An Automated Pipeline for the 3D Structural Modelling & Docking of ssDNA Probes for Pathogen Detection

Beatriz Pereira[1,2,3], Romeu Fernandes[4], Pedro Soares[5,6], Diogo Pratas[7,8,9,10], Sérgio F. Sousa[11], João Carneiro[3]

[1]Department of Computer Science, Faculty of Sciences, University of Porto, Porto, Portugal; [2]Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal; [3]Universidade Católica Portuguesa, CBQF - Centro de Biotecnologia e Química Fina – Laboratório Associado, Escola Superior de Biotecnologia, Rua Diogo Botelho 1327, 4169-005 Porto, Portugal; [4]Informatics Department, University of Minho, 4710-057 Braga, Portugal; [5]Centre of Molecular and Environmental Biology, Department of Biology, University of Minho, Braga, Portugal; [6]Institute of Science and Innovation for Bio-Sustainability, University of Minho, Braga, Portugal; [7]Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Aveiro, Portugal; [8]LASI – Intelligent System Associate Laboratory, Portugal; [9]Department of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal; [10]Department of Virology, University of Helsinki, Helsinki, Finland; [11]LAQV/REQUIMTEBioSIM - Department of Biomedicine, Faculty of Medicine, University of Porto, Porto, Portugal;

CATÓLICA

CBQF · CENTRE FOR BIOTECHNOLOGY AND FINE CHEMISTRY ASSOCIATE LABORATORY CBQF

PORTO

Why Structural Probe Modelling?

BACKGROUND & MOTIVATION

Rapid **pathogen identification** is critical in clinical diagnostics, yet conventional methods remain slow and resource demanding. **ssDNA** probes offer a sequence-specific alternative, but hybridisation efficiency is governed by probe conformation.

PROBLEM: Pathogen detection (virus & bacteria) | **RISK:** Antimicrobial resistance

LIMITATION: Diagnostic methods are slow | **NEED:** Rapid, specific diagnostics

↘ **SOLUTION: DNA Aptamers – rapid & specific molecular probes**

Secondary structures such as **stem-loops** or **hairpins** can sequester the binding region, compromising assay sensitivity. Accurate 3D structural **modelling** is therefore essential prior to experimental validation.

Objectives

- Develop **an automated** pipeline for 3D prediction of ssDNA probes targeting bacterial & viral pathogens;
- Predict stable **secondary structures** via RNAfold with DNA-specific Mathews 2004 parameters;
- Generate 3D conformations via **knowledge-based fragment assembly** (NSP toolkit);
- Apply conformational **clustering** & AMBER OL15 **minimisation** for refined, ranked 3D models;
- Run the same pipeline on the **complement strand** (A↔T, G↔C) for duplex characterisation;
- Perform HADDOCK3 *ab initio* **docking** between probe & complement — full DNA duplex assembly.

Methodology 3D Prediction + HADDOCK3 Docking

I. NUCLEOFOLD3D — 3D PREDICTION

5 STEPS PER STRAND · ×2 FOR COMPLEMENT

INPUT — Probe Sequence
Single probe, Excel batch or CSV. Auto-detects DNA / RNA. Validates ACGT / ACGU
`--seq --xlsx --csv`

SECONDARY STRUCTURE — RNAfold + DNA params
Mathews 2004 for DNA, default for RNA. Returns dot-bracket + MFE
ViennaRNA `(((...))) · kcal/mol`

3D ASSEMBLY — NSP Fragment Build
Knowledge-based fragment assembly. Quality check: perfect → direct ×5, else opt + cluster
NSP `pred1–5.pdb`

ENERGY MINIMISATION — Amber OL15
CG → all-atom (NSP fit), tleap topology, sander minimisation. 5 parallel threads
RMSD `salt 0.15 M · maxcyc 1000`

CLUSTERING — Conformational Grouping
Trajectory clustered into k=5 representative centroids. Sequence reapplied via NSP mutate
igb=5 `cluster.raw.1–5`

COMPLEMENT RUN `make_complement(seq)` A↔T G↔C — The complete five-step pipeline above repeats on the complementary strand, producing both strands ready for duplex docking. Outputs land alongside the original in the same results folder.

Lowest E — `job_minenerg.pdb` — Lowest-energy 3D model of the original strand

Complement — `job_complement_minenerg.pdb` — Lowest-energy complement strand model

All Models — `final_structures/[1–5].pdb` — All five minimised conformations per strand

Metrics — `quality_report.xlsx` — MFE . Rg . clash . planarity . C2'-endo % . RMSD

II. HADDOCK3 — AB INITIO DNA:DNA DOCKING

PREP — Chain Assign
original → A
complement → B
`mol1_min · mol2_min`

RESTRAINTS — act_pass
All residues active — *ab initio* mode
`act_pass_mol{1,2}.txt`

AIRS — ambig.tbl
restraints active passive to ambig
`AIR restraints`

TOPOLOGY — topoaa
All-atom AMBER topology build
`~ 1 s`

RIGID BODY — rigidbody
400 models . ε=78 . w_desolv=0 . cdie
`slowest ~9 min`

FLEX REFINE — flexref
Semi-flexible SA in torsion space
`dnarest_on · T=300K`

EM REFINE — emref
Final energy minimisation
`cartesian space`

CLUSTER — clustfcc
FCC clustering . strictness 0.6
`min_pop 2`

DNA-TUNED PARAMETERS | **ε** = 78.0 **w_desolv** = 0 **dielec** = cdie **dnarest_on** = true **tadfactor** = 4 **temp_cool3_init** = 300 K **sampling** = 400 **ncores** = 10

HADDOCK score — Primary ranking — composite energy + restraints — `lower = better`

E_{elec} — Electrostatics — dominant for DNA phosphates — `kcal/mol`

E_{vdw} — Van der Waals — stacking & backbone contacts — `kcal/mol`

BSA — Buried surface area — duplex interface size — `600–1500 Å typical`

iRMSD — Interface RMSD — cluster geometric consistency — `< 2 Å excellent`

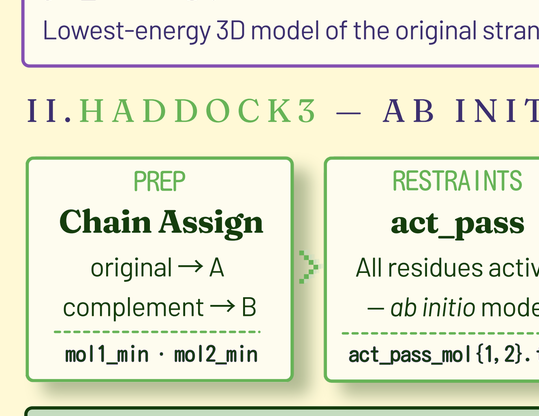
lRMSD — Ligand RMSD — pose reproducibility — `< 5 Å acceptable`

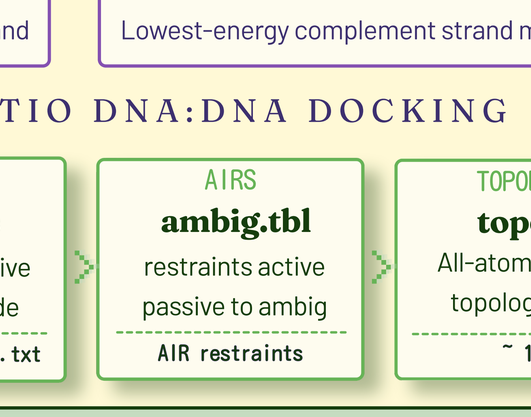
Cluster size — Models per cluster — pose confidence — `> 4 reproducible`

N clusters — Total clusters — landscape diversity — `1–3 convergent`

Results

We benchmarked 97 DNA probes spanning 14 species and 3 antimicrobial-resistance (AMR) gene targets through the pipeline. The **DockQ** distribution is bimodal (43% achieve DockQ > 0.8), while **HADDOCK scores** cluster around –36 kcal/mol.





Best probe per target - quality, folding and binding metrics

Probe	Score	MFE (kcal/mol)	Rg (Å)	HADDOCK score	DockQ	iRMSD (Å)	BSA (Å²)
Paeroginosa	+1.15	-3.50	15.68	-58.8	0.90	0.53	1507
Spneumoniae	+1.15	-1.10	16.08	-50.5	0.93	0.42	1643
Saureus	+1.05	0.00	17.70	-47.0	0.82	0.97	1729
Pfalciparum	+1.05	-0.10	19.58	-47.6	0.95	0.29	1652
H1N1	+0.94	-0.20	17.90	-48.1	0.85	0.61	1568
Hsapiens	+0.76	0.00	18.53	-36.3	0.93	0.41	1637
Kpneumoniae	+0.66	0.00	12.29	-51.9	0.84	0.85	1181
Hinfluenzae	+0.58	0.00	19.30	-36.4	0.89	0.62	1449
Afumigatus	+0.51	0.00	24.52	-33.6	0.89	0.60	1465
Syphilis	+0.41	0.00	17.00	-37.1	0.78	1.24	1368
Gonorhea	+0.39	-0.40	19.92	-35.9	0.78	1.29	1523
hsv1	+0.39	0.00	19.89	-35.3	0.79	1.51	1409
hsv2	+0.20	0.00	12.20	-39.2	0.86	0.69	1092
mecA	-0.00	0.00	45.75	-21.2	0.72	2.60	1638
kpc	-0.15	-6.80	14.30	-39.4	0.51	2.50	1205
Chlamydia	-0.22	0.00	13.79	-32.4	0.21	5.49	1612
oxa	-0.46	-1.10	16.84	-31.7	0.24	5.06	1390

Score = weighted z-score of Haddock, DockQ. iRMSD, BSA and clash (higher = better). MFE = secondary structure stability; Rg = compactness of the 3D model.

A composite **z-score** identifies the best probe per species : 13/17 targets yield an above-average candidate, 9 of them clearly high-quality. The 4 below-average picks had only 2 candidates each → more sequences are needed.

Conclusions

- The pipeline generates realistic, **energy-minimised** 3D conformations of ssDNA probes from sequence input;
- Five independent predictions per probe ensure **sampling** of the conformational landscape;
- The pipeline descriptors form a structured **feature** set for **probe selection**;
- Complement strand modelling enables HADDOCK3 **docking** with DNA-optimised parameters.

Acknowledgements

This work is carried out within the scope of contract ref. 2023.15056.TENURE.043, funded by national funds through the FCT – Foundation for Science and Technology, I.P. This work is also funded by national funds through the FCT – Foundation for Science and Technology within the scope of UID/50016/2025 and LA/P/0076/2020 (https://doi.org/10.54499/LA/P/0076/2020).

Fundação para a Ciência e a Tecnologia · ciimar · P.PORTO · ESCOLA SUPERIOR DE SAÚDE · U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO