 μm. b) Bar plot showing cell viability measured by CCK-8 assay under the same conditions, n = 3 independent experiments. Data are presented as mean ± SD.



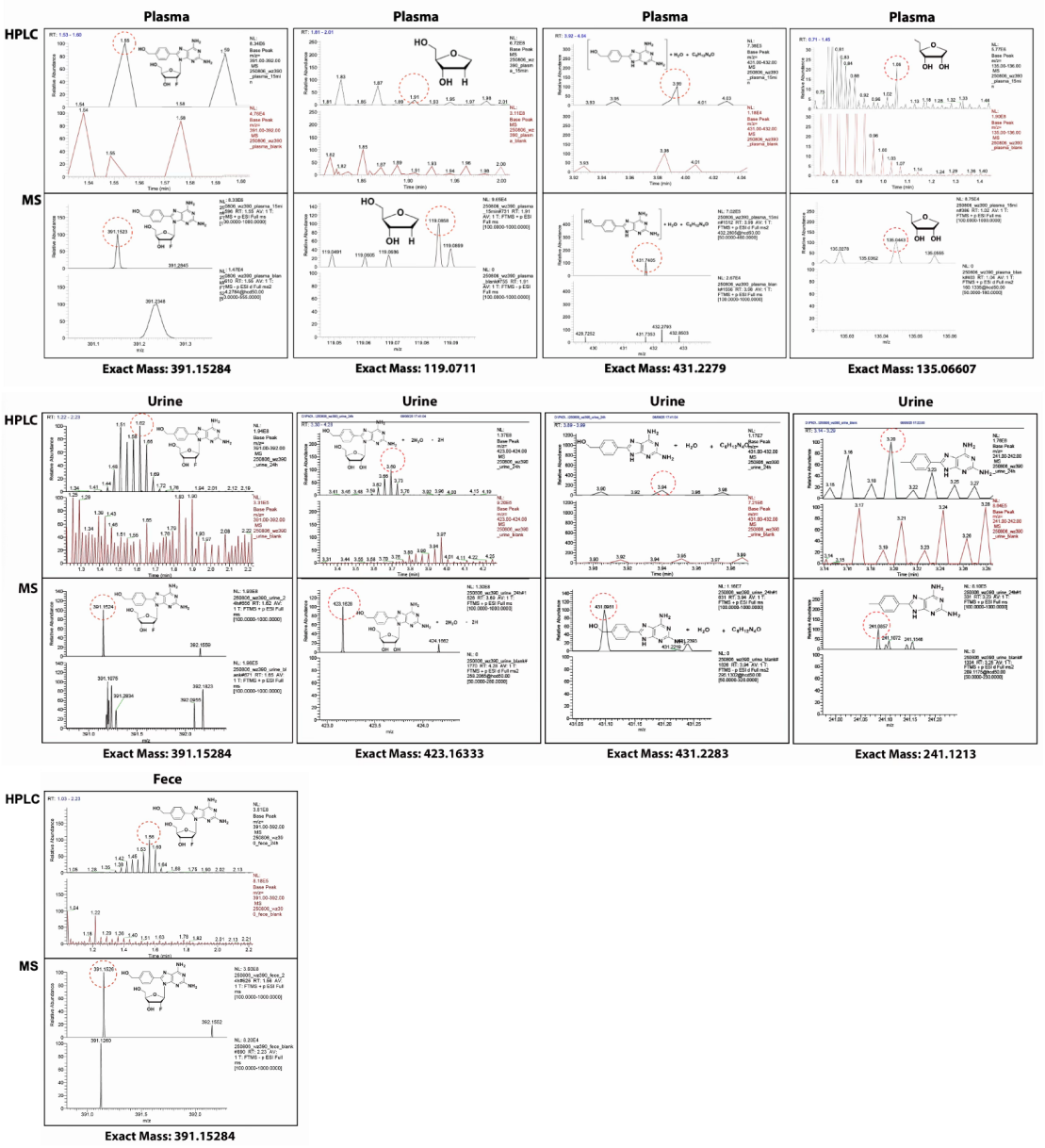
Supplementary Fig. 22. Hemocompatibility evaluation of PAMAM-G3 and 8BA at various concentrations. a) Representative images of hemolysis assay after incubation of red blood cells with PBS (negative control), Triton X-100 (positive control), PAMAM-G3 at indicated concentrations (0.05, 0.1, 0.5, 1, 2, 5 mM), and 8BA at indicated concentrations (0.05, 0.1, 0.5, 1, 2, 5 mM). b) Bar plot showing hemolysis percentage of red blood cells after incubation with PBS (negative control), PAMAM-G3 at indicated concentrations, and 8BA at indicated concentrations. Hemoglobin release was measured by OD at 545 nm. PBS group was set as 0% hemolysis reference (or 100% viability). Data are presented as mean ± SD, n=3.



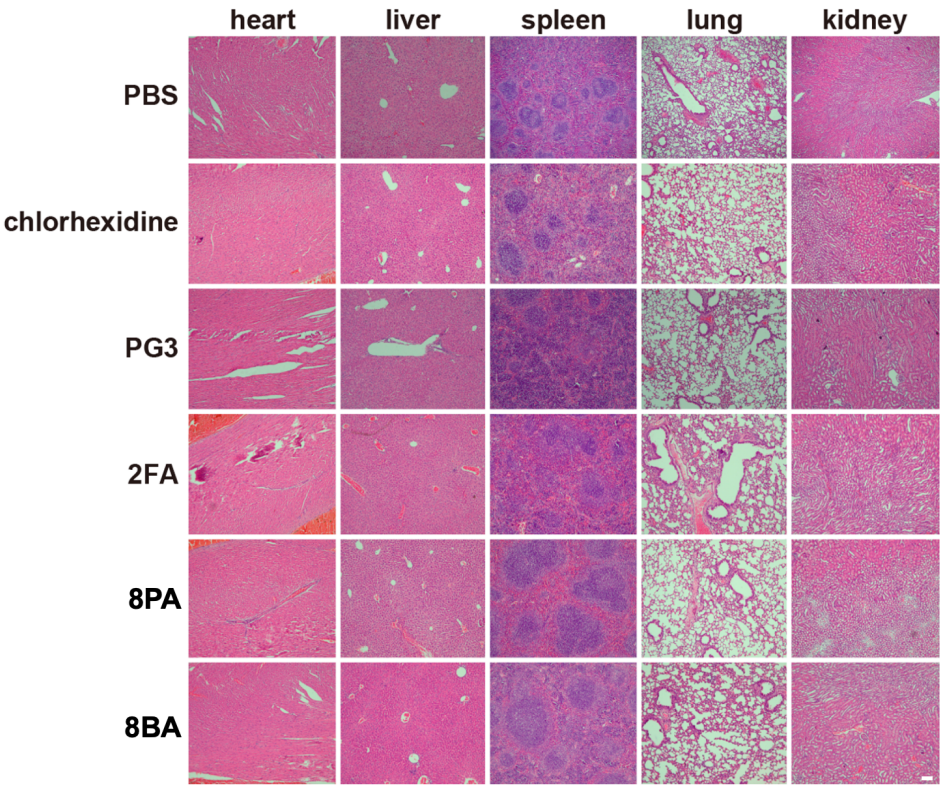
Supplementary Fig. 23. Pharmacokinetic profiles of 8BA after intravenous (i.v.) and oral administration (o.p.) in mice. a) Bar plot showing plasma concentration of 8BA (5 mg/kg) over time following intravenous (i.v.) administration at 0.083, 0.25, 0.5, 1, 2, 4, 6, 8, 10 and 24 h post-injection. b) Bar plot showing plasma concentration of 8BA (10 mg/kg) over time following oral (p.o.) administration at the same time points. Data are presented as mean ± SD, n = 3.



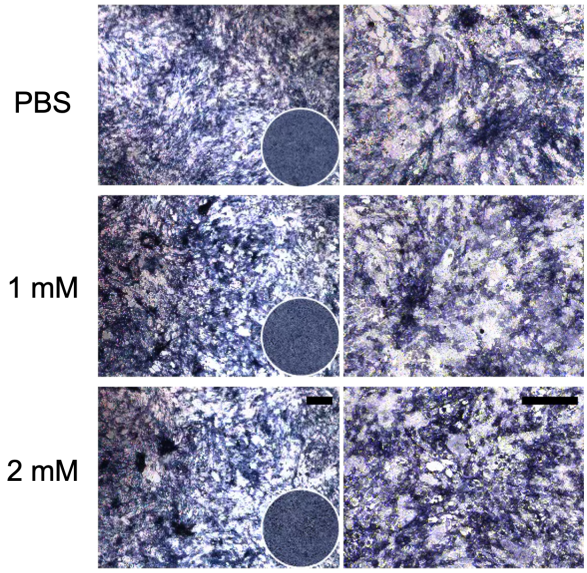
Supplementary Fig. 24. Identification and proposed metabolic pathway of 8BA *in vivo* after intravenous administration. a) Predicted chemical structures of all metabolites identified in plasma (1 h), feces (48 h), and urine (48 h) samples after intravenous administration of 8BA. b) A schematic summary of 8BA metabolism *in vivo* is presented.



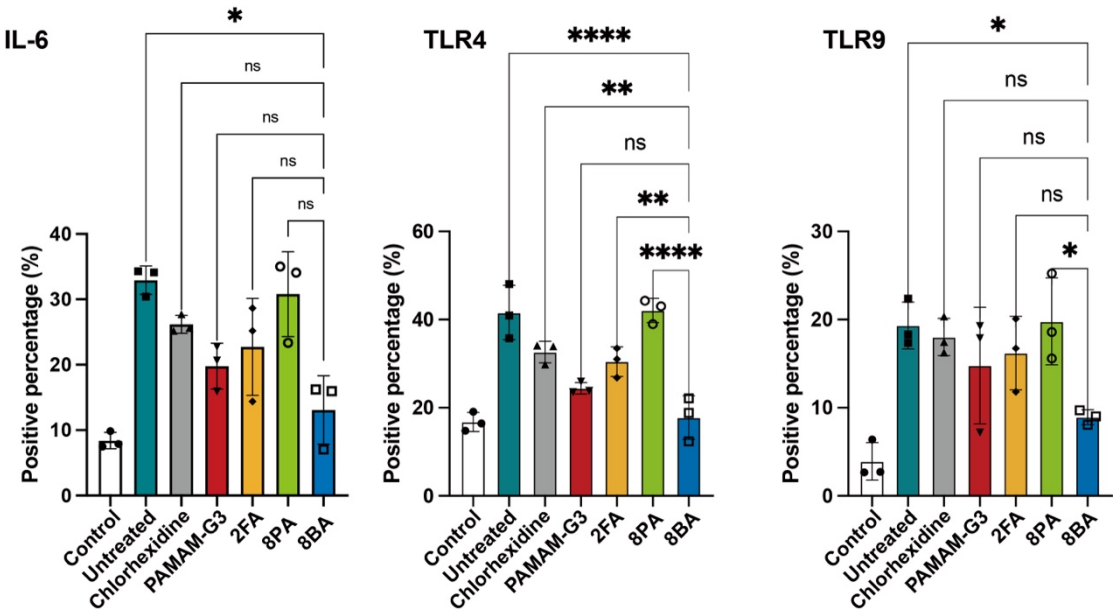
Supplementary Fig. 25. The main metabolites of the *in vivo* metabolism of 8BA. For HPLC, the black one indicates the 8BA treated group and the red one indicates the control group. For MS, the top one indicates the 8BA treated group and the bottom one indicates the control group.



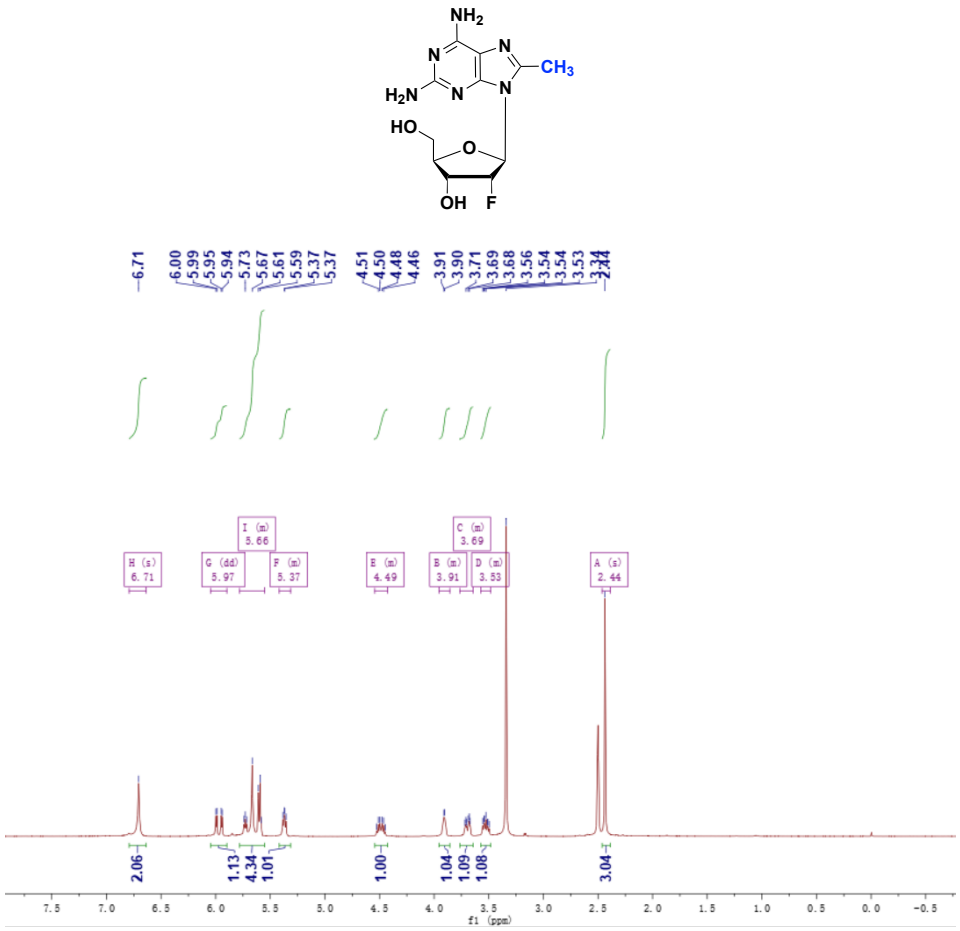
Supplementary Fig. 26. *In vivo* biocompatibility. Paraffin-embedded heart, liver, spleen, lung and kidney of mice on 14 d post-treatment were sectioned and stained by H&E. Scale bars: 100 μm.



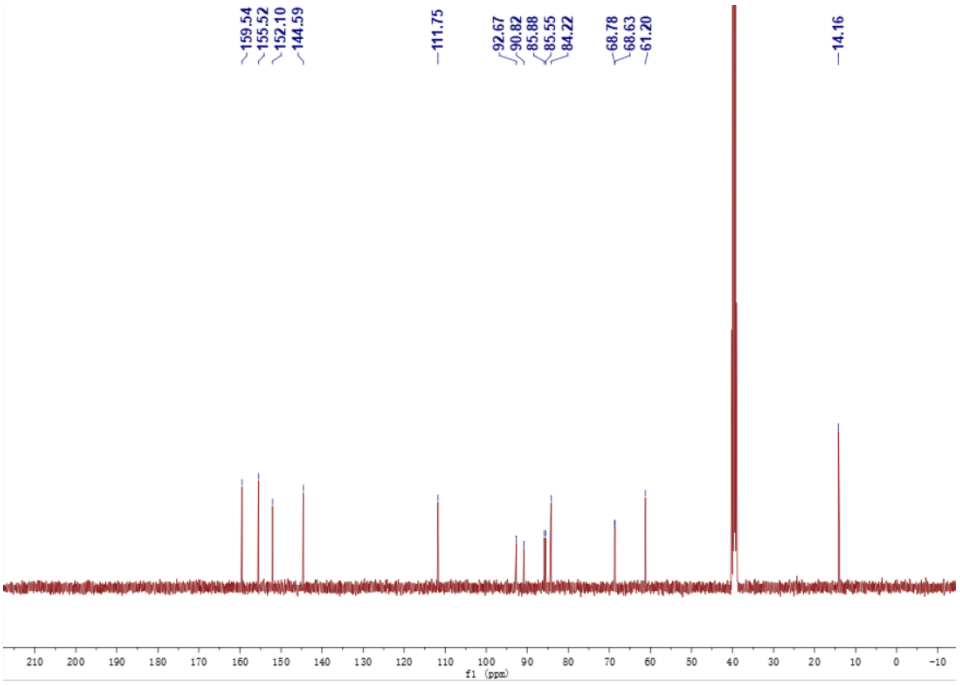
Supplementary Fig. 27. Osteogenic differentiation of BMSCs treated with PBS, 1 mM 8BA or 2 mM 8BA. Representative images of ALP showing osteogenic differentiation of BMSCs after culture with PBS (control), 1 mM 8BA or 2 mM 8BA for 7 days. Scale bar: 75 μm.



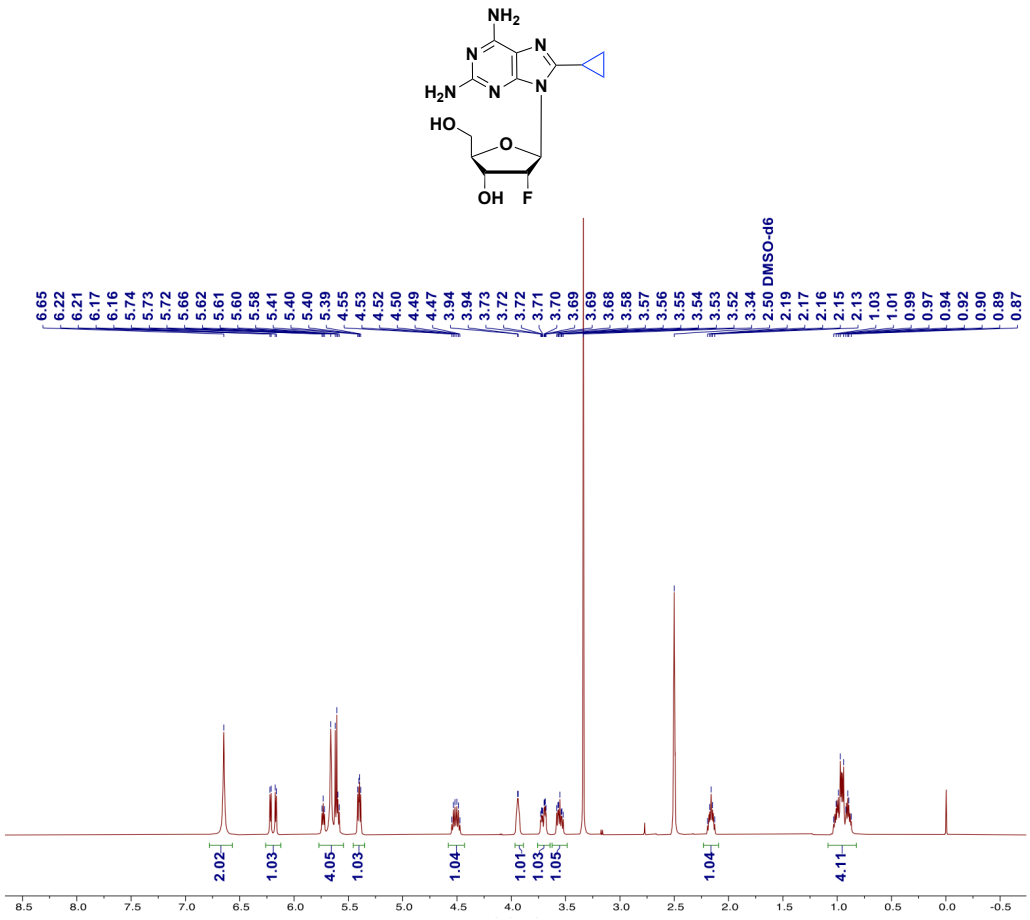
Supplementary Fig. 28. IHC analysis of IL-6, TLR4 and TLR9 expression in mouse tissues across different treatment groups. Bar plot showing the positive rate of IL-6, TLR4 and TLR9 levels in IHC staining of mice in different groups. Statistical analyses were performed using one-way ANOVA with Tukey's multiple comparisons test. Bar graphs are shown as mean ± SD, n = 3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.



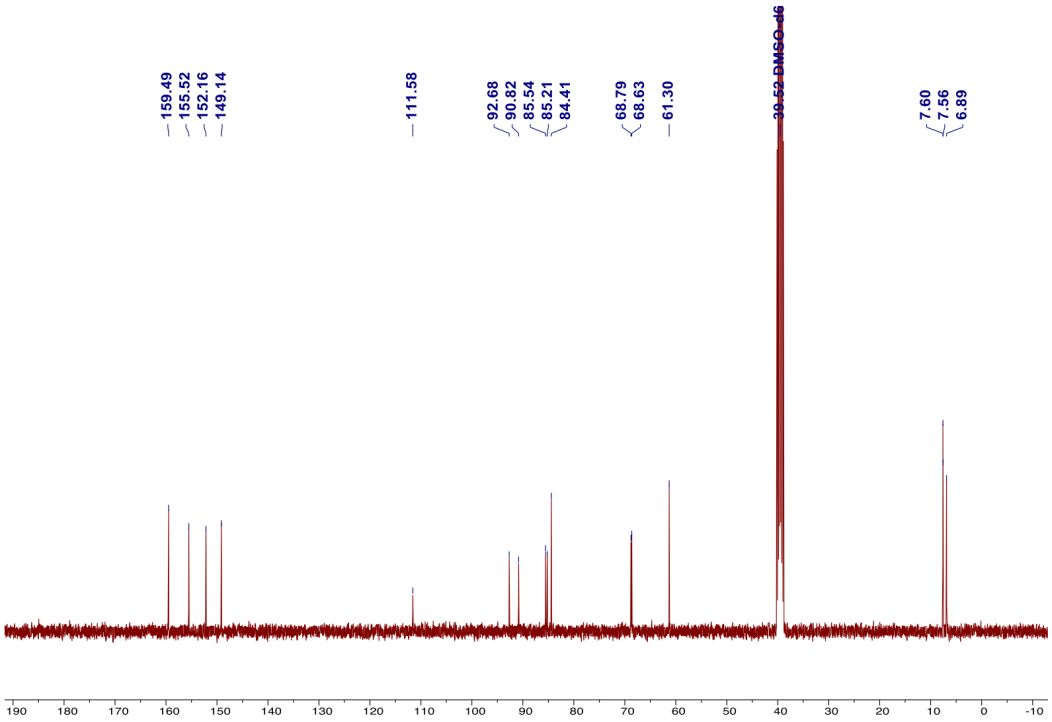
Supplementary Fig. 29. ^{1}H NMR spectra of compound **2**



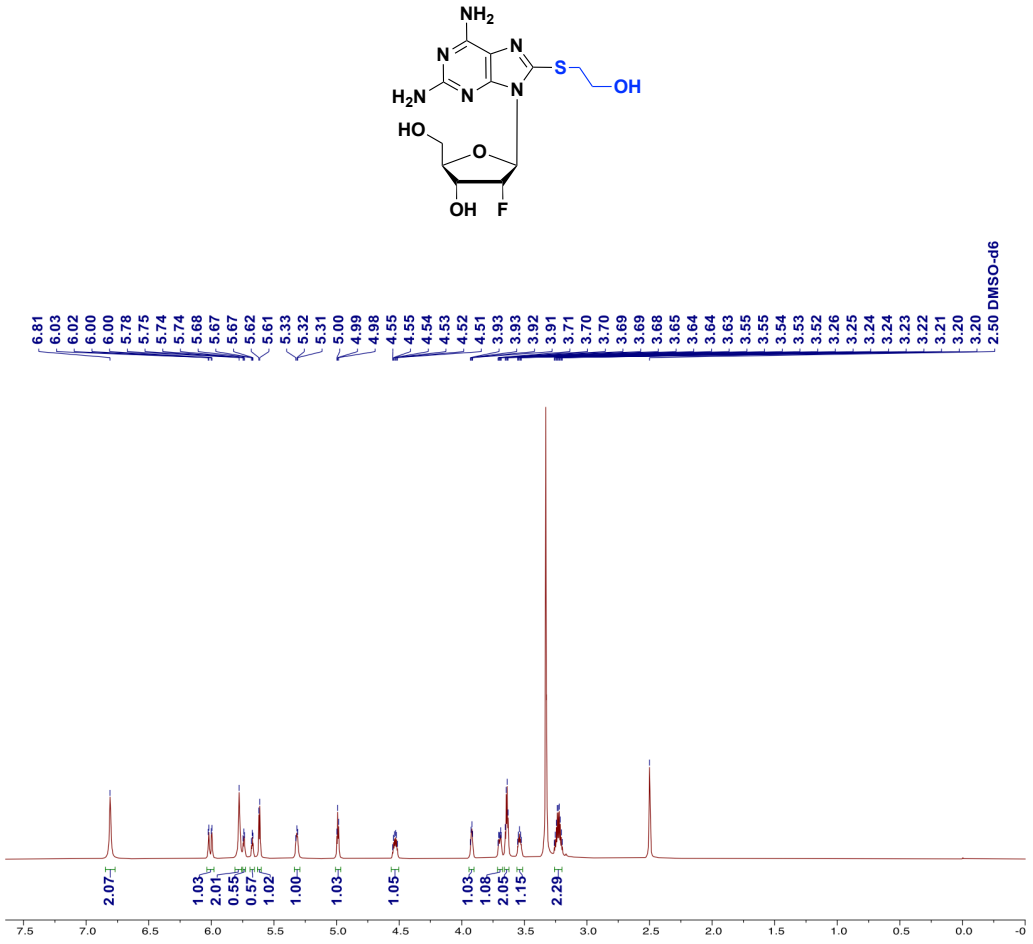
Supplementary Fig. 30. ^{13}C NMR spectra of compound **2**



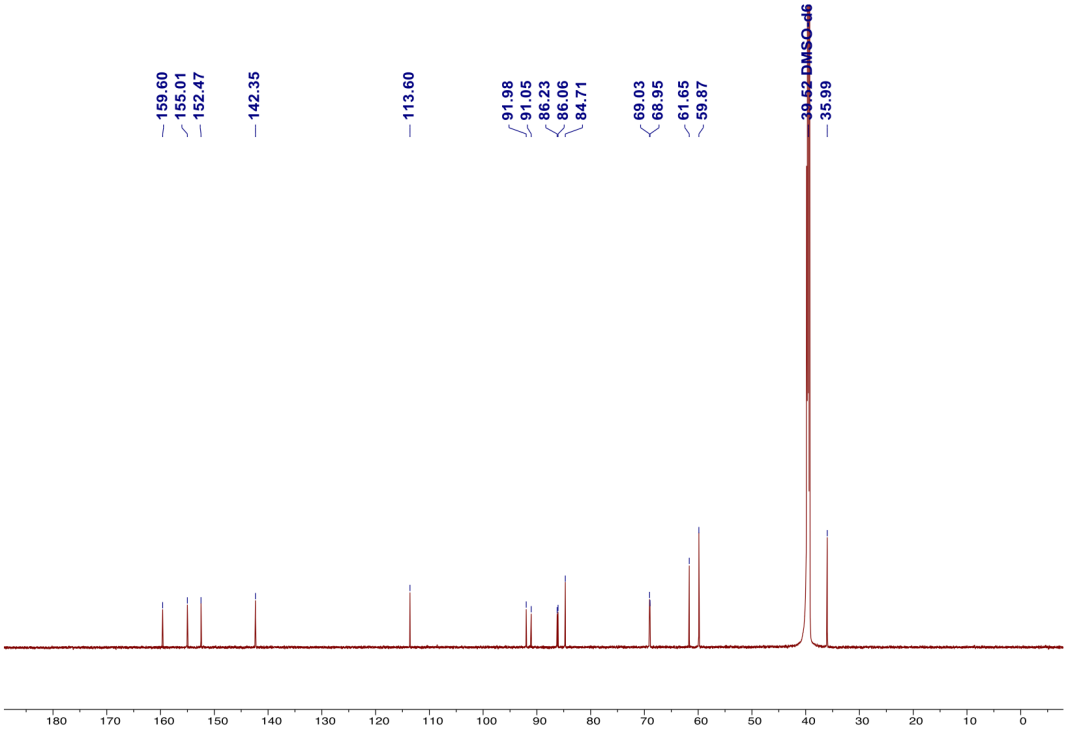
Supplementary Fig. 31. ^{1}H NMR spectra of compound **3**



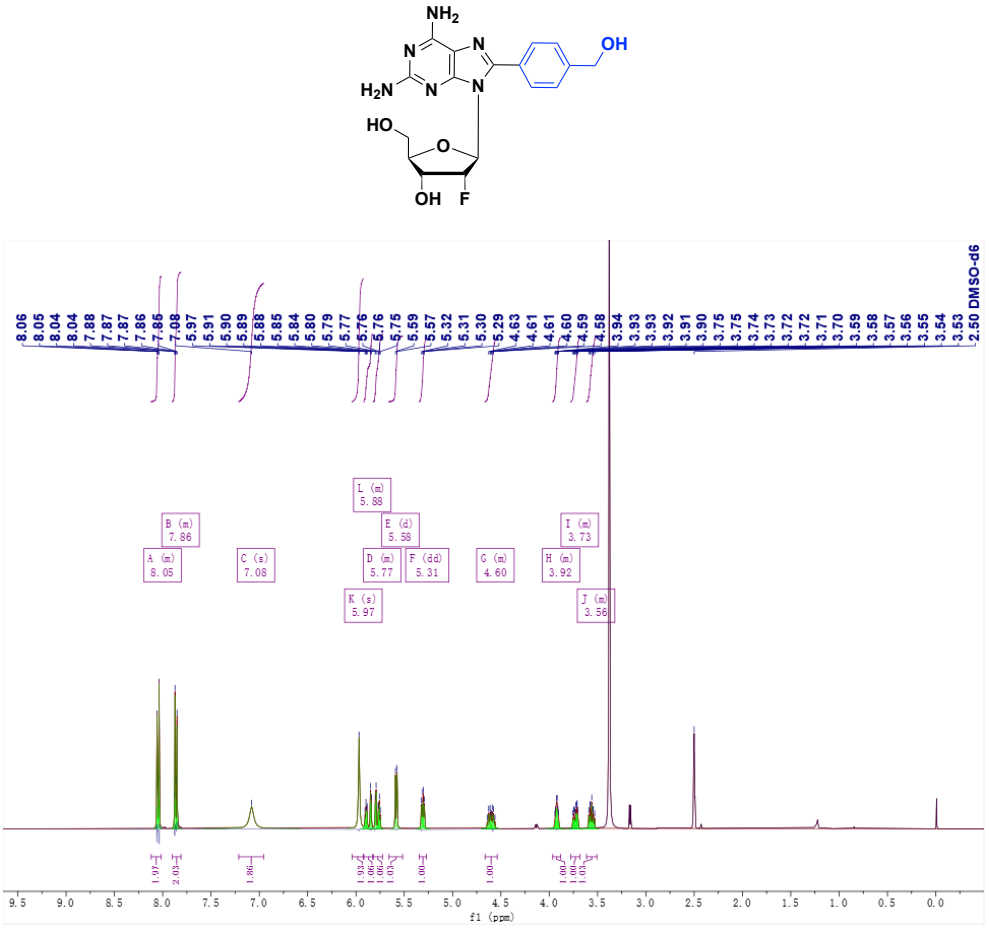
Supplementary Fig. 32. ^{13}C NMR spectra of compound **3**



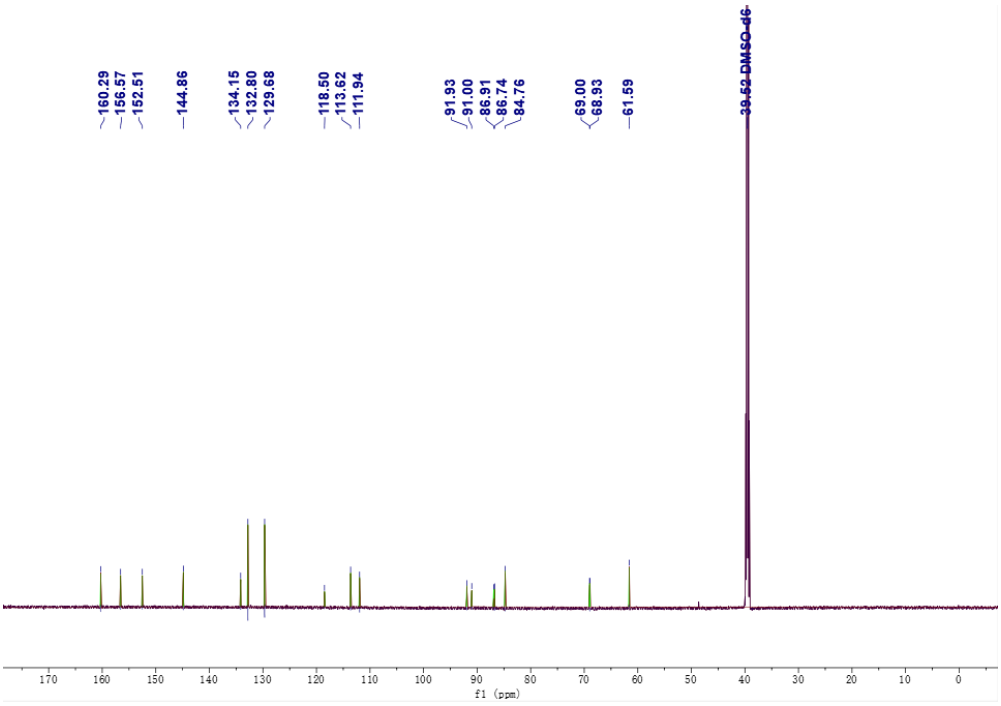
Supplementary Fig. 33. ^{1}H NMR spectra of compound **4**



Supplementary Fig. 34. ^{13}C NMR spectra of compound **4**



Supplementary Fig. 35. ^{1}H NMR spectra of compound **6**



Supplementary Fig. 36. ^{13}C NMR spectra of compound **6**

3. Supplementary Tables

Supplementary Table 1. The total computational time for all predicted structures.

Molecule	Molecule Cost (h)*
Target 1	15.49
Target 2	17.68
Target 3	24.11
Target 4	18.06
Target 5	10.81
Target 6	7.05
Target 7	7.97
Target 8	16.55
Target 9	9.59
Target 10	10.76
Target 11	18.15
Target 12	24.58
N1	21.15
N2	33.59
N5	35.33
N6	32.34
N7	37.53

*The above time was reported in our computing environment (Supplementary Material Section 1.4)

Supplementary Table 2. Comparison of the experimental structure and matching predictions of Targets **1-12**, in terms of the relative deviation in lattice parameters, volume and density: ((pred.−expt.)/expt.) × 100%. The $RMSD_{15}$ and $RMSD_1$ are also given in Å. Experimental values for lattice parameters and density are reported in Å, °, and g/cm³, respectively.

	a	b	c	β	ρ	$RMSD_{15}$ (Å)	Packing Number	$RMSD_1$ (Å)
Experiment (Target 1)	7.7046	10.6062	9.3384	95.033	1.338			
SE3CSP	5.24	-3.33	52.91	18.53	4.29	1.296	8	0.050
Genarris 3.0	0.60	-16.55	15.67	29.83	22.71	1.334	15	0.055
Experiment (Target 2)	7.2643	10.6393	15.6331	90	1.606			
SE3CSP	5.62	-0.69	-15.69	0	13.07	0.939	15	0.093
Genarris 3.0	-2.41	-16.46	-2.77	0	26.14	1.277	15	0.304
Experiment (Target 3)	8.2506	8.9643	15.087	91.21	1.484			
SE3CSP	35.44	-21.77	-9.61	-12.60	6.08	1.762	5	1.269
Genarris 3.0	-7.96	21.98	-17.81	-24.07	15.85	1.642	5	1.442
Experiment (Target 4)	12.5793	4.8531	17.2663	99.1624	1.5279			
SE3CSP	-31.33	56.72	18.08	13.12	-16.17	1.710	8	0.401
Genarris 3.0	-9.94	48.98	-18.92	-50.32	16.55	2.262	7	0.446
Experiment (Target 5)	9.3538	12.1757	7.2286	104.593	1.669			
SE3CSP	7.92	-4.44	5.04	-46.00	7.15	0.88	15	0.030
Genarris 3.0	11.91	0.63	-21.38	-22.00	10.49	1.419	10	0.038
Experiment (Target 6)	6.97	9.514	9.752	90	1.152			
SE3CSP	-0.25	0.11	-8.14	0	10.33	0.351	15	0.027
Genarris 3.0	37.05	28.64	-10.20	0	37.51	1.198	7	0.622

$RMSD_1$ represents the root-mean-square deviation between predicted and experimental molecular conformations, whereas $RMSD_{15}$ represents the root-mean-square deviation of overlaid 15-molecule packing clusters.

Supplementary Table 2 (continued). Comparison of the experimental structure and matching predictions of Targets **1-12**, in terms of the relative deviation in lattice parameters, volume and density: ((pred.−expt.)/expt.) × 100%. The $RMSD_{15}$ and $RMSD_1$ are also given in Å. Experimental values for lattice parameters and density are reported in Å, °, and g/cm³, respectively.

	a	b	c	β	ρ	$RMSD_{15}$ (Å)	Packing Number	$RMSD_1$ (Å)
Experiment (Target 7)	3.8943	13.5109	14.4296	93.636	2.528			
SE3CSP	6.11	-4.64	1.58	-21.89	1.84	0.900	15	0.022
Genarris 3.0	-18.24	-2.15	-9.75	-8.32	38.59	0.740	15	0.036
Experiment (Target 8)	13.06	9.738	9.335	105.8	1.479			
SE3CSP	23.10	-11.41	-7.58	12.58	9.34	1.302	11	0.151
Genarris 3.0	-13.19	-5.96	5.10	-18.82	12.45	0.798	12	0.308
Experiment (Target 9)	9.6451	7.381	16.185	90	1.385			
SE3CSP	11.01	0.83	-13.83	0	3.75	1.947	9	0.035
Genarris 3.0	-11.51	-12.73	-3.55	0	34.26	1.090	15	0.078
Experiment (Target 10)	12.6386	5.9794	11.422	96.807	1.837			
SE3CSP	14.52	-3.79	-9.78	10.49	1.74	0.977	8	0.298
Genarris 3.0	-36.71	125.89	-41.50	-4.54	18.83	1.525	5	0.746
Experiment (Target 11)	9.8888	8.887	24.969	90	1.566			
SE3CSP	67.30	-2.72	-33.22	13.93	-5.73	0.785	5	0.648
Genarris 3.0	78.03	-32.78	-23.60	0	9.38	1.644	2	0.734
Experiment (Target 12)*	14.4724	8.7359	21.9759	96.236	1.478			
SE3CSP	21.36	-9.93	-1.22	4.11	-6.46	1.991	9	1.118
Genarris 3.0	-17.35	7.57	1.44	5.98	12.68	1.395	7	1.605

*This experimentally observed structure was assigned to space group No. 14, whereas the best-matched predicted structure was identified in space group No. 15, likely owing to subtle differences in symmetry and unit-cell setting. The unit-cell parameters reported here were therefore transformed into comparable settings for direct comparison.

Supplementary Table 3. Optimization of density tolerance windows for prospective crystal structure identification

Molecule	ρ_T = 0.10 g/cm^3	ρ_T = 0.15 g/cm^3	ρ_T = 0.20 g/cm^3
Target 1	OOW	14	14
Target 2	OOW	20	20
Target 4	OOW	>500	>500
Target 5	OOW	1	1
Target 6	1	1	1
Target 7	8	8	8
Target 8	OOW	4	23
Target 9	>500	>500	>500
Target 10	1	1	1
Target 12	OOW	>500	>500
N1	32	230	>500
N2	OOW	35	35
N5	23	41	>500
N6	2	2	2

*OOW: Outside the predicted density window ($\rho_{predict} \pm \rho_T$).

Supplementary Table 4. Nucleoside structural parameters.

Name	**2**
Conformer Torsion angle χ (O4'-C1'-N9-C4)	χ=47.8(7)°, syn
Sugar puckering	C3'-endo, P=28.7(2)°, τ_m=36.1(9)°
Torsion angle γ (O5'-C5'-C4'-C3')	γ=54.2(0)°
Name	**3**
Conformer Torsion angle χ (O4'-C1'-N9-C4)	χ=39.5(9)°, syn
Sugar puckering	C3'-endo, P=27.3(8)°, τ_m=36.1(8)°
Torsion angle γ (O5'-C5'-C4'-C3')	γ=50.7(2)°
Name	**4**
Conformer Torsion angle χ (O4'-C1'-N9-C4)	χ=64.6(9)°, syn
Sugar puckering	C2'-endo, P=163.0(1)°, τ_m=39.8(5)°
Torsion angle γ (O5'-C5'-C4'-C3')	γ=49.3(7)°
Name	**8BA (6)**
Conformer Torsion angle χ (O4'-C1'-N9-C4)	χ=45.6(4)°, syn
Sugar puckering	C2'-endo, P=162.0(0)°, τ_m=39.7(1)°
Torsion angle γ (O5'-C5'-C4'-C3')	γ=47.2(6)°

Supplementary Table 5. Pharmacokinetic parameters of 8BA in mouse plasma after intravenous administration

Parameters	Unit	No.1	No.2	No.3	Mean	SD
$AUC_{(0-t)}$	ug/L*h	671.60	450.46	578.68	566.91	111.04
$AUC_{(0-\infty)}$	ug/L*h	671.82	450.72	579.10	567.21	111.03
C_{max}	ug/L	2578.55	1672.05	2204.95	2151.85	455.58
$MRT_{(0-t)}$	h	0.18	0.18	0.19	0.18	0.01
$MRT_{(0-\infty)}$	h	0.19	0.18	0.20	0.19	0.01
$T_{1/2z}$	h	0.98	1.19	1.45	1.21	0.23
T_{max}	h	0.08	0.08	0.08	0.08	0.00

Supplementary Table 6. Pharmacokinetic parameters of 8BA in mouse plasma after oral administration

Parameters	Unit	No.1	No.2	No.3	Mean	SD
$AUC_{(0-t)}$	ug/L*h	19.26	14.73	21.33	18.44	3.38
$AUC_{(0-\infty)}$	ug/L*h	22.23	18.93	24.73	21.96	2.91
C_{max}	ug/L	19.52	6.99	18.60	15.03	6.99
$MRT_{(0-t)}$	h	2.08	2.51	2.33	2.31	0.21
$MRT_{(0-\infty)}$	h	3.22	4.90	3.82	3.98	0.85
$T_{1/2z}$	h	2.75	3.79	3.29	3.28	0.52
T_{max}	h	0.08	0.25	0.08	0.14	0.10

$AUC_{(0-t)}$, area under the time-concentration curve from the time of dosing to the last measurable concentration; **$AUC_{(0-\infty)}$**, AUC from dosing time extrapolated to infinity; **C_{max}**, maximum observed concentration; **$MRT_{(0-t)}$**, mean residence time from the time of dosing to the time of the last measurable concentration; **$MRT_{(0-\infty)}$**, MRT extrapolated to infinity; **$T_{1/2z}$**, terminal half-life, **T_{max}**, time for maximal concentration.

Supplementary Table 7. The hydrogen-bond geometry for **1** (Å, °)

D-H	d(D-H)	d(H···A)	<DHA	d(D···A)
O5'-H3···N2	0.82	1.993	165.23	2.793
O3'-H3'···O5'	0.82	1.839	163.51	2.636
N2-H2A···O4'	0.86	2.266	160.39	3.09
N2-H2B···O3'	0.86	2.25	168.66	3.097
C1'-H1···Br8	0.98	2.768	113.65	3.286
C5'-H5'A···N1	0.97	2.583	163.89	3.526
C3'-H3'···N2	0.98	2.674	125.7	3.342
C2'-H2'···O3'	0.98	2.581	142.14	3.408
N6-H6A···N1	0.86	2.213	169.45	3.063
N6-H6B···N1	0.86	2.175	161.02	3.001

Supplementary Table 8. The hydrogen-bond geometry for **2** (Å, °)

D-H	d(D-H)	d(H···A)	<DHA	d(D···A)
O5'-H5'···N3	0.82	1.946	164.85	2.746
O5'-H5'···O4'	0.82	2.49	110.02	2.876
O3'-H3'···O5'	0.82	1.825	171.59	2.639
N2-H2A···O4'	0.86	2.276	164.76	3.114
N2-H82···O3'	0.86	2.232	166.47	3.075
C5'-H5'A···N1	0.97	2.528	165.57	3.476
C2'-H2'A···N3	0.98	2.679	125.91	3.349
C2'-H2'···O2'	0.98	2.565	142.92	3.399
C10-H10B···O5'	0.96	2.534	166.61	3.475
N6-H6A···N7	0.86	2.212	167.21	3.057
N6-H6B···N1	0.86	2.176	161.63	3.004

Supplementary Table 9. The hydrogen-bond geometry for **3** (Å, °)

D-H	d(D-H)	d(H···A)	<DHA	d(D···A)
N6-H6A···N7	0.86	2.204	173.3	3.06
N6-H6B···N1	0.86	2.215	159.58	3.036
O5'-H5'···N3	0.82	1.943	165.82	2.745
O3'-H3'···O5'	0.82	1.86	173.05	2.676
C5'-H5'A···N6	0.97	2.698	159.92	3.624
N2-H2A···O3'	0.86	2.382	167.22	3.227
N2-H2B···O4'	0.86	2.433	145.2	3.177
C2'-H2'A···O5	0.98	2.628	134.13	3.384
C10-H10···O5'	0.98	2.378	164.88	3.334

Supplementary Table 10. The hydrogen-bond geometry for **4** (Å, °)

D-H	d(D-H)	d(H···A)	<DHA	d(D···A)
O5'-H5'···N3	0.82	1.985	168.83	2.794
O3'-H3'···O8	0.82	2.075	151.01	2.82
O1A-H1A···O5'	0.85	2.036	156.25	2.835
O1A-H1B···N1	0.85	2.037	174.32	2.884
O8-H8···N7	0.82	1.964	167.32	2.77
C8B-H8BA···F2'	0.97	2.547	142.3	3.367
C8A-H8AB···F2'	0.97	2.6	115.98	3.148
C1'-H1'···S8	0.98	2.756	112.16	3.255
N2-H2A···O1A	0.86	2.411	156.36	3.218
N2-H2B···N2	0.86	2.362	158.87	3.179
C2'-H2'···O4'	0.98	2.523	159.91	3.46
C2'-H2'···O5'	0.98	2.581	114.42	3.116
C2'-H2'···N3	0.98	2.575	124.92	3.238
N6-H6A···O3'	0.855	2.31	129.94	2.933
N6-H6B···O8	0.855	2.242	166.58	3.08

Supplementary Table 11. The hydrogen-bond geometry for **6** (Å, °)

D-H	d(D-H)	d(H···A)	<DHA	d(D···A)
O2-H2A···O1	0.86	1.902	178.62	2.762
O2-H2B···N7	0.86	1.98	156.52	2.79
O5'-H5'···N3	0.83	1.958	177.02	2.787
O16-H16···O3'	0.83	1.955	171.86	2.779
O3'-H3'···N1	0.83	1.952	149.89	2.702
N6-H6A···O3'	0.87	2.634	123.66	3.199
N6-H6B···O2	0.87	2.09	165.81	2.941
N2-H2A···O16	0.87	2.139	158.01	2.963
O1-H1A···O5'	0.86	1.941	167.31	2.787
O1-H1B···O2	0.86	1.931	174.33	2.788
C4'-H4'···O5'	0.99	2.544	116.21	3.11
C2'-H2'···O5'	0.99	2.57	113.79	3.104
C2'-H2'···O4'	0.99	2.642	163.39	3.602
C2'-H2'···N3	0.99	2.687	120.74	3.305
C3'-H3'···O16	0.99	2.385	153.27	3.299

Supplementary Table 12. Summary of crystallographic data of compounds **1** and **2.**

Identification code	Compound **1**	Compound **2**
Empirical formula	$C_{10}H_{12}BrFN_6O_3$	$C_{11}H_{15}FN_6O_3$
Formula weight	361.17	298.29
Temperature/K	303.0	302.0
Crystal system	orthorhombic	orthorhombic
Space group	$P2_12_12_1$	$P2_12_12_1$
a/Å	7.5369(3)	7.5582(15)
b/Å	12.4354(7)	12.115(3)
c/Å	14.8372(7)	14.783(3)
α/°	90	90
β/°	90	90
γ/°	90	90
Volume/Å^3	1390.61(12)	1353.6(5)
Z	4	4
ρ_{calc} g/cm^3	1.725	1.464
2θ range for data collection/°	4.274 to 55.01	4.346 to 55.008
Index ranges	$-9 \le h \le 9$, $-16 \le k \le 16$, $-19 \le l \le 19$	$-9 \le h \le 9$, $-15 \le k \le 15$, $-17 \le l \le 19$
Reflections collected	20341	10203
Independent reflections	3182 [R_{int} = 0.0596, R_{sigma} = 0.0453]	3105 [R_{int} = 0.0493, R_{sigma} = 0.0468]
Data/restraints/parameters	3182/0/192	3105/0/193
Goodness-of-fit on F^2	0.947	1.070
Final R indexes [I>=2σ (I)]	R_1 = 0.0382, wR_2 = 0.0887	R_1 = 0.0399, wR_2 = 0.0909
Final R indexes [all data]	R_1 = 0.0529, wR_2 = 0.0966	R_1 = 0.0577, wR_2 = 0.1006
CCDC	2328041	2328038

Supplementary Table 13. Summary of crystallographic data of compounds **3** and **4.**

Identification code	Compound **3**	Compound **4**
Empirical formula	$C_{13}H_{17}FN_6O_3$	$C_{12}H_{19}FN_6O_5S$
Formula weight	324.32	378.39
Temperature/K	303.0	302.0
Crystal system	orthorhombic	orthorhombic
Space group	$P2_12_12_1$	$P2_12_12_1$
a/Å	7.9695(6)	4.89990(10)
b/Å	12.4446(12)	10.4794(3)
c/Å	14.9827(13)	31.2171(9)
α/°	90	90
β/°	90	90
γ/°	90	90
Volume/Å^3	1485.9(2)	1602.94(7)
Z	4	4
ρ_{calc} g/cm^3	1.450	1.568
2θ range for data collection/°	4.254 to 55.126	4.1 to 54.978
Index ranges	$-10 \le h \le 10$, $-16 \le k \le 16$, $-19 \le l \le 18$	$-6 \le h \le 6$, $-13 \le k \le 13$, $-40 \le l \le 40$
Reflections collected	24498	24029
Independent reflections	3431 [R_{int} = 0.0564, R_{sigma} = 0.0360]	3687 [R_{int} = 0.0675, R_{sigma} = 0.0389]
Data/restraints/parameters	3431/0/210	3687/4/240
Goodness-of-fit on F^2	1.042	1.031
Final R indexes [I>=2σ (I)]	R_1 = 0.0389, wR_2 = 0.0886	R_1 = 0.0341, wR_2 = 0.0754
Final R indexes [all data]	R_1 = 0.0559, wR_2 = 0.0961	R_1 = 0.0470, wR_2 = 0.0806
CCDC	2551697	2551685

Supplementary Table 14. Summary of crystallographic data of compound **6.**

Identification code	8BA (**6**)
Empirical formula	$C_{17}H_{23}FN_6O_6$
Formula weight	426.41
Temperature/K	222.0
Crystal system	triclinic
Space group	P1
a/Å	4.8809(2)
b/Å	9.9286(4)
c/Å	10.3953(4)
α/°	103.1050(10)
β/°	98.6040(10)
γ/°	101.1790(10)
Volume/Å^3	471.35(3)
Z	1
ρ_{calc} g/cm^3	1.502
2θ range for data collection/°	4.108 to 55.026
Index ranges	$-5 \leq h \leq 6$, $-12 \leq k \leq 12$, $-13 \leq l \leq 13$
Reflections collected	11232
Independent reflections	3995 [R_{int}= 0.0338, R_{sigma}= 0.0347]
Data/restraints/parameters	3995/3/280
Goodness-of-fit on F^2	1.049
Final R indexes [I>=2σ (I)]	R_1= 0.0296, wR_2= 0.0704
Final R indexes [all data]	R_1= 0.0337, wR_2= 0.0726
CCDC	2300863

Supplementary Table 15. Summary of crystallographic data of 8BA in Stage I and II**.**

Identification code	8BA (**6**) (Stage I)	8BA (**6**) (Stage II)
Empirical formula	$C_{17}H_{23}FN_6O_6$	$C_{17}H_{20}FN_6O_4$
Formula weight	426.41	391.39
Temperature/K	303.0	303.0
Crystal system	triclinic	triclinic
Space group	P1	P1
a/Å	4.9124(8)	5.3349(10)
b/Å	9.9387(16)	9.7052(18)
c/Å	10.4294(18)	9.9046(18)
α/°	103.156(6)	65.757(6)
β/°	98.338(6)	79.472(6)
γ/°	101.193(6)	75.882(6)
Volume/Å^3	476.76(14)	451.49(15)
Z	1	1
ρ_{calc} g/cm^3	1.485	1.439
2θ range for data collection/°	4.092 to 55.032	4.686 to 55.194
Index ranges	$-6 \le h \le 6$, $-12 \le k \le 12$, $-13 \le l \le 13$	$-6 \le h \le 6$, $-12 \le k \le 12$, $-12 \le l \le 12$
Reflections collected	8691	9223
Independent reflections	4109 [R_{int} = 0.0398, R_{sigma} = 0.0557]	3949 [R_{int} = 0.0457, R_{sigma} = 0.0604]
Data/restraints/parameters	4109/3/280	3949/3/257
Goodness-of-fit on F^2	1.077	1.028
Final R indexes [I>=2σ (I)]	R_1 = 0.0447, wR_2 = 0.0884	R_1 = 0.0490, wR_2 = 0.1121
Final R indexes [all data]	R_1 = 0.0758, wR_2 = 0.1042	R_1 = 0.0790, wR_2 = 0.1307
CCDC	2559152	2559153

Supplementary Table 16. Summary of crystallographic data of 8BA (back to Stage I)**.**

Identification code	8BA (**6**) (back to Stage I)
Empirical formula	$C_{17}H_{23}FN_6O_6$
Formula weight	426.41
Temperature/K	303.0
Crystal system	triclinic
Space group	P1
a/Å	4.9126(7)
b/Å	9.9394(15)
c/Å	10.4255(15)
α/°	103.125(5)
β/°	98.340(5)
γ/°	101.211(5)
Volume/Å^3	476.67(12)
Z	1
ρ_{calc} g/cm^3	1.485
2θ range for data collection/°	4.092 to 55.056
Index ranges	$-6 \leq h \leq 6$, $-12 \leq k \leq 12$, $-13 \leq l \leq 13$
Reflections collected	11045
Independent reflections	4237 [R_{int} = 0.0462, R_{sigma} = 0.0558]
Data/restraints/parameters	4237/3/280
Goodness-of-fit on F^2	1.055
Final R indexes [I>=2σ (I)]	R_1 = 0.0450, wR_2 = 0.0865
Final R indexes [all data]	R_1 = 0.0787, wR_2 = 0.1013
CCDC	2559156

4. Supplementary Algorithms

Algorithm S1 Find parent indices algorithm.

Require: the DSMILES token sequence T, positional pointer sequence P
Ensure: Parent indices R_1, grandparent indices R_2, great-grandparent indices R_3

Initialize R_1, R_2, R_3 as zero arrays of length $|T|$
for $i \leftarrow 0$ to $|T| - 1$ **do**
    $r_1 \leftarrow \text{FindParent}(T, P, i)$ ▷ Get parent's index
    $R_1[i] \leftarrow r_1$
    $r_2 \leftarrow R_1[r_1]$
    $R_2[i] \leftarrow r_2$ ▷ Get grandparent's index
    $r_3 \leftarrow R_1[r_2]$
    $R_3[i] \leftarrow r_3$ ▷ Get great-grandparent's index
end for
return R_1, R_2, R_3

Reference

1 Brehmer, J., de Haan, P., Behrends, S. & Cohen, T. Geometric Algebra Transformer. arXiv:2305.18415 (2023). <https://ui.adsabs.harvard.edu/abs/2023arXiv230518415B>.

2 Wang, Z. *et al.* A Water Channel-like Structure Self-Assembled by Nucleosides. *Chemistry-A European Journal* **31**, e202404045 (2025). https://doi.org/10.1002/chem.202404045

3 Münzel, M., Szeibert, C., Glas, A. F., Globisch, D. & Carell, T. Discovery and Synthesis of New UV-Induced Intrastrand C(4−8)G and G(8−4)C Photolesions. *Journal of the American Chemical Society* **133**, 5186–5189 (2011). https://doi.org/10.1021/ja111304f