UMONS
Université de Mons

Faculté des Sciences

Deciphering sequence-controlled macromolecules –

From sequence to 3D structures through modeling approaches

A thesis submitted in fulfillment of the requirements for the degree of Doctor in Sciences

David Dellemme

Laboratory for Chemistry of Novel Materials (CMN)

University of Mons, Belgium

Doctoral commission:

Prof. Sylvain Gabriele	University of Mons, Belgium – President
Prof. David Beljonne	University of Mons, Belgium – Secretary
Dr. Quentin Duez	University of Mons, Belgium – Examiner
Dr. Nezha Badi	Ghent University, Belgium – Examiner
Dr. Mohsen Sadeghi	Free University of Berlin, Germany – Examiner
Prof. Mathieu Surin	University of Mons, Belgium – Promoter

Academic year 2025-2026

CMN

CIRMAP
Center of Innovation and Research in Materials and Polymers

“It's not just about doing bucket list stuff and doing massive things, it's just about appreciating the mundane fun of life, the mundane elements of life which can be wonderful that you don't necessarily appreciate when you're on this treadmill of next, next, next, what we're doing tomorrow? Never mind tomorrow, enjoy today!”

– Chris Hoy

Abstract

Sequence-controlled macromolecules (SCMs) are polymeric or oligomeric systems in which the sequence of monomers is partially or totally regulated. When the control is absolute, *i.e.* when all chains contain the exact same sequence and number of monomers, the material is classified as a sequence-defined macromolecule (SDM), a specific subclass of SCMs. SCMs are ubiquitous in Nature and perform specific biological functions, DNA and proteins being prime examples. Proteins, in particular, have the ability to fold into specific 3D structures, executing functions with very high selectivity. This remarkable structural control is encoded in their sequence of monomers, the amino acids, which governs the folding process. The discovery of the importance of monomer sequence in natural macromolecules sparked a considerable interest among researchers, fuelling the desire to produce human-made SCMs. The recent advances in polymer synthesis have enabled the design of a wide range of fully synthetic SCMs, building on the virtually unlimited library of monomers available to a polymer chemist. However, this diversity of structures, while very attractive, brings considerable challenges: how to rationalize the design of synthetic SCMs? Which backbone and side-chains to use for a given application? What will be the 3D structure of the system in solution? Currently, synthetic SCMs are designed following "chemical intuition" rather than sound guidelines.

The aim of this thesis is to address these questions. The 3D structures of various natural and synthetic SCMs are investigated using the tools of molecular modeling. Molecular dynamics (MD) simulations, a computational method based on classical mechanics, allow us to predict the 3D structure and dynamics of (macro)molecular systems at the atomistic scale, using only their chemical structure as input. Our results are systematically compared to experimental data, to provide a better understanding of the links between sequence of monomers, 3D structure, and function.

The first part of the thesis focuses on biorecognition applications, one system targeting a protein, the other DNA. The simulations give insights on the mechanisms of assembly and the interactions at the molecular level, helping to understand experimental results.

The second part concerns the study of a supramolecular catalyst made by the assembly of two complementary SDMs, functionalized with nucleobases for the recognition between the chains, and catalytic units. MD simulations and network representations are used to elucidate the formation and dynamics of the catalytic duplex, and help to rationalize the experimental results of catalytic activity.

In the last part of the thesis, MD simulations are combined with small-angle X-ray scattering (SAXS) experiments to reveal the 3D structure of purely synthetic copolymers, in the context of single-chain polymeric nanoparticles. The folding in

water is studied for two different copolymer designs, showing how the nature of the hydrophilic grafts can influence the resulting nanostructures.
Globally, our thesis provides insights into the sequence-structure-property relationships in SCMs, towards a rational design of functional macromolecular systems.

Résumé

Les macromolécules de séquence contrôlée (SCMs) sont des systèmes polymériques ou oligomériques au sein desquels la séquence de monomères est partiellement ou totalement régulée. Quand le contrôle est absolu, c'est-à-dire quand toutes les chaînes contiennent exactement la même séquence et le même nombre de monomères, le matériau est classé comme macromolécule de séquence définie (SDM), une sous-classe au sein des SCMs. Les SCMs sont omniprésentes dans la Nature et exercent des fonctions biologiques spécifiques, l'ADN et les protéines étant des exemples typiques. Les protéines, en particulier, ont la capacité de se replier en des structures 3D spécifiques, exécutant des fonctions précises avec une très haute sélectivité. Cet exceptionnel contrôle structurel est encodé dans leur séquence de monomères, les acides aminés, qui gouvernent leur processus de repliement. La découverte de l'importance de la séquence de monomères au sein des macromolécules naturelles a engendré un grand intérêt parmi les chercheurs, nourrissant le désir de produire des SCMs artificielles. Les avancées récentes en synthèse des polymères ont permis la création d'une large gamme de SCMs synthétiques, s'appuyant sur une bibliothèque de monomères virtuellement illimitée à disposition des chercheurs. Cependant, cette diversité de structures, bien que très attrayante, soulève plusieurs questions : comment rationnaliser le design des SCMs synthétiques ? Quel squelette moléculaire et chaînes latérales utiliser pour une application donnée ? Quelle sera la structure 3D du système en solution ? Actuellement, les SCMs synthétiques sont construites sur base de « l'intuition chimique » plutôt qu'en suivant des principes bien établis.

Le but de cette thèse est de s'adresser à ces questions. Les structures 3D de diverses SCMs naturelles et synthétiques sont étudiées en utilisant des outils de modélisation moléculaire. Les simulations de dynamique moléculaire (MD), une méthode computationnelle basée sur les lois de la mécanique classique, nous permettent de prédire la structure 3D et la dynamique de systèmes (macro)moléculaires à l'échelle atomique, en utilisant uniquement leur structure chimique comme point de départ. Nos résultats sont systématiquement comparés à des données expérimentales, afin de fournir une meilleure compréhension des liens qui unissent séquence de monomères, structure 3D, et fonction.

La première partie de la thèse se concentre sur des applications de bioreconnaissance, un système ciblant une protéine, l'autre l'ADN. Les simulations nous donnent un aperçu des mécanismes d'assemblage et des interactions au niveau moléculaire, nous aidant à mieux comprendre des résultats expérimentaux. La deuxième partie concerne l'étude d'un catalyseur supramoléculaire obtenu par assemblage de deux SDMs complémentaires, fonctionnalisées avec des nucléobases pour la reconnaissance entre

les chaînes, et des unités catalytiques. Les simulations de MD et des représentations en réseau sont utilisées pour élucider la formation et la dynamique du duplexe catalytique, et nous aident à rationnaliser des mesures expérimentales d'activité catalytique. Dans la dernière partie, des simulations de MD sont combinées avec des expériences de diffusion de rayons X aux petits angles pour révéler la structure 3D de copolymères purement synthétiques, dans le contexte des nanoparticules polymériques à chaîne unique. Le repliement dans l'eau est étudié pour différents copolymères, montrant comment la nature des groupements hydrophiles peut influencer les nanostructures obtenues.

Globalement, notre thèse apporte des pistes pour mieux comprendre les liens séquence-structure-fonction au sein des SCMs, afin d'évoluer vers une conception rationnelle de systèmes macromoléculaires fonctionnels.

Remerciements

Me voilà face à l'immense défi des remerciements ! Une partie que je redoute un peu d'écrire, par peur de ne pas rendre correctement hommage à toutes les personnes envers lesquelles je suis reconnaissant. En même temps, je ne suis pas très démonstratif : c'est donc l'occasion idéale de mettre par écrit ce que je n'exprime peut-être pas assez.

First of all, I sincerely thank the members of the jury, Prof. Sylvain Gabriele, Prof. David Beljonne, Dr. Quentin Duez, Dr. Nezha Badi and Dr. Mohsen Sadeghi. I am really happy to have such a diverse jury, with experts in various domains. Many thanks for the time you spent reading my manuscript and for your comments.

Ensuite, je tiens à remercier mon promoteur, Mathieu Surin. Ça a été un vrai bonheur pour moi de travailler sous ta supervision, depuis mon arrivée au labo en Master 1, où j'ai été attiré par d'intrigants quadruplexes d'ADN… J'ai toujours ressenti beaucoup de confiance de ta part, j'ai pu travailler avec beaucoup de liberté et d'autonomie, mais tu as toujours été disponible pour discuter de mes résultats. Ton soutien et ta positivité ont été très importants tout au long de mes années de mémoire et de thèse, pour dissiper les doutes – inévitables – qui accompagnent la recherche. Merci !
Plus généralement, j'ai eu énormément de chance d'arriver au CMN, qui est un environnement de travail absolument parfait pour moi. En ce sens, je remercie toutes les personnes qui ont contribué et qui contribuent encore à ce que l'ambiance au labo soit aussi bonne. Merci à Roberto Lazzaroni, pour toute la bonne humeur et l'énergie que tu insuffles, pour tous les évènements conviviaux que tu incites à réaliser et qui permettent de créer un vrai esprit de groupe au sein du labo. Merci aussi à tous les autres « chefs », Jérôme Cornil et ses blagues qui font toujours *(parfois ?)* mouche, David Beljonne pour sa gentillesse et sa disponibilité, Claudio Quarti pour ses bons conseils et sa pédagogie, ainsi que Patrick Brocorens pour ses anecdotes de voyage, parfois invraisemblables mais toujours passionnantes.
Merci aussi à Laura, pour la bonne humeur que tu apportes, tous les évènements que tu contribues à organiser et toutes ces notes de frais traitées. Encore félicitations à toi et Alex pour la naissance d'Eliott, beaucoup de bonheur à vous quatre (j'inclus Pippa, bien sûr) ! Je remercie également Pocket, le poète du labo. Sans tes précieux conseils et ton expérience, bien des mémorant(e)s et doctorant(e)s seraient perdu(e)s ; mais surtout, ils manqueraient de café, carburant indispensable à la production scientifique.
Je remercie également tous les membres du bureau 157, un véritable concentré d'excellence, l'élite de l'élite ! Merci au chef de bureau, Vincent, d'avoir toléré ma

présence dans ce haut lieu de la Science montoise. J'ai en tête tant de magnifiques instants où la « solidarité de bureau » – une solidarité, il faut le dire, toute relative – a triomphé au Dalmuti… Plus sérieusement, j'ai tout de suite eu un bon contact avec toi, tu m'as vraiment aidé à m'intégrer au labo quand je suis arrivé, et tu as toujours été disponible lorsque j'avais une question ou un problème (pour des scans d'angles de torsion par exemple…), et pour tout ça je te remercie sincèrement. Je remercie également Corentin, membre d'honneur du bureau. De Templeuve au bureau 157 (c'est d'ailleurs toi qui m'avais transmis l'invitation officielle, quel honneur), j'ai décidément toujours suivi ton chemin. Merci pour toute l'aide apportée quand j'ai débuté ma thèse ! Ari, tu es arrivé en digne successeur, comblant les dettes qui t'avaient été léguées en relançant la tradition des fèves (bon, il faudra essayer de re-relancer cette tradition…), avec en prime le prix de la SRC, et surtout un article dans Chimie Nouvelle. La grande classe ! And of course, a word for Sarajit. It has been a pleasure to meet you, and to share the good, and sometimes more difficult, moments of the thesis. I will always remember our trip in Warwick; the pictures of you enjoying life in the park will stay in my mind. Thanks to your advice, I discovered good Indian dishes in England!
Pour continuer au CMN (ça en fait du monde, quand même), je dois remercier quelques « anciens » (avec tout le respect). Merci à Mathieu Fossépré, pour m'avoir initié et formé à la modélisation moléculaire. J'ai eu énormément de chance de démarrer ma thèse en pouvant m'appuyer sur les travaux que tu avais déjà réalisés, ça m'a beaucoup aidé pour débuter. Merci aussi à Sinan Kardas, j'ai beaucoup apprécié les moments passés avec toi et j'aurais aimé te côtoyer un peu plus longtemps. Comme Mathieu, tes travaux et ton aide ont été très précieux pour « lancer » ma thèse. Je remercie aussi Alexandre Remson, le prodige du Pays Blanc. Je garde en mémoire ton franc-parler et ton humour, et la belle collaboration que tu nous as permis de réaliser ! J'étais un peu dubitatif au départ, pas sûr du tout qu'on arriverait à quelque chose, mais finalement ça a donné de belles choses ! Tu as bien fait d'insister. Merci à Maxime Leclercq, c'est toi qui m'as superbement formé aux mesures CD, et plus généralement qui m'as initié à la recherche. C'était au cours d'une sombre période de coronavirus, où le labo était bien vide. Heureusement que tu étais là pour répondre à mes questions et m'aiguiller ! Je remercie aussi les mémorant(e)s et doctorant(e)s avec qui j'ai pu travailler. Maxime, ça a toujours été très cool de travailler avec toi et de partager des moments ensemble. Même si tu es désormais blacklisté des boîtes de Rouen… On se croisera, je l'espère, encore pour des courses à pied ; au moins sur la ligne de départ, après je risque d'avoir un peu de mal à te suivre. Julien, c'était un plaisir de pouvoir discuter de modélisation avec toi. Même si c'est sans doute toi qui m'a appris des choses plutôt que l'inverse ! Je suis devenu un peu meilleur réalisateur avec PyMOL grâce à tes astuces. Louis, le roi des protéines et véritable fast learner ; il a suffi de te montrer les commandes une fois,

et puis tu étais lancé. C'est vraiment cool qu'on ait pu collaborer ! Pauline, qui entame un périlleux voyage dans le monde de la modélisation. Pas simple avec toutes les techniques que tu dois maîtriser, mais je suis sûr que tu arriveras à assembler le « puzzle » de ta thèse ! Et Cassandra, c'était pas évident pour toi avec un « encadrement à distance », mais tu t'en es super bien sortie pour ton mémoire, avec beaucoup d'autonomie. Bon courage pour le FRIA, je croise les doigts pour que ça passe ! J'ai toujours eu la chance de travailler avec des gens très à l'écoute, attentifs et désireux d'apprendre, donc c'était un vrai plaisir pour moi, merci à tous !
Encore un mot pour mes camarades de fin de thèse, Antoine et Florian, indissociables jusque dans mes remerciements. Antoine, je me souviens avec émotion des cours du CECI, où nous avons tant appris sur le codage en Python… Florian, je ne pourrai jamais oublier un certain spectacle de Noël, où tes talents de DJ (ainsi que tes muscles saillants) ont été révélés au grand jour. Je suis vraiment, vraiment content de vous avoir rencontrés et j'espère que l'on gardera contact !
Et tous les autres que je n'ai pas encore cités, la « nouvelle génération » du CMN (mais pas que) : Alexis, grâce à toi le mot « spintronique » signifie quelque chose pour moi. Cristina, thanks for the amazing energy that you bring to the lab! Guillermo, your moves are always perfect, whether on the dance floor or on the chessboard. Isaac, l'homme qui alterne les masterclass (souvent) et les « masterclaques » (parfois). Loïc, déjà un papier, quelle classe ! Bon courage pour ta thèse, et laisse un peu de répit à Pocket, pense à sa santé. Louis (Duhayon), sage chercheur le jour et fêtard invétéré la nuit. Mariano, dont l'envie de courir (voire de rouler en vélo) est proportionnelle à l'alcoolémie, un véritable triathlète dans l'âme. Mohamed, beaucoup trop gentil et trop honnête, surtout au grand Dalmuti. Nico, j'espère que tu auras cette thèse en Espagne ; si pas, tu es plein de ressources et je ne doute pas que tu trouveras une voie qui te plaira. Tudor, avec toi on ne sait jamais à quoi s'attendre : MMA, béhourd, retraites spirituelles… C'est trop cool, reste comme tu es ! Zoé, après la Suède, les États-Unis : c'est sympa la chimie, quand même. Profite bien de ton voyage ! Vous contribuez tous et toutes à créer cette superbe ambiance qui m'a donné envie de venir au labo jour après jour, et je garde vraiment de beaux souvenirs avec chacun de vous. Je suis heureux de vous avoir rencontrés.
Pour enfin conclure avec la partie CMN, je souhaite adresser un immense merci à Sébastien Kozlowskyj. Merci pour ta gentillesse, ta disponibilité et ton efficacité pour résoudre nos problèmes informatiques divers et variés. Tu es vraiment indispensable au bon fonctionnement du labo, et ça a été un plaisir de croiser ton chemin durant ces quelques années.

J'aimerais ensuite remercier l'ensemble des personnes avec lesquelles j'ai eu la chance de collaborer durant ma thèse. Ce n'est pas un hasard si chacun des chapitres de résultats met en avant une collaboration différente : cette thèse, ce n'est pas moi, seul dans mon coin. Elle n'aurait jamais pu voir le jour sans l'aide et les contributions de toute une série de personnes, et je suis heureux de présenter ce travail comme un véritable effort collectif. Ces collaborations m'ont aussi permis de me plonger dans des domaines très variés, en me nourrissant de l'expertise d'autres chercheurs, ce qui a été véritablement passionnant et très enrichissant d'un point de vue scientifique. Merci aux groupes d'Alain Jonas et Karine Glinel de l'UCLouvain : je me souviens être venu répéter ma présentation pour le FRIA chez vous, et vos conseils ont été précieux pour l'améliorer. Many thanks to the group of Dr. Andres de la Escosura, in particular to Noemi Nogal, for her nice spectroscopic results and the energy that she brought in Mons! Merci au groupe de Sylvain Gabriele, en particulier Alex que j'ai déjà cité, pour la belle collaboration et pour m'avoir encouragé et poussé à approfondir mes résultats. And many thanks to the group of Anja Palmans, it was an immense pleasure and honor to work with you. Your mails always brought a smile to my face, your joy and visible enthusiasm are truly contagious. And I particularly thank Stefan Wijker. This is the project where the comparisons between simulations and experiments were the most "direct" and probably the most satisfying. You have always been very available to answer my questions, and the discussions about our results were very enriching. I also wish to thank Patrick Norman and Mathieu Linares, for the wonderful welcome that I received in Sweden. I learned a lot in an amazing environment, and discovered a truly beautiful country. Je remercie également particulièrement Fabrice Saintmont, pour toute l'aide que tu m'as apportée là-bas, au labo et en dehors !

Et puis, ces années à Mons auront quand même été marquées par plusieurs personnes exceptionnelles ; je vais essayer de faire par ordre de connaissance, en espérant que mes souvenirs ne me trahissent pas. Quentin, rencontre incroyable, deux glandus arrivés aux TPs de BAC 1 sans avoir de groupe, on s'est retrouvés ensemble par hasard. Et bah, quelle chance d'être tombé sur toi. On n'était pas les étudiants les plus brillants, mais c'était rigolo. Ça a tout de suite super bien accroché de mon côté, on avait déjà des références communes et le contact a été naturel. Avoir immortalisé notre amitié au stand Notélé vaut tout l'or du monde… Louis, j'ai en mémoire les temps de midi à jouer au Mao, les cours suivis de manière trop peu sérieuse (encore désolé que tu te sois attiré les foudres de Coulembier…), les TPs de physique (bravo coach). Et puis toutes les soirées chez toi, surtout en BAC 2 / BAC 3, mémorables. Je n'oublierai jamais ce qu'on a vécu, un soir, au Connemara… Et puis, c'est toi qui as lancé l'idée de la coloc' : une masterclass, merci. Pierre, tah l'époque, comme disent les jeunes, où la moindre

taquinerie t'énervait. La première fois que je t'ai vu, tu as tout de suite su me mettre à l'aise (non) grâce à une bonne blague bien trash comme tu les aimes et que je ne pourrais pas relater ici. Et puis on ne s'est plus quittés, même quand tu es devenu physicien : six ans de vie commune, quand même… Barnabé, tu es resté moins longtemps à la coloc', mais ça ne t'a pas empêché de réaliser quelques coups d'éclat, avec bien souvent Pierre comme victime malheureuse (et parfois agresseur, en retour…). Les deux années ensemble étaient super drôles, pleines de superbes souvenirs. Je retiens notre duo à la manille, avec un talent particulier pour remporter les « parties en or ». Amandine, malgré tes passions pour certains groupes de rock énervés et les chats noirs, tu es tout le contraire de « satanique », n'écoute pas les haters. Par contre, à un moment, il faudra que tu acceptes de jouer à Galérapagos avec nous… Et à l'autre jeu là, le genre de « Loups-Garous » avec les libéraux… Nathan, merci d'avoir été là pour ramener un peu d'ordre dans le kot ! Grâce à toi et Clara, je me suis enfin intéressé au cyclisme, ce qui m'a permis de vibrer pour le triplé de Remco aux championnats du monde. Les soirées Top Chef et tes petits desserts me manqueront ; mais j'ai entendu dire que tu ouvrais bientôt ta boulangerie… Clara, la cinquième coloc', seule vraie Montoise ! Merci d'avoir égayé les soirées du kot par ta présence. Je continuerai à suivre, admiratif, tes sorties sur Strava (en espérant que tu ne rencontres plus de pavés de trop près…). Très content pour toi et Nathan que tout se passe bien pour vous, bien installés dans votre superbe appart' ! Je vais aussi avoir un petit mot pour Thomas, respect au deuxième plus grand fan de la Royale Fanfare Communale de Huissignies. Et pour finir, Benjamin, « parrain » comme dirait l'autre (et bientôt « papa » maintenant, encore félicitations à Lia et toi), toujours un vrai plaisir de te voir. La petite réunion à la Lorgnette pendant le Doudou est déjà un incontournable.
En écrivant tout ça, je me rends compte que neuf ans, quand même, ça passe vite… Les premières années d'unif' semblent si proches, et en même temps tout a tellement changé depuis, certains souvenirs paraissent si lointains. J'espère qu'à l'image des Montois, la « bande à Pierrick » ne périra pas… Vous méritez bien plus que quelques lignes dans les remerciements de ma petite thèse, mais le plus important sera de ne pas perdre le contact, malgré la distance, où que la vie nous emmène. On pourra toujours se retrouver pour une Taverne. Je vous aime.

Sur ce, je quitte Mons pour retrouver mes origines templeuvoises. Dimitri, on se connaît depuis qu'on est des petits bézots et on ne se lâche pas. Je sais que je pourrai toujours compter sur toi. Merci pour toutes ces soirées gaming, ces week-ends de Mario Kart (entraîne-toi un peu par contre), de Smash Tennis, et de cyclisme acharné (avec la frite qui suit, à ne surtout pas commander sur Alezy, évidemment), les 100 bornes

pour aller à la mer, les vacances au lac de Garde, et j'attends impatiemment la suite ! Bastien, tu es un jeune cadre dynamique maintenant, mais tu restes au fond le petit Satcheu que j'ai toujours connu. Je me souviendrai toujours des innombrables voyages à Bellewaerde, des matchs de badminton (il faudra qu'on remette ça un jour…), des sessions gamings nocturnes sur Pokémon (les souterrains sur Diamant/Perle, c'était le feu). Et comme Dimitri, je sais que je peux te vouer une confiance absolue. Et puis Adrien, le petit bricoleur. Les vacances avec toi ont été une régalade. Je suis sûr que tu seras prêt pour la prochaine lan Smash Bros, un jeu qui permet de mettre en valeur ta grande sérénité et ton calme à toute épreuve (FAUX). Mais reste comme ça, t'es un boss ! Je vous aime les gars.

Je vais conclure ces remerciements par le plus important. Merci papa et maman, pour l'éducation et tout l'amour que vous m'avez donnés. J'ai eu la chance incroyable de toujours pouvoir faire ce que je voulais, j'ai pu étudier dans les meilleures conditions, je n'ai eu à me soucier de rien. Durant toutes ces années à Mons, revenir à la maison le week-end était toujours un vrai bonheur (pas seulement pour faire mes lessives…). Merci d'avoir toujours été là pour moi, je ne pourrai jamais vous rendre tout ce que vous m'avez donné, mais je vais faire de mon mieux pour être quelqu'un de bien et vous rendre fiers. Je t'aime papa. Je t'aime maman, j'aurais tant aimé que tu sois là.
Merci Luca, je n'imagine pas la vie sans toi. Rentrer le week-end, c'était aussi te retrouver. Merci pour tous ces matchs de ping-pong (même quand la balle ne rebondit pas), pour tous ces vendredis où tu es venu me chercher à la gare, pour toutes les fois où tu m'as demandé « un p'tit jeu ? », pour m'avoir montré que tout est possible avec un peu de volonté (même finir un marathon sans entraînement). Je suis fier de toi, chapeau l'artiste. Je t'aime.
Merci mamie, je suis heureux que tu sois là. C'est grâce à ton bon pain, tous tes gâteaux et ton incroyable tiramisu que j'ai eu l'énergie de terminer cette thèse. Encore un peu de temps, et je connaîtrai toute l'histoire d'Evregnies et de la « vie d'avant ». Je t'aime mamie, ne change pas, tu es géniale.

Ces remerciements sont beaucoup trop longs mais tant pis, c'est la moindre des choses que d'exprimer un peu de gratitude envers tous ceux et toutes celles qui ont contribué à ce que ces années de mémoire et de thèse soient aussi formidables. Je vous suis tous et toutes vraiment redevables. Merci, c'était cool !

List of abbreviations

AA	Amino acid
AMBER	Assisted model building with energy refinement
aMD	Accelerated molecular dynamics
AM1-BCC	Austin Model 1 with bond charge correction
BO	Born-Oppenheimer
BTA	Benzenetricarboxamide
CASP	Critical Assessment of protein Structure Prediction
CD	Circular dichroism
CG	Coarse-grained
CMP	Collagen-mimetic peptide
DFT	Density functional theory
DOSY	Diffusion-ordered spectroscopy
DLS	Dynamic light scattering
DNA	Deoxyribonucleic acid
DNP	Dinitrophenyl
DP	Degree of polymerization
dsDNA	Double-stranded deoxyribonucleic acid
ECM	Extracellular matrix
FF	Force field
GAFF	General AMBER force field
GB	Generalized Born
HDD	Hard disk drive
HMR	Hydrogen mass repartitioning
HSA	Human serum albumin
ICD	Induced circular dichroism
K_a	Equilibrium association constant
LCPO	Linear combination of pairwise overlaps
MD	Molecular dynamics
MIDAS	Metal-ion dependent adhesion site
ML	Machine learning
MM	Molecular mechanics
MMPBSA	Molecular mechanics Poisson-Boltzmann surface area
mRNA	Messenger ribonucleic acid
MS/MS	Tandem mass spectrometry
NAB	Nucleic acid builder

NMR	Nuclear magnetic resonance
PB	Poisson-Boltzmann
PBC	Periodic boundary conditions
PDB	Protein Data Bank
PEG	Polyethylene glycol
PME	Particle mesh Ewald
PPII	Polyproline type II
PTM	Post-translational modification
PNA	Peptide nucleic acid
QM	Quantum mechanics
RDF	Radial distribution function
RESP	Restrained electrostatic potential
R_G	Radius of gyration
R_H	Hydrodynamic radius
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNA	Ribonucleic acid
SANS	Small-angle neutron scattering
SASA	Solvent-accessible surface area
SAXS	Small-angle X-ray scattering
SCM	Sequence-controlled macromolecule
SCPN	Single-chain polymeric nanoparticle
SDM	Sequence-defined macromolecule
SEC	Size exclusion chromatography
SSD	Solid-state drive
ssDNA	Single-stranded deoxyribonucleic acid
TAMRA	5-Carboxytetramethylrhodamine
TEMPO	2,2,6,6-tetramethyl-1-piperidinyloxyl
TD-DFT	Temperature-dependent density functional theory
THF	Tetrahydrofuran
TOF	Turnover frequency
tRNA	Transfer ribonucleic acid
UAA	Unnatural amino acid
UV-Vis	Ultraviolet-visible
VdW	Van der Waals
XNA	Xenonucleic acid

Table of contents

I. Overview and aim of the thesis

Nature has always been a major source of inspiration for humanity. The design of the ornithopter by Leonardo da Vinci, directly inspired by the wings of birds. Self-cleaning superhydrophobic surfaces, reproducing the microstructure of the lotus leaf. The aerodynamic nose of the Shinkansen, Japan's high-speed train, mimicking the beak of the kingfisher. The list could continue with countless examples: Nature has developed many highly efficient architectures over millions of years of evolution. At the molecular scale, Nature also carries out fascinating biochemical processes, relying on highly sophisticated chemical structures. Major examples include nucleic acids and proteins. Among all biomacromolecules, proteins constitute the class operating the broadest range of functions. Enzymes, for instance, constitute a subclass of proteins dedicated to catalysis, and display remarkably selective binding and efficient activity, even within the complex and crowded cellular environment. This efficiency is inscribed in their highly defined 3D structures. Starting from disordered conformations, many proteins will spontaneously fold back into their native state in water. This naturally raises a question: how is the structure of a protein controlled? The answer began to emerge in the early 20th century, when their chemical structure was elucidated. Proteins are polymers: long macromolecular chains constituted by the covalent linkage of smaller molecular units – the monomers. The building blocks of proteins are the amino acids (AAs), which comprise a relatively limited set of 22 monomers, including two non-standard residues (selenocysteine and pyrrolysine). A striking particularity of proteins, shared by other functional biopolymers, is that their sequence of monomers is precisely controlled. Each protein is characterized by a unique sequence of AAs. This sequence is the code governing their folding, thus their 3D structure. Beyond the well-known structure – function, there exists sequence – structure relationships. These links are not completely understood yet, but many protein structures have been elucidated over time and are regrouped in sequence – structure databases. The Protein Data Bank, for example, comprises more than 200,000 experimentally determined structures.[1] These extensive datasets are particularly valuable for training machine learning (ML) algorithms, which are among the most powerful tools available for uncovering complex relationships between large sets of inputs and outputs. This approach led to the development of ML models, such as AlphaFold, capable of efficiently predicting the 3D structure of proteins from their AA sequence alone.[2] It is important to note that ML algorithms do not "understand" the physico-chemical laws governing folding; they only extract statistical patterns, relying sequences to structures, from vast datasets. The achievements of AlphaFold, however, clearly demonstrate that the AA sequence encodes the information required to determine the 3D structure of a protein.

Understanding that highly functional biomacromolecules are characterized by a precisely controlled sequence of monomers, giving rise to a well-organized 3D structure, inspired researchers to develop artificial systems based on the same principle. The tools of polymer chemistry offer an ideal platform for such applications. Since the discovery of the chemical structure of polymers by Staudinger in 1920, the field has undergone tremendous development.[3] Polymers were initially synthesized as a disperse, heterogeneous mixture of chains, without control on their individual length or monomer sequence. Nowadays, many synthesis pathways are available to produce polymer samples with a very low dispersity, and where the incorporation of monomer units is partially or totally regulated.[4] These systems are defined as sequence-controlled macromolecules (SCMs). When the control over sequence and chain length is absolute, attaining a dispersity of one, the term sequence-defined macromolecule (SDM) can be used.[5] Natural biopolymers such as proteins and nucleic acids belong to this subclass of SCMs. Synthetic SCMs provide opportunities to go beyond these natural examples, offering a virtually unlimited number of possibilities in terms of chemical diversity – not only with the choice of the side-chains, but also with the nature of the backbone. The influence of the solvent and stereochemistry can be further taken into account, allowing researchers to finely tune the properties of their systems. While this wide chemical space constitutes a fascinating playground, it significantly complexifies the establishment of sequence – structure relationships; a task already challenging for proteins, which have a unique backbone and a limited number of side-chains. Additionally, although synthesis pathways continue to improve, it remains difficult to obtain synthetic SCMs – let alone SDMs – with both high yields and sufficient chain lengths. Consequently, it can be very time-consuming and cost-intensive to produce these macromolecules, especially if one wants to screen a broad range of sequences for a given system. These elements bring interrogations: is there really a significant advantage to precisely control the sequence of monomers for an artificial SCM? Will a defined sequence always translate into a controlled 3D structure? How to rationalize the design of SCMs, in view of specific applications?

The aim of this thesis is to address these questions. Our methodology relies on molecular modeling, a computational method bypassing the constraints of synthesis. The molecules of interest are built *in silico*, and their conformations and dynamics are simulated, helping researchers to identify the most promising compounds for a given application. The atomistic view offered by molecular modeling provides precious information on the 3D structure, folding dynamics, and interactions inside single-chain systems or supramolecular assemblies. This molecular-level knowledge serves to better understand experimental properties and can be used to guide the design of more

efficient systems. Our results are systematically compared to experimental measurements, bringing insights on sequence – structure – function relationships for various SCMs, including natural and synthetic structures **(Figure I.1)**. In parallel, this approach is a step towards the establishment of sequence – structure databases for synthetic SCMs, analogous to the Protein Data Bank for proteins. Ideally, such databases could feed ML algorithms and support the development of predictive models for artificial systems.

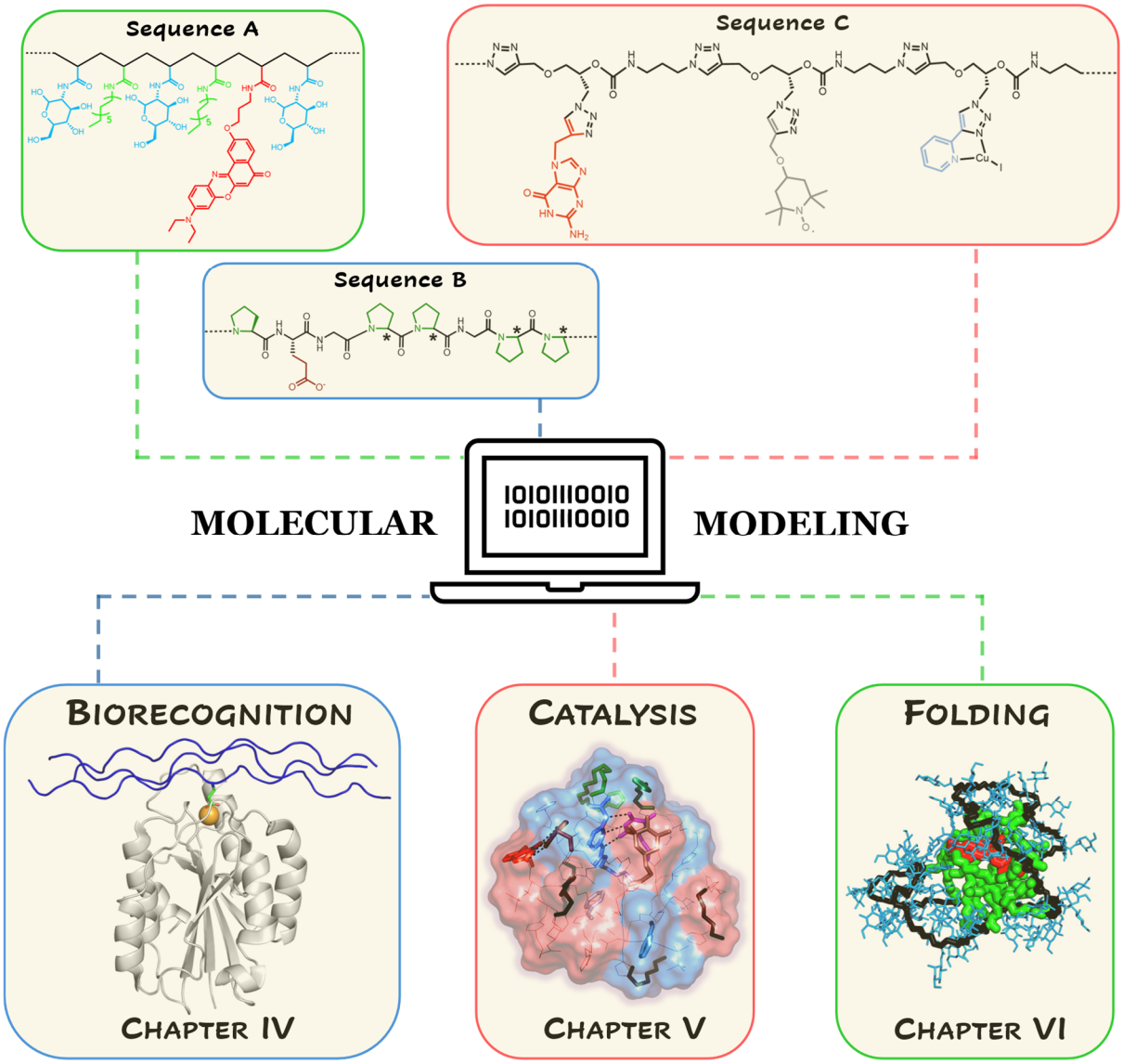


Figure I.1. Schematic representation of different chemical structures investigated in the thesis and their associated applications.

After this general introduction, the main concepts related to the field of SCMs are reviewed along with state-of-the-art examples in **Chapter II**. The discussion progresses from natural biomacromolecules and the lessons they provide, towards their translation into human-made SCMs, based on synthetic oligomers and polymers.

In **Chapter III**, the fundamentals of our computational approach are explained. This chapter reviews the basics of quantum mechanics (QM), molecular mechanics (MM), and the workflow of molecular dynamics (MD) simulations. The limits of the method and advanced approaches are also described. Finally, the descriptors and tools mainly used to analyze the molecular conformations and their dynamics are detailed.
Then, **Chapter IV** regroups results obtained for biorecognition applications for two distinct systems. The first part concerns the interaction between peptides and an integrin, studied in the framework of cellular migration. Our results show that subtle changes in stereochemistry modulate peptide – protein binding. The second part describes the modeling of supramolecular complexes between DNA and photoswitchable ligands. The simulations reveal that the *trans* to *cis* isomerization of the ligands impacts the assembly with DNA.
Chapter V focuses on SDMs targeting applications in supramolecular catalysis. Two complementary strands are functionalized with nucleobases and catalytic moieties. The mechanisms of recognition between the two strands and the dynamics of the duplex, responsible for the observed catalytic activity, are investigated by a combination of MD simulations and network representations.
In **Chapter VI**, the folding dynamics and 3D structures of purely synthetic copolymers in water are studied. The atomistic picture given by the simulations is compared to small-angle X-ray scattering (SAXS) experimental spectra. Our results show that, depending on the nature of the hydrophilic grafts, different nanostructures can be obtained in solution, varying in shape and folding properties.
To conclude, a summary of our results is presented in **Chapter VII**. Insights collected on the various systems are regrouped, and our findings concerning the establishment of sequence – structure – function relationships for synthetic SCMs are discussed. Finally, perspectives for future research in the field of SCMs, and the role that molecular modeling could play, are mentioned.

References

[1] H. M. Berman. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[3] H. Staudinger. Über Polymerisation. *Berichte der deutschen chemischen Gesellschaft (A and B Series)* **1920**, *53*, 1073–1085.

[4] K. Hakobyan, B. B. Noble, J. Xu. The current science of sequence-defined macromolecules. *Prog. Polym. Sci.* **2023**, *147*, 101754.

[5] J. Lutz. Defining the Field of Sequence-Controlled Polymers. *Macromol. Rapid Commun.* **2017**, *38* (24), 1700582.

II. Sequence control in macromolecules – From natural inspiration to the design of original systems

Sequence-controlled macromolecules (SCMs) constitute an emerging research area, and at the same time have always been exploited by living organisms. The main concepts related to sequence and structural control in biomacromolecules are reviewed along this introductory chapter **(Section II.1)**. Proteins and nucleic acids are taken as examples, as they are either directly involved in the systems studied during this thesis (see **Chapters IV-A and IV-B** for proteins and DNA, respectively), or served as a major source of inspiration for the design of novel macromolecules (see **Chapters V and VI**, for DNA- and protein-inspired systems, respectively). These amazing compounds can be used in their native form, but researchers have also explored ways to introduce small modifications to their scaffold and to design artificial biomimetic systems for targeted applications **(Section II.2)**. Nowadays, the knowledge acquired on biomacromolecules and the improvements in controlled polymer synthesis allow to go even further, with the design of entirely original systems **(Section II.3)**. SCMs are already demonstrating their interest for various applications. Here, a focus is made on three areas of research, namely biorecognition, catalysis, and information storage.

II.1. Biomacromolecules: a lesson on sequence and structural control

II.1.1. Protein

II.1.1.1. A brief history and the different structural levels of proteins

The first documented usage of the word “protein” is attributed to Jöns Jacob Berzelius, in a letter addressed to Gerardus Johannes Mulder, and dates back to 1838, less than 200 years ago.[1] Mulder, later that same year, published the paper “Sur la composition de quelques substances animales”, in the *Bulletin des Sciences Physiques et Naturelles en Néerlande*. In this work, he described fibrin, albumin and gelatin as essential organic substances found in the animal and vegetal bodies, and formally coined the term “protein”. However, despite the understanding that these compounds could be regrouped in a new, particular molecular class, there was limited knowledge on their exact chemical nature. The suggestion that proteins are made of an ensemble of covalently linked small molecular fragments – the amino acids (AAs) – was put forward around 1902 by the groups of Emil Fischer and Franz Hofmeister.[2] Hydrolysis experiments revealed that whole proteins could be decomposed into smaller AAs, which was a major discovery at the time. A total of 22 genetically encoded AAs are known to be incorporated into proteins, including two non-standard residues

(selenocysteine and pyrrolysine), which are rarely found. With the exception of glycine, all of them are chiral, *i.e.* non-superimposable to their mirror image. Consequently, each AA (except glycine) can exist in two forms, sharing the same chemical composition but differing in their 3D arrangement, called enantiomers. Natural proteins are nearly exclusively composed of L-enantiomers, which also makes them chiral structures. The sequence of AAs constituting a polypeptide chain is called its *primary structure*. Decoding the sequence of a protein seemed unattainable in 1948, when Raymond M. Fuoss wrote: "We may, for instance, never learn the detailed sequence of amino acids in a protein molecule [...]".[3] Only seven years later, Sanger published his work on the sequencing of insulin, a chain of 51 AAs, for which he was awarded his first Nobel Prize in Chemistry in 1958.[4] In parallel to the discoveries related to their primary structure, polypeptides were investigated on a structural level, notably using X-ray diffraction. At first, short peptides were studied, which allowed to elaborate models for what is now known as the *secondary structures* of proteins. These are local folded motifs, resulting from a particular organization of the AAs, such as the α-helix or the β-sheet. In 1932, William T. Astbury detected two forms – that he named

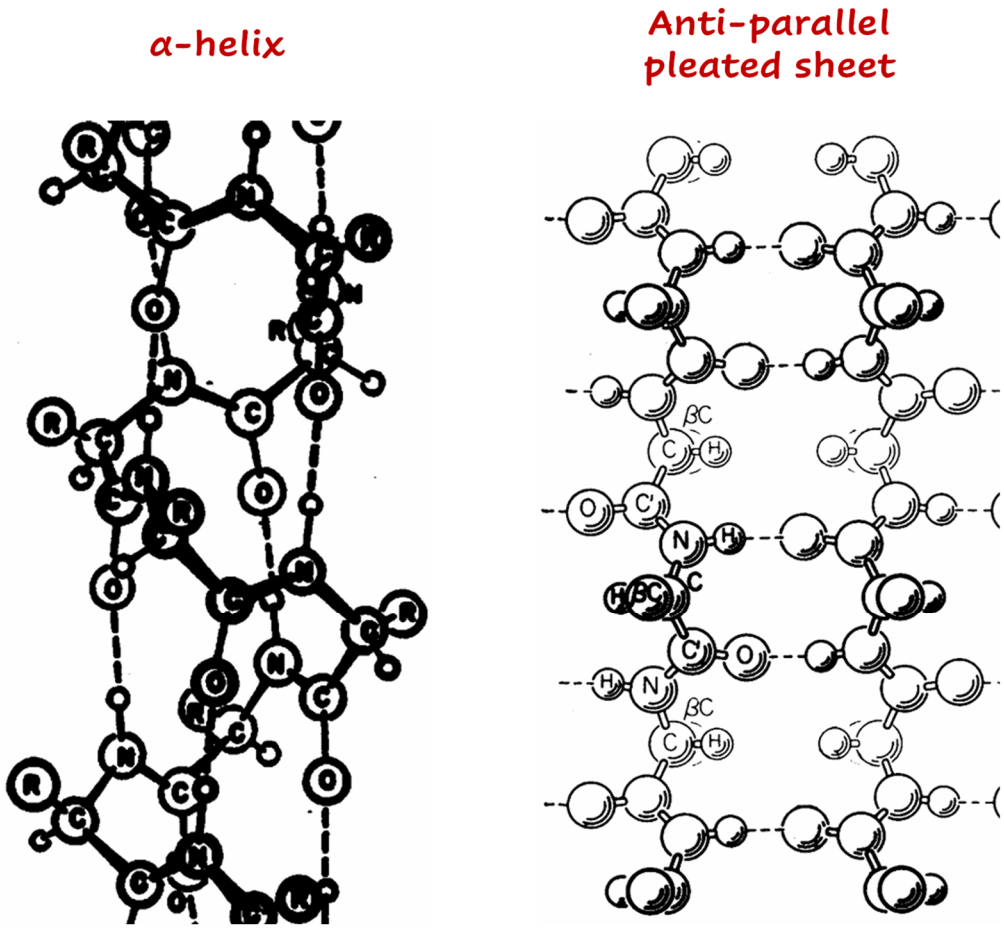


Figure II.1. Drawings of the α-helix and the anti-parallel pleated sheet (the β-sheet) structures published by Pauling and Corey in 1951. Hydrogen bonding interactions are represented as dotted lines. Adapted from Refs. 6 and 8.

α and β – for various fibres, depending on their stretching.[5] With more accurate information on bond distances and angles, Pauling and Corey established, in 1951, a precise picture of the α-helix as well as the existence of parallel and anti-parallel pleated sheets.[6–8] These structures are maintained by well-organized intramolecular hydrogen bonding interactions **(Figure II.1)**. When discussing about secondary structures, it is important to cite the work of Ramachandran and his famous diagram, published in 1963.[9] He further refined the so-called "Pauling-Corey coordinates" by defining the authorized boundaries for the backbone dihedral angles Φ and Φ' **(Figure II.2)**. His diagram reveals that the conformational space available to the peptide backbone is quite restricted, many regions being inaccessible due to steric hindrance. The highly ordered secondary structures combine with one another, sometimes including more flexible regions, to constitute the *tertiary structure*, *i.e.* the entire 3D structure of the protein. Myoglobin was the first protein whose 3D structure was elucidated, in 1958.[10] Interestingly, the researchers noted that the model was "more complicated than has been predicated by any theory of protein structure", and were surprised by the lack of symmetry and regularities along the chain. This marked an early encounter with the ever-present challenge of predicting the 3D structure of a protein. The same group published a higher-resolution model in 1960, demonstrating clearly for the first time that α-helices exist inside globular proteins.[11] The last structural level of proteins, the *quaternary structure*, designs functional complexes made by the assembly of several polypeptide chains. A prime example is hemoglobin, a four-protein complex, which was the first experimentally determined quaternary structure.[12] These remarkable works on the structure of myoglobin and hemoglobin, which were pivotal in the history of protein science, were awarded the Nobel Prize in Chemistry in 1962.

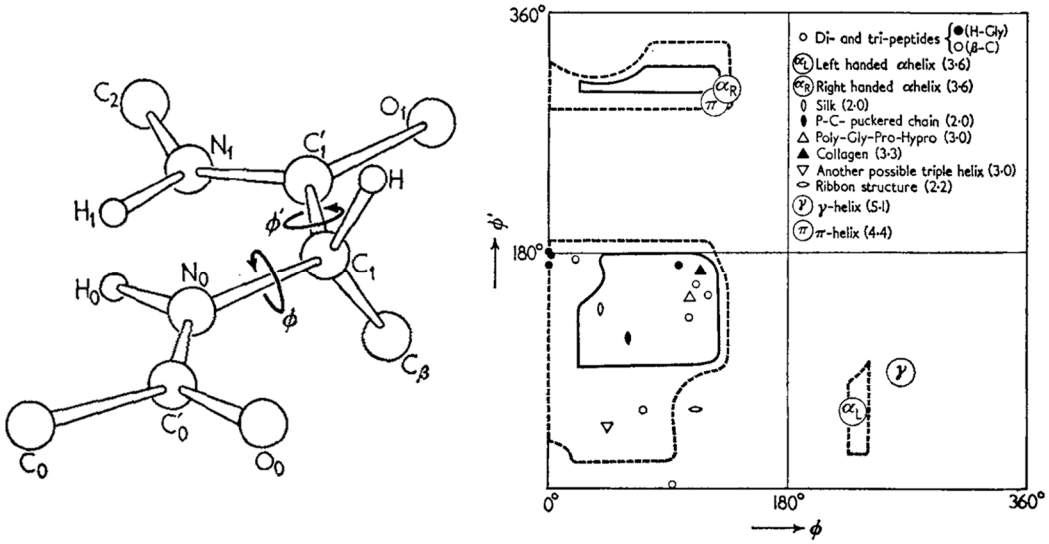


Figure II.2. Illustration of a polypeptide chain (left) and Ramachandran plot (right), showing the allowed values of the Φ and Φ' dihedral angles. Adapted from Ref. 9.

II.1.1.2. Sequence – structure – function relationships in proteins

Most proteins display the exceptional ability to fold into a unique and well-defined 3D structure, known as the native structure. The folding process must be extremely efficient, as a protein's ability to perform its biological functions is rooted in its 3D structure. Yet, proteins remain highly flexible entities and are essentially stabilized by non-covalent interactions. This structural flexibility is essential to many of their biological functions, and proteins often undergo conformational changes in response to external stimuli, such as the binding to another receptor or to a substrate.[13] Consequently, native structures cannot be too thermodynamically stable, which opens the door to failures of the folding process. Misfolded proteins are not only inactive: they are prone to interact with other chains, potentially leading to the formation of very stable aggregates, which play a role in several neurodegenerative diseases.[14] Intramolecular folding and intermolecular aggregation are two competing phenomena. The former ended to be a very complex process, optimized during thousands of years of evolution, to circumvent the latter. A whole family of proteins is even dedicated to assisting folding: the molecular chaperones.[15]

The search for the native structure is anything but random, as illustrated by the famous Levinthal's paradox.[16] However, there is no clear consensus about the exact folding mechanisms, although several important principles are generally accepted. The first step of a protein folding would involve the formation of the secondary structures, driven by nonspecific and local interactions, essentially hydrophobic effects and hydrogen bonds.[17] These structures are conformationally restricted to the regions shown in the Ramachandran plot, due to steric hindrance. Further organization of the secondary structures would form a network of longer-range intramolecular interactions.[18] The protein would then sample several intermediate states (sometimes described as "molten globule" states), until finally reaching its most stable, native structure. The folding process would follow a "funnel-shaped" pathway in the potential energy of the system, where the formation of partially folded and compact intermediates would reduce the conformational space and quicken the search towards the lowest energy structure **(Figure II.3)**. Conflicting views address the sampling of these intermediates, sometimes with different interpretations of the same experimental results. Some argue that folding occurs through a unique pathway, were intermediate states are sampled in a specific order through cooperative processes **(Figure II.3 A)**.[19] Others express the view of a rather chaotic process, where multiple pathways can lead to the native structure from a wide conformational ensemble of compact intermediates **(Figure II.3 B)**.[18,20]

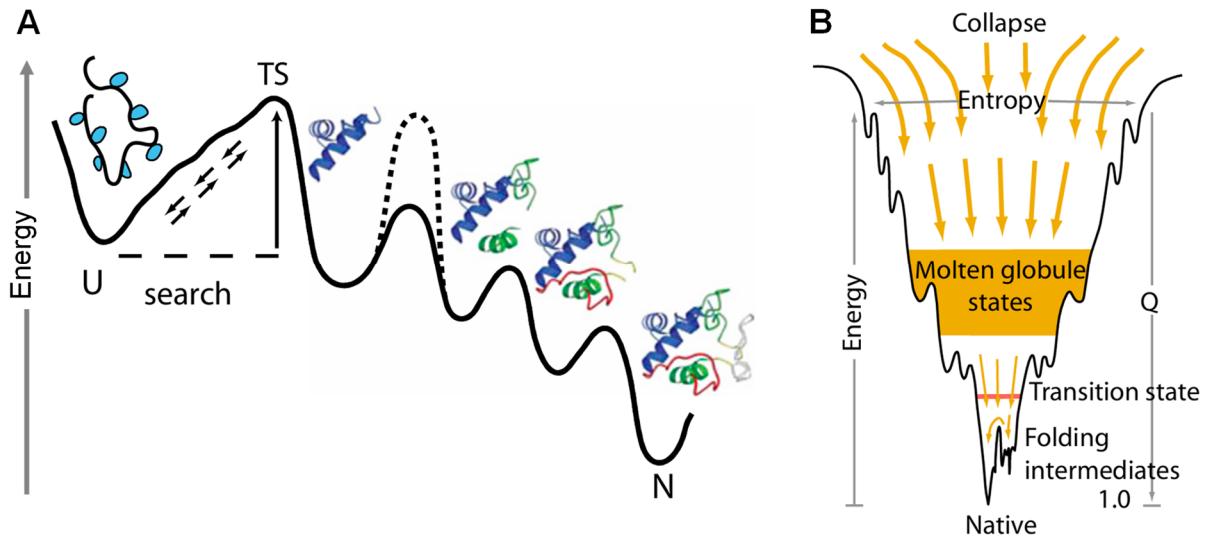


Figure II.3. Simplified funnel-shaped representations of the folding mechanism. **(A)** The protein would go from its unfolded (U) to its native (N) state by sampling a series of intermediates in a defined order, through the cooperative sequential formation of secondary structures. **(B)** Other view of the folding process, where the fast formation of the secondary structures would lead to compact, partially folded states (*i.e.* molten globules). The protein can fold into its native structure through multiple pathways. Adapted from Ref. 19.

A crucial point is that the formation of the secondary structures is mainly driven by local effects, thus strongly depends on the sequence order of the AAs. This well-organized, unique suite of monomers encodes all the information governing the folding and the formation of the native structure.[21] Replacing or modifying as little as one AA can sometimes strongly impact the folding of a protein and its biological functions.[22–24] Despite all the work carried out over the years and the understanding of some key steps of protein folding, researchers have not yet unraveled all the mysteries relating a given sequence of AAs to a given native structure. Establishing sequence – structure – function relationships remains challenging. However, a lot of data has been acquired and huge databases, such as the Protein Data Bank (PDB), regroup the sequence of proteins and their associated 3D structure.[25] This is particularly useful to feed machine learning (ML) algorithms, which constitute a method of choice to uncover relationships between a series of inputs (sequences) and outputs (structures). In 2020, AlphaFold, an ML algorithm developed by Google DeepMind, entered the Critical Assessment of protein Structure Prediction (CASP), a competition testing the efficiency of different methods to predict the structure of proteins.[26] During CASP, researchers receive the AA sequences of several proteins whose experimental structures have recently been established but are not yet public. AlphaFold won the competition, displaying an outstanding accuracy in its predictions. The success of AlphaFold, which is able to predict the structure of a protein solely from its AA sequence and a large

dataset of known sequence – structure pairs, provides further evidence that the sequence of monomers dictates the 3D structure of proteins. An updated version, AlphaFold3, was released in 2024.[27] This new iteration enables structural predictions of various biomacromolecules, small ligands, and supramolecular complexes made by the assembly of several components. Unfortunately, AlphaFold only brings us from point A (sequence) to point B (structure), without giving any information on the pathway connecting them.

The precise sequence – structure control found in proteins allows the emergence of remarkable functions. The chaperones, mentioned earlier, are an interesting example. Enzymes constitute another impressive family of proteins, dedicated to catalysis. They display exceptional selectivity towards their substrates, even within the crowded and complex cellular environment.[28] Substrate recognition by the active site of the enzyme occurs through shape complementarity and the formation of stabilizing interactions. Among all potential substrates, those most stably bound are selectively transformed into the desired product. Similarly, antibodies display a specific binding site, enabling the selective recognition and neutralization of pathogens. Biological processes rely on a myriad of host – guest interactions that depend directly on the 3D structures of the proteins involved; structures that are themselves encoded in the AA sequence. Therefore, the biosynthesis of proteins – a process called *translation* – must be perfectly controlled. To this end, Nature developed a complex machinery involving ribonucleic acids (RNAs).

II.1.2. Ribonucleic acid

RNAs constitute another major class of sequence-defined biomacromolecules. They belong to the group of nucleic acids, along with deoxyribonucleic acids (DNAs). Their polymeric backbone is made of a sugar – the ribose for RNA – and a phosphate moiety. Like proteins, natural nucleic acids are chiral, the sugar component existing exclusively in its D-enantiomeric form. Each ribose is linked to a nitrogenous base, or nucleobase. While proteins are built on 22 different AAs, RNAs rely on a smaller set of four different monomers, distinguished by the nature of the nucleobase, which can be the adenine (A), uracil (U), cytosine (C) or guanine (G). The sugar – phosphate – nucleobase triad constitutes the nucleotides, *i.e.* the monomers forming nucleic acids, which are linked together by covalent phosphodiester bonds. The nucleobases are complementary by pairs, forming A---U and C---G dimers through hydrogen bonding interactions, known as the Watson-Crick pairing. Although this allows the hybridization of two RNA strands to form a double helix, single-stranded structures are more frequent in living organisms. RNA is chemically nearly identical to DNA, but in terms of folding and

functions, it is much closer to proteins. However, RNAs are more flexible, making the determination of their 3D structures even more challenging. This underlines once more the complexity of establishing sequence – structure relationships, even for polymers made with only four different monomers. ML models are currently investigated for predicting the 3D structures of RNAs, although the available data is much sparser (RNA-only structures account for less than 1 % of the PDB).[29] Even AlphaFold3 displays important errors on some RNA structures, especially on less common motifs.[30] More generally, ML models are far from accurately predicting the conformational landscape of nucleic acids, as illustrated in **Figure II.4.**

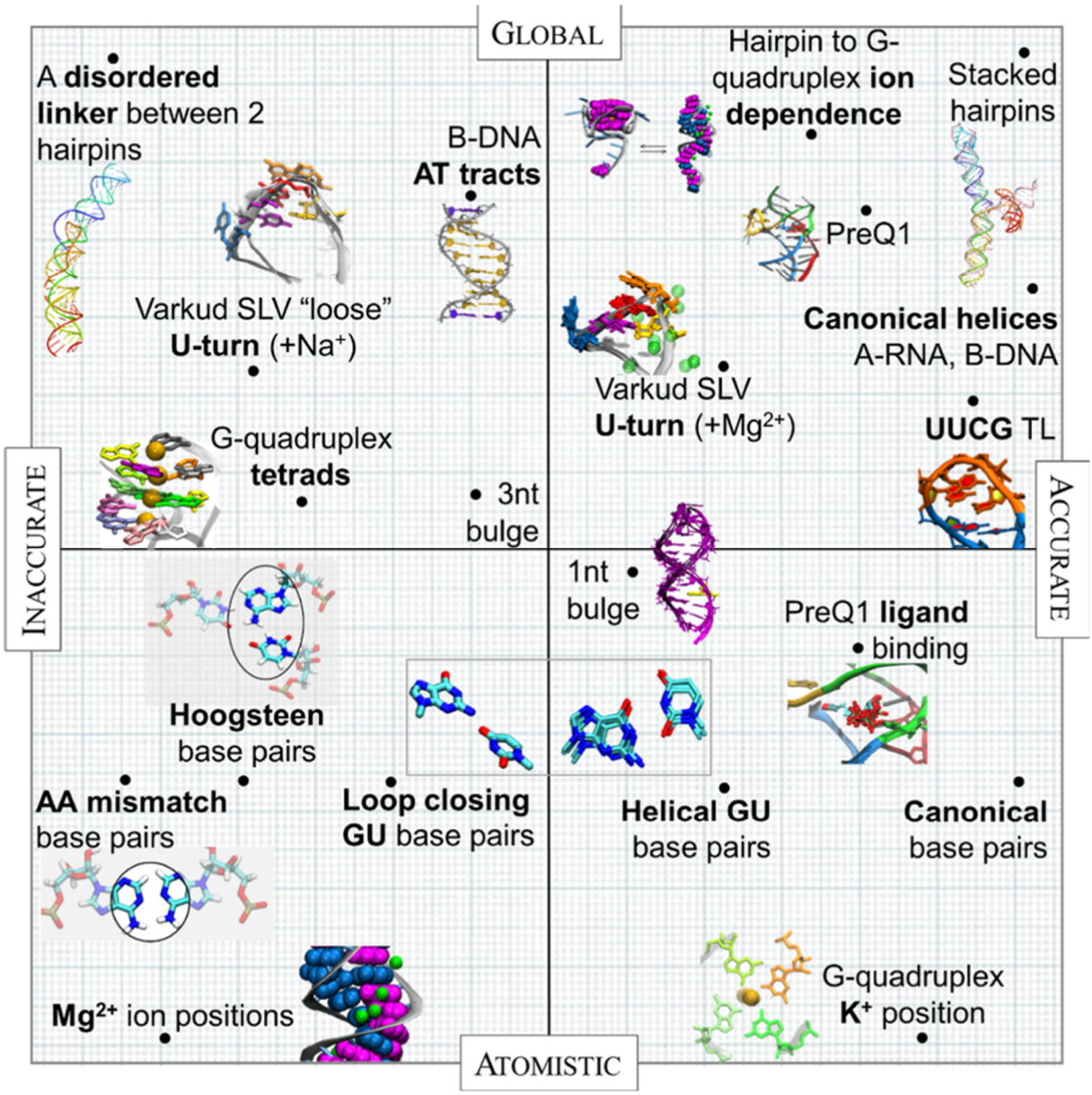


Figure II.4. Overview of the accuracy of ML models in predicting nucleic acid structures, as discussed in Ref. 30. The structures are ordered along the horizontal axis from inaccurately (left) to accurately (right) predicted by ML models. The vertical axis spans from atomistic details (bottom) to global shape and conformation (top). Reproduced from Ref. 30.

RNAs are able to fold into a wide variety of conformations, stabilized by intramolecular stacking interactions and hydrogen bonds involving the nucleobases, allowing it to perform various biological functions.[31] A well-known example is the synthesis of proteins, which involves different RNA species. As mentioned at the end of the previous section, this process is called translation. The information required to build a specific protein is encoded in the sequence of nucleobases of a single-stranded messenger RNA (mRNA). This sequence is deciphered by the ribosome, a complex macromolecular machine made of proteins and RNA. Inside the ribosome, the mRNA sequence is decoded three nucleotides at a time – these triplets are called *codons* – by other polynucleotides called transfer RNAs (tRNAs). Each tRNA contains a recognition site made of a sequence of three nucleotides, able to bind only to the matching codon through the complementary hydrogen bonding pattern of the nucleobases. At its opposite end, the tRNA carries an AA. Therefore, to each codon corresponds one specific AA. During translation, tRNAs bind to the successive codons of mRNA, bringing the AAs one after another to form the polypeptide chain. After complete decoding of the mRNA strand, the synthesized protein is released into the cytoplasm. Its structure can subsequently be modified through post-translational modifications (PTMs), *i.e.* the chemical attachment of a functional group to the protein after its biosynthesis.
RNAs are another example of the prime importance of the sequence control in biomacromolecules. As for proteins, their biosynthesis must be perfectly controlled and cannot bear mistakes: an error in the mRNA sequence would lead to an erroneous codon, possibly translating into the wrong AA. The biosynthesis of RNA is templated by DNA itself, along a process called *transcription*.

II.1.3. Deoxyribonucleic acid

When mentioning the molecules of life, DNA, the carrier of genetic information, often comes first to mind. The chemical structure of this biopolymer is nearly identical to the one of RNA, with two notable differences. First, the sugar in DNA is deoxyribose. Then, the uracil nucleobase of RNA is replaced by thymine (T). As U, T is complementary to A. In terms of structure, however, DNA displays a significantly different behaviour. In eukaryote cells, it is essentially found in a double helix conformation, formed by the supramolecular assembly of two complementary DNA strands. This structure was officially elucidated in 1953 by Watson and Crick, with inputs from many other researchers, notably X-ray experimental data from Rosalind Franklin.[32] The DNA double helix is an extremely conserved and protected structure, as it contains all the information necessary to the proper functioning of cells: its sequence of nucleotides is the code governing protein biosynthesis. During transcription, the DNA double helix

is locally unwound at a specific site, making the targeted sequence of nucleotides accessible. One of the two strands then serves as a template for synthesizing the complementary RNA sequence, following base pairing rules. The synthesized RNA can then undergo maturation steps to produce the mRNA, which will be used to synthesize proteins. The process of transcription is catalyzed by RNA polymerase, assisted by a variety of enzymes and complex molecular machineries. Along transcription and translation, the control of sequence is propagated from DNA to proteins, through RNA intermediates. This process is extremely complex and involves three different biological "languages": the DNA tetrad A-T-C-G; the RNA tetrad A-U-C-G; and finally, the 22 amino acids of proteins. It opens the door to dramatic butterfly effects, as one single mutation in the DNA code, *i.e.* the insertion, deletion or substitution of one nucleotide, may disrupt this flow of information and lead to inactive, misfolded, and potentially harmful proteins. Such complexity explains the highly evolved machinery that Nature has built to synthesize proteins.

As described along this entire first section, sequence control is a central characteristic of the highly functional biomacromolecules of life, translating into an absolute control of their 3D structures and functions. Their fascinating properties naturally triggered a major interest for many researchers, eager to explore the possibilities offered by their particular chemistry and seeking ways to expand their use beyond natural contexts. This idea is not new: in 1902, Emil Fischer stated in his Nobel Lecture: "To equal Nature here, the same means have to be applied, and I therefore foresee the day when physiological chemistry will not only make extensive use of the natural enzymes as agents, but when it will also prepare synthetic ferments for its purposes." As will be covered in the next section, the future proved him right.

II.2. Mimicking or modifying Nature's building blocks

II.2.1. Chemical synthesis of (modified) biopolymers

Before using biopolymers for specific applications or simply to study their sequence – structure – function relationships, they must be produced. To this end, one must either extract them from their environment, or directly synthesize them. The latter offers the advantage that any polypeptide or polynucleotide can be formed, without being limited by the sequences biologically available.[33,34] The chemical synthesis of biopolymers has greatly benefitted from advances in solid-phase synthesis, notably with Merrifield's work on polypeptides in the 1960s and the development of phosphoramidite chemistry for polynucleotides in the 1980s.[35,36] In these approaches, the growing chain is covalently linked to a solid support, typically a column, and monomers functionalized

with a protecting group are added iteratively. Between each addition, the column is washed to remove the excess of reagents. A deprotection step is then performed before coupling the next monomer, guaranteeing sequence control. Once the desired chain has been assembled, it is cleaved from the solid support and purified. These methods are still routinely used to produce tailor-made biomacromolecules and are now commonly automated. In general, solid-phase approaches allow the formation of relatively short biopolymers (around 50-100 monomers), which can subsequently be coupled through ligation steps to create longer chains. Nevertheless, several recent advances in biopolymer synthesis have been made to improve standard protocols. One trend concerns the miniaturization of the synthesis sites and devices, with the advent of microarray-based methods,[37–41] or the use of microfluidic technologies.[40,42,43] Flow chemistry also constitutes a promising area. Automated protocols have recently been optimized for the synthesis of long polypeptides,[44] enabling the incorporation of site-specific modifications,[45] or the formation of synthetic covalent dimers that mimic the quaternary structure of complex natural dimers.[45] Another interesting way of making proteins is to divert their biosynthesis pathway, by directly modifying the DNA sequence through genetic engineering. These approaches have become extremely powerful since the emergence of genetic edition tools such as the CRISPR/Cas9 enzyme, which allow scientists to bring changes on precise locations of the genome.[46] An example used this process to biosynthesize proteins functionalized with fluorescent tags, to facilitate their localization and study their functions in the cell.[47] Concerning DNA synthesis, an elegant approach harnesses a natural enzyme belonging to the class of DNA polymerases, the terminal deoxynucleotidyl transferase.[48] This enzyme enables the controlled addition of nucleotides to a growing chain and even tolerates modified nucleotides, allowing the formation of tailored DNA strands.[49]

II.2.2. Twisting proteins to create artificial systems

As mentioned in the previous section, the synthesis pathways developed for biopolymers can be adapted to incorporate (stereo)chemical modifications or unnatural monomers into their structures. This approach is extremely attractive, as it allows site-specific modifications on known and efficient scaffolds. Many different unnatural AAs have been studied to bring new side-chains while keeping intact the natural peptide backbone.[50] An interesting application consists in the modification of proteins with stimuli-responsive AAs. For example, one group introduced a photo-crosslinkable AA into a protein to identify its interaction partners.[51] Upon irradiation with UV light, a covalent bond was formed between the protein and its binders, permitting their isolation and characterization. Another group added a photolabile protecting group to a tyrosine residue in an interleukin receptor.[52] The binding of the

interleukin to its receptor was significantly lowered before irradiation. After exposure to UV light, thus removal of the photo-responsive protecting group, the complex could be formed with normal affinity. Here, a slight modification on a single AA allowed researchers to modulate an interaction, having repercussions on a whole phosphorylation cascade, which offers interesting perspectives for therapeutic applications. Another important application of unnatural AAs concerns the improvement of the activity and stability of natural enzymes, or even the addition of catalytic functions to non-enzymatic proteins. These systems constitute an ideal starting point to design biocatalysts with novel properties. For example, a protein was transformed into an artificial endonuclease – an enzyme able to cleave phosphodiester bonds within nucleic acids – by incorporating an unnatural AA containing a bipyridyl moiety **(Figure II.5)**.[53] The naturally occurring protein recognizes and binds short double-stranded RNAs of 19 to 25 nucleotides in length in a size-selective and sequence-independent manner. After the site-specific introduction of the bipyridyl moiety and in the presence of copper, the artificial enzyme was able to cleave short non-coding RNAs with high specificity, a function not observed in any known natural endonuclease. Another work attempted to slightly modify a histidine – an AA involved in the catalytic center of many biocatalysts – by attaching a vinyl moiety to one of its nitrogen atoms.[54] This modification increases the electron-withdrawing behaviour and lowers the pKa of the imidazole ring, leading to improved catalytic properties at pH = 5.5. These impressive examples confirm the possibility to create systems with novel or improved properties, by modifying enzymes at the level of a single AA. Once the role of each monomer unit in the sequence has been understood and related to its position in the 3D structure, thus when sequence – structure information has been deciphered, it becomes possible to precisely engineer remarkable functional systems. Many successful examples make use of unnatural AAs, but not all changes on the precise monomer sequence of proteins are tolerated. An ML model was recently proposed to rationalize the design of proteins containing unnatural AAs, by identifying positions in the primary structure that would be likely to tolerate substitutions.[55] At the time of publication, the training dataset included 1221 unnatural AA substitution sites, with a marked imbalance between the number of successful and unsuccessful cases (1064 and 157, respectively). The ML model would undoubtedly benefit from more examples of failed modifications, to better capture statistical trends underlying prohibited substitutions. Unfortunately, successful results are more prone to be published, which is detrimental to the development of reliable ML algorithms. Nevertheless, the model achieved reasonable predictive accuracy and was experimentally validated in a test case.

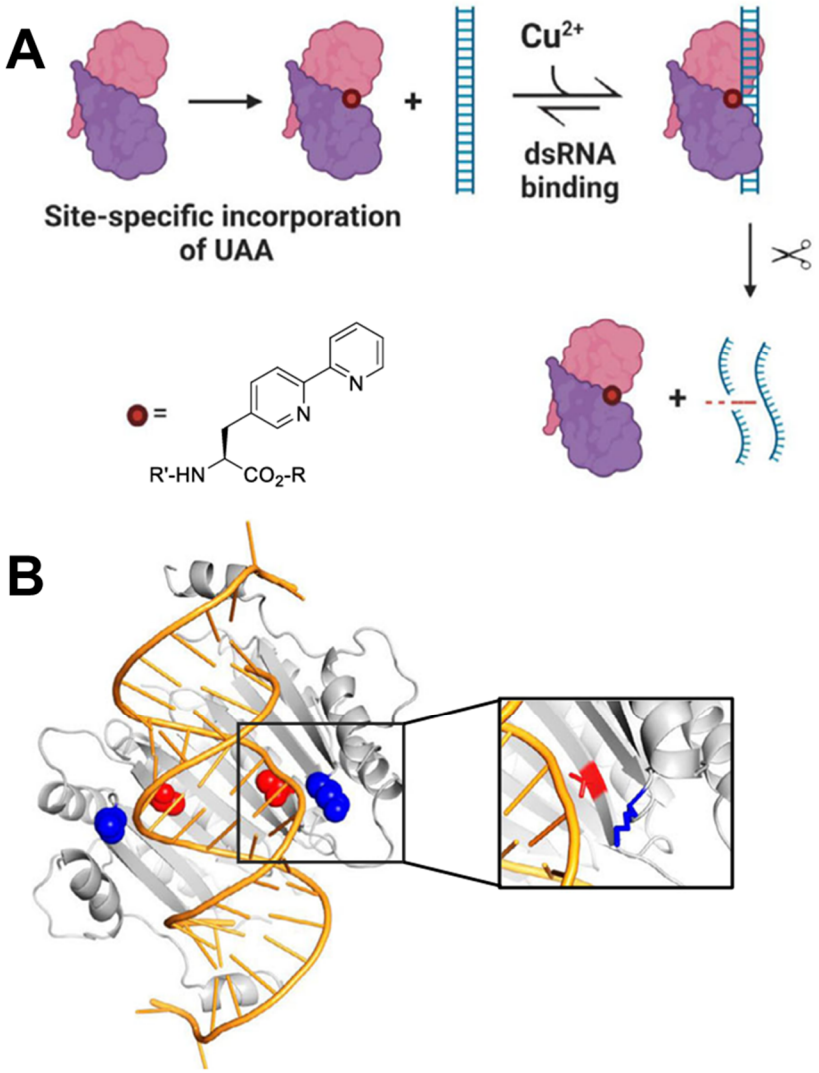


Figure II.5. Design of an artificial endonuclease through insertion of an unnatural AA (UAA) into a non-enzymatic protein. **(A)** Schematic representation of the process; after the site-specific insertion of a bipyridyl moiety (depicted as the small red ball) the protein is able to cut an RNA strand. **(B)** Crystal structure of the binding pocket of the protein before modification, assembled with a short double-stranded RNA (PDB ID: 1RPU). The AA in red is the one that is replaced by the unnatural AA. Adapted from Ref. 53.

Other SCMs were designed with the idea of modifying the natural peptide backbone. These peptidomimetic systems constitute a broad range of chemical structures. Their synthesis generally relies on solid-phase protocols and iterative approaches, ensuring sequence control. Note that some of these structures are purely artificial, despite their resemblance to natural peptides. They could as well have been discussed in **Section II.3**, which focuses on synthetic SCMs. However, we made the choice to integrate them in this section due to their strong biomimetic character. One of the most minimal structural modifications to a peptide backbone is the inversion of the stereochemistry of its building blocks, the L-AAs. Their D-enantiomers are exploited in diverse

applications, such as the development of "mirror-image life" and racemic protein crystallography.[56] More interestingly, D-AAs are less efficiently recognized by natural biomolecules, such as proteases – enzymes specialized in protein degradation – which essentially interact with natural AAs. This property is exploited in the design of peptide drugs, notably through mirror-image phage display.[57] This strategy was recently followed to isolate D-peptides able to disrupt the activity of the epidermal growth factor, a protein associated to the uncontrolled proliferation of tumour cells.[58] However, mirror-image phage display has had few practical applications, essentially because it requires the synthesis of long D-polypeptide targets, which remains challenging. The recent developments of automated flow chemistry protocols (mentioned in the previous section) could renew interest for this technique and allow the emergence of more protease-resistant D-peptide drugs.[59] The influence of peptide chirality is investigated in the framework of cellular migration in our thesis, see **Chapter IV-A**. Among peptidomimetic SCMs, peptoids constitute a molecular class that attracted great attention in recent years. In contrast to peptides, the side-chain is carried by the nitrogen atom instead of the α-carbon. This apparently trivial modification implies two major changes: the peptoid backbone is achiral and does not possess hydrogen bond donors, which prevents intramolecular interactions in the backbone. Therefore, the folding and secondary structures of these SCMs are essentially dictated by the nature of their side-chains. This simpler network of interactions makes peptoids ideal targets to study sequence – structure relationships.[60] For instance, the placement of hydrophobic units in the sequence was shown to affect the dynamics of hydration water in short amphiphilic polypeptoids.[61] The researchers even found that changes in the sequence had more impact than the peptoid conformation on water behaviour. They attributed this observation to the inability of the chains to bury water molecules even when being compact, due to their small size, and the stronger impact of local chemical environment on water. Sequence – structure relationships were studied on longer amphiphilic polypeptoids by varying the position of hydrophobic units in the chains.[62] The results, combining experiments and simulations, showed that different conformational ensembles were obtained depending on the distribution of hydrophobic moieties in the primary structure. Beyond single-chain systems, the supramolecular assembly of small peptoids into nanohelices was also demonstrated.[63] Impressively, the nanohelix handedness could be controlled by the incorporation of a single chiral side-chain. The versatility of peptoids makes these SCMs attractive for various applications. An interesting example concerns the storage of solar energy, using photoswitchable azobenzene compounds as side-chains.[64] Azobenzene molecules can undergo *trans* to *cis* photoisomerization upon irradiation with UV-Vis light, and spontaneously revert back to their more stable

trans form after some time, releasing energy in the form of heat during the process. In this example, the position of the azobenzene side-chain in the sequence affects its spectroscopic properties and the kinetics of retro-isomerization, which are crucial parameters for the storage of solar energy. Another promising area for peptoids is artificial catalysis, where the formation of specific secondary structures can advantageously be exploited to build well-defined catalysts. Peptoids functionalized with chiral substituents were shown to fold into helices of preferred handedness, and displayed enantioselective catalysis when covalently bound to an achiral 2,2,6,6-tetramethyl-1-piperidinyloxyl (TEMPO) catalytic unit.[65] Interestingly, when the TEMPO moiety was grafted on the center of the peptoid rather than at its extremity, the enantioselective behavior was nearly fully lost, demonstrating an important effect of sequence.

II.2.3. Exploiting and modifying nucleic acids for novel applications

As the relation between the sequence of AAs and the resulting 3D structure of a protein is not always straightforward, despite the advances of predictive ML models, full tailor-made proteins are rarely built. Instead, new functions are introduced into known scaffolds through site-specific modifications, or simpler mimetic systems are designed. On the other hand, nucleic acids are much easier to program, therefore much easier to use "as is". Not at the single-chain level, because RNAs and DNAs are also able to form various secondary structures, but at the level of their assemblies, which is much more easily programmable. Two complementary strands will form a double helix, following the simple rules of Watson-Crick pairing. Therefore, high order self-assembled architectures can be engineered through a perfect control over the sequence of nucleotides **(Figure II.6)**. A beautiful example is shown by DNA origamis.[66] These structures are formed by combining a long single-stranded DNA (ssDNA) scaffold with many short oligonucleotides playing the role of staples **(Figure II.6 A)**. The long ssDNA (typically extracted from a virus) can be folded into a variety of nanostructures through complementary base pairing on specific locations. Recent improvements in the protocols of assembly make possible the formation of many controlled DNA nanostructures, including nanogrids and very complex 3D shapes.[67,68] A freely available software allows users to determine the sequence of the oligonucleotides necessary to build their desired 2D or 3D shape, without even requiring the use of an ssDNA scaffold.[69] Beyond being beautiful scientific accomplishments, these nucleic acid-based nanostructures are envisioned for various applications, such as templating of nanomaterials, drug delivery, nanophotonics, etc.[70] An interesting example concerns the development of a DNA tweezer, able to reversibly control the activity of an enzyme **(Figure II.6 B)**.[71] The structure, composed of two DNA arms

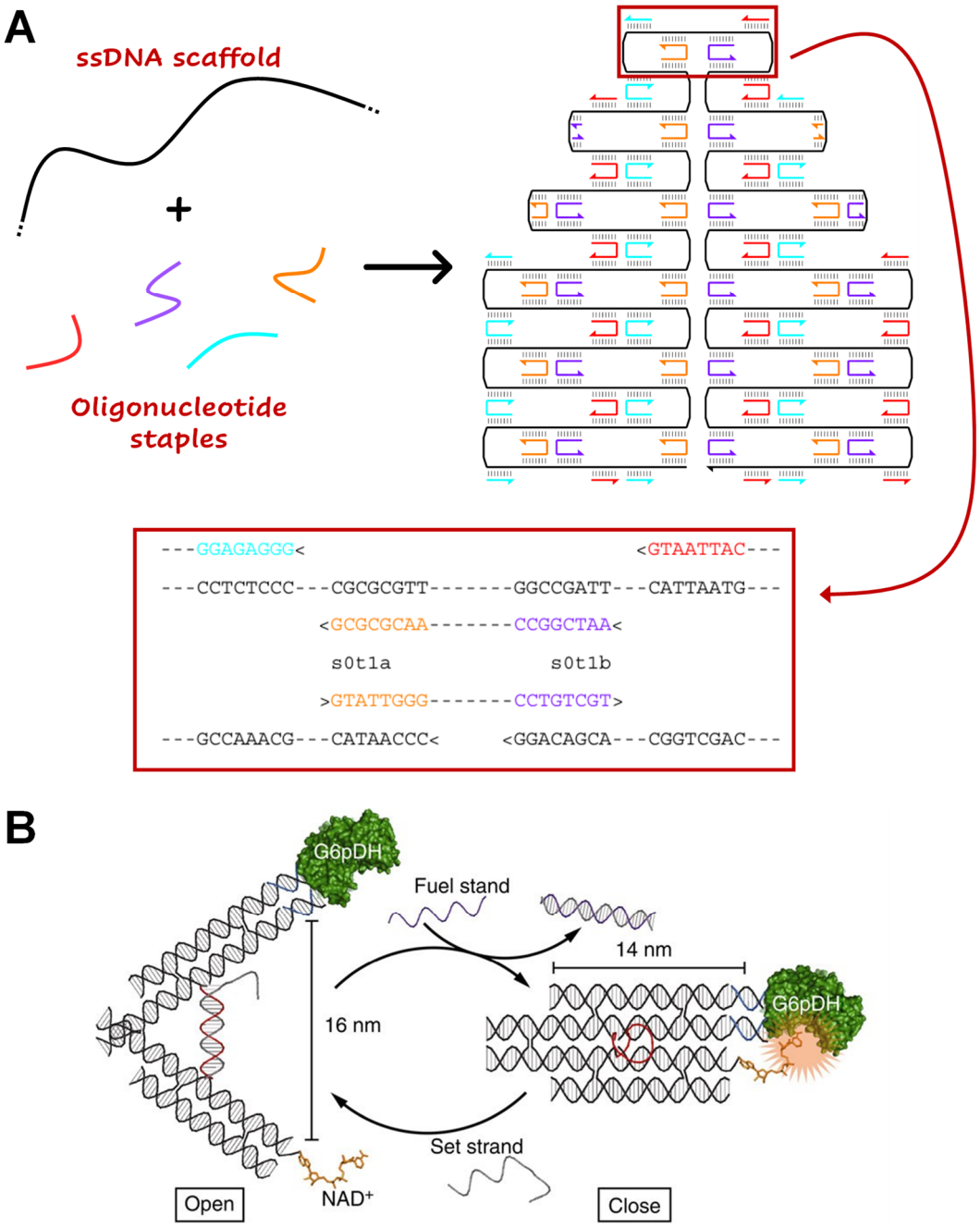


Figure II.6. Examples of ordered DNA nanostructures. **(A)** Representation of the functioning of DNA origamis, where the folding of a long ssDNA is guided by short oligonucleotides. The area in the red rectangle is shown with the detailed nucleobase sequence below, highlighting the complementary A---T and C---G pairings. Adapted from Ref. 66. **(B)** Functional DNA tweezer able to reversibly switch between open (left) and closed (right) forms, depending on the conformation adopted by the central oligonucleotide, in red. Reproduced from Ref. 71.

functionalized with an enzyme and its cofactor, can switch between closed (active) and open (inactive) forms. This switching is governed by a central regulatory oligomer. In the absence of a complementary strand, the oligomer folds into a hairpin structure, bringing the two arms into proximity and favoring the formation of the active enzyme-

cofactor complex. Upon hybridization with a strand of complementary sequence, the conventional double helix conformation is retrieved, which spatially separates the arms and disables enzymatic activity. Reversible activation / deactivation cycles were demonstrated, confirming the possibility to regulate the activity of an enzyme through a precisely engineered DNA nanomachine.
Researchers also aimed at expanding the functions of nucleic acids by modifying their structure, either the ribose-phosphate backbone or the nucleobases. These synthetic systems are called xenonucleic acids (XNAs). Similarly to unnatural AAs for proteins, a variety of new nucleobases were designed to enrich the A-C-G-T/U biological alphabet.[72] For instance, an ssDNA was functionalized with a pH-responsive artificial nucleobase for targeting cancer cells.[73] The lower pH of the microenvironment of these cells triggers a switch of the nucleobase, which becomes able to recognize and inhibit receptors involved in cell migration. In our thesis, an unnatural nucleobase was investigated for the supramolecular assembly of a catalytic complex, see **Chapter V**. Another example of XNA is the peptide nucleic acid (PNA), an interesting hybrid structure between a peptide-like backbone and nucleobases as side-chains.[74] PNA-DNA complexes were shown to be more stable than their DNA-DNA counterpart because of the absence of electrostatic repulsion between the phosphate groups, as PNAs are not negatively charged.

The various examples shown throughout this section illustrate the interest of using the scaffold of well-defined natural SCMs, which are ideal targets for site-specific modifications and constitute important inspirations for the design of biomimetic compounds. The next step towards artificial systems is to apply the fantastic lesson taught by Nature into fully human-made macromolecules.

II.3. Lesson learned! Applying sequence control to synthetic macromolecules

II.3.1. Controlling polymer synthesis towards artificial SCMs

Advances in polymer synthesis in the past 15-20 years have allowed researchers to go beyond the heterogeneous mixture of chains associated with polymer chemistry. As a reminder, there is a distinction between “sequence control” and “sequence definition”: the latter refers to perfectly uniform samples in which all chains share the exact same ordering and number of monomers (SDMs), whereas the former also includes samples with low dispersity and partial sequence regulation (SCMs) **(Figure II.7)**.[75]

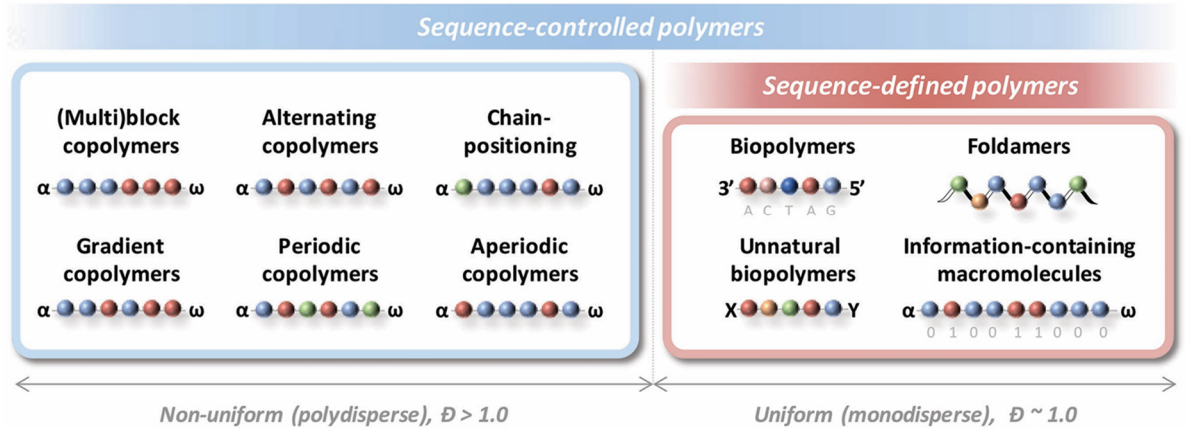


Figure II.7. Illustration of the meaning of the terms "sequence control" and "sequence definition". "Polymer" is replaced by the more general term "macromolecule" in our thesis. Adapted from Ref. 75.

Traditional step-growth and chain-growth polymerization pathways have been considerably improved, notably through so-called "living" and controlled radical approaches.[76] By reversibly modulating the reactivity of the growing chains, a better control on monomer incorporation is gained, leading to SCMs with lower dispersity and the formation of complex multiblock systems. However, these methods still follow statistical rules: their improvement narrows the gaussian distribution of chain lengths within a sample, but the control over sequence and degree of polymerization (DP) is still not absolute.

The synthesis of perfectly controlled macromolecules generally relies on iterative approaches and the stepwise incorporation of monomer units.[77] While these are efficient for controlling the sequence, they generally only give access to oligomers of limited length. For example, if each monomer addition is realized with a yield of 99 %, the overall yield to reach a 16-mer is about 86 %. If the yield of each step decreases to 95 %, which remains very high, the overall yield drops to approximately 46 %. Therefore, synthesis protocols must be extremely well optimized to reach high DP; they often exploit orthogonal reactions and click-chemistry. Nevertheless, the SDMs formed by such step-by-step approaches remain generally limited to around 20 monomer units.[78]

This introductory section does not aim to thoroughly cover synthetic strategies to reach SCMs; the interested reader is directed to excellent papers reviewing this topic.[76,77,79,80] Two key points should be highlighted: first, the synthesis of SCMs and SDMs remains challenging and their large-scale production is still limited; second, SCMs are attainable through various pathways, allowing the use of many diverse backbones and side-chains.[78] While it is already difficult to predict the 3D structure of proteins, made with a unique backbone and a limited number of different side-

chains, the huge chemical space offered by synthetic SCMs brings considerable challenges.[81] In addition to sequence, stereochemistry can also be controlled to further increase an already complex conformational landscape.[82] Despite these challenges, the influence of sequence has already been demonstrated for SCMs in various applications, reinforcing the promise of sequence control in synthetic chains.

II.3.2. Synthetic SCMs for biorecognition

Many biological processes rely on specific ligand – receptor interactions. The use of synthetic SCMs as ligands emerged rapidly, due to their conceptual resemblance with biomacromolecules. Precisely controlling the position of each monomer unit onto a polymer backbone is very attractive to establish well-defined interaction networks and to develop selective systems towards biological receptors of interest.
An original approach used the sequence-defined peptide backbone of the human serum albumin (HSA) as a scaffold to precisely introduce functional groups **(Figure II.8)**.[83] The researchers exploited the presence of a unique cysteine residue on the outer surface of the protein in its native state to site-specifically substitute it by a biotin. Then, long polyethylene glycol (PEG) chains were introduced to replace surface-accessible carboxylate moieties, leading to the formation of a brush polymer. Subsequently, the biotin was used to recruit streptavidin, a protein having a very high affinity towards biotin. The streptavidin could, in turn, recruit other units, such as an antibody (as shown in **Figure II.8**). This work demonstrates the possibility to design a precisely functionalized brush polymer by exploiting the defined sequence of a protein scaffold.

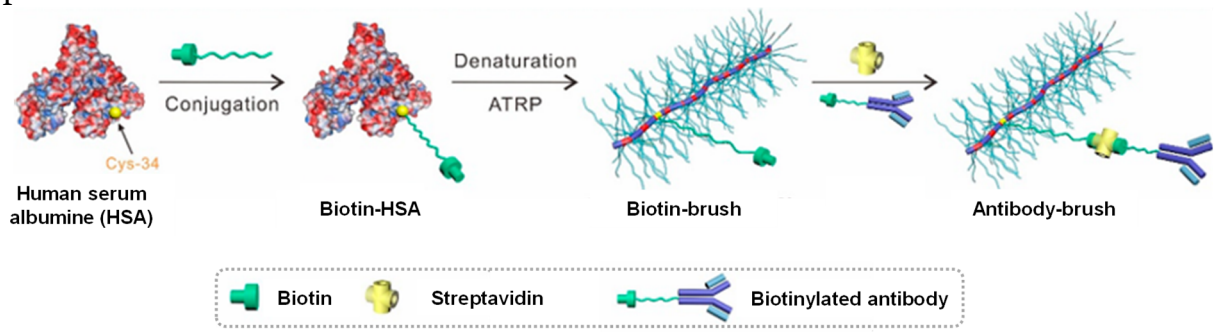


Figure II.8. Protein scaffold functionalized with a biotin grafted to the residue cysteine 34, before the addition of long PEG chains, leading to the formation of the brush polymer. The biotin, through interactions with streptavidin, can be used to recruit different compounds, such as an antibody. Adapted from Ref. 83.

Another intensive research area concerns the targeting of lectins, *i.e.* protein receptors able to specifically bind glycans and involved in the regulation of many biological processes. Multivalency, *i.e.* the presence of multiple interaction sites between a ligand and a receptor, was shown to be advantageous in lectin binding. Therefore, using

polymer backbones able to carry multiple copies of a glycan ligand is thought to be a promising approach to interact with lectin receptors. However, traditional disperse glycopolymers may lack selectivity in the binding due to their heterogeneous nature, preventing the specific targeting of one lectin over another. For this reason, precision glycopolymers with an absolute control of sequence and stereochemistry were designed and tested against eight kinds of lectins.[84] Interestingly, it was shown that the equilibrium association constant (K_a) for a given receptor varies by one order of magnitude between two stereoisomers. Glycopolymers containing alternating stereoisomers were, in most cases, more efficient than their full R or full S isotactic counterparts. This behavior was attributed to their ability to sample more diverse conformational landscapes, as suggested by MD simulations. An important contribution of this work is the demonstration that lectin binding is strongly influenced by small conformational and stereochemical modifications of the glycopolymer carrier. Such fine structure – function relationships could only be accessed through SDMs, devoid of the heterogeneity of traditional samples. Another research group attempted a similar work, playing with tacticity, but without an absolute control on the length of the produced glycopolymers.[85] While an effect of stereochemistry was again observed on the binding to lectins, the lack of absolute sequence control, leading to differences in DP between the stereoisomers, made the comparison more difficult. In general, the longer chains were more efficient, as expected due to their higher multivalency. Another group aimed to target galectin-3, a particular lectin recognizing β-galactosides.[86] To this end, they synthesized SDMs functionalized with several copies of a sugar, ensuring multivalency, but also with nonglycosidic moieties. In particular, the incorporation of aromatic motifs between the glycans was shown to improve the binding to the targeted receptor, galectin-3, while decreasing the affinity towards a similar receptor, galectin-1. It shows that site-specific modifications of SDMs constitute an interesting tool to modulate the selectivity of interactions.

Sequence control was also exploited to design synthetic mimics of antibodies, dedicated to peptide recognition.[87] Natural antibodies are protein complexes presenting specific recognition sites towards antigens, *i.e.* any foreign body identified as harmful for the organism, and are able to trigger their elimination after binding. Here, poly(*N*-isopropyl acrylamide) nanoparticles were functionalized with sequence-defined oligomers designed to recognize melittin, a peptide found in bee venom, and, upon binding, inhibit its hemolytic activity. Satisfyingly, the nanoparticles decorated with the SDM showed much stronger affinity for melittin than nanoparticles functionalized with randomly incorporated monomers or truncated oligomers. Furthermore, slight modifications in the AA sequence of the melittin led to a significant decrease in the binding of the nanoparticles, demonstrating that sequence

complementarity improves the recognition process. Another strategy to stimulate the immune system consists of using SDMs to recruit antibodies.[88] Recently, sequence-defined heptamers functionalized with three copies of a dinitrophenyl (DNP) ligand, able to interact with anti-DNP antibodies, were designed. Three SDMs were compared, by varying the number of spacing units between the DNPs (zero, one or two). Initially, the addition of spacing units was envisioned as a way to promote multivalent binding, by making the ligands accessible to several antibodies. Counter-intuitively, the most efficient antibody-recruiting molecule was the one without spacers between the DNPs. MD simulations revealed that the three molecules formed similar globular folded structures, and that the addition of spacers led to the burying of the DNPs inside the core, making them less accessible. This example shows again the importance of sequence control, and the complexity of designing efficient systems following chemical intuition. MD simulations were of utmost importance to understand and rationalize the experimental behavior. We also studied biomolecular SDMs in our thesis, exploiting sequence-specific interactions for biorecognition (see **Chapter IV-A and IV-B**).

II.3.3. Synthetic SCMs for catalysis

Artificial enzymes have already been mentioned in this thesis, notably through the incorporation of site-specific modifications to protein backbones or with peptidomimetic systems. Synthetic SCMs are promising in the field of catalysis, as controlled folding and well-organized 3D structures could be attained through the control of sequence, as observed for natural enzymes, with the versatility of polymer chemistry.

Impressive sequence effects were exhibited by trifunctional oligomers dedicated to the aerobic oxidation of alcohols.[89] The chains carry a TEMPO unit, an imidazole, and a copper complex: they constitute a catalytic triad of interest, where all three functional groups must be spatially close for an efficient catalysis.[90] Therefore, their incorporation onto the same scaffold should increase their probability of encounter compared to free monomers in solution. Two trimers were synthesized, differing only by the position of two monomers in their primary structure. They were densely grafted on a surface, to promote cooperative interchain interactions and reduce folding and conformational flexibility. The best oligomer displayed a turnover frequency (TOF) five times higher than the other, a remarkable difference given the very small sequence modification. The effect of sequence was markedly less important for the oligomers diluted in solution, where the folding probably blurred the role of primary structure. Following this work, MD simulations and network representations were applied on very similar catalytic trimers to rationalize their activity **(Figure II.9)**.[91] The results

indicated that all trimers adopted similar globular yet very flexible conformations in acetonitrile, regardless of the sequence. However, network representations helped to rationalize the measured catalytic activities by revealing a higher and more efficient intrachain connectivity for the most efficient system. In contrast, in the less active catalyst, the interactions between the functional groups were hindered by non-catalytic units, typically backbone atoms. Impressively, despite the flexibility of the chains, the influence of the sequence was still apparent in the intrachain connectivity patterns.

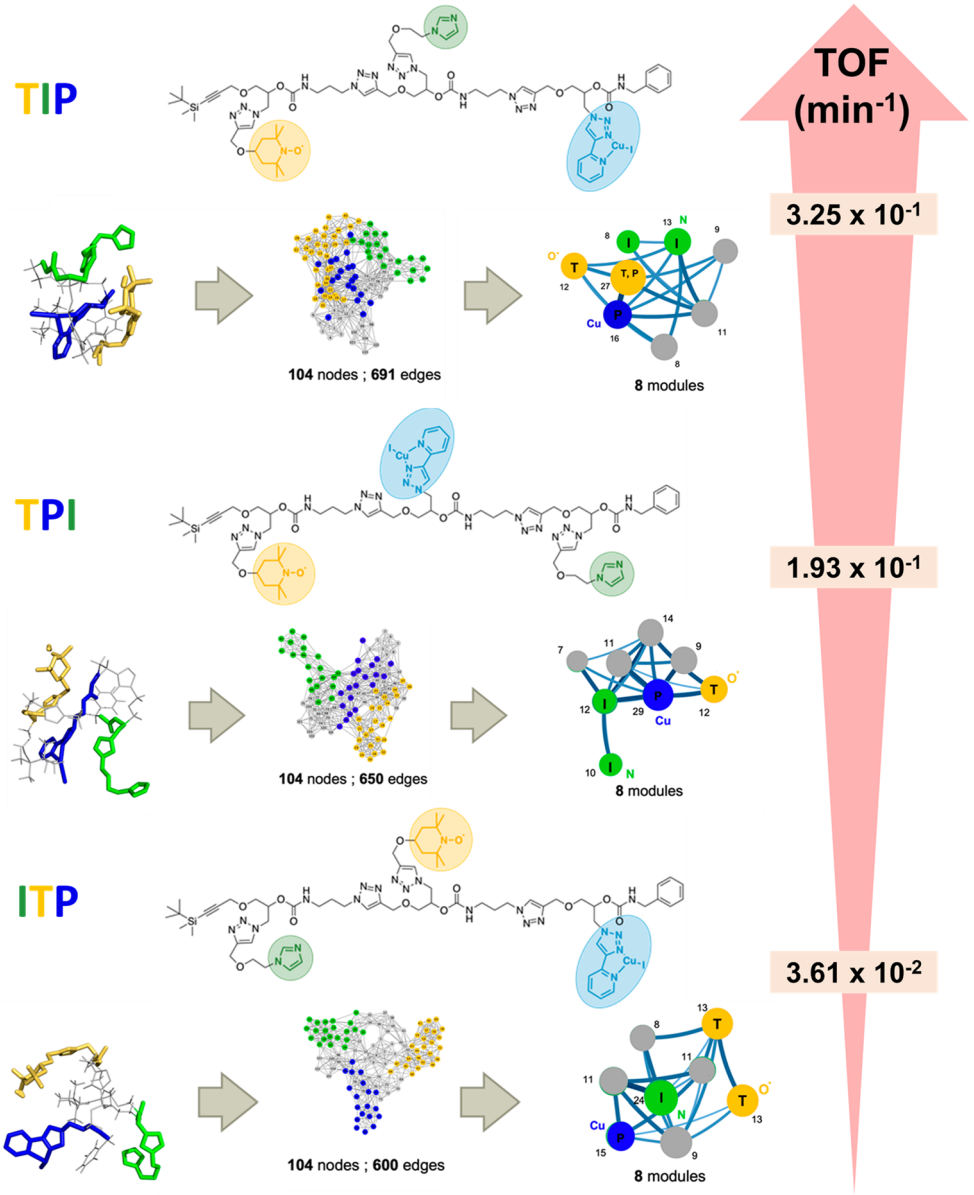


Figure II.9. Chemical structures, final MD snapshots, network representations and modularizations of the investigated catalytic trimers. The sequence of the chains is given by the order of the letters T (TEMPO), I (Imidazole) and P (Copper complex). The number of edges in the network representations, related to intrachain connectivity, follows the same trend than the experimental catalytic activities. The TOFs are given for a catalyst concentration of 5 mol % (relative to substrate concentration). Adapted from Ref. 91 and reproduced from Ref. 81.

In another work, a catalytic trimer was shown to catalyze the elongation of a polymer bearing a complementary substituent.[92] Here, two monomers bearing complementary recognition units – named "A" and "D", for "acceptor" and "donor", respectively – were mixed with a linker moiety, following a combinatorial approach. The monomers can reversibly oligomerize through the formation of dynamic covalent imine bonds. Researchers thought that, by adding a trimer of sequence "AAA" to the mixture, the complementary trimer "DDD" would preferentially form. Instead, they found that "AAA" catalyzed the polymerization of pure "D_n" oligomers. Interestingly, a dimer of sequence "AA" did not show any catalytic activity. It seems that the trimer was able to bind to the extremity of a "D_n" growing chain through complementary A-D interactions, and that this binding facilitated further oligomerization. This discovery shows that the design of SCMs bearing recognition units at precise positions allows the emergence of remarkable catalytic effects.

All these works highlight the interest of precisely engineered polymer chains for catalytic applications, especially in the case of multifunctional catalysts, where important sequence effects were demonstrated. Computational approaches able to predict the folding of synthetic SCMs are particularly needed for this kind of applications, where the 3D structure strongly impacts the efficiency of the system. Such an approach was undertaken recently on polyurethanes, where the design of the chain was optimized through MD simulations before synthesizing the most promising sequence.[93] Catalysis is also explored in our thesis, with the formation of a sequence-defined supramolecular duplex (see **Chapter V**). We also used MD simulations to predict the single-chain folding of different polymers, an important step towards the design of efficient catalytic systems (see **Chapter VI**).

II.3.4. Synthetic SDMs for information storage

Information storage is a topic of intense research. Nowadays, most of the data generated is stored digitally. "Information" can be viewed as a simple binary sequence of "0" and "1", which can represent any kind of data: audio files, images, text, etc. The amount of data produced every day is estimated to be around 10^{11} gigabytes, a number that increases at an uncontrollable pace. The current devices used to store data, such as hard disk drives (HDDs) and solid-state drives (SSDs), may become insufficient in terms of storage density. Additionally, their stability over time and energy consumption constitute other improvable factors. Pursuing an ideal of more stable and compact storage devices, SDMs have emerged as a pertinent alternative. Indeed, the "0" and "1" of a binary code can be represented by two monomers in a primary structure. DNA, for instance, stores all the genetic information in its sequence of nucleotides and is considered as a viable platform, especially for long-term storage

applications.[94] The storage density of DNA is considerably higher than current methods, with a maximum of 2 bits per nucleotide. Theoretically, all the information produced in the world in one year could be stored in some grams of DNA. Logically, synthetic SDMs are also envisioned for information storage. This is the most obvious and probably simpler application, as there is no problematic of controlled folding or precise engineering of interactions with complex binding sites, in complex environments. Here, everything is related to the primary structure itself, which implies that the sequence must be perfectly controlled. New methods to sequence SDMs, *i.e.* to decode their primary structure, need to be developed. Currently, tandem mass spectrometry (MS/MS) is the most commonly used technique. Briefly, the idea is to break the polymer chain into a series of fragments, which are subsequently put in order based on their fragmentation patterns, to reconstruct the whole sequence. The advantage of synthetic polymers envisioned for data storage is that their chemical structure can be optimized to contain predictable fragmentation sites, facilitating the readout.[95] For example, researchers developed an algorithm to automatically sequence oligo(amide-urethane)s from their MS/MS spectra.[96] To simplify the readout, both extremities of the chain are decorated with a different moiety, allowing a software to easily understand the sense of reading. The efficiency of the algorithm was demonstrated by its ability to decode a sentence written with several oligomers **(Figure II.10)**. In the same work, another software was used to write and read a QR-code. A QR-code can be seen as a binary sequence of "0" and "1", and was converted into a series of sequence-defined oligomers. After synthesis of the library of oligomers, the software was able to re-convert them into a binary sequence, thus to rebuild the QR-code. Here, an advantage of synthetic SDMs is the possibility to use a broad range of monomers, which enables a dense storage capacity despite limited chain length.

Following the idea of maximizing storage density, "dual" SDMs, storing information not only in the side-chains, but also in the backbone, were designed.[97] This method significantly increases the storage capacity, as a "dual" pentamer contains nearly as much information than a decamer without information in its backbone. Another group followed the opposite approach and managed to synthesize very long SDMs, up to 256 units with satisfying yield, incorporating only two different co-monomers.[98] Impressively, these systems possess a density of information storage 50 % higher than that of DNA.

Information-containing macromolecules can also be used for cryptographic applications. One group designed "molecular keys" using SDMs.[99] In this practical example, a molecule was adsorbed onto paper, here an envelope, containing a coded message. The molecule acts as a password; using MS/MS, the specific sequence of the molecule could be deciphered and converted into digital information, enabling

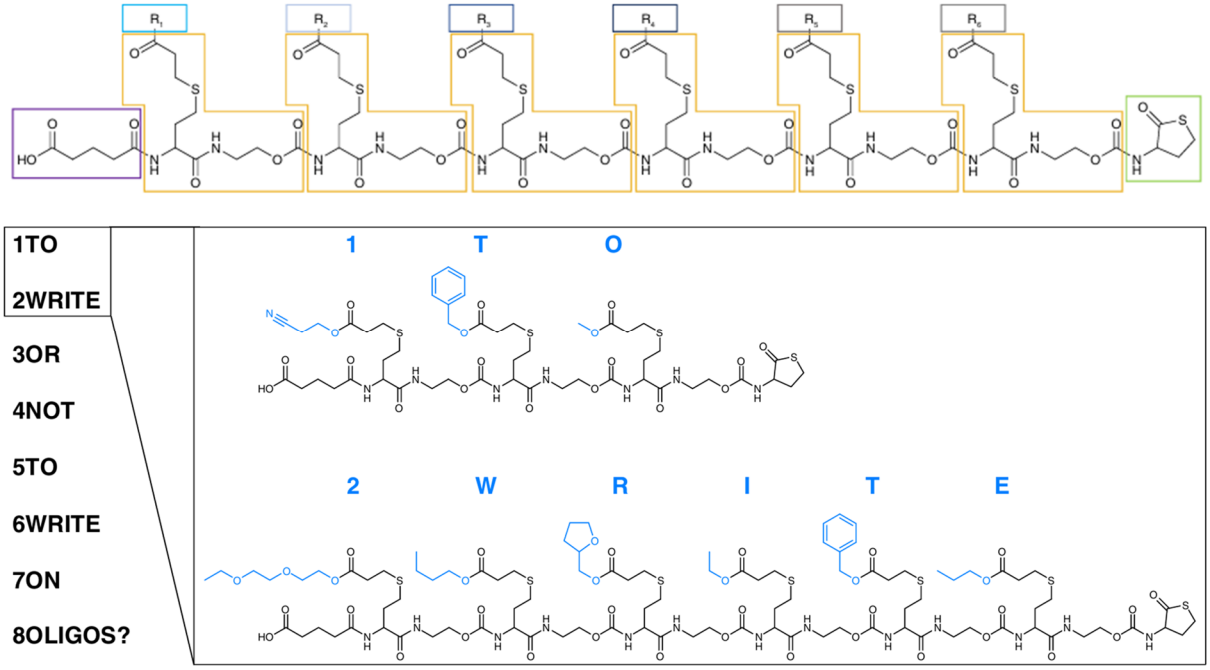


Figure II.10. General chemical structure of SDMs dedicated to information storage and practical example with the writing of a sentence. Each monomer is decorated with a functional group, and each functional group is associated to a letter. The oligomers can be sequenced by MS/MS, showing that it is possible to store and extract information using SDMs. Adapted from Ref. 96.

decryption of the message. This proof-of-concept illustrates the potential of SDMs in anti-counterfeiting applications.

SDMs for information storage applications are not yet a completely mature field and will particularly benefit from improvements in the synthesis pathways, especially in terms of increased DP and speed of the reading/writing processes. In this respect, a fully automated protocol has recently been proposed to synthesize and sequence oligourethanes.[100] It seems likely that precision macromolecules will find practical applications in the medium term, notably for long-term data storage, where their very high storage density should be a highly valuable advantage.

II.4. Conclusion

Less than 200 years ago, practically nothing was known about proteins or nucleic acids – there were not even words to describe them. Nowadays, researchers are not only able to build tailor-made biopolymers, fulfilling the dream of artificial enzymes mentioned by Fischer in 1902, but also to functionalize them with many unnatural substituents. This is due in large part to the knowledge gained on sequence – structure relationships. By “simply” controlling the order in which the monomers are inserted into the chains, complex 3D structures and supramolecular assemblies were designed. Since about 15

years ago, advances in polymer synthesis have enabled the emergence of a new field of research, dedicated to purely synthetic SCMs. The exquisite control of sequence displayed by functional biomacromolecules, combined with the chemical diversity offered by polymer chemistry, is seen as a way to design novel highly performant nanomaterials. The field is still very young and SCMs, while they carry a lot of promises, have to go beyond proofs-of-concept and to demonstrate their suitability for practical applications. However, it is now clear that playing with the order of monomers impacts the properties of the chains, sometimes even within flexible systems. Information storage will probably be the first area to benefit from SCMs, as it does not directly depend on a fine 3D organization or the establishment of specific interactions. For applications having these requirements, such as catalysis or biomolecular recognition, a more fundamental understanding of sequence – structure – function relationships is required. Computational approaches such as molecular modeling will be essential to reach this goal. This is the approach followed in this thesis, and the details of our methodology are explained in the next chapter. In parallel to this fundamental understanding, we can hope that expanding the database of known sequence – structure pairs will lead to the development of predictive ML models, which could help to rationalize the design of SCMs, despite the immensity of the chemical space that has been opened.

References

[1] H. Hartley. Origin of the Word 'Protein'. *Nature* **1951**, *168*, 244–244.

[2] H. B. Vickery, T. B. Osborne. A review of hypotheses of the structure of proteins. *Physiol. Rev.* **1928**, *8*, 393–446.

[3] R. M. Fuoss. Polyelectrolytes. *Science (1979)* **1948**, *108*, 545–550.

[4] F. Sanger, E. O. P. Thompson, R. Kitai. The amide groups of insulin. *Biochemical Journal* **1955**, *59*, 509–518.

[5] W. T. Astbury, A. Street. X-ray studies of the structure of hair, wool, and related fibres.- I. General. *Philos. Trans. R. Soc., A* **1931**, *230*, 75–101.

[6] L. Pauling, R. B. Corey, H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205–211.

[7] L. Pauling, R. B. Corey. Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 235–240.

[8] L. Pauling, R. B. Corey. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 729–740.

[9] G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95–99.

[10] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, D. C. Phillips. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181*, 662–666.

[11] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, V. C. Shore. Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. *Nature* **1960**, *185*, 422–427.

[12] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, A. C. T. North. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* **1960**, *185*, 416–422.

[13] J. Zahradník, L. Kolářová, Y. Peleg, P. Kolenko, S. Svidenská, T. Charnavets, T. Unger, J. L. Sussman, B. Schneider. Flexible regions govern promiscuous binding of IL-24 to receptors IL-20R1 and IL-22R1. *FEBS J.* **2019**, *286*, 3858–3873.

[14] G. B. Irvine, O. M. El-Agnaf, G. M. Shankar, D. M. Walsh. Protein Aggregation in the Brain: The Molecular Basis for Alzheimer's and Parkinson's Diseases. *Mol. Med.* **2008**, *14*, 451–464.

[15] F. U. Hartl, A. Bracher, M. Hayer-Hartl. Molecular chaperones in protein folding and proteostasis. *Nature* **2011**, *475*, 324–332.

[16] R. Zwanzig, A. Szabo, B. Bagchi. Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20–22.

[17] V. E. Bychkova, G. V. Semisotnov, V. A. Balobanov, A. V. Finkelstein. The Molten Globule Concept: 45 Years Later. *Biochemistry (Moscow)* **2018**, *83*, S33–S47.

[18] P. Malhotra, J. B. Udgaonkar. How cooperative are protein folding and unfolding transitions?. *Protein Sci.* **2016**, *25*, 1924–1941.

[19] S. W. Englander, L. Mayne. The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 15873–15880.

[20] S. Bhatia, J. B. Udgaonkar. Understanding the heterogeneity intrinsic to protein folding. *Curr. Opin. Struct. Biol.* **2024**, *84*, 102738.

[21] C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science* **1973**, *181*, 223–230.

[22] U. Arnold, M. P. Hinderaker, J. Köditz, R. Golbik, R. Ulbrich-Hofmann, R. T. Raines. Protein Prosthesis: A Nonnatural Residue Accelerates Folding and Increases Stability. *J. Am. Chem. Soc.* **2003**, *125*, 7500–7501.

[23] M. F. AlAjmi, S. Khan, A. Choudhury, T. Mohammad, S. Noor, A. Hussain, W. Lu, M. S. Eapen, V. Chimankar, P. M. Hansbro, S. S. Sohal, A. M. Elasbali, Md. I. Hassan. Impact of Deleterious Mutations on Structure, Function and Stability of Serum/Glucocorticoid Regulated Kinase 1: A Gene to Diseases Correlation. *Front. Mol. Biosci.* **2021**, *8*, 780284.

[24] L. Groignet, D. Dellemme, Q. Duez, A. Fizazi, J.-M. Colet, P. Brocorens, M. Surin, P. Gerbaux, J. De Winter. Impact of Post-Translational Succination on Small Ubiquitin-Like Modifier 1

Structure: A Dual Approach Combining Gas Phase and Solution Studies. *ACS Pharmacol. Transl. Sci.* **2025**, 8, 2683–2693.

[25] H. M. Berman. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[26] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[27] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.

[28] L. Hedstrom. "*Enzyme Specificity and Selectivity*" in *Encyclopedia of Life Sciences*, Wiley, **2010**.

[29] T. Shen, Z. Hu, S. Sun, D. Liu, F. Wong, J. Wang, J. Chen, Y. Wang, L. Hong, J. Xiao, L. Zheng, T. Krishnamoorthi, I. King, S. Wang, P. Yin, J. J. Collins, Y. Li. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods* **2024**, *21*, 2287–2298.

[30] C. Bergonzo, A. Grishaev. Critical Assessment of RNA and DNA Structure Predictions via Artificial Intelligence: The Imitation Game. *J. Chem. Inf. Model.* **2025**, *65*, 3544–3554.

[31] L. R. Ganser, M. L. Kelly, D. Herschlag, H. M. Al-Hashimi. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 474–489.

[32] J. D. Watson, F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171*, 737–738.

[33] S. Kosuri, G. M. Church. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **2014**, *11*, 499–507.

[34] B. L. Nilsson, M. B. Soellner, R. T. Raines. Chemical Synthesis of Proteins. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 91–118.

[35] R. B. Merrifield. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *J. Am. Chem. Soc.* **1963**, *85*, 2149–2154.

[36] S. L. Beaucage, M. H. Caruthers. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **1981**, *22*, 1859–1862.

[37] X. Gao, J. P. Pellois, Y. Na, Y. Kim, E. Gulari, X. Zhou. High density peptide microarrays. In situ synthesis and applications. *Mol. Diversity* **2004**, *8*, 177–187.

[38] N. Tang, S. Ma, J. Tian. "*New Tools for Cost-Effective DNA Synthesis*", Elsevier, **2013**, pp. 3–21.

[39] D. S. Mattes, N. Jung, L. K. Weber, S. Bräse, F. Breitling. Miniaturized and Automated Synthesis of Biomolecules—Overview and Perspectives. *Adv. Mater.* **2019**, *31*, 1806656.

[40] J. Li, S. Zhao, G. Yang, R. Liu, W. Xiao, P. Disano, K. S. Lam, T. Pan. Combinatorial Peptide Microarray Synthesis Based on Microfluidic Impact Printing. *ACS Comb. Sci.* **2019**, *21*, 6–10.

[41] M. Yu, X. Tang, Z. Li, W. Wang, S. Wang, M. Li, Q. Yu, S. Xie, X. Zuo, C. Chen. High-throughput DNA synthesis for data storage. *Chem. Soc. Rev.* **2024**, *53*, 4463–4489.

[42] D. S. Kong, P. A. Carr, L. Chen, S. Zhang, J. M. Jacobson. Parallel gene synthesis in a microfluidic device. *Nucleic Acids Res.* **2007**, *35*, e61.

[43] C.-C. Lee, T. M. Snyder, S. R. Quake. A microfluidic oligonucleotide synthesizer. *Nucleic Acids Res.* **2010**, *38*, 2514–2521.

[44] N. Hartrampf, A. Saebi, M. Poskus, Z. P. Gates, A. J. Callahan, A. E. Cowfer, S. Hanna, S. Antilla, C. K. Schissel, A. J. Quartararo, X. Ye, A. J. Mijalis, M. D. Simon, A. Loas, S. Liu, C. Jessen, T. E. Nielsen, B. L. Pentelute. Synthesis of proteins by automated flow chemistry. *Science* **2020**, *368*, 980–987.

[45] A. Charalampidou, T. Nehls, C. Meyners, S. Gandhesiri, S. Pomplun, B. L. Pentelute, F. Lermyte, F. Hausch. Automated Flow Peptide Synthesis Enables Engineering of Proteins with Stabilized Transient Binding Pockets. *ACS Cent. Sci.* **2024**, *10*, 649–657.

[46] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **2012**, *337*, 816–821.

[47] M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, , E3501-E3508.

[48] I. Sarac, M. Hollenstein. Terminal Deoxynucleotidyl Transferase in the Synthesis and Modification of Nucleic Acids. *ChemBioChem* **2019**, *20*, 860–871.

[49] J. Ashley, I. M. Potts, F. J. Olorunniji. Applications of Terminal Deoxynucleotidyl Transferase Enzyme in Biotechnology. *ChemBioChem* **2023**, *24*, e202200510.

[50] A. Adhikari, B. R. Bhattarai, A. Aryal, N. Thapa, P. KC, A. Adhikari, S. Maharjan, P. B. Chanda, B. P. Regmi, N. Parajuli. Reprogramming natural proteins using unnatural amino acids. *RSC Adv.* **2021**, *11*, 38126–38145.

[51] S. Wagner, B. Sudhamalla, P. Mannes, S. Sappa, S. Kavoosi, D. Dey, S. Wang, K. Islam. Engineering bromodomains with a photoactive amino acid by engaging 'Privileged' tRNA synthetases. *Chem. Commun.* **2020**, *56*, 3641–3644.

[52] P. N. Pham, J. Zahradník, L. Kolářová, B. Schneider, G. Fuertes. Regulation of IL-24/IL-20R2 complex formation using photocaged tyrosines and UV light. *Front. Mol. Biosci.* **2023**, *10*, 1214235.

[53] N. Ahmed, N. Ahmed, D. A. Bilodeau, J. P. Pezacki. An unnatural enzyme with endonuclease activity towards small non-coding RNAs. *Nat. Commun.* **2023**, *14*, 3777.

[54] H. Huang, T. Yan, C. Liu, Y. Lu, Z. Wu, X. Wang, J. Wang. Genetically encoded Nδ-vinyl histidine for the evolution of enzyme catalytic center. *Nat. Commun.* **2024**, *15*, 5714.

[55] H. Zhang, Z. Zheng, L. Dong, N. Shi, Y. Yang, H. Chen, Y. Shen, Q. Xia. Rational incorporation of any unnatural amino acid into proteins by machine learning on existing experimental proofs. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4930–4941.

[56] A. J. Lander, Y. Jin, L. Y. P. Luk. D-Peptide and D-Protein Technology: Recent Advances, Challenges, and Opportunities. *ChemBioChem* **2023**, *24*, e202200537.

[57] T. N. M. Schumacher, L. M. Mayr, D. L. Minor, M. A. Milhollen, M. W. Burgess, P. S. Kim. Identification of D-Peptide Ligands Through Mirror-Image Phage Display. *Science* **1996**, *271*, 1854–1857.

[58] C. Díaz-Perlas, M. Varese, S. Guardiola, M. Sánchez-Navarro, J. García, M. Teixidó, E. Giralt. Protein Chemical Synthesis Combined with Mirror-Image Phage Display Yields D-Peptide EGF Ligands that Block the EGF–EGFR Interaction. *ChemBioChem* **2019**, *20*, 2079–2084.

[59] A. J. Callahan, S. Gandhesiri, T. L. Travaline, R. M. Reja, L. Lozano Salazar, S. Hanna, Y.-C. Lee, K. Li, O. S. Tokareva, J.-M. Swiecicki, A. Loas, G. L. Verdine, J. H. McGee, B. L. Pentelute. Mirror-image ligand discovery enabled by single-shot fast-flow synthesis of D-proteins. *Nat. Commun.* **2024**, *15*, 1813.

[60] R. K. Spencer, G. L. Butterfoss, J. R. Edison, J. R. Eastwood, S. Whitelam, K. Kirshenbaum, R. N. Zuckermann. Stereochemistry of polypeptoid chain configurations. *Biopolymers* **2019**, *110*, e23266.

[61] S. Jiao, D. M. Rivera Mirabal, A. J. DeStefano, R. A. Segalman, S. Han, M. S. Shell. Sequence Modulates Polypeptoid Hydration Water Structure and Dynamics. *Biomacromolecules* **2022**, *23*, 1745–1756.

[62] A. J. DeStefano, S. D. Mengel, M. W. Bates, S. Jiao, M. S. Shell, S. Han, R. A. Segalman. Control over Conformational Landscapes of Polypeptoids by Monomer Sequence Patterning. *Macromolecules* **2024**, *57*, 1469–1477.

[63] R. Zheng, M. Zhao, J. S. Du, T. R. Sudarshan, Y. Zhou, A. K. Paravastu, J. J. De Yoreo, A. L. Ferguson, C.-L. Chen. Assembly of short amphiphilic peptoids into nanohelices with controllable supramolecular chirality. *Nat. Commun.* **2024**, *15*, 3264.

[64] B. Tassignon, Z. Wang, A. Galanti, J. De Winter, P. Samorì, J. Cornil, K. Moth-Poulsen, P. Gerbaux. Site Selectivity of Peptoids as Azobenzene Scaffold for Molecular Solar Thermal Energy Storage. *Chem. Eur. J.* **2023**, *29*, e202303168.

[65] G. Maayan, M. D. Ward, K. Kirshenbaum. Folded biomimetic oligomers for enantioselective catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 13679–13684.

[66] P. W. K. Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature* **2006**, *440*, 297–302.

[67] C. Rossi-Gendron, F. El Fakih, L. Bourdon, K. Nakazawa, J. Finkel, N. Triomphe, L. Chocron, M. Endo, H. Sugiyama, G. Bellot, M. Morel, S. Rudiuk, D. Baigl. Isothermal self-assembly of multicomponent and evolutive DNA nanostructures. *Nat. Nanotechnol.* **2023**, *18*, 1311–1318.

[68] J. M. Weck, A. Heuer-Jungemann. Fully addressable designer superstructures assembled from one single modular DNA origami. *Nat. Commun.* **2025**, *16*, 1556.

[69] W. Wang, A. Mohammed, R. Chen, A. Elonen, S. Chen, M. Tian, J. Yang, Y. Xiang, P. Orponen, B. Wei. Automated design of scaffold-free DNA wireframe nanostructures. *Nat. Commun.* **2025**, *16*, 4666.

[70] P. Zhan, A. Peil, Q. Jiang, D. Wang, S. Mousavi, Q. Xiong, Q. Shen, Y. Shang, B. Ding, C. Lin, Y. Ke, N. Liu. Recent Advances in DNA Origami-Engineered Nanomaterials and Applications. *Chem. Rev.* **2023**, *123*, 3976–4050.

[71] M. Liu, J. Fu, C. Hejesen, Y. Yang, N. W. Woodbury, K. Gothelf, Y. Liu, H. Yan. A DNA tweezer-actuated enzyme nanoreactor. *Nat. Commun.* **2013**, *4*, 2127.

[72] A. Berdis. Nucleobase-modified nucleosides and nucleotides: Applications in biochemistry, synthetic biology, and drug discovery. *Front. Chem.* **2022**, *10*, 1051525.

[73] Y. Chen, K. Morihiro, Y. Nemoto, A. Ichimura, R. Ueki, S. Sando, A. Okamoto. Selective inhibition of cancer cell migration using a pH-responsive nucleobase-modified DNA aptamer. *Chem. Sci.* **2024**, *15*, 17097–17102.

[74] F. Pellestor, P. Paulasova. The peptide nucleic acids (PNAs), powerful tools for molecular genetics and cytogenetics. *Eur. J. Hum. Genet.* **2004**, *12*, 694–700.

[75] J. Lutz. Defining the Field of Sequence-Controlled Polymers. *Macromol. Rapid Commun.* **2017**, *38*, 1700582.

[76] J. De Neve, J. J. Haven, L. Maes, T. Junkers. Sequence-definition from controlled polymerization: the next generation of materials. *Polym. Chem.* **2018**, *9*, 4692–4705.

[77] M. Nerantzaki, J. Lutz. Multistep Growth 'Polymerizations'. *Macromol. Chem. Phys.* **2022**, *223*, 2100368.

[78] J.-F. Lutz. The future of sequence-defined polymers. *Eur. Polym. J.* **2023**, *199*, 112465.

[79] P. Nanjan, M. Porel. Sequence-defined non-natural polymers: synthesis and applications. *Polym. Chem.* **2019**, *10*, 5406–5424.

[80] S. C. Solleder, R. V. Schneider, K. S. Wetzel, A. C. Boukis, M. A. R. Meier. Recent Progress in the Design of Monodisperse, Sequence-Defined Macromolecules. *Macromol. Rapid Commun.* **2017**, *38*, 1600711.

[81] D. Dellemme, S. Kardas, C. Tonneaux, J. Lernould, M. Fossepre, M. Surin. From sequence definition to structure-property relationships in discrete synthetic macromolecules: insights from molecular modeling. *Angew. Chem. Int. Ed.* **2025**, *64*, e202420179.

[82] R. Szweda. Sequence- and stereo-defined macromolecules: Properties and emerging functionalities. *Prog. Polym. Sci.* **2023**, *145*, 101737.

[83] C. Chen, K. Wunderlich, D. Mukherji, K. Koynov, A. J. Heck, M. Raabe, M. Barz, G. Fytas, K. Kremer, D. Y. W. Ng, T. Weil. Precision Anisotropic Brush Polymers by Sequence Controlled Chemistry. *J. Am. Chem. Soc.* **2020**, *142*, 1332–1340.

[84] M. Hartweg, Y. Jiang, G. Yilmaz, C. M. Jarvis, H. V.-T. Nguyen, G. A. Primo, A. Monaco, V. P. Beyer, K. K. Chen, S. Mohapatra, S. Axelrod, R. Gómez-Bombarelli, L. L. Kiessling, C. R. Becer, J. A. Johnson. Synthetic Glycomacromolecules of Defined Valency, Absolute Configuration, and Topology Distinguish between Human Lectins. *JACS Au* **2021**, *1*, 1621–1630.

[85] J. Becker, R. Terracciano, G. Yilmaz, R. Napier, C. R. Becer. Step-Growth Glycopolymers with a Defined Tacticity for Selective Carbohydrate–Lectin Recognition. *Biomacromolecules* **2023**, *24*, 1924–1933.

[86] T. Freichel, V. Heine, D. Laaf, E. E. Mackintosh, S. Sarafova, L. Elling, N. L. Snyder, L. Hartmann. Sequence-Defined Heteromultivalent Precision Glycomacromolecules Bearing Sulfonated/Sulfated Nonglycosidic Moieties Preferentially Bind Galectin-3 and Delay Wound Healing of a Galectin-3 Positive Tumor Cell Line in an In Vitro Wound Scratch Assay. *Macromol. Biosci.* **2020**, *20*, 2000163.

[87] Y. Saito, R. Honda, S. Akashi, H. Takimoto, M. Nagao, Y. Miura, Y. Hoshino. Polymer Nanoparticles with Uniform Monomer Sequences for Sequence-Specific Peptide Recognition. *Angew. Chem. Int. Ed.* **2022**, *61*, e202206456.

[88] R. Aksakal, C. Tonneaux, A. Uvyn, M. Fossépré, H. Turgut, N. Badi, M. Surin, B. G. De Geest, Filip. E. Du Prez. Sequence-defined antibody-recruiting macromolecules. *Chem. Sci.* **2023**, *14*, 6572–6578.

[89] P. Chandra, A. M. Jonas, A. E. Fernandes. Sequence and Surface Confinement Direct Cooperativity in Catalytic Precision Oligomers. *J. Am. Chem. Soc.* **2018**, *140*, 5179–5184.

[90] J. M. Hoover, S. S. Stahl. Highly Practical Copper(I)/TEMPO Catalyst System for Chemoselective Aerobic Oxidation of Primary Alcohols. *J. Am. Chem. Soc.* **2011**, *133*, 16901–16910.

[91] J. Li, Q. Qin, S. Kardas, M. Fossépré, M. Surin, A. E. Fernandes, K. Glinel, A. M. Jonas. Sequence Rules the Functional Connections and Efficiency of Catalytic Precision Oligomers. *ACS Catal.* **2022**, *12*, 2126–2131.

[92] L. Gabrielli, C. A. Hunter. Supramolecular catalysis by recognition-encoded oligomers: discovery of a synthetic imine polymerase. *Chem. Sci.* **2020**, *11*, 7408–7414.

[93] S. Samokhvalova, J. Lutz, I. Coluzza. Precision Design of Sequence-Defined Polyurethanes: Exploring Controlled Folding Through Computational Design. *Macromol. Chem. Phys.* **2024**, *225*, 2400223.

[94] M. G. T. A. Rutten, F. W. Vaandrager, J. A. A. W. Elemans, R. J. M. Nolte. Encoding information into polymers. *Nat. Rev. Chem.* **2018**, *2*, 365–381.

[95] K. Launay, J. Amalian, E. Laurent, L. Oswald, A. Al Ouahabi, A. Burel, F. Dufour, C. Carapito, J. Clément, J. Lutz, L. Charles, D. Gigmes. Precise Alkoxyamine Design to Enable Automated Tandem Mass Spectrometry Sequencing of Digital Poly(phosphodiester)s. *Angew. Chem. Int. Ed.* **2021**, *60*, 917–926.

[96] S. Martens, A. Landuyt, P. Espeel, B. Devreese, P. Dawyndt, F. Du Prez. Multifunctional sequence-defined macromolecules for chemical data storage. *Nat. Commun.* **2018**, *9*, 4451.

[97] K. S. Wetzel, M. Frölich, S. C. Solleder, R. Nickisch, P. Treu, M. A. R. Meier. Dual sequence definition increases the data storage capacity of sequence-defined macromolecules. *Commun. Chem.* **2020**, *3*, 63.

[98] J. M. Lee, M. B. Koo, S. W. Lee, H. Lee, J. Kwon, Y. H. Shim, S. Y. Kim, K. T. Kim. High-density information storage in an absolutely defined aperiodic sequence of monodisperse copolyester. *Nat. Commun.* **2020**, *11*, 56.

[99] A. C. Boukis, K. Reiter, M. Frölich, D. Hofheinz, M. A. R. Meier. Multicomponent reactions provide key molecules for secret communication. *Nat. Commun.* **2018**, *9*, 1439.

[100] J. R. Shuluk, C. D. Wight, J. R. Howard, M. E. King, S. R. Moor, R. J. DeHoog, S. D. Dahlhauser, L. S. Eberlin, E. V. Anslyn. A Workflow Enabling the Automated Synthesis, Chain-End Degradation, and Rapid Mass Spectrometry Analysis for Molecular Information Storage in Sequence-Defined Oligourethanes. *JACS Au* **2025**, *5*, 1232–1242.

III. Methodology

This chapter aims at providing the reader with a brief yet comprehensive introduction to the computational methods employed in this thesis. The first section is mainly dedicated to non-experts and beginners in the field, starting with a quick review of the theoretical foundations governing the behavior of atoms and molecules, *i.e.* quantum physics. Then, the principles of molecular mechanics (MM) and molecular dynamics (MD) simulations are concisely explained, focusing on the main concepts and the steps that one has to follow to predict the time evolution of a molecular system **(Section III.1)**. Then, more advanced topics are introduced, including the limitations of MD simulations and methodological advances designed to enhance the speed or accuracy of conventional approaches. Recent literature examples illustrating the use of some of these methods are discussed, showing their successes but also, in some cases, their failures **(Section III.2)**. Finally, several descriptors and tools used throughout this thesis to characterize and analyze molecular conformations are explained **(Section III.3)**. The detailed simulation protocols and parameters will be further developed in their corresponding chapters.

III.1. Basics of Molecular Dynamics Simulations

III.1.1. Fundamentals of quantum chemistry

The ability to model molecular systems through computer simulations resides in the existence of physical models capable of predicting their properties. The most accurate mathematical framework currently available to describe the behavior of the infinitesimally small components of matter, such as atoms and molecules, relies on the laws of quantum mechanics (QM), established about a century ago.[1] All the information about a chemical system in a stationary state, *i.e.* one whose observable properties remain constant over time, can be accessed by solving the time-independent Schrödinger equation **(Equation III.1)**.

$$\hat{H}\psi = E\psi \qquad \textbf{(III.1)}$$

Where $\hat{H}$ is the Hamiltonian operator, Ψ is the wavefunction of the system and E is the corresponding energy eigenvalue.

Unfortunately, solving the Schrödinger equation exactly is impossible for most chemical systems due to its mathematical complexity.[1,2] Analytical solutions are only available for very simple systems, such as hydrogenoid species containing a single electron.[3] The only way to use the Schrödinger equation to get knowledge on chemical systems is to introduce approximations. The Born-Oppenheimer (BO) approximation is probably the most well-known, and consists of decoupling the movements of the

electrons from that of the nuclei.[4] Since electrons are much lighter and thus much faster than the nuclei, the coordinates of the latter can be considered fixed. This leads to the so-called electronic Schrödinger equation. This first step simplifies the problem, but obtaining exact solutions to the electronic Schrödinger equation remains extremely challenging in most cases. To address this and reach approximate solutions, several methods have been developed, such as Hartree-Fock,[5] post-Hartree-Fock,[6] density functional theory,[7] or hybrid approaches.[8] Each of these is based on diverse assumptions, with several levels of approximations. They provide a flexible theoretical framework to study a given system, depending on the properties of interest, the targeted accuracy, and the computational resources at hand.[9] Nonetheless, using the mathematical framework of QM remains computationally demanding and limits its use to relatively small systems, typically ranging between a few atoms to a few thousand.[2]

III.1.2. The simpler framework of molecular mechanics

Chemical systems can also be described using a much simpler approach based on classical mechanics, known as molecular mechanics (MM), in which atoms and molecules are treated as classical particles. This allows the study of much larger systems, with up to several million atoms.[10] Using MM models, molecules are represented as balls connected by sticks, or springs **(Figure III.1)**. Obviously, chemical systems cannot be described as simple hard spheres: each ball, *i.e.* each atom, is characterized by an *atom type* and a *partial charge*. The atom types serve to identify the chemical elements and also account for their bonding environment. For instance, carbon atoms in a carbonyl group and in a phenyl ring will have different atom types. Then, the partial charges describe the electrostatic properties of the system. In MM models, contrary to quantum chemical approaches, the electrons are not explicitly represented. Their effect, and the electrostatic potential that they generate, is implicitly taken into account through point charges, directly located on the nuclei. Therefore, MM does not give access to the electronic properties of materials and does not allow the formation or breaking of covalent bonds.

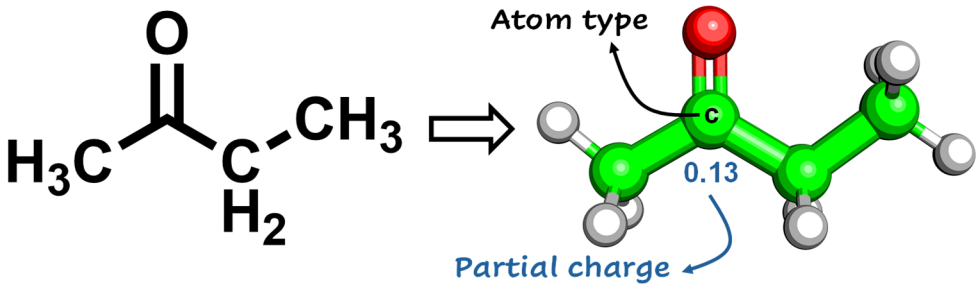


Figure III.1. Chemical structure (left) and "ball and sticks" representation (right) of a molecule. An example of atom type (in black) and partial charge (in blue) is shown for one carbon atom on the 3D representation.

Knowing the atomic coordinates of a system, all the information needed to compute its potential energy and related properties using MM is provided by a *force field* (FF). A force field contains both the functional form and an ensemble of parameters required for these calculations **(Figure III.2)**. The philosophy behind FFs is that the potential energy of the whole system can be decomposed into a sum of independent terms, each described by a specific mathematical expression. For example, bond stretching is usually modelled as a harmonic potential: deviation from the equilibrium bond length will result in an energy penalty, proportional to the square of the deviation (as shown in the red box in **Figure III.2**). To evaluate each term, the atomic coordinates and a set of parameters, which are tabulated in the FF (see k_b and r_{eq} in **Figure III.2**), are required. These parameters generally come from QM calculations or experimental data. Many different FFs exist, each distinguished by its set of atom types, parameters, and the mathematical expressions of the potential energy and its individual terms. Consequently, selecting an appropriate FF depends on the molecular class under study, as different FFs will lead to different levels of performance and accuracy.

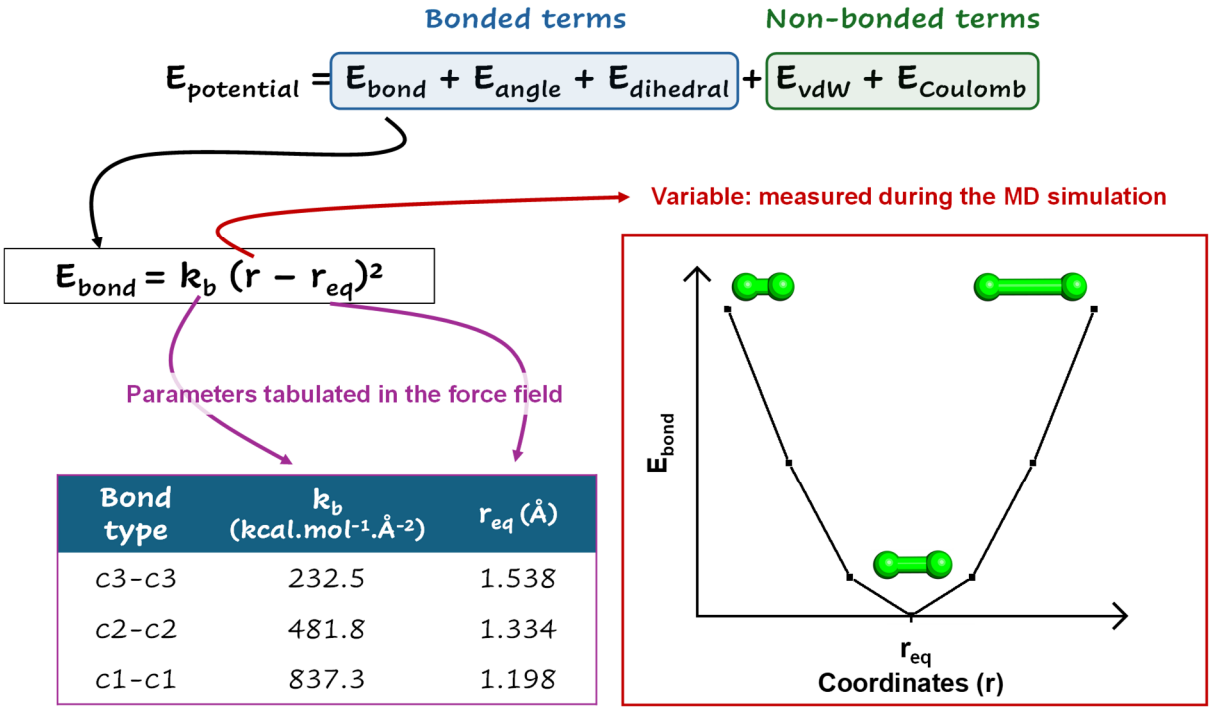


Figure III.2. General expression of the potential energy, decomposed into a sum of terms, as expressed by a force field. A detailed example is provided for the bonding energy, E_{bond}, here described as a harmonic term. Examples of k_b (the force constant) and r_{eq} (the equilibrium bond length), two FF parameters, are shown in the purple box. The terms c3, c2 and c1 refer to the atom types of sp^3, sp^2 and sp^1 carbon atoms, respectively. The variation of the bonding energy with respect to the bond length, r, is displayed in the red box.

III.1.3. The workflow of molecular dynamics simulations

Molecular dynamics (MD) simulations build upon MM models, using force fields to calculate the potential energy and the forces acting on the atoms, in order to predict the temporal evolution of chemical systems. Note that, in reality, MD simulations following the laws of quantum mechanics are possible – they are called *ab initio* MD simulations.[11] In practice, they are limited to very small systems and timescales, and are generally not adapted to the study of (bio)macromolecules.

At the beginning of an MD simulation, only the initial structure of the system is known, *i.e.* a set of atomic coordinates r(0). The purpose of the simulation is to predict the suite of conformations that the system will adopt over time – its *trajectory*. As the real, continuous trajectory cannot be solved analytically, it is approximated by a series of discrete states, separated by a timestep Δt. The movements of the atoms, treated as classical particles, are computed by integrating Newton's equations of motion using a numerical integrator such as the Verlet algorithm.[12] This integrator estimates the next set of coordinates, r(0 + Δt), through a Taylor series expansion around r(0) **(Equation III.2)**.

$$r(0 + \Delta t) = r(0) + v(0)\Delta t + a(0)\frac{\Delta t^2}{2} \qquad \textbf{(III.2)}$$

Where *v(0)* and *a(0)* are the initial sets of velocities and accelerations, respectively.

In this equation, the coordinates r(0) are known. The velocities v(0) are initialized "randomly", following a Maxwell-Boltzmann distribution (thus depending on the temperature of the system) **(Equation III.3)**.

$$f(v) = \left(\frac{m}{2\pi kT}\right)^{1/2} e^{-mv^2/2kT} \qquad \textbf{(III.3)}$$

Where *f(v)* is the probability density function describing the likelihood of finding a particle with a given velocity v, m is the mass, k is the Boltzmann constant, and T is the temperature.

The last unknown part in **Equation III.2** is the acceleration, a(0). From Newton's second law of motion and the fact that a force can be expressed as the derivative of the potential energy with respect to the coordinates, we can link the acceleration a, the forces F and the potential energy E_P **(Equation III.4)**.

$$F = ma = -\frac{dE_P}{dr} \qquad \textbf{(III.4)}$$

Therefore, after computing the potential energy of the system using the force field, the corresponding set of forces – and hence the atomic accelerations – can be determined. Knowing r(0), v(0) and a(0), it is then possible to solve **Equation III.2** and obtain r(0 + Δt). With this new set of coordinates, the accelerations a(0 + Δt) can be computed, as the potential energy only depends on the atomic positions (and the parameters defined in the force field). Finally, **Equation III.5** can be used to compute v(0 + Δt).

$$v(0+\Delta t) = v(0) + [a(0) + a(0+\Delta t)]\frac{\Delta t^2}{2} \qquad \textbf{(III.5)}$$

The iterative process supporting MD simulations is summarized in **Figure III.3**. This procedure, which decomposes the continuous trajectory into a series of discrete states, would only yield the exact coordinates for infinitesimally small timesteps, Δt. However, using longer timesteps is desirable, as it allows to reach a given simulation time with fewer steps, thus at lower computational cost. In practice, Δt is often set to 1 fs (10^{-15} s), about ten times smaller than the timescale of the fastest motions in the system – typically, vibrations involving hydrogen atoms. Nowadays, various methods enable the use of longer timesteps, as will be discussed in **Section III.2.3**.

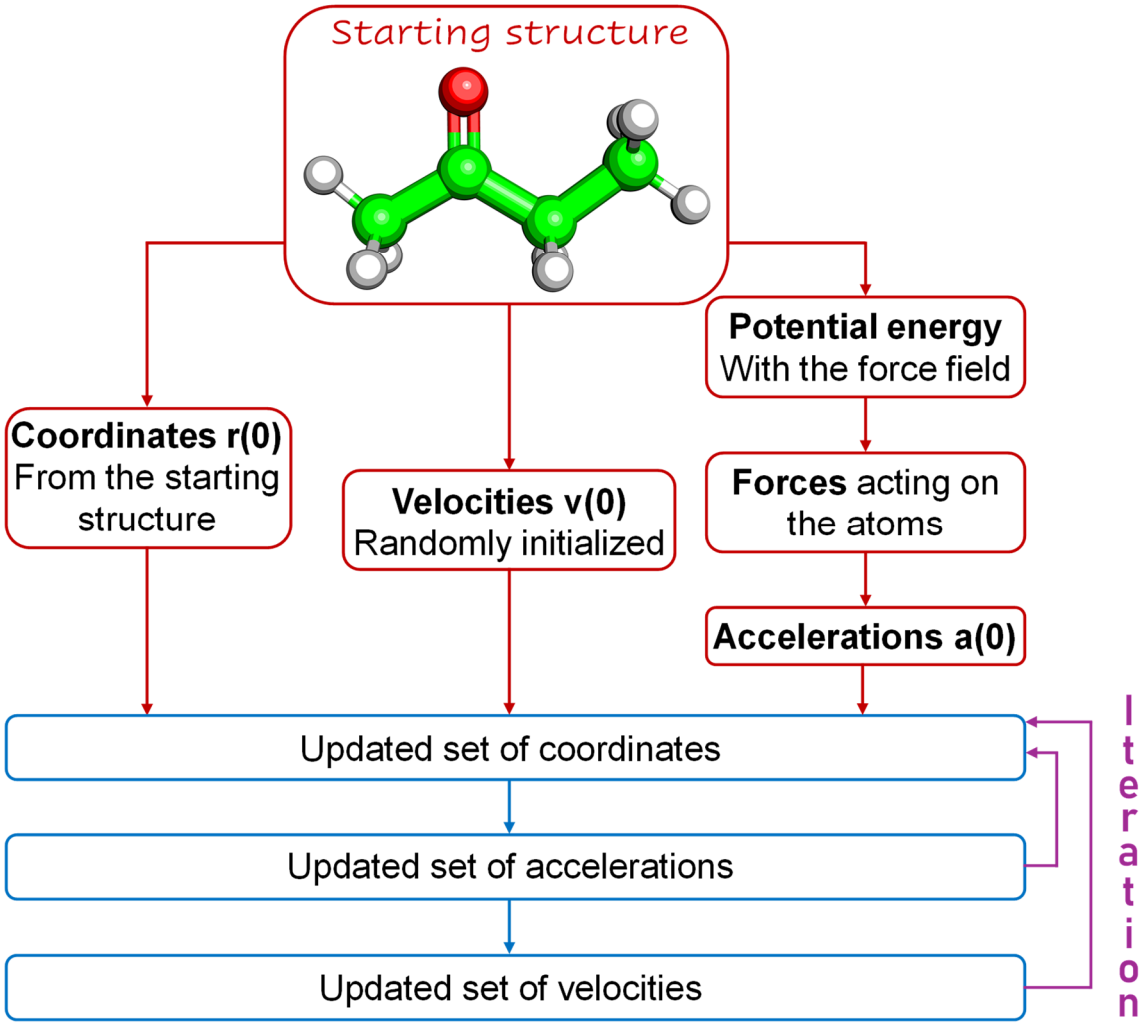


Figure III.3. Workflow of a MD simulation. Based on the structure provided by the user, the first sets of coordinates, velocities and accelerations are initialized. They allow the computation of a new set of coordinates (*i.e.* the next molecular conformation), from which the accelerations can be calculated with the information contained in the force field. These accelerations, in turn, serve to calculate the velocities.

III.1.4. MD simulations are an ideal tool to model (bio)macromolecular systems

The much lower computational cost required to run classical MD simulations, in comparison to QM approaches, makes it a method of choice to treat macromolecular systems. In particular, several MD engines were initially developed for studying biomolecules, especially proteins and nucleic acids. The software used throughout this thesis, the Assisted Model Building with Energy Refinement (AMBER), follows this trend.[13] AMBER contains a variety of FFs dedicated to the modelling of proteins, nucleic acids, sugars, solvents, and so on. It also includes tools to build custom organic molecules. The general amber force field (GAFF), implemented in 2004, provides parameters for most organic compounds, and is compatible with the other AMBER FFs.[14] GAFF is parametrized against a wide range of molecular structures commonly found in ligands. An updated version of the force field, GAFF 2, was released in 2015 (and, later, GAFF 2.11 released in May 2016) and seems to display slightly improved performances than the original version.[15] When building a custom molecule, it is also necessary to compute its partial charges. Two different models were used in this thesis: the restrained electrostatic potential (RESP),[16] and the Austin Model 1 with bond charge correction (AM1-BCC).[17] They both aim to reproduce the electrostatic potential of the molecules calculated at the Hartree-Fock/6-31G* level of theory (a QM method), against which GAFF was parametrized. The set of partial charges used, and its accuracy to represent the true electrostatic potential, can have a dramatic influence on the conformations adopted by a chemical system, as will be discussed in **Chapter VI**. In addition, AMBER offers very practical tools to build oligomers and polymers, as it is possible to constitute its own library of custom monomeric units. The monomers can then be combined in any desired sequence to constitute a tailor-made oligomer or polymer, just as one would build a protein or a nucleic acid by inputting its sequence of amino acids or nucleotides, respectively. This methodology has often been used by our group, as described in a recent review.[18]

Another important aspect of (bio)molecular simulations concerns the treatment of the solvent. It can be described implicitly, *i.e.* without including the solvent molecules in the system. The Generalized Born model,[19] an approximated version of the Poisson-Boltzmann equation, is commonly used. This model treats the solvent as a continuum, whose screening effect on the electrostatic interactions depends on its dielectric constant and the degree of burial of atoms in the 3D structure. The other possibility is to describe the solvent molecules explicitly. This approach, more accurate, allows to directly probe the interactions between the solvent and solute molecules. However, it significantly increases the number of atoms in the system and, therefore, the

computational cost. It also imposes the use of periodic boundary conditions (PBC) **(Figure III.4)**. The principle of PBC is to replicate the simulation box in all directions by creating mirror images, approximating an infinite system. It minimizes edge effects at the boundaries of the box, preventing molecules at the edges from being exposed to vacuum. It also ensures that a molecule diffusing out of the box is replaced by another molecule from a mirror image, as illustrated in **Figure III.4**.

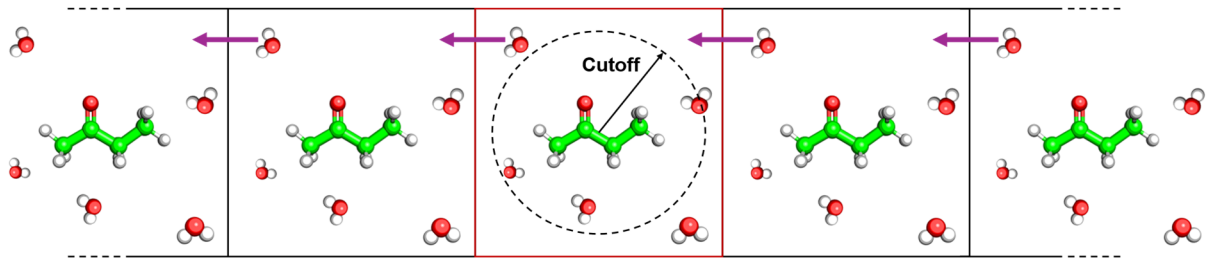


Figure III.4. Illustration of the PBC in two dimensions. The central simulation box, in red, contains a solute and water molecules. It is replicated in the left and right directions. If the water molecule in the upper left corner leaves the simulation box to the left (purple arrow), it is replaced by its mirror image from the opposite side. The cutoff, *i.e.* the distance at which the models used to compute the non-bonded interactions (van der Waals and Coulomb) change, is also illustrated (further details are provided in the text).

In MD simulations, the heavier burden on the computational cost consists in treating the non-covalent interactions, *i.e.* van der Waals (vdW) and electrostatic terms. Their number typically scales with the square of the number of atoms, which can become very large, especially in the case of simulations in explicit solvent. Therefore, non-covalent interactions are generally truncated at a threshold distance, known as the cutoff (as illustrated in **Figure III.4**). The van der Waals terms are computed using a Lennard-Jones potential for pairs of particles whose distance is within the cutoff, and a continuum model is applied as a correction for long-range interactions. Electrostatic terms are calculated through the particle mesh Ewald (PME) scheme, which treats the electrostatic interactions as a sum of short-range and long-range terms.[20] Electrostatic forces for particles whose distance is within the cutoff are calculated in the real space, using direct summation. Long-range interactions are computed in the reciprocal space, using Fourier transforms.

These concepts constitute a basis to understand the functioning of computational simulations, in particular classical MD simulations. In the next part, recent developments in the field, aiming to improve the accuracy of the simulations and the exhaustivity of the sampling, are discussed.

III.2. Limits of MD simulations and how to push their boundaries

III.2.1. MD simulations are based on many (many, many) approximations

The "computational microscope" offered by MD simulations is an extremely powerful tool, giving researchers precious insights on the 3D structure and dynamics of (bio)molecular systems.[21] The field has considerably evolved since the first biomolecular simulation published, in 1997, which studied the dynamics of the bovine pancreatic trypsin inhibitor during 8.8 ps.[22] 20 years later, in 2007, the first microsecond timescale simulation of B-DNA was published.[23] The significant improvements in MD algorithms,[24] the optimization of GPU codes,[25] and the remarkable evolution of the computational resources have made the simulation of hundreds of thousands of atoms on the microsecond timescale routine. MD simulations have achieved many successes over the years and, more generally, the field of computational chemistry is now well established, as evidenced by the 1998, 2013, and 2024 Nobel Prizes in Chemistry.

However, simulations are far from infallible. Their classical nature already represents a considerable approximation. The accuracy of a simulation is dictated by the quality of the parameters contained in its FFs, which may, in some cases, completely fail to describe certain properties of a particular molecular system. A striking example is the case of TIP3P, one of the most widely used water model in biomolecular simulations.[26] It happens that TIP3P (and other common water models) is unable to correctly reproduce the behavior of bulk water, predicting several macroscopic properties with significant errors **(Figure III.5)**.[27] A new water model called OPC was proposed in 2014, and seems to constitute a significant improvement.[28]

It shows the complexity of finding good FF parameters, even for water. Another lesson can be drawn from this example: a simulation does not need to be perfectly accurate to give meaningful results. The failure of TIP3P to reproduce several properties of bulk water does not mean that it is completely unsuitable to study the folding and dynamics of biomacromolecules. It is the role of the researcher to keep a critical eye on the outcome of a simulation, to know the limits of its model, and to take them into account when drawing conclusions. Nevertheless, it is highly desirable to develop more accurate force fields. Improved versions are continuously released, even for well-known biomolecules – see for example ff19SB (for proteins)[29] and tumuc1[30] or OL24[31] (for DNA), available since 2020, 2021 and 2025, respectively. Finding valid FF parameters for synthetic systems is even more challenging, given their enormous chemical diversity. Most FFs dedicated to synthetic molecules are rather general: they

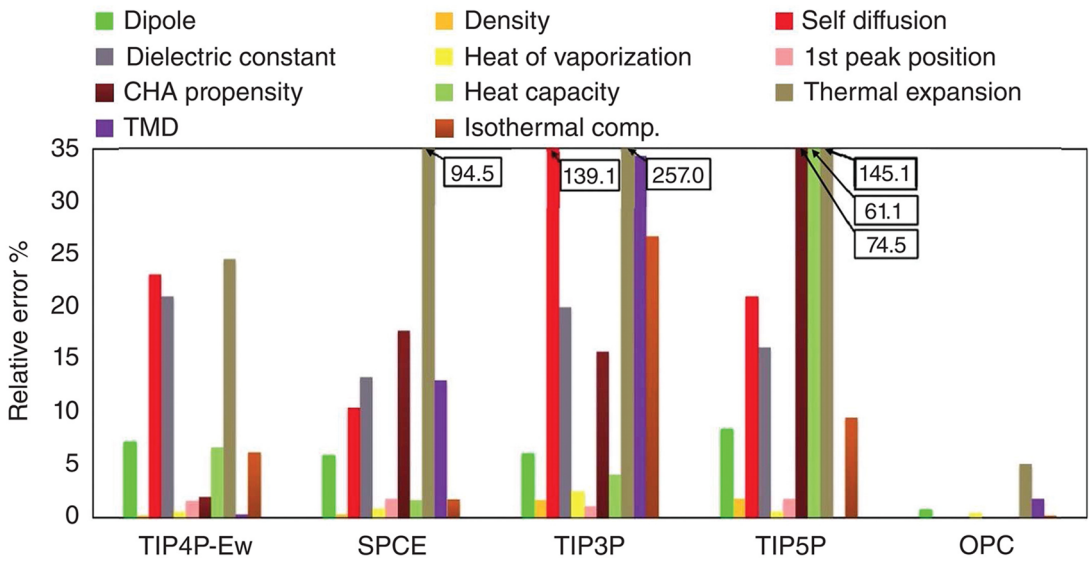


Figure III.5. Relative errors between the predicted theoretical and experimental values of several macroscopic properties of water for five widely used water models. Reproduced from Ref. 27.

contain parameters to describe a wide range of atoms, bonds, angles and torsions. However, their accuracy is limited, as one single FF cannot perfectly reproduce every chemical system. Consequently, when modeling a synthetic compound, it is good practice (and sometimes mandatory) to reparametrize the FF, *i.e.* to adjust its parameters to describe more accurately the molecule of interest. In general, the parameters are refined by comparison with QM calculations. This task can be very tedious, although several tools exist to simplify and automatize it. Within the AMBER suite of programs, mainly used in this thesis, we can cite *Paramfit*[32] or *mdgx*.[33]

III.2.2. Enhancing the accuracy of MD simulations

Beyond parameters adjustment, other methods exist to enhance the accuracy of MD simulations, *i.e.* to bridge the gap between classical and QM approaches. Polarizable FFs, for instance, aim at improving the representation of the electrostatic properties of the molecules, which are modeled as fixed atomic charges in classical approaches.[34] The most intuitive way to account for the deformability of the electronic cloud is to introduce extra particles carrying a fraction of the atomic charges. This is the basis of the Drude oscillator model.[35] Each atom is assigned an additional pseudo-particle, called the Drude particle. The atomic partial charge is distributed between the nucleus and its Drude oscillator, which is free to move, allowing the atom to respond to its electrostatic environment. This improved description of the charge distribution has been successful in different biomolecular applications, including studies of base-flipping in DNA,[36] the structural dynamics of RNA hairpins[37] and mannose disaccharides,[38] as well as for computing the free energy of hydration for most amino

acids.[39] However, the model failed to correctly describe a highly flexible RNA structure, probably due to an overstabilization of hydrogen bonding interactions.[40]
Machine learning (ML) FFs are also becoming a common tool in computational chemistry. Their purpose, similarly to traditional FFs, is to establish relationships between a set of atomic coordinates and the potential energy of the system, the forces acting on the atoms, or both. However, the construction of the mathematical model is completely different. An MLFF does not need to decompose the potential energy of the system into a sum of individual terms (as expressed in **Figure III.2**): it can build any mathematical model. The MLFF is trained on an extensive database of molecular structures with known potential energy, computed at the QM level.[41] From this dataset, a functional representation of the potential energy surface is extracted. The MLFF must then be tested against a validation set, *i.e.* another ensemble of structures with known potential energy. If the error of the MLFF on the validation set is sufficiently low, it can be used for MD simulations. The QM method used to compute the potential energy fixes the upper limit of accuracy of the ML algorithm, as it will, at best, perfectly reproduce the method used for training. MLFFs can be viewed as an intermediate between QM methods and traditional FFs, being more accurate (if correctly trained) but less computationally efficient (due to their more complex functional form) than the latter.[42] Very promising results were published recently for the use of MLFFs in biomolecular simulations, to compute the relative energy,[43] or even to perform MD simulations on proteins.[44] In that study, the model was trained on small fragments, and then applied to simulate a 46-residue protein in explicit water (more than 25,000 atoms). Although limited to the nanosecond timescale, it demonstrates that MD simulations with MLFFs, reaching *ab initio* accuracy, are achievable on macromolecular systems.
Finally, another way to improve the accuracy of MD simulations is to use hybrid QM/MM approaches, treating a small part of the system at the QM level and describing the rest with MM.[45] This is particularly useful in biomolecular simulations, where the computational cost of running pure *ab initio* MD simulations would be prohibitive, while a sub-part of the system requires QM accuracy. A typical example of these multiscale approaches is host-guest chemistry, such as enzymatic catalysis or ligand binding. The active site, where the reaction or binding occurs, is treated at the QM level, while the surrounding protein scaffold and the solvent are described using FFs. The main challenge associated to QM/MM simulations lies in computing the interactions at the interface between atoms in the "MM region" and atoms in the "QM region". A recent example made use of QM/MM methods to correctly estimate the unbinding rate constant of a ligand.[46] Classical FFs from AMBER were accurate to describe the stable bound state, but overestimated the potential energy of the

transition state, thus overestimating the unbinding rate constant. The QM/MM approach was necessary to correctly describe the change in electrostatic properties between the bound and transition states. It illustrates a disadvantage of the fixed point charges used by classical FFs. Another example took advantage of QM/MM simulations to investigate the reaction mechanisms of four inhibitors of a SARS-CoV-2 enzyme, the main protease, which plays a part in viral replication.[47] The four ligands are able to form a covalent bond with a cysteine in the active site of the enzyme – a chemical reaction that cannot be captured with classical MD simulations – thereby inhibiting its activity.

III.2.3. Enhancing the speed of MD simulations

In addition to the efforts aimed at improving the accuracy of MD simulations, several approaches have been developed to accelerate conformational sampling. It is now routine to run microsecond-long simulations, but this timescale remains too short to probe many biomolecular processes. One way to increase the speed of a simulation is to increase the timestep, *i.e.* to reduce the frequency at which the equations of motions need to be solved. The problem in increasing the timestep is that it quickly leads to numerical instabilities, when the forces change importantly between two steps. The typical value of the timestep is 1 fs, as mentioned earlier. However, algorithms are commonly applied to freeze the vibrations including hydrogen atoms, such as SHAKE,[48] SETTLE,[49] or LINCS,[50] allowing the use of a 2-fs timestep. More recently, the hydrogen mass repartitioning (HMR) scheme was used in MD simulations, enabling the use of timesteps of 4 fs.[51] The principle of HMR is to redistribute the mass of the heavy atoms to their bonded hydrogen atoms, such as to increase their mass and to slow down their vibration frequency. This strategy proved effective to simulate biological membranes.[52] The increased timestep did not alter the computed properties, in comparison to simulations using a 2-fs timestep, and brought a speedup comprised between 40 to 90 %, depending on the system and the computational architecture. However, the method may apparently slow down protein-ligand binding.[53] Here, the increased timestep led to faster diffusion of the ligand and increased protein dynamics compared to classical MD. This apparently complicated the stabilization of metastable states encountered in the binding process. Although appealing, increasing the timestep may not be an ideal choice to reduce the computational time in every case, especially when key binding intermediates must be sampled.

Different methods were also developed to improve the sampling efficiency, to find more quickly local minima on the potential energy surface.[54] In particular, accelerated molecular dynamics (aMD) simulations use a bias potential to lower the energy

barriers between minima.[55] When the potential energy of the system falls below a threshold, a boost potential is applied. The idea is to artificially flatten the potential energy surface, preventing the system from being trapped for long times in local minima. Accelerated simulations were performed on ligand-protein binding studies and permitted to identify the same binding sites than those detected by conventional MD simulations in much shorter simulation time.[56] A similar approach was followed to compute the thermodynamic and kinetic properties of binding of several guests to a cyclodextrin host.[57] The accelerated protocol allowed to sample several binding/unbinding events in 300 ns, while several microseconds were required with conventional MD. These two methods (HMR and aMD) were exploited in this thesis to study the dynamics of large heteropolymers in a reasonable computational time **(Chapter VI)**.

Finally, we can cite coarse-grained (CG) approaches, which reduce the computational cost – thus the computational time – by decreasing the number of particles in the system. To do so, several atoms are merged into the same particle – for example, one amino acid could be represented by only one particle, instead of taking into account all its atoms. Consequently, the system is described with a lower resolution than in all-atom MD simulations. CG models necessarily miss finer atomic details, such as directionality of H-bonds[58] or interactions with the solvent,[58,59] and lack flexibility in the description of the secondary structures of proteins.[59] Still, this method is popular to study biomolecular systems,[58,60] notably with the well-known MARTINI force field.[59] Originally developed to model lipids, its scope has been expanded over the years, making it a general force field, even applicable to model organic polymers. CG methods provide a way to study systems of a size that all-atom MD simulations could never reach. A particularly striking example is the building of a whole cell using tools from the MARTINI ecosystem.[61] Although no MD simulations were actually launched, being able to develop a computational model for an entire cell with all its components, containing more than six billion atoms, remains remarkable.

III.3. Common molecular descriptors and analyses

Several descriptors are used throughout this thesis to characterize the conformations, dynamics and interactions of molecular systems. They are explained hereafter to clarify their meaning to the reader. Most of the analyses were performed with the *cpptraj* module of AMBER and in-house scripts.[62]

The root mean square deviation (RMSD) is the average deviation of the atomic coordinates of a structure by comparison to the atomic coordinates of a reference structure **(Equation III.6)**. The reference structure is often chosen as the initial

conformation of the MD simulation. The evolution of the RMSD over time gives insights on the flexibility of the system, and its convergence can indicate the stabilization of a given conformation.

$$RMSD(t) = \sqrt{\frac{\sum_{i=1}^{N}[x_i(t) - x_i(ref)]^2}{N}} \quad \textbf{(III.6)}$$

Where N is the number of atoms, and $x_i(t)$ and $x_i(ref)$ are the coordinates of atom i in the structure generated at time t and in the reference structure, respectively.

The root mean square fluctuation (RMSF) is a similar measurement, but is more local. The RMSF computes the positional fluctuations of an atom around its average coordinates **(Equation III.7)**. It is often averaged over groups of atoms, typically monomer units, providing insights into the distribution of flexible and rigid regions within a molecule.

$$RMSF = \sqrt{\frac{\sum_{t_j=1}^{M}(x_i(t_j) - \langle x_i \rangle)^2}{M}} \quad \textbf{(III.7)}$$

Where M represents the number of input structures, $x_i(t_j)$ represents the coordinates of atom i in the structure generated at time t_j, and $\langle x_i \rangle$ denotes the average coordinates of atom i, computed over all input structures. Note that the RMSD involves a sum on the number of atoms, thus being a descriptor of the global structure, while the RMSF involves a sum on all snapshots for one atom (or one group of atoms), thus being an average local descriptor.

The radius of gyration (R_G) is a measure of the size and compactness of a system. It computes the average distance of the atoms from their geometric center **(Equation III.8)**. The evolution of the R_G as a function of time can be tracked to follow the folding of a molecule.

$$R_G = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_{center})^2}{N}} \quad \textbf{(III.8)}$$

Where N is the number of atoms, x_i represents the coordinates of atom i, and x_{center} denotes the coordinates of the geometric center.

The solvent-accessible surface area (SASA) measures the extent of a molecular surface that can be probed by a solvent molecule **(Figure III.6)**. It reflects the exposure of a molecule to its surrounding environment. The SASA can be decomposed by sub-parts of the whole system, such as monomeric units, revealing which regions are buried and which ones are exposed. The SASA was determined using the linear combination of pairwise overlaps (LCPO) method.[63] In this model, atoms are approximated as perfect spheres, with a radius equal to their van der Waals radius plus that of a solvent probe (typically 1.4 Å, to represent a water molecule).

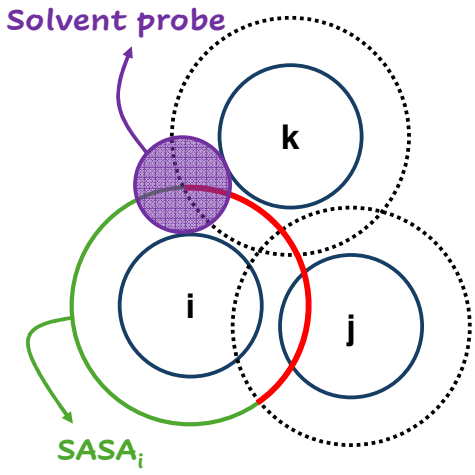


Figure III.6. 2D schematic representation of the SASA of an atom i. The solvent probe (in purple) defines the accessible region (in green), while the occluded area (in red) is inaccessible due to overlaps with neighboring atoms j and k.

Analyses can also be carried out in the form of images, such as "heatmaps", which were often used in this thesis to highlight and localize relationships between pairs of variables. Many kinds of data can be represented, such as distances, interactions, free enthalpy of binding, and so on. A very simple example is shown in **Figure III.7**.

Finally, network representations were used to investigate the connections inside supramolecular assemblies, using the *Cytoscape 3.9.1* software.[64] The 3D conformations generated during the MD simulations are converted into 2D networks, where each heavy atom, *i.e.* any atom except hydrogen, constitutes a node **(Figure III.8)**. Two nodes are connected, *i.e.* linked to each other by an edge, if their distance

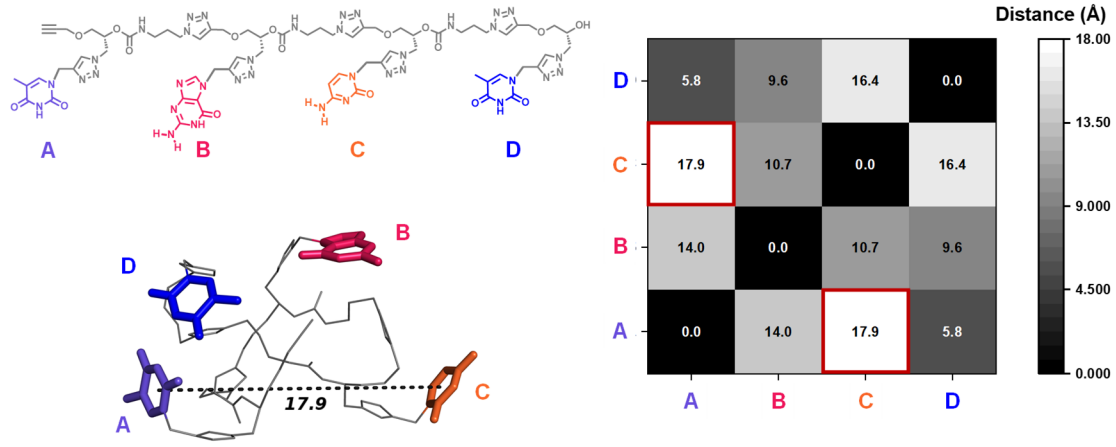


Figure III.7. Illustration of a distance heatmap. The molecule investigated is a tetramer bearing four functional units, represented by the letters A-D. The heatmap is built based on the 3D structure shown on the left. At the crossing of two letters, on the x and y axes, is found a square, whose color indicates the distance between the two units (see the scale on the right). As an example, the distance between the units A and C was measured on the 3D structure and corresponds to the squares circled in red on the heatmap, which is symmetric with respect to the diagonal.

in the 3D structure is inferior than a cutoff distance typical of short-range interactions (around 5 Å). Network representations are a useful visualization tool, and an interesting way to investigate the connectivity and recognition inside molecular systems. It has been commonly applied to biomolecular systems or in materials science.[65,66] Two descriptors are used in this thesis to mathematically describe the connectivity inside a network. The betweenness centrality of a node describes the number of shortest paths, *i.e.* the shortest sequence of nodes that must be traversed to go from one node to another, involving this node. The closeness centrality indicates how long are the paths connecting one node to all the others. A node able to reach all the others with short paths will have a high closeness centrality. Networks were further studied with the partition algorithm *Infomap*,[67] which detects groups of highly connected nodes. These nodes are regrouped into the same particle, called "module" or "community". This approach allows a coarse-grained representation of the network, simplifying the visualization of the most important connections between specific moieties.

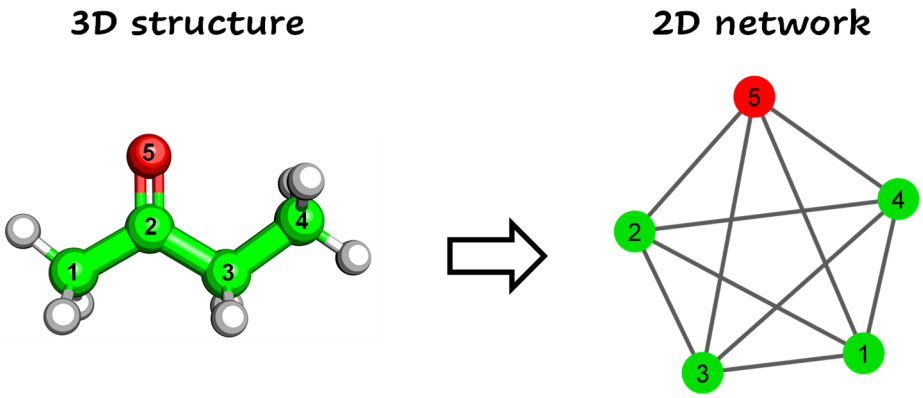


Figure III.8. Representation of the conversion of a molecular 3D structure into a 2D network. Only the heavy atoms, *i.e.* not hydrogen atoms, are conserved in the network representation.

References

[1] P. Dirac. Quantum mechanics of many-electron systems. *Proc. R. Soc. London, Ser. A* **1929**, *123*, 714–733.

[2] L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella, L. Genovese. Challenges in large scale quantum mechanical calculations. *WIREs Comput. Mol. Sci.* **2017**, *7*, e1290.

[3] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142–10186.

[4] M. Born, R. Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. Phys.* **1927**, *389*, 457–484.

[5] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24*, 89–110.

[6] Chr. Møller, M. S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

[7] P. Hohenberg, W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

[8] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[9] P. Norman, K. Ruud, T. Saue. "*Principles and Practices of Molecular Properties*", Wiley, **2018**, pp. 1–9.

[10] S. Antolínez, P. E. Jones, J. C. Phillips, J. A. Hadden-Perilla. AMBERff at Scale: Multimillion-Atom Simulations with AMBER Force Fields in NAMD. *J. Chem. Inf. Model.* **2024**, *64*, 543–554.

[11] E. Paquet, H. L. Viktor. Computational Methods for Ab Initio Molecular Dynamics. *Adv. Chem.* **2018**, *2018*, 1–14.

[12] L. Verlet. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98–103.

[13] R. Salomon-Ferrer, D. A. Case, R. C. Walker. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210.

[14] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.

[15] D. Vassetti, M. Pagliai, P. Procacci. Assessment of GAFF2 and OPLS-AA General Force Fields in Combination with the Water Models TIP3P, SPCE, and OPC3 for the Solvation Free Energy of Druglike Organic Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1983–1995.

[16] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[17] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[18] D. Dellemme, S. Kardas, C. Tonneaux, J. Lernould, M. Fosséprè, M. Surin. From sequence definition to structure-property relationships in discrete synthetic macromolecules: insights from molecular modeling. *Angew. Chem. Int. Ed.* **2025**, *64*, e202420179.

[19] A. Onufriev, D. Bashford, D. A. Case. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

[20] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[21] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, D. E. Shaw. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

[22] J. A. McCammon, B. R. Gelin, M. Karplus. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.

[23] A. Pérez, F. J. Luque, M. Orozco. Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

[24] P. Larsson, B. Hess, E. Lindahl. Algorithm improvements for molecular dynamics simulations. *WIREs Comput. Mol. Sci.* **2011**, *1*, 93–108.

[25] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, R. C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

[26] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[27] A. V. Onufriev, S. Izadi. Water models for biomolecular simulations. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1347.

[28] S. Izadi, R. Anandakrishnan, A. V. Onufriev. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **2014**, *5*, 3863–3871.

[29] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, C. Simmerling. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528–552.

[30] K. Liebl, M. Zacharias. Tumuc1: A New Accurate DNA Force Field Consistent with High-Level Quantum Chemistry. *J. Chem. Theory Comput.* **2021**, *17*, 7096–7105.

[31] M. Zgarbová, J. Šponer, P. Jurečka. Refinement of the Sugar Puckering Torsion Potential in the AMBER DNA Force Field. *J. Chem. Theory Comput.* **2025**, *21*, 833–846.

[32] R. M. Betz, R. C. Walker. Paramfit: Automated optimization of force field parameters for molecular dynamics simulations. *J. Comput. Chem.* **2015**, *36*, 79–87.

[33] K. T. Debiec, D. S. Cerutti, L. R. Baker, A. M. Gronenborn, D. A. Case, L. T. Chong. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **2016**, *12*, 3926–3947.

[34] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal, P. Ren. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* **2019**, *48*, 371–394.

[35] J. A. Lemkul, J. Huang, B. Roux, A. D. MacKerell. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983–5013.

[36] J. A. Lemkul, A. Savelyev, A. D. MacKerell. Induced Polarization Influences the Fundamental Forces in DNA Base Flipping. *J. Phys. Chem. Lett.* **2014**, *5*, 2077–2083.

[37] M. Y. Sengul, A. D. MacKerell. Accurate Modeling of RNA Hairpins Through the Explicit Treatment of Electronic Polarizability with the Classical Drude Oscillator Force Field. *J. Comput. Biophys. Chem.* **2022**, *21*, 461–471.

[38] A. Ruda, A. H. Aytenfisu, T. Angles d'Ortoli, A. D. MacKerell, G. Widmalm. Glycosidic α-linked mannopyranose disaccharides: an NMR spectroscopy and molecular dynamics simulation study employing additive and Drude polarizable force fields. *Phys. Chem. Chem. Phys.* **2023**, *25*, 3042–3060.

[39] V. A. Ngo, J. K. Fanning, S. Y. Noskov. Comparative Analysis of Protein Hydration from MD simulations with Additive and Polarizable Force Fields. *Adv. Theory Simul.* **2019**, *2*, 1800106.

[40] L. Winkler, T. E. Cheatham. Benchmarking the Drude Polarizable Force Field Using the r(GACC) Tetranucleotide. *J. Chem. Inf. Model.* **2023**, *63*, 2505–2511.

[41] V. Botu, R. Batra, J. Chapman, R. Ramprasad. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121*, 511–522.

[42] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142–10186.

[43] Y. Han, Z. Wang, Z. Wei, J. Liu, J. Li. Machine learning builds full-QM precision protein force fields in seconds. *Briefings Bioinf.* **2021**, *22*, bbab158.

[44] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. Medrano Sandonas, J. T. Berryman, A. Tkatchenko, K.-R. Müller. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Sci. Adv.* **2024**, *10*, eadn4397.

[45] G. Groenhof. *"Introduction to QM/MM simulations"*, Humana Press, **2013**, pp. 43–66.

[46] R. Capelli, W. Lyu, V. Bolnykh, S. Meloni, J. M. H. Olsen, U. Rothlisberger, M. Parrinello, P. Carloni. Accuracy of Molecular Simulation-Based Predictions of k_{off} Values: A Metadynamics Study. *J. Phys. Chem. Lett.* **2020**, *11*, 6373–6381.

[47] B. L. Grigorenko, I. V. Polyakov, M. G. Khrenova, G. Giudetti, S. Faraji, A. I. Krylov, A. V. Nemukhin. Multiscale Simulations of the Covalent Inhibition of the SARS-CoV-2 Main Protease: Four Compounds and Three Reaction Mechanisms. *J. Am. Chem. Soc.* **2023**, *145*, 13204–13214.

[48] J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

[49] S. Miyamoto, P. A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962.

[50] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[51] C. W. Hopkins, S. Le Grand, R. C. Walker, A. E. Roitberg. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874.

[52] C. Balusek, H. Hwang, C. H. Lau, K. Lundquist, A. Hazel, A. Pavlova, D. L. Lynch, P. H. Reggio, Y. Wang, J. C. Gumbart. Accelerating Membrane Simulations with Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2019**, *15*, 4673–4686.

[53] M. Sahil, S. Sarkar, J. Mondal. Long-time-step molecular dynamics can retard simulation of protein-ligand recognition process. *Biophys. J.* **2023**, *122*, 802–816.

[54] D. Perez, B. P. Uberuaga, Y. Shim, J. G. Amar, A. F. Voter. Chapter 4 Accelerated Molecular Dynamics Methods: Introduction and Recent Developments. *Annu. Rep. Comput. Chem.* **2009**, 5, 79–98.

[55] D. Hamelberg, J. Mongan, J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

[56] K. Kappel, Y. Miao, J. A. McCammon. Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor. *Q. Rev. Biophys.* **2015**, *48*, 479–487.

[57] Y. Miao, A. Bhattarai, J. Wang. Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics. *J. Chem. Theory Comput.* **2020**, *16*, 5526–5547.

[58] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, S. J. Marrink. The power of coarse graining in biomolecular simulations. *WIREs Comput.l Mol. Sci.* **2014**, *4*, 225–248.

[59] S. J. Marrink, L. Monticelli, M. N. Melo, R. Alessandri, D. P. Tieleman, P. C. T. Souza. Two decades of Martini: Better beads, broader scope. *WIREs Comput. Mol. Sci.* **2023**, *13*, e1620.

[60] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, 090901.

[61] J. A. Stevens, F. Grünewald, P. A. M. van Tilburg, M. König, B. R. Gilbert, T. A. Brier, Z. R. Thornburg, Z. Luthey-Schulten, S. J. Marrink. Molecular dynamics simulation of an entire cell. *Front. Chem.* **2023**, *11*, 1106495.

[62] D. R. Roe, T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

[63] J. Weiser, P. S. Shenkin, W. Clark Still. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217230.

[64] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.

[65] M. Fossépré, L. Leherte, A. Laaksonen, D. P. Vercauteren. *"Understanding the Structure and Dynamics of Peptides and Proteins Through the Lens of Network Science"*, Wiley, **2018**, pp. 105–161.

[66] L. S. G. Leite, S. Banerjee, Y. Wei, J. Elowitt, A. E. Clark. Modern chemical graph theory. *WIREs Comput. Mol. Sci.* **2024**, *14*, e1729.

[67] L. Bohlin, D. Edler, A. Lancichineei, M. Rosvall. "*Community detection and visualization of networks with the map equation framework"*, Springer, Cham, **2014**, pp. 3–34**.**

IV. Exploiting sequence-controlled architectures to master biorecognition

The results section of this thesis starts with systems involved in biorecognition applications, where the defined sequence and programmable folding of biopolymers help us to target them. This section is divided in two parts. The first one is dedicated to the study of the interactions between an integrin and different peptides. We will see how subtle stereochemical modifications can influence the intrinsic conformation of the peptides, and how it affects their binding to a protein receptor involved in cellular migration **(Section IV-A)**. In the second part, we will investigate the supramolecular assembly of small organic molecules around a DNA template. The synthetic molecules, functionalized with nucleobases, are able to recognize a single-stranded DNA bearing the complementary unit. Furthermore, these synthetic molecules are photoisomerizable, and the switch between *trans* and *cis* configurations allows us to modulate the assembly with DNA **(Section IV-B)**.

IV-A. Chiral mismatch in collagen-mimetic peptides modulates cell migration through integrin-mediated molecular recognition

Part of this work is reported in: *Chiral mismatch in collagen-mimetic peptides modulates cell migration through integrin-mediated recognition*.
A. Remson, D. Dellemme, M. Luciano, M. Surin, S. Gabriele. Deposited on bioRxiv (author preprint): https://doi.org/10.1101/2024.07.23.604866.

IV-A.1. Introduction

Many cellular processes rely on interaction cascades, where specific host-guest recognitions trigger conformational changes that propagate from one molecular species to the next.[1,2] These interactions take place in the crowded cellular environment, involving many proteins and small molecules, yet they are extremely well regulated: the same receptor, when activated by different ligands, can trigger different responses.[3] The specificity of these recognition events arises from the highly controlled 3D structures of proteins, whose sequence-encoded folding generates well-defined binding sites. All these interactions mediate the behavior of cells, which respond to the stimuli that they perceive when probing their environment. As a consequence, the physicochemical parameters characterizing the cell environment, *i.e.* the extracellular matrix (ECM), can strongly influence the internal organization of cells and the processes in which they are involved. For instance, matrix rigidity was shown

to impact cell shape, polarization, adhesion and migration.[4,5] Modifications on the microstructure of the matrix, such as the presence of curvatures or other patterns, also modify cell properties.[5,6] A recent example showed that matrix viscoelasticity and stiffness influence cell spreading and migration, and that spatial confinement can alter the way cells respond to the mechanical properties of their environment.[7] Another characteristic of biological components is their chirality. This property is found at all scales in living matter, from proteins and nucleic acids, built on chiral monomers, to organs such as the heart. Cells contain intrinsically chiral components, such as actin filaments, helical supramolecular polymers that are part of the cytoskeleton. Interestingly, the dynamic network of actin filaments was shown to self-organize into chiral motifs, twisted radial fibres rotating in a counter-clockwise manner, when cells are confined on circular micropatterns.[8] This chiral organization could even induce the chiral motion of other cellular components. Impressively, the sense of rotation could be reversed, from counter-clockwise to clockwise, through the overexpression of α-actinin-1, a protein involved in the crosslinking of actin filaments. Another group found that cell aggregates embedded within a 3D hydrogel environment spontaneously exhibit rotational motions, the sense of rotation being regulated by the same mechanism involving actin filaments and α-actinin-1.[9] Chirality was also demonstrated to propagate from the molecular scale to an entire organism.[10] The localized overexpression of myosin 1D, a molecular motor, was sufficient to induce a complete twist of the body of a larva and perturb its movements. Despite these examples and the known importance of chirality in living systems, the impact of the stereochemistry of ECM components on cell behavior has remained largely unexplored until now.

To bridge this gap, three peptide-coated surfaces of varying chirality were engineered as representative models of the ECM components, and exploited to study the adhesion and migration of epithelial keratocytes, cells derived from fish scales. Collagen, the most abundant component of the ECM, was chosen as the natural coating, acting as a control.[11] Then, two collagen-mimetic peptides (CMPs), able to reproduce the triple helix structure of collagen, were designed.[12] The two CMPs share the same sequence of AAs and differ only in their stereochemistry. The first CMP contains only L-AAs, while the second consists of a block of L-AAs followed by a block of D-AAs; they are referred to as homochiral and heterochiral CMPs, respectively. Experiments showed differences in cell adhesion and migration on these substrates of opposite chirality. In particular, the heterochiral CMP, which contains a “chiral mismatch” at the junction between its natural L and unnatural D-AAs, displayed a lower ability to support cell adhesion and migration. These results reveal that the stereochemistry of the ECM components impacts cell behavior. At the molecular level, integrins, a family of

transmembrane proteins expressed by cells, are known to mediate cellular migration through interactions with the ECM components.[13] Therefore, inhibition experiments were carried out and identified the $\alpha_1\beta_1$ integrin, a well-known collagen receptor, as sensitive to the ECM chirality. It led us to investigate the behavior of collagen and the CMPs in interaction with this integrin at the atomic scale, by means of molecular dynamics (MD) simulations. Our results suggest that the chiral mismatch in the heterochiral CMP destabilizes its triple helix conformation, reducing its interactions with the binding site of the integrin. This perturbation at the molecular level could contribute to the decreased cell adhesion on this substrate. All the experimental results presented in this chapter were obtained by Alexandre Remson.[14]

IV-A.2. Design of the peptide substrates and simplified models for MD simulations

Collagen I being the major component of the ECM, it was chosen as the control substrate to probe the effect of ECM chirality on cell migration.[11] Collagen forms a left-handed helix at the single-chain level, with a conformation known as polyproline type II (PPII), but self-assembles into a supramolecular right-handed triple helix **(Figure IV.1 A)**.[12] To mimic collagen, the two CMPs incorporate an AA sequence $(PPG)_{10}$, P and G refer to as proline and glycine, respectively. A long sequence with this triplet was shown to reproduce a conformation similar to PPII.[15] The effect of chirality is incorporated in this sequence: the proline residues have the natural L chirality in the homochiral CMP, while they have the artificial D chirality in the heterochiral CMP. The formation of PPII conformations with opposite chirality for the two CMPs, both in solution and in the solid state, was confirmed by circular dichroism (CD) spectroscopy.[14] A sequence $(PEG)_2$ precedes the $(PPG)_{10}$ part, E refers to as glutamate. This AA was shown to play an important role in the recognition between collagen and α integrins, and was therefore integrated into the CMPs.[16] Before the $(PEG)_2$ residues, four lysines were added to improve solubility in water. Finally, a glycine unit links the CMPs to a fluorescent dye commonly used with peptides (5-Carboxytetramethylrhodamine, TAMRA). All lysines and the $(PEG)_2$ residues have the natural L chirality in both CMPs. Therefore, the only difference between the homochiral and the heterochiral CMPs is the stereoinversion in the $(PPG)_{10}$ section **(Figure IV.1 B)**. For the MD simulations, simplified peptide models were employed, retaining only the AAs relevant to the interaction with the integrin and to reduce computational cost. Collagen type I was represented by a sequence of 21 AAs containing the binding motif "GFOGER", known to be involved in the interaction with collagen-binding integrins **(Figure IV.1 C)**.[17] For the CMPs, only the $(PEG)_2$ part and a $(PPG)_5$ section were considered, as this minimal model contains the glutamate

residue, necessary for the binding, and the (PPG) triplets which induce the formation of the supramolecular triple helix and contain the chiral information distinguishing the two CMPs **(Figure IV.1 D)**.

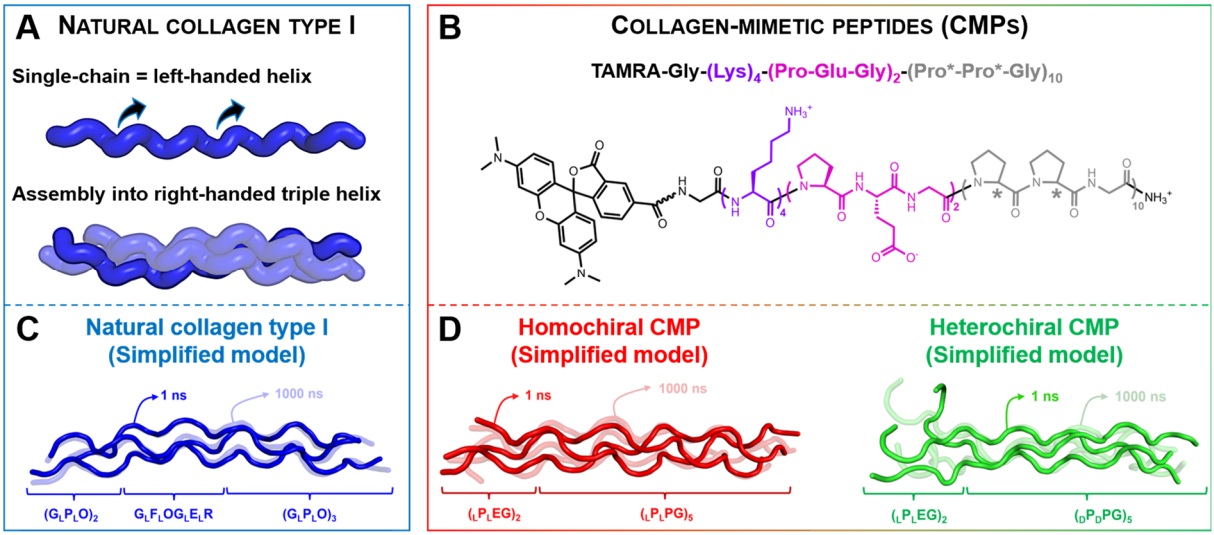


Figure IV.1. Structures of the peptides investigated. **(A)** Cartoon representation of collagen type I, highlighting the formation of a supramolecular right-handed triple helix from left-handed single-chains. **(B)** AA sequence and chemical structure of the CMPs. "Pro*" indicates prolines that have the opposite chirality between the two CMPs. **(C), (D)** Cartoon representation of the simplified models of the peptides, shown as triple helices, used for the MD simulations. Two snapshots, illustrating the first (at 1 ns) and last (at 1000 ns) conformations, are superimposed for each system. The AA sequence is given below (G: glycine; P: proline; O: hydroxyproline; F: phenylalanine; E: glutamate; R: arginine). The letters "L" and "D" preceding the AA letters indicate their chirality.

IV-A.3. Cell migration involves peptide-integrin interactions mediated by a glutamate residue

The migration of epithelial cells was experimentally studied on the three substrates, and a significant impact of the chirality on cell migration speed was observed. Similar speed values were found on collagen and the homochiral CMP (8.78 ± 3.01 µm/min and 8.64 ± 2.67 µm/min, respectively), but the migration was significantly slower on the heterochiral CMP (6.50 ± 2.14 µm/min). In addition, cells performed less focal adhesions on this substrate. These results indicate that cells are sensitive to the chirality of their matrix. We hypothesized that this behavior could be attributed to interactions between collagen (or the CMPs) and integrins, which are protein receptors expressed by the cell to mediate adhesion and migration through specific interactions with components of the ECM **(Figure IV.2 A)**.[4] Integrins constitute a class of 24

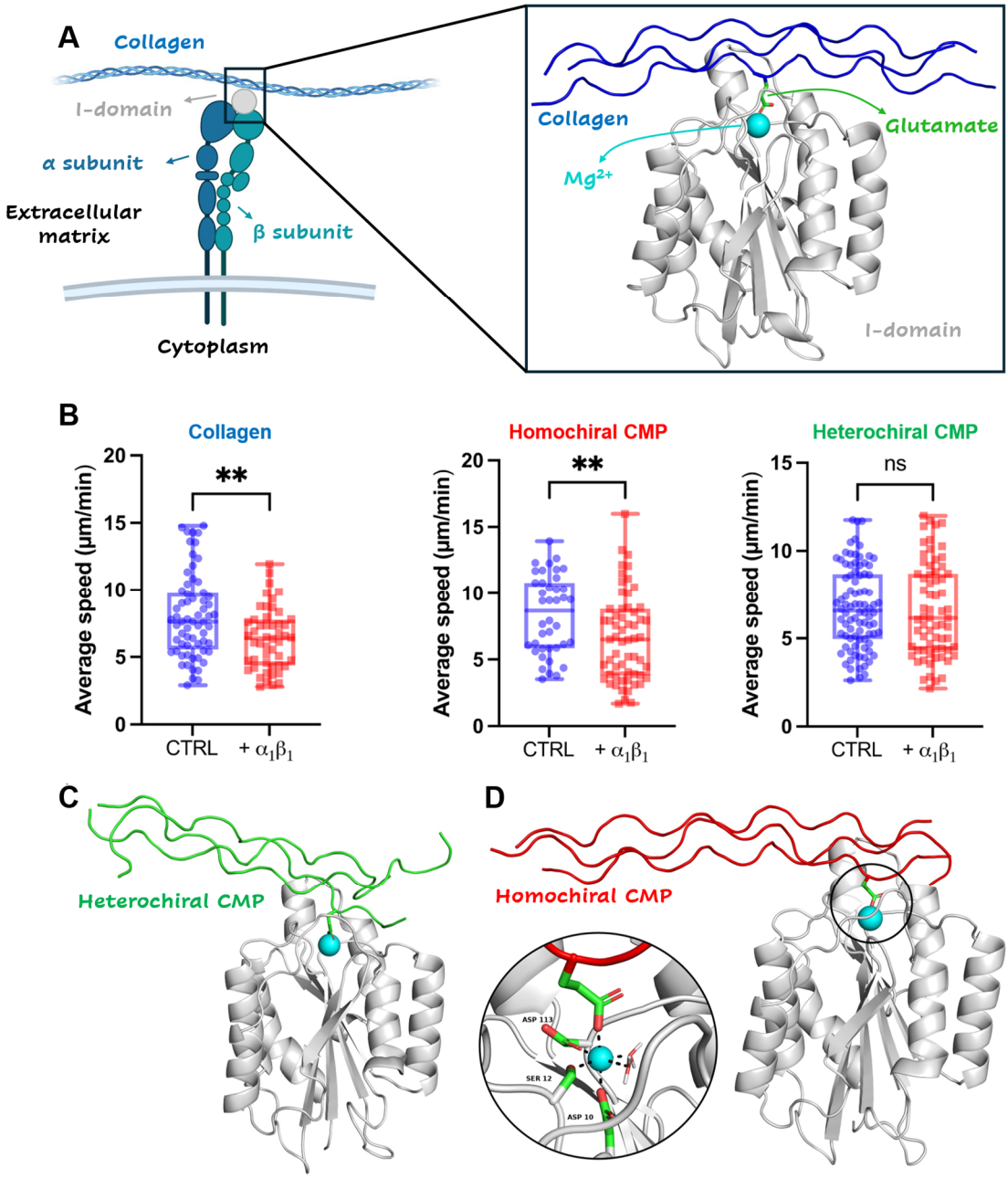


Figure IV.2. Summary of the investigated cell-matrix interactions. **(A)** Schematic representation of the interaction between collagen, found in the ECM, and the I-domain of the collagen-binding integrin $\alpha_1\beta_1$, expressed by the cell. The zoom shows the binding site, where the collagen triple helix coordinates the divalent cation (Mg^{2+}, represented as a cyan sphere) in the MIDAS with a glutamate residue. The left part of the figure was created with BioRender.com. **(B)** Results of inhibition experiments performed by Alexandre Remson, showing the modulation of cell migration speed before (in blue) and after (in red) inhibition of the $\alpha_1\beta_1$ integrin on the collagen (left), homochiral CMP (middle) and heterochiral CMP (right) substrates. **(C), (D)** Final MD snapshots of the heterochiral CMP and homochiral CMP, in interaction with the I-domain of the $\alpha_1\beta_1$ integrin. In **(D)**, a zoom is made on the binding site to show the Mg^{2+} complexation, made by AAs of the I-domain, the glutamate residue of the CMP and completed by water molecules.

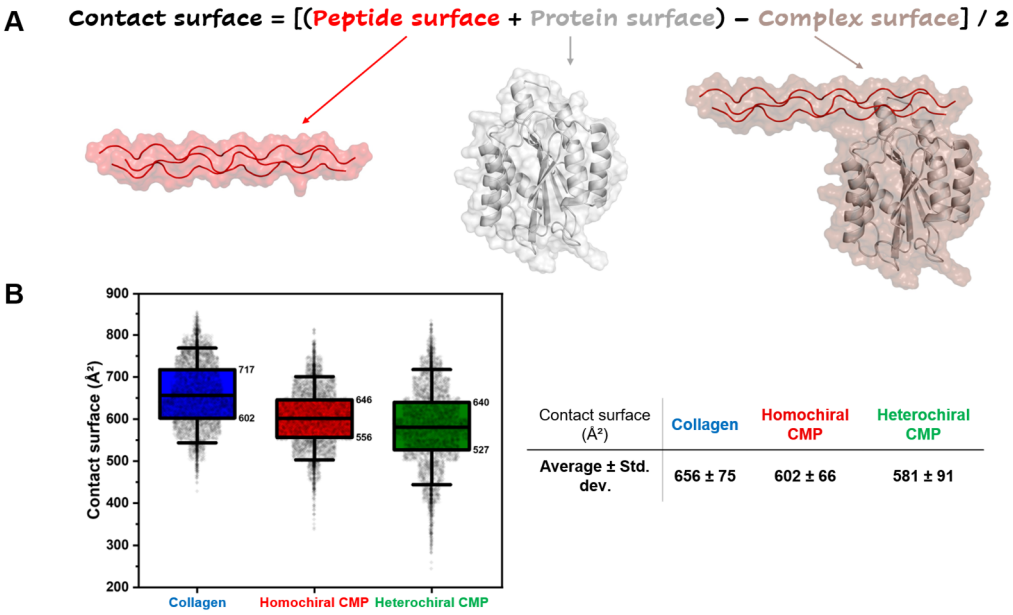
heterodimeric proteins, composed of the association between an α and a β subunit.[13] There exists a subclass of integrins dedicated to collagen-binding ($\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_{10}\beta_1$ and $\alpha_{11}\beta_1$) whose selectivity depends on the collagen type.[18] However, they all share the presence of a particular I-domain, carried by the α subunit, responsible for binding.[13] This domain contains a cavity hosting a divalent cation, known as the metal-ion dependent adhesion site (MIDAS).[19] Binding to MIDAS was shown to be strongly influenced by the presence of a glutamate residue (which is present in the GFOGER motif in collagen), able to coordinate the central cation (see the zoom in **Figure IV.2 A**). Inhibition experiments were thus carried out to identify the cell receptors involved in the migration of our keratocytes. Antibodies were added to block the binding site of collagen-binding integrins, preventing their interactions with the peptides. These experiments showed a particularly strong response from the $\alpha_1\beta_1$ integrin, in agreement with other studies highlighting its role in cell migration.[20,21] Once again, the results were similar for collagen and the homochiral CMP, with a cell speed decrease of 20 to 25 % upon $\alpha_1\beta_1$ inhibition. In contrast, the migration speed on the heterochiral substrate was not significantly affected **(Figure IV.2 B)**.

Based on these results, MD simulations were carried out to better understand the impact of chirality on the intermolecular interactions between the $\alpha_1\beta_1$ integrin and the peptides. The simulations were realized with the AMBER suite of programs, by placing either collagen, the homochiral CMP or the heterochiral CMP in the binding site of the I-domain of the $\alpha_1\beta_1$ integrin (see **Figure IV.2 A, C, D**), in explicit water boxes containing Na^+ and Cl^- ions (see details of the simulation protocol in **Section IV-A.5**).[22] As with collagen, the CMPs contain glutamate residues, that serve as coordinating units for a divalent cation located in the MIDAS **(Figure IV.2 C, D)**. Here, glutamate can coordinate a Mg^{2+} ion in the binding site, thus completing its coordination sphere (see the zoom in **Figure IV.2 D**). In all cases, the glutamate maintains its interaction with the ion during the whole simulations, ensuring that the triple helices remained bound to the domain. The fact that the two CMPs did not disassemble from the MIDAS shows that the full AA sequence "GFOGER" is not mandatory for the binding. This agrees with several studies suggesting that recognition could occur with other, similar motifs of the general type GXX'GEX'', although with generally lower affinity.[23] Our CMPs contain the $(PEG)_2$ triplets which, if written reversely as (GEPGEP), match this binding pattern. There is not a unique sequence that fits the binding site, but a common feature of the ligands is the presence of a glutamate moiety. For comparison, we realized simulations on pure L- or D-$(PPG)_{10}$ triple helices, which lack the (PEG) triplets, in interaction with the I-domain. The peptides were markedly less stabilized, displaying very few intermolecular hydrogen bonding interactions (see **Figure IV.5**), with possible complete unbinding. In line

with our results, the absence of binding of $(PPG)_{10}$ sequences to the similar I-domain of the $\alpha_2\beta_1$ integrin was demonstrated experimentally by others.[23] Our simulations thus highlight the crucial role of the glutamate moiety, without which the recognition seems unlikely. This observation confirms the findings of other studies.[16] Even a single AA mutation, replacing the glutamate by an aspartate, which is only one methylene shorter, strongly weakens the binding.[17,24]

IV-A.4. Interactions with the integrin are perturbed by the presence of a chiral mismatch

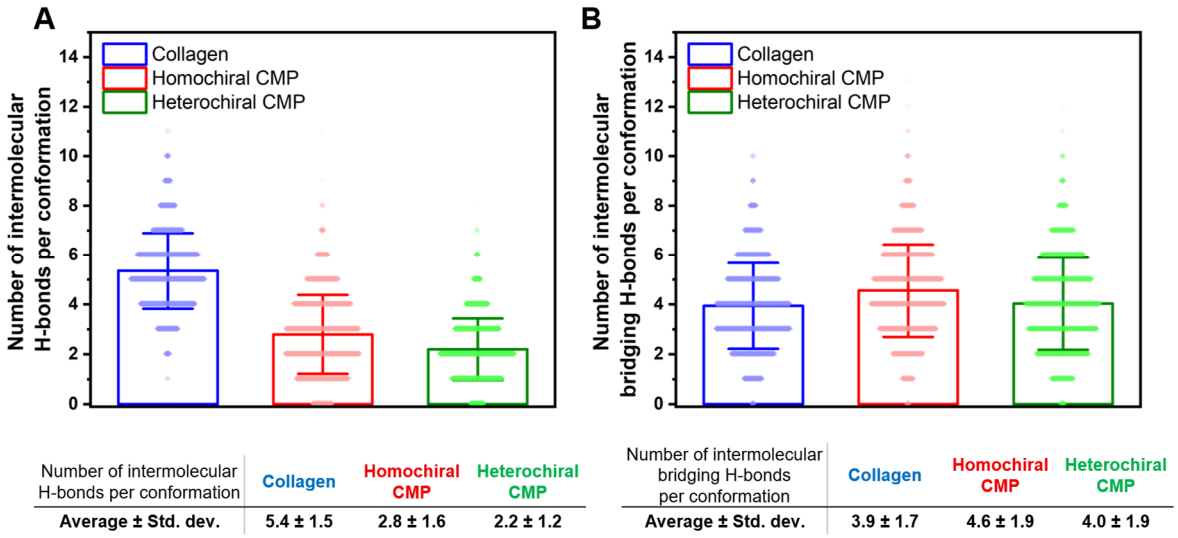
Based on the previous observations, the presence of D-AAs causing a chiral mismatch in the heterochiral CMP does not seem to completely prevent the interaction with integrin, thanks to the glutamate residue. This is not unexpected, as cells are able to adhere and spread on this substrate, although with less efficiency than on its homochiral counterpart or on collagen. The stereochemical modification seems to act more as a modulation of the interactions than as an ON/OFF switch. We therefore investigated the peptide-integrin interactions during the simulations in more details. The anchoring of the peptide in the binding site was evaluated by measuring the contact surface between the triple helix and the I-domain **(Figure IV.3)**.



Contact surface (Å²)	Collagen	Homochiral CMP	Heterochiral CMP
Average ± Std. dev.	656 ± 75	602 ± 66	581 ± 91

Figure IV.3. Measurements of the contact surface between the peptides and the I-domain of the integrin. **(A)** Equation used to compute the contact surface, and surface representation of the different species involved. **(B)** Distribution of the contact surface values for each system. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. The average values and standard deviations are given in the table.

The heterochiral CMP displays the lowest value, although the three peptides show similar levels overall, which is consistent with the fact that binding is maintained in all cases. Similarly, the number of intermolecular hydrogen bonds, both direct and water-mediated (so-called "bridging" H-bonds, *i.e.* between two species via a water molecule) was found to be slightly higher for the homochiral CMP than for its heterochiral counterpart **(Figure IV.4)**.



Number of intermolecular H-bonds per conformation	Collagen	Homochiral CMP	Heterochiral CMP
Average ± Std. dev.	5.4 ± 1.5	2.8 ± 1.6	2.2 ± 1.2

Number of intermolecular bridging H-bonds per conformation	Collagen	Homochiral CMP	Heterochiral CMP
Average ± Std. dev.	3.9 ± 1.7	4.6 ± 1.9	4.0 ± 1.9

Figure IV.4. Number of **(A)** intermolecular H-bonds and **(B)** intermolecular bridging H-bonds per conformation measured during the whole simulations between the three peptides and the I-domain of the integrin. Data is shown as mean ± standard deviation. The pale lines represent the distribution of the measurements during the whole simulations. Statistics are given in the tables below.

Collagen is involved in more intermolecular H-bonds, which is expected due to the presence of the full GFOGER binding motif, as well as the presence of other hydroxyproline units in the collagen sequence. In contrast, the CMPs can only form H-bonds through their glutamate residues and the backbone amide bonds. To localize the AAs of the I-domain interacting with the peptides, a heatmap of the intermolecular H-bonds was drawn **(Figure IV.5)**. On the x-axis on the heatmap are represented the AAs of the binding site involved in H-bonds with the peptides, while the different peptides studied are displayed on the y-axis. At the crossing of the axes are found colored rectangles, whose color indicates the number of H-bonds detected between a peptide and the corresponding AA of the binding site. In addition to collagen and the two CMPs, the heatmap features the pure L- and D-$(PPG)_{10}$ chains. The first striking observation is that the $(PPG)_{10}$ peptides form very few interactions, regardless of their stereochemistry. The addition of the glutamate units in the CMPs significantly increases their anchoring in the binding site, thus their number of intermolecular H-bonds. The pattern of interactions of the CMPs approaches that of collagen, although

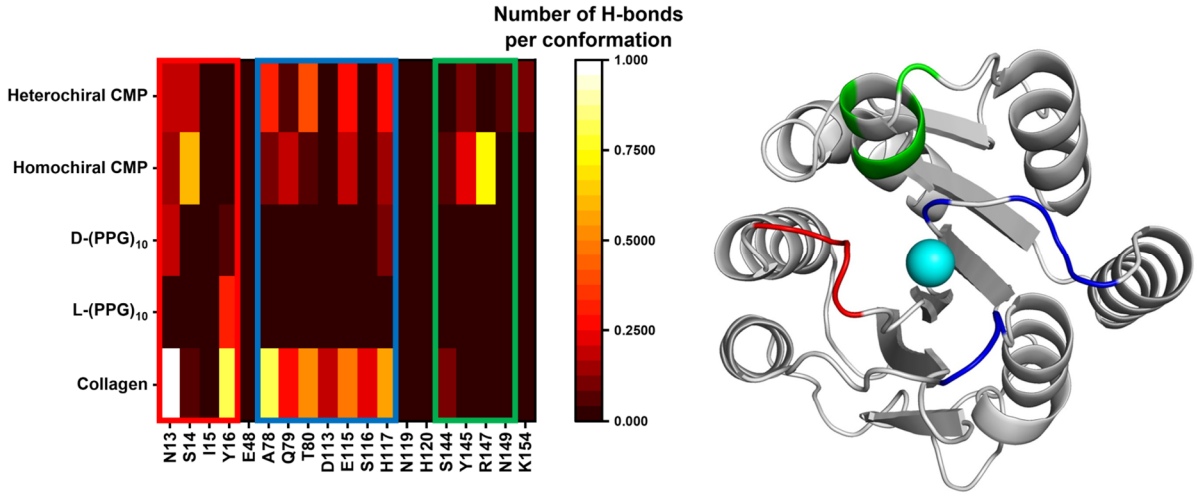


Figure IV.5. Heatmap of the intermolecular H-bonds between the five simulated peptides and the I-domain of the integrin (left), and top view of the I-domain (right). The heatmap displays the AAs of the binding site (represented by their one-letter code and their residue number) on its x-axis, and the five peptides on its y-axis. At the crossing of the x and y axes is found a rectangle, whose color indicates the frequency of the intermolecular H-bonds (the brightest the color, the more frequent the interaction). The red, blue and green rectangles on the heatmap highlight the position of the AAs in the binding site (see associated colored areas on the top view of the I-domain).

there are some differences. Three groups of AAs can be distinguished, based on their position in the binding site (see red, blue and green colored rectangles in the heatmap and associated areas on the snapshot in **Figure IV.5**). The heterochiral CMP essentially interacts with AAs located in the blue area, and performs much fewer H-bonds with the other parts of the I-domain. Furthermore, it does not feature very persistent interactions, its most frequent hydrogen bond (with tyrosine T80) occurring less than 40 % of the time. In contrast, the homochiral CMP shows interactions in all three areas and forms persistent interactions with a serine (S14) and an arginine (R147). Interestingly, this H-bond is located in an area where the collagen does not really interact. This could suggest that the CMP is stabilized by interactions with other AAs than natural collagen, despite their identical stereochemistry. This is reasonable, as collagen mainly interacts through its polar side-chains, while the CMPs essentially contain apolar substituents (glycine and proline residues), thus mainly interact through their backbone amides and free glutamate residues. Therefore, the similar but not identical behavior of cells on collagen and the homochiral substrates observed experimentally could be explained by such differences in the sequence of AAs, the CMP containing the necessary units to interact, but finding other mechanisms to stabilize in the binding site of the integrin. Additionally, the homochiral CMP displays a very persistent intermolecular bridging H-bond, occurring between one of its free glutamate

and another glutamate residue on the I-domain **(Figure IV.6)**. This bond is found more than 90 % of the time, and could be another interaction helping the homochiral CMP to stabilize in the binding site. In comparison, the most persistent bridging intermolecular H-bonds for collagen and the heterochiral CMP are found less than 50 % of the time.

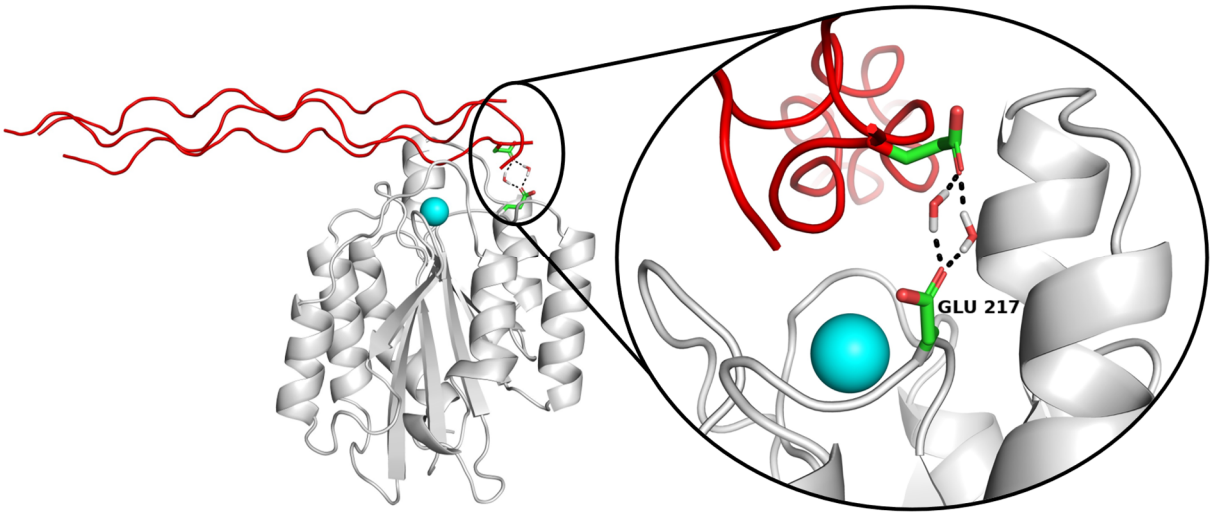


Figure IV.6. Final snapshot of the MD simulation of the homochiral CMP, showing its most persistent bridging intermolecular H-bond, which occurs between one of its glutamate residue and another glutamate (GLU 217) exposed on the I-domain, through interaction with two water molecules. H-bonds are represented as black dots.

Overall, the heterochiral CMP displays less interactions with the binding site than its homochiral counterpart, and more importantly, does not feature persistent interactions, in marked contrast to the homochiral CMP.

Based on these atomic-scale investigations, our hypothesis to explain the less efficient interactions observed between the heterochiral CMP and the integrin is related to an increased internal flexibility caused by the chiral mismatch in this peptide. The stereochemical inversion of the AAs in the $(PPG)_5$ section induces the formation of a left-handed triple helix, while the $(PEG)_2$ segment, composed of L-prolines, cannot follow this handedness. This brings conformational disorder in the supramolecular assembly of the triple helix at the junction between the L- and D-AAs, located right in the binding site. This disorder is detrimental to the recognition, especially because the triple helix structure was deemed essential for the binding, as GFOGER-containing peptides lacking a triple helix structure were shown to be unable to interact with the I-domain of $\alpha_1\beta_1$ or support cell adhesion.[17] The higher flexibility of the $(PEG)_2$ part in the heterochiral CMP, compared to the homochiral CMP and collagen, is shown by RMSF measurements **(Figure IV.7)**. In contrast, the $(PPG)_5$ part is similarly stable in both CMPs, indicating that stereoinversion alone does not compromise the integrity of the triple helix, as destabilization is localized at the chiral mismatch.

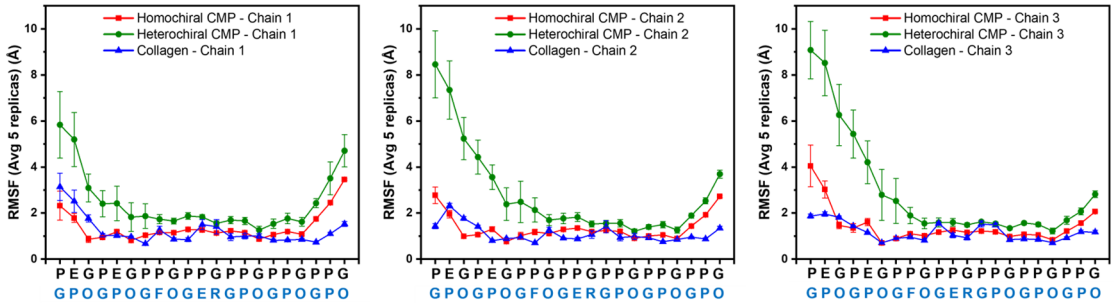


Figure IV.7. RMSF measurements for each AA in the first (left), second (middle) and third (right) chains forming the triple helix, for all systems. The higher the RMSF value, the more flexible is the residue. The sequences of the CMPs and collagen are written in black and blue on the x-axis, respectively.

This can be observed on the superimposed first and final snapshots obtained for all peptides, presented in **Figure IV.1 D**, where the $(PEG)_2$ part of the heterochiral CMP is visibly disordered. This lack of stability is also reflected by the number of H-bonds inside the supramolecular triple helix, between the three peptide chains, measured for all systems **(Figure IV.8)**. In the last 12 AAs, *i.e.* inside the $(PPG)_5$ triple helix, the chiral inversion does not seem to impact the network of H-bonds (around 6 H-bonds per conformation for both CMPs). However, the heterochiral CMP displays much less interactions within its first nine AAs (around 1 H-bond per conformation), *i.e.* in the $(PEG)_2$ part, compared to the homochiral CMP (around 4 H-bonds per conformation), see **Figure IV.8 C**. The increased flexibility of the heterochiral CMP and the partial loss of the triple helix conformation could explain the less efficient interactions with the integrin, thus the lower affinity of cells for this substrate. As cells have less adhesions with this matrix, their ability to exert contractile forces necessary for their displacement will be reduced, leading to a slower migration speed.

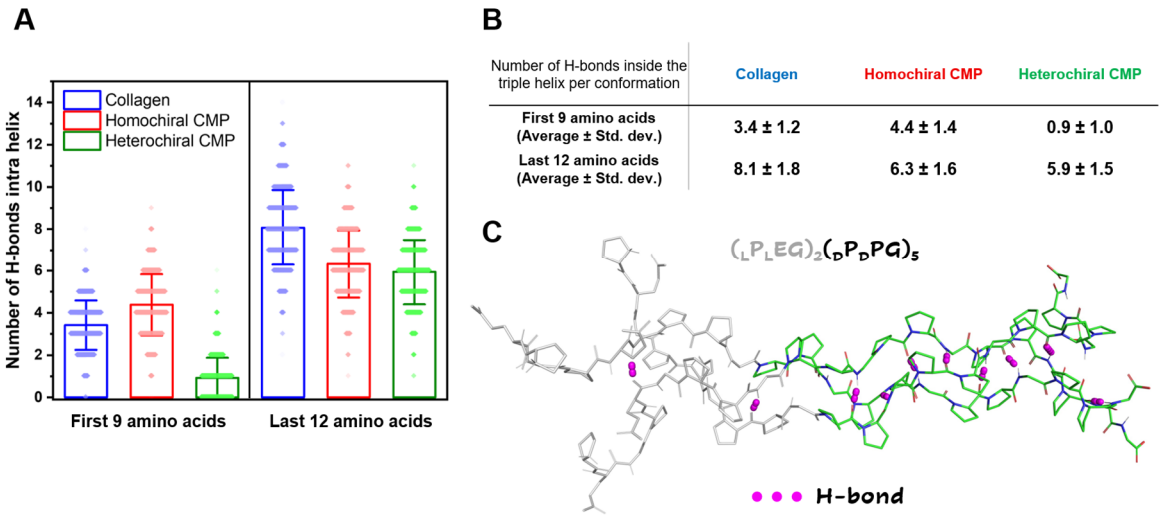


Figure IV.8. Estimates of the number of H-bonds inside the supramolecular triple helix. **(A)** Number of H-bonds between the first nine (left) and last twelve (right) AAs of the three peptide strands inside the triple helix for collagen (blue), homochiral CMP (red), and heterochiral CMP (green). Data is shown as mean ± standard deviation. The pale lines represent the distribution of the measurements during the whole simulations. **(B)** Table summarizing the data. **(C)** Final MD snapshot of the heterochiral CMP showing that H-bonds (displayed as pink dots) are maintained in the $(PPG)_5$ part, while they vanish in the more disordered $(PEG)_2$ part (represented in gray).

IV-A.5. Conclusion

In agreement with the experimental results, our simulations indicate that the heterochiral CMP is not able to interact with the $\alpha_1\beta_1$ integrin as efficiently as natural collagen and the homochiral CMP. We attribute this behavior to the chiral mismatch present in the sequence of the heterochiral CMP, where the triple helix conformation is disorganized at the junction between the L- and D-AAs. This could explain the lower number of cell-matrix adhesions on the heterochiral substrate, thus resulting in slower migration. It is remarkable that a small stereochemical perturbation has such an impact on the supramolecular assembly of two peptides sharing the exact same sequence of monomers, leading to major differences in their interactions with important cellular receptors. Our simulations also confirmed the major role of the glutamate moiety, without which binding in the MIDAS cavity is not possible.

Of course, our simulations only capture a minor fraction of the complexity of cell migration, which involves much more interaction partners and components of the ECM, and cannot entirely explain the differences between the two CMPs. However, we believe that the information brought by the MD simulations is particularly valuable to better understand the impact of small changes in the primary structure on (supra)molecular conformations and cellular processes. This work highlights clearly

the sensitivity of cells to their environment, and how modifications at the atomic scale can lead to perturbations in cell behavior. Chirality is an important factor that researchers can use to modulate the physicochemical properties of the ECM and better understand cell behavior in response to perturbations in their environment.

IV-A.6. Simulation protocol

MD simulations were carried out with the AMBER package.[22] The structures of the I-domain of the $\alpha_1\beta_1$ integrin and of the $(PPG)_{10}$ triple helix were directly taken from the Protein Data Bank (PDB), with PDB ID: 1qcy[25] and 1k6f,[15,25] respectively. To build the collagen-mimetic peptides $(PEG)_2$-$(PPG)_5$, the backbone of the $(PPG)_{10}$ triple helix was reproduced, and its length and AA composition was subsequently adapted with the *LEaP* module of AMBER. The D-enantiomers were obtained by creating the mirror image of the L-enantiomers using *LEaP*. The structure of collagen was extracted from a crystal structure of its complex with the I-domain of the $\alpha_2\beta_1$ integrin (PDB ID: 1dzi).[19] This binding mode was reproduced for the simulations of collagen and the CMPs in interaction with the I-domain of $\alpha_1\beta_1$. This assumption seems reasonable, given the high structural similarity between the I-domains of $\alpha_1\beta_1$ and $\alpha_2\beta_1$.[26] Collagen, the homochiral and heterochiral CMPs, the L- and D-$(PPG)_{10}$ triple helices and the I-domain of $\alpha_1\beta_1$ were described with the ff19SB force field.[27] The peptide – I-domain complexes were solvated in explicit water boxes, using the 4-point OPC water model.[28] NaCl ions were added at a concentration of 0.15 M, using the "SPLIT" method.[29] The simulations started with a geometry optimization performed by molecular mechanics to get a stable starting structure. A first phase served to stabilize the solvent molecules and the Na^+ and Cl^- ions, which underwent 1,000 steps of steepest descent followed by 9,000 steps of conjugated gradient, with restraints on the solute atoms. The second phase of optimization was performed with the same protocol, without any constraints. Next, a heating step of 2 ns was performed in the NVT ensemble. The system was brought to a temperature of 300 K in 1 ns, and was maintained at this temperature for a further 1 ns (and for the rest of the simulation) using a Langevin thermostat, with a collision frequency of 1 ps^{-1}. Positional restraints were applied to all solute atoms during heating, with a force-constant of 10 $kcal.mol^{-1}.Å^{-2}$. Then, the system was equilibrated during 10 ns in the NPT ensemble at a pressure of 1 bar using a Monte Carlo barostat, with a pressure relaxation time of 2 ps. Finally, the production phase of 1 μs was launched in the NPT ensemble. Five independent replicas were launched for each peptide – I-domain complex, starting from the same structure optimized by molecular mechanics. A timestep of 2 fs was used with the SHAKE algorithm to constrain bonds involving hydrogen atoms. A cutoff of 12.0 Å was used for non-bonded interactions and the particle mesh Ewald method was used to

treat long-range electrostatic interactions. A snapshot was extracted each ns of the production phase for further analyses (5,000 conformations for each system when considering the five replicas). The *cpptraj* module of AMBER and in-house scripts were used to analyze the simulations.[30] The solvent-accessible surface area (SASA) values were computed using the LCPO algorithm, using a van der Waals radius of 1.4 Å for the solvent probe. These values were injected in the equation shown in **Figure IV.3 A** to determine the contact surfaces. RMSF values were computed for each amino acid of the triple helices, after removal of the translational and rotational movements. Hydrogen bonds were detected using geometric criteria: the distance between the acceptor and the donor heavy atom must be ≤ 3.0 Å, and the angle between the donor, the hydrogen atom and the acceptor must be ≥ 135°. The *PyMOL 2.5.4* software was used to produce the snapshots.[31] The solvent water molecules were hidden on the snapshots, for the sake of clarity. Statistics given in the tables are always calculated on the 5,000 conformations for each system.

References

[1] W. Zhang, H. T. Liu. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* **2002**, *12*, 9–18.

[2] P. Sassone-Corsi. The Cyclic AMP Pathway. *Cold Spring Harbor Perspect. Biol.* **2012**, *4*, a011148.

[3] B. N. Kholodenko. Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 165–176.

[4] M. Riaz, M. Versaevel, D. Mohammed, K. Glinel, S. Gabriele. Persistence of fan-shaped keratocytes is a matrix-rigidity-dependent mechanism that requires α5β1 integrin engagement. *Sci. Rep.* **2016**, *6*, 34141.

[5] T. Tzvetkova-Chevolleau, A. Stéphanou, D. Fuard, J. Ohayon, P. Schiavone, P. Tracqui. The motility of normal and cancer cells in response to the combined influence of the substrate rigidity and anisotropic microstructure. *Biomaterials* **2008**, *29*, 1541–1551.

[6] M. Luciano, S.-L. Xue, W. H. De Vos, L. Redondo-Morata, M. Surin, F. Lafont, E. Hannezo, S. Gabriele. Cell monolayers sense curvature by exploiting active mechanics and nuclear mechanoadaptation. *Nat. Phys.* **2021**, *17*, 1382–1390.

[7] G. Ciccone, M. Azevedo Gonzalez-Oliva, M. Versaevel, M. Cantini, M. Vassalli, M. Salmeron-Sanchez, S. Gabriele. Epithelial Cell Mechanoresponse to Matrix Viscoelasticity and Confinement Within Micropatterned Viscoelastic Hydrogels. *Adv. Sci.* **2025**, *12*, 2408635.

[8] Y. H. Tee, T. Shemesh, V. Thiagarajan, R. F. Hariadi, K. L. Anderson, C. Page, N. Volkmann, D. Hanein, S. Sivaramakrishnan, M. M. Kozlov, A. D. Bershadsky. Cellular chirality arising from the self-organization of the actin cytoskeleton. *Nat. Cell Biol.* **2015**, *17*, 445–457.

[9] A. S. Chin, K. E. Worley, P. Ray, G. Kaur, J. Fan, L. Q. Wan. Epithelial Cell Chirality Revealed by Three-Dimensional Spontaneous Rotation. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 12188–12193.

[10] G. Lebreton, C. Géminard, F. Lapraz, S. Pyrpassopoulos, D. Cerezo, P. Spéder, E. M. Ostap, S. Noselli. Molecular to organismal chirality is induced by the conserved myosin 1D. *Science* **2018**, *362*, 949–952.

[11] J. Khoshnoodi, J.-P. Cartailler, K. Alvares, A. Veis, B. G. Hudson. Molecular Recognition in the Assembly of Collagens: Terminal Noncollagenous Domains Are Key Recognition Modules in the Formation of Triple Helical Protomers. *J. Biol. Chem.* **2006**, *281*, 38117–38121.

[12] M. D. Shoulders, R. T. Raines. Collagen Structure and Stability. *Annu. Rev. Biochem.* **2009**, *78*, 929–958

[13] M. Barczyk, S. Carracedo, D. Gullberg. Integrins. *Cell Tissue Res.* **2010**, *339*, 269–280.

[14] A. Remson, D. Dellemme, M. Luciano, M. Surin, S. Gabriele. Chiral mismatch in collagen-mimetic peptides modulates cell migration through integrin-mediated molecular recognition. Biorxiv, **2024**, 604866, *author preprint*.

[15] R. Berisio, L. Vitagliano, L. Mazzarella, A. Zagari. Crystal structure of the collagen triple helix model $[(\text{Pro-Pro-Gly})_{10}]_3$. *Protein Sci.* **2002**, *11*, 262–270.

[16] J. Bella, H. M. Berman. Integrin–collagen complex: a metal–glutamate handshake. *Structure* **2000**, *8*, R121–R126.

[17] C. G. Knight, L. F. Morton, A. R. Peachey, D. S. Tuckwell, R. W. Farndale, M. J. Barnes. The Collagen-binding A-domains of Integrins α1β1 and α2β1 Recognize the Same Specific Amino Acid Sequence, GFOGER, in Native (Triple-helical) Collagens. *J. Biol. Chem.* **2000**, *275*, 35–40.

[18] M. Tulla, O. T. Pentikäinen, T. Viitasalo, J. Käpylä, U. Impola, P. Nykvist, L. Nissinen, M. S. Johnson, J. Heino. Selective Binding of Collagen Subtypes by Integrin α1I, α2I, and α10I Domains. *J. Biol. Chem.* **2001**, *276*, 48206–48212.

[19] J. Emsley, C. G. Knight, R. W. Farndale, M. J. Barnes, R. C. Liddington. Structural Basis of Collagen Recognition by Integrin α2β1. *Cell* **2000**, *101*, 47–56.

[20] L. M. Moir, J. L. Black, V. P. Krymskaya. TSC2 modulates cell adhesion and migration via integrin-α1β1. *Am. J. Physiol.: Lung Cell. Mol. Physiol.* **2012**, *303*, L703–L710.

[21] E. C. Reilly, K. Lambert Emo, P. M. Buckley, N. S. Reilly, I. Smith, F. A. Chaves, H. Yang, P. W. Oakes, D. J. Topham. T_{RM} integrins CD103 and CD49a differentially support adherence and motility after resolution of influenza virus infection. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 12306–12314.

[22] D. A. Case, T. E. Cheatham Iii, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[23] N. Raynal, S. W. Hamaia, P. R.-M. Siljander, B. Maddox, A. R. Peachey, R. Fernandez, L. J. Foley, D. A. Slatter, G. E. Jarvis, R. W. Farndale. Use of Synthetic Peptides to Locate Novel Integrin α2β1-binding Motifs in Human Collagen III. *J. Biol. Chem.* **2006**, *281*, 3821–3831.

[24] W.-M. Zhang, J. Käpylä, J. S. Puranen, C. G. Knight, C.-F. Tiger, O. T. Pentikäinen, M. S. Johnson, R. W. Farndale, J. Heino, D. Gullberg. α11β1 Integrin Recognizes the GFOGER Sequence in Interstitial Collagens. *J. Biol. Chem.* **2003**, *278*, 7270–7277.

[25] Y. Nymalm, J. S. Puranen, T. K. M. Nyholm, J. Käpylä, H. Kidron, O. T. Pentikäinen, T. T. Airenne, J. Heino, J. P. Slotte, M. S. Johnson, T. A. Salminen. Jararhagin-derived RKKH Peptides Induce Structural Changes in α1I Domain of Human Integrin α1β1. *J. Biol. Chem.* **2004**, *279*, 7962–7970.

[26] M. Nolte, R. B. Pepinsky, S. Yu. Venyaminov, V. Koteliansky, P. J. Gotwals, M. Karpusas. Crystal structure of the α1β1 integrin I-domain: insights into integrin I-domain function. *FEBS Lett.* **1999**, *452*, 379–385.

[27] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, C. Simmerling. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528–552.

[28] S. Izadi, R. Anandakrishnan, A. V. Onufriev. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **2014**, *5*, 3863–3871.

[29] M. R. Machado, S. Pantano. Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 1367–1372.

[30] D. R. Roe, T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

[31] Schrödinger LLC, *The PyMOL Molecular Graphics System, Version 2.5.4*, **2015**.

IV-B. Selectivity in the chiral self-assembly of nucleobase-arylazopyrazole photoswitches along DNA templates

Part of this work is reported in: *Selectivity in the chiral self-assembly of nucleobase-arylazopyrazole photoswitches along DNA templates.*
N. Nogal, S. Guisán, D. Dellemme, M. Surin, A. de la Escosura, *J. Mater. Chem. B*, **2024**, 12, 3703-3709.

IV-B.1. Introduction

The possibility to control the nucleobase sequence of DNA allows the emergence of programmable assemblies, such as DNA origamis and various 2D and 3D nanostructures.[1–4] Its controllable architecture can be functionalized with the covalent attachment of synthetic components, to expand the range of applications.[5,6] The grafted units can promote the formation of networks of non-covalent interactions, allowing the formation of complex supramolecular assemblies involving DNA strands.[7–13] The programmable structure of DNA can also be exploited without modifications, to template the organization of molecules at the nanoscale. These pure supramolecular approaches are attractive, as they make use of readily available DNA strands, avoiding the challenges associated with the synthesis of modified polynucleotides. For instance, DNA can guide the supramolecular assembly of chromophores, modulating donor-acceptor coupling to control energy transfer.[14–16] Specific secondary structures of DNA can be targeted by tailor-made ligands, in view of biosensing applications.[17–23] The templating effect of DNA can also be used to direct supramolecular polymerization, to preorganize monomers before their coupling into a covalent polymer of defined sequence, or to build various highly controlled nanostructures.[1,24–28] Another advantage of using supramolecular interactions is their dynamic and reversible nature, meaning that external stimuli can be exploited to design responsive and adaptable systems. Light is a particularly attractive stimulus, providing fine spatial and temporal resolution without introducing contaminants in the system. Light-responsive molecules able to reversibly change their configuration upon irradiation, such as azobenzene derivatives, have emerged as promising systems for a wide range of applications, including energy harvesting, catalysis or bioimaging.[29–32] Combining light-responsive components with programmable DNA templates opens new avenues for the development of functional nanomaterials.[33,34] Several examples have demonstrated the possibility to control the supramolecular assembly of photoswitchable ligands with DNA and to selectively stabilize specific DNA conformations using light.[35–37] For instance, an azobenzene derivative was

incorporated inside a gene to regulate its expression *in vitro*.[38] In the *trans* configuration, the molecule intercalates into the DNA double helix, blocking RNA polymerase binding and thereby inhibiting transcription. Upon photoisomerization to the *cis* configuration, transcription resumes, demonstrating the possibility of temporally regulating gene expression. More recently, a short oligothymine single-strand was combined with a complementary photoswitchable molecule bearing two adenine bases to develop a photo-responsive hydrogel.[39] The molecule can form H-bonds with two ssDNAs simultaneously, leading to the formation of a network composed of large twisted fiber bundles after several weeks of equilibration. Upon irradiation with UV light, which alters the photoswitch conformation, local shrinking of the hydrogel was observed on the illuminated area. These examples, among many others, demonstrate the interest of combining DNA and stimuli-responsive components to develop adaptable nanomaterials.

In our work, novel photoswitches based on an arylazopyrazole unit were designed. Arylazopyrazoles are easier to photoisomerize and exhibit greater thermal stability than azobenzenes.[40,41] Our compounds are decorated with a nucleobase, either thymine or adenine, with the goal to assess their supramolecular organization templated by complementary oligonucleotides. This design is motivated by previous works that made use of short ssDNAs, such as oligoadenine (dA_n) or oligothymine (dT_n), to template the self-assembly of molecules bearing a complementary recognition unit.[25,39,42–44] Chiroptical spectroscopy experiments revealed that both arylazopyrazoles can bind to their complementary DNA strand and adopt a chiral organization in their *trans* configuration, while partial disassembly occurs upon photoisomerization into the *cis* configuration **(Figure IV.9)**. Molecular dynamics (MD) simulations were carried out to shine light on the binding modes of the arylazopyrazoles in their *trans* and *cis* configurations with their DNA partner. Our results show that the *trans* form allows the emergence of stabilizing π-type interactions between the molecules wrapped around DNA, which helps to maintain the supramolecular assembly. In the *cis* configuration, these interactions are partially lost, leading to the formation of more disordered aggregates and less persistent H-bonds with the template. All the experimental results presented in this chapter were obtained by Noemí Nogal,[45] and the compounds were synthesized by Santiago Guisán, in the frame of a collaboration with the group of Dr. A. de la Escosura at Universidad Autónoma de Madrid.

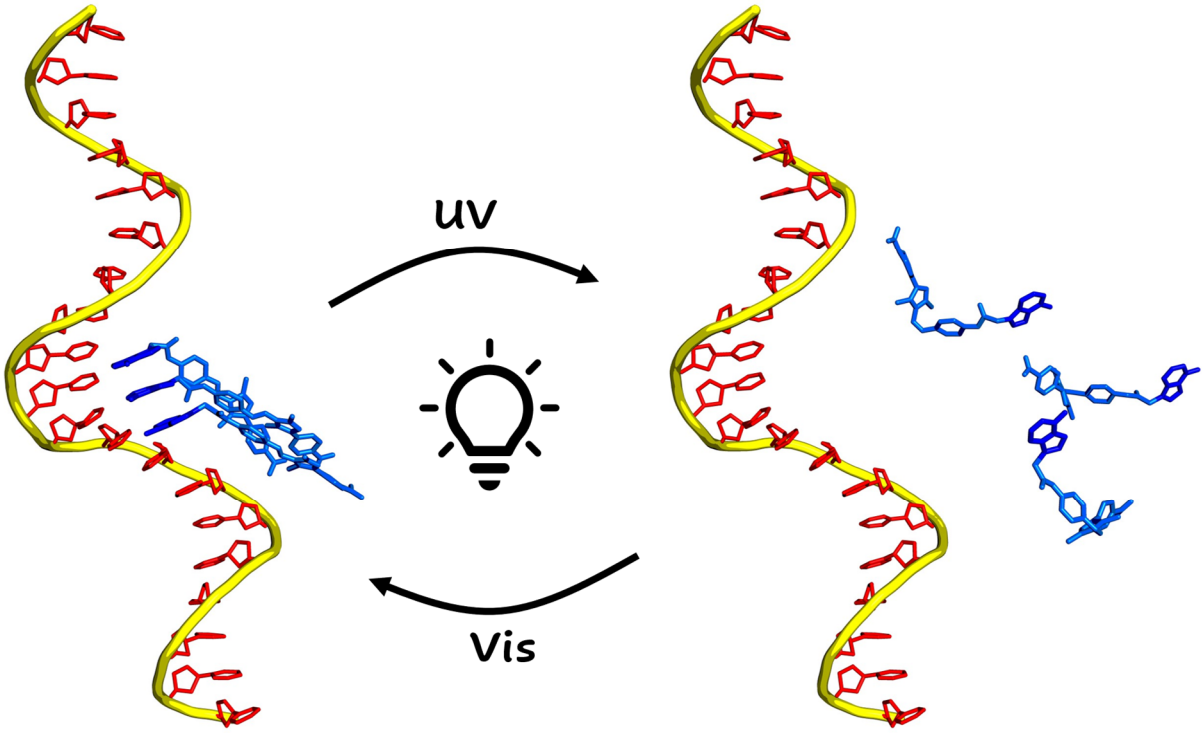


Figure IV.9. Cartoon representation of the investigated systems. Partial disassembly occurs upon irradiation with UV light and isomerization of the molecules into their *cis* configuration. Ligands are colored in blue, while DNA nucleosides are shown in red. The DNA backbone is represented as a yellow tube.

IV-B.2. Design of the ligands and reparametrization of the force field for the MD simulations

Two arylazopyrazole derivatives were designed, having the same conjugated region and functionalized with either the adenine (Azo-A) or the thymine (Azo-T) nucleobase **(Figure IV.10 A)**. Upon irradiation with UV light (365 nm), these molecules undergo *trans* to *cis* photoisomerization. The reversible reaction occurs upon irradiation with visible light (465 nm). The photoswitches are combined with a complementary ssDNA, with which they are able to assemble through hydrogen bonding interactions **(Figure IV.10 B)**. MD simulations were performed using the AMBER package on assemblies of 10 arylazopyrazole molecules (Azo-A or Azo-T) with their complementary oligonucleotide template of 20 nucleobases (dT_{20} or dA_{20}), reproducing the experimental stoichiometry, in explicit water boxes containing Na^+ and Cl^- ions.[46] The carboxylate group at the end of each azo compound was modeled in its deprotonated form, as the pKa value of benzoic acid is around four.[47] Independent simulations were run for each assembly (Azo-A/dT_{20} and Azo-T/dA_{20}), with all azo compounds either in the *trans* or *cis* configuration preorganized along the DNA template. Two replicas were run for each condition, giving a total of eight simulations. Complementary H-bonds

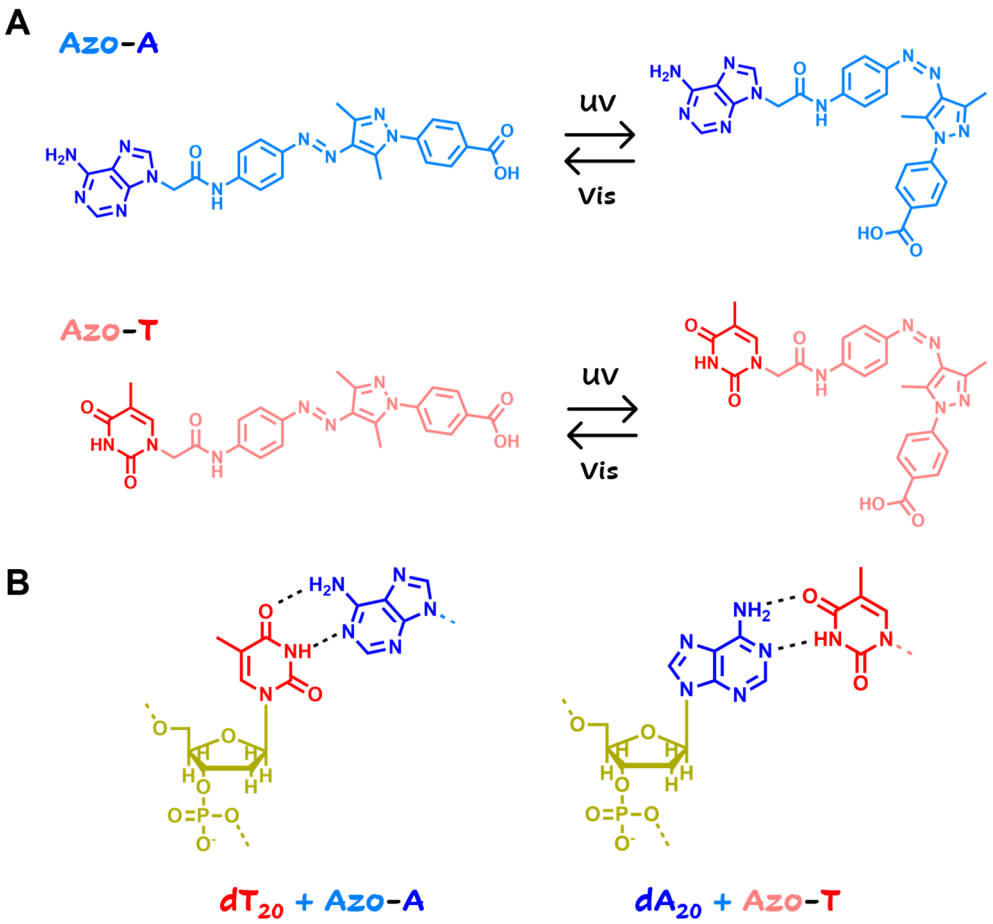


Figure IV.10. Chemical structures of the systems studied. **(A)** Structure of Azo-A and Azo-T and illustration of their photoswitchable *trans* and *cis* configurations, upon irradiation with visible (465 nm) or UV (365 nm) light. Adenine and thymine nucleobases are represented in blue and red, respectively, and the conjugated region is depicted in lighter colors. **(B)** Representation of the supramolecular assemblies studied, involving oligonucleotides comprising 20 nucleobases (dT_{20} or dA_{20}) and the complementary arylazopyrazole molecule.

between the nucleobases (as represented in **Figure IV.10 B**) were constrained for 250 ns to allow equilibration of the supramolecular complex without ligand unbinding, followed by 1 µs of unrestrained simulation (see full details of the protocol in **Section IV-B.6**).

Given the particular structure of the azo compounds, which contain an extended conjugated region including several heteroatoms, we carried out a reparametrization of the torsional parameters of the GAFF 2.11 force field **(Figure IV.11)**. The reparametrization was done with the *mdgx* module implemented in AMBER. In short, two fragments containing the dihedral angles of interest, Φ_1 to Φ_4, were built **(Figure IV.11 A)**. Hundreds of conformers were generated and optimized with the default GAFF 2.11 parameters to sample the torsions in the interval [-180 ° ; 180 °]. Then, the energy of these conformers was obtained at the MM and QM levels, with the QM energy

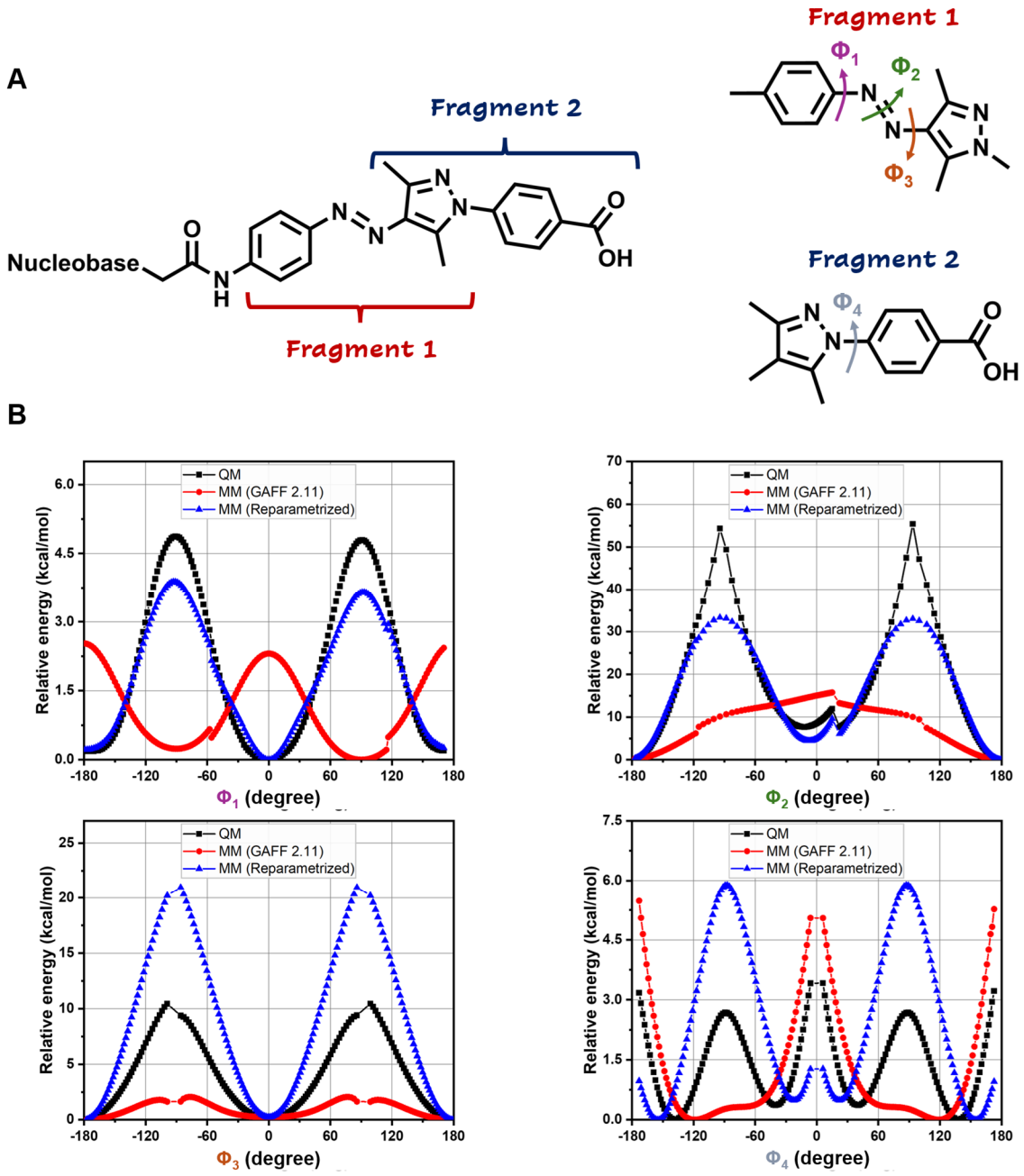


Figure IV.11. Reparametrization of the four dihedral angles, Φ_1 to Φ_4. **(A)** Chemical structure of the two fragments used for the reparametrization. **(B)** Relative potential energy curves, in kcal/mol, as a function of the dihedral angle for the four torsions. The energy curves obtained at the QM level, with the default GAFF 2.11 parameters and the modified parameters are shown in black, red and blue, respectively.

serving as the reference data. Finally, new dihedral parameters were generated in order to reduce the gap between the MM and QM energies. The results of the reparametrization of the four dihedral angles are displayed in **Figure IV.11 B**. The curves generated with the refined force field parameters (in blue in **Figure IV.11 B**) are not perfectly matching the QM reference curves (in black in **Figure IV.11 B**),

especially for the torsion Φ_4. The height of the barrier is also too high for Φ_3. However, the new profiles represent a considerable improvement compared to the curves obtained with the initial force field parameters (in red in **Figure IV.11 B**), coming from GAFF 2.11. It reveals significant flaws in this force field's ability to accurately reproduce the conformations of extended conjugated systems, especially when they involve heteroatoms and combinations of aromatic cycles and double bonds. In particular, the parameters describing Φ_1 were completely erroneous, with minima and maxima of the potential energy surface inverted compared to the QM reference. Although not perfect, our modifications will ensure that the conjugated region keeps its planarity and will prevent spontaneous *trans-cis* isomerization (with a barrier of more than 30 kcal/mol between the *trans* and *cis* states, see Φ_2 curves), which is the expected behavior according to the QM torsional profiles. Further refinement of these parameters may be needed in view of more sophisticated analyses sensitive to small conformational changes, such as the simulation of CD spectra. However, these approaches were not undergone in this work and are envisioned as perspectives: preliminary attempts have been realized and will be discussed in **Section IV-B.5**.

IV-B.3. DNA templating organizes the stacking of the *trans* isomers and requires high ionic strength

Experimental CD spectra indicate that both azo compounds in their *trans* configuration are able to interact with their complementary DNA template **(Figure IV.12)**. This can be stated by the appearance of a strong induced CD (ICD) signal between around 325 and 450 nm. Only the achiral ligands absorb light in this region,

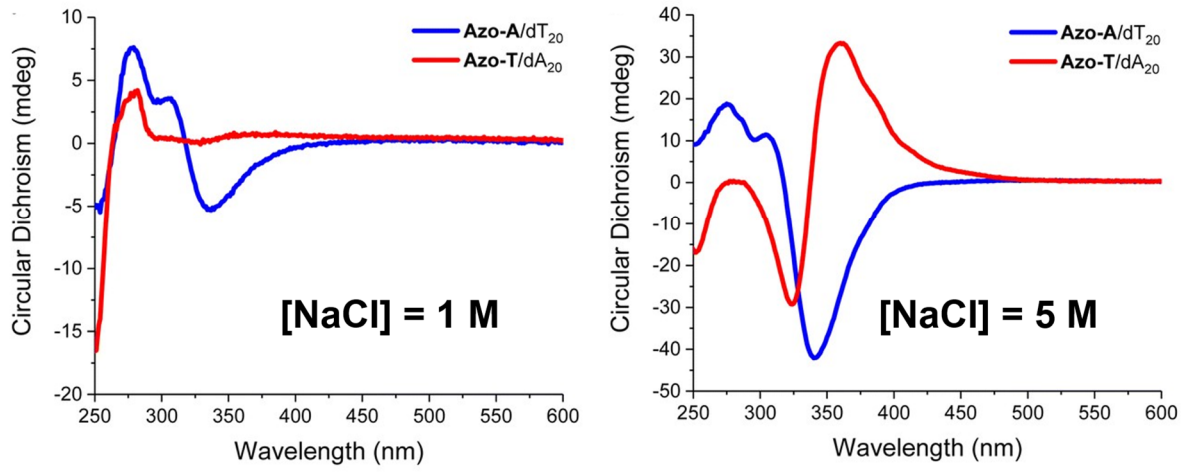
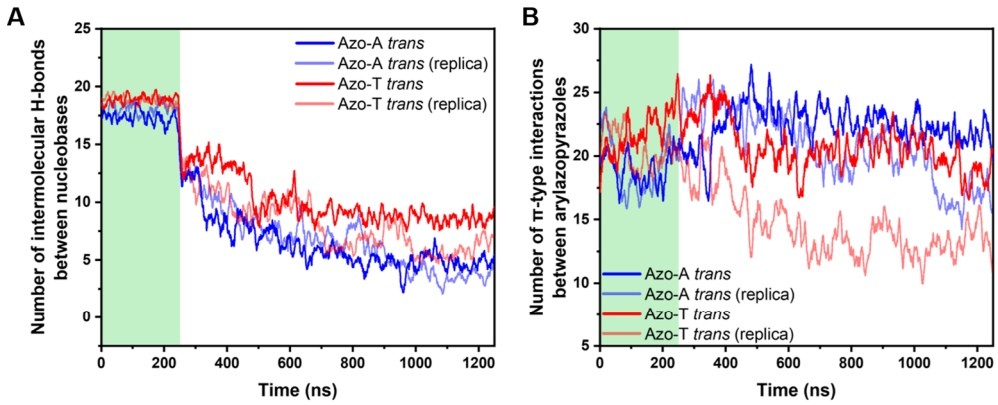


Figure IV.12. Experimental CD spectra of the Azo-A/dT_{20} and Azo-T/dA_{20} complexes, represented by the blue and red curves, respectively. The spectra were measured at NaCl concentrations of 1 M (left) and 5 M (right).

and they do not present a chiral signature alone in solution: this signal means that they acquire a chiral organization upon interaction with the ssDNA template. The chiroptical spectra provide two further valuable insights. Firstly, a high ionic strength promotes the supramolecular assembly, as shown by the significantly stronger intensity of the ICD signals at a NaCl concentration of 5 M compared to 1 M. Secondly, the signs of the ICD signals are opposite for the two azo compounds, which is surprising given that they are expected to bind to ssDNAs with the same helical sense. MD simulations were therefore carried out on assemblies of the azo compounds in their *trans* configuration with their complementary ssDNA, to better understand the interactions stabilizing the supramolecular complexes. In all cases, the number of H-bonds between the complementary nucleobases of the ligands and of their template instantaneously decreased after removal of the restraints **(Figure IV.13 A)**. The azo compounds quickly reorganized along the DNA strand, although some ligands are still H-bonded at the end of the simulation, with around 5 to 9 H-bonds remaining between complementary nucleobases. In marked contrast, the number of π-type interactions[1] stayed generally stable after removal of the restraints **(Figure IV.13 B)**.



C

= Constrained H-bonds (first 250 ns)

Average ± Std. dev. (last 500 ns)	Azo-A	Azo-A (replica)	Azo-T	Azo-T (replica)
H-bonds (nucleobases)	4.8 ± 1.8	4.7 ± 2.0	8.7 ± 1.8	6.3 ± 1.8
π-type interactions (arylazopyrazoles)	22.2 ± 2.4	18.9 ± 2.9	20.1 ± 2.6	13.2 ± 2.3

Figure IV.13. Interactions between the *trans* azo compounds and their complementary DNA template. **(A)** Evolution of the number of H-bonds between the complementary nucleobases of the arylazopyrazoles and the DNA template. **(B)** Evolution of the number of π-type interactions between the arylazopyrazoles. Blue and red curves represent the Azo-A/dT_{20} and Azo-T/dA_{20} systems, respectively. The running average including the five previous and five subsequent conformations is displayed, for ease of visualization. **(C)** Table summarizing the average number of interactions during the last 500 ns of the simulation.

[1] π-type interactions are counted between aromatic cycles following these geometric criteria: the distance between their centers of mass must be ≤ 5 Å, and the angle between their planes must be < 45 ° or > 135 °.

The organization of the *trans* azo compounds can be observed on the final snapshot of the MD simulation of the Azo-A/dT_{20} assembly, showing the molecules wrapped around the oligonucleotide **(Figure IV.14 A)**. Although several ligands have lost their H-bonds, π-type interactions between their large conjugated regions stabilize the formation of a well-organized stack around the template, as shown by the zoom in **Figure IV.14 B**. Additionally, the H-bonding and π-type interactions occurring during the last 500 ns of the simulation were localized using heatmaps (**Figure IV.14 C, D)**. The heatmap of H-bonds represents the azo compounds and the nucleotides of

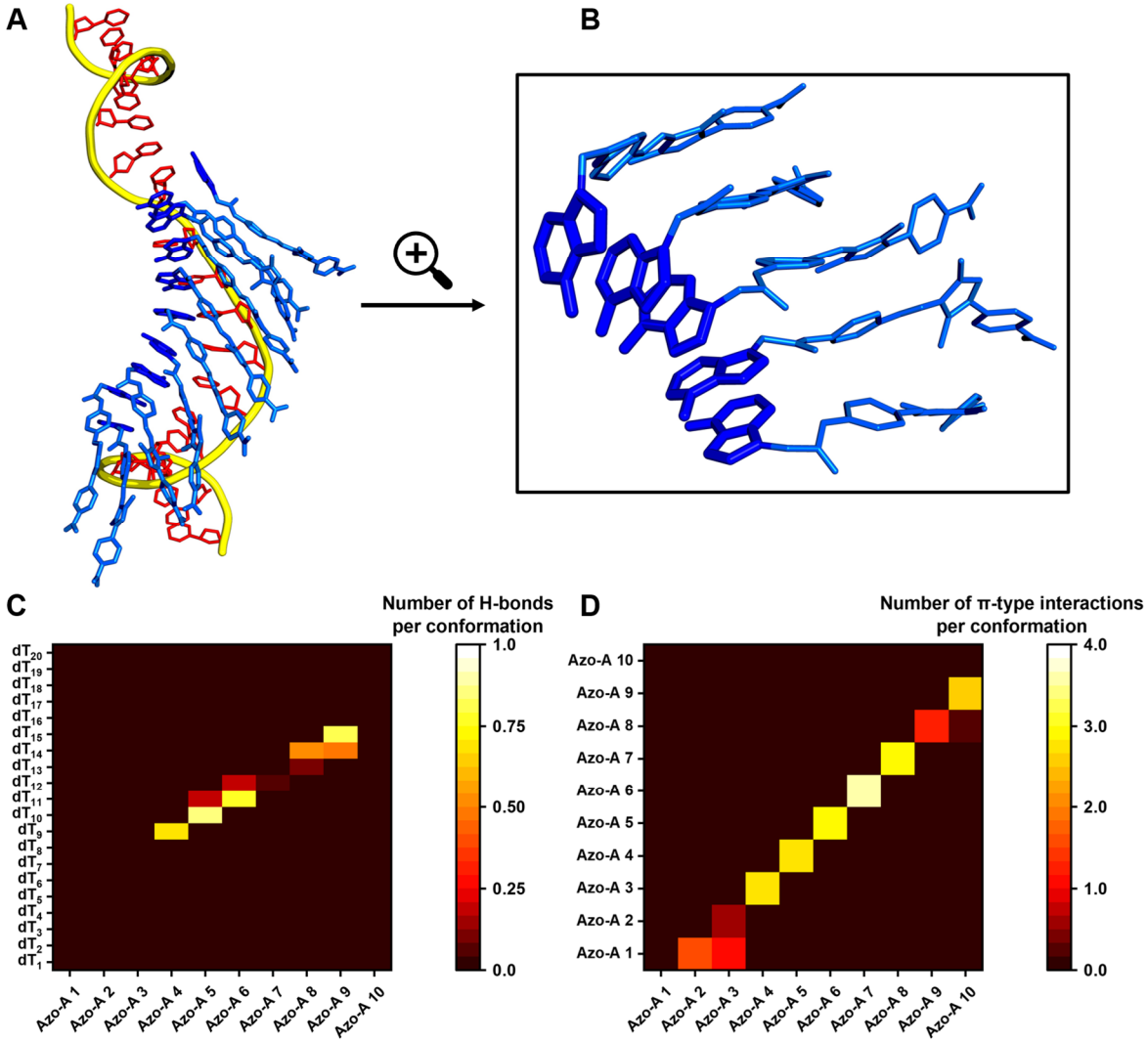
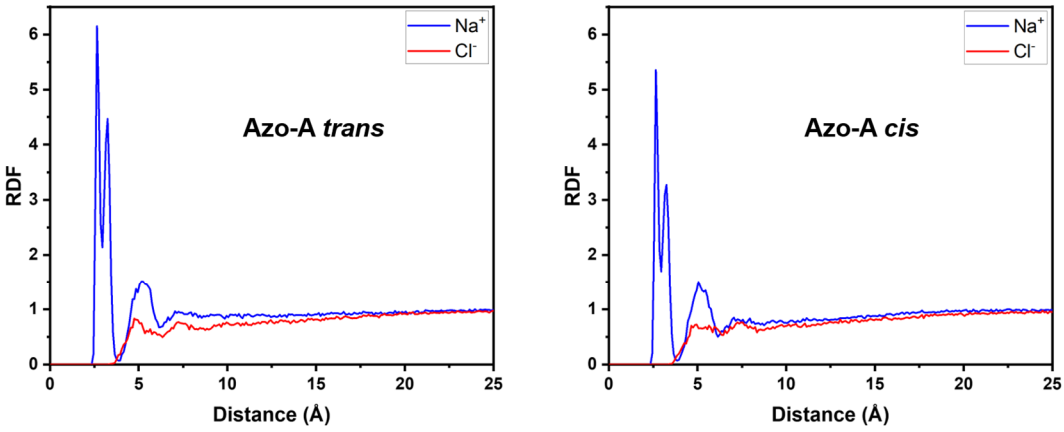


Figure IV.14. Overview of the simulation of the Azo-A *trans*/dT_{20} complex (first replica). **(A)** Final MD snapshot showing the 10 azo compounds (in blue) wrapped around the DNA strand (represented as a yellow tube with the nucleosides in red). **(B)** Zoom on five arylazopyrazoles forming a well-ordered stack. **(C)** Heatmap of H-bonds between the complementary nucleobases of the 10 azo compounds (represented on the x-axis, from Azo-A 1 to Azo-A 10) and of the DNA template (the 20 nucleotides are depicted in the y-axis, from dT_1 to dT_{20}). **(D)** Heatmap of π-type interactions between the 10 arylazopyrazole molecules (represented on the x- and y-axes).

the template on the x- and y-axes in **Figure IV.14 C**, respectively. It clearly shows that five ligands remain engaged in interactions with nucleobases along the dT_{20} strand. Simultaneously, the organized packing of the molecules is maintained at the microsecond timescale, with persistent π-type interactions occurring between neighboring arylazopyrazoles (displayed on the x- and y-axes in **Figure IV.14 D**), as indicated by the colored squares all along the diagonal of the heatmap. A similar behavior was observed for the four simulations of the *trans* isomers (see the left column of **Figures IV.S1 and IV.S2** for the replica of Azo-A *trans*/dT_{20} and for the simulations of Azo-T *trans*/dA_{20}). These heatmaps confirm that molecules dissociated from the template can remain efficiently stacked; see for example Azo-A 7, which forms very few H-bonds with the oligonucleotide but maintains π-type interactions with its neighbors Azo-A 6 and Azo-A 8. Overall, the molecules seem to be stabilized essentially by their π-type interactions, whereas H-bonds with the template help to order the stacks of azo compounds. The ligands that bind efficiently to the oligonucleotide can serve as "anchors" for the stacking of other units, which may remain within the chiral templated stacks even without forming direct H-bonds with the template. This tendency to favor π-type interactions over H-bonds with a DNA template was observed for other compounds presenting an extended conjugated region.[48–50] The stacking mode of the azo compounds in their *trans* configuration may also explain the important role played by the salt concentration. Within the stacks, the nucleobases are directed towards the template, which brings the negatively charged carboxylate groups at the other end of the molecules closer together. Increasing salt concentration helps to decrease the electrostatic repulsion, thus stabilizing the assemblies. The attraction of Na^+ ions towards the carboxylate moieties is illustrated by radial distribution functions (RDFs) **(Figure IV.15)**. Interestingly, the density of Na^+ ions close to the carboxylates is slightly higher for the molecules in their *trans* configuration than for those in their *cis* configuration, with higher RDF values for the first three peaks. This could arise from weaker π-type interactions between the *cis* azo compounds, as will be discussed in the next section, leading to less proximity between the carboxylates, thus a lower local density of cations. Finally, the organization of the *trans* azo compounds was investigated by measuring the rotation between consecutively stacked units. To this end, a vector was defined along the conjugated region (see illustration in **Figure IV.16 A**), and the angle between consecutive vectors was measured during the first 250 ns, *i.e.* when the H-bonds are constrained, to analyze the organization of "ideal" stacks. The average angle between the conjugated region of two stacked molecules is 23.9 ± 17.6 ° (with a median value of 20 °). In comparison, the helical twist between consecutive nucleobases in dsDNA in its B-form is about 34°. This shows that the rotation around the azo units does not strictly follow the natural helical twist of the

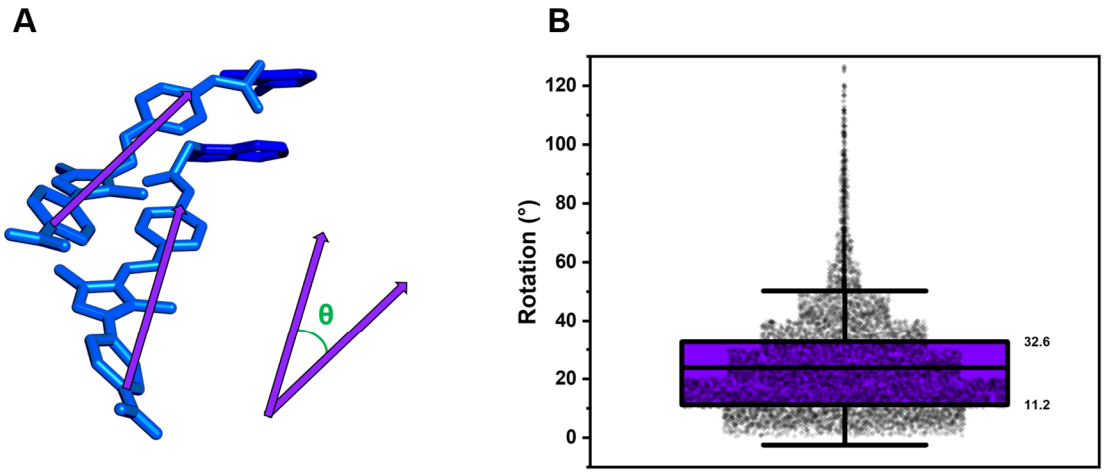


Average ± Std. dev. (RDF Na^+, all replicas)	Azopyrazoles *trans*	Azopyrazoles *cis*
First peak (2.65 Å)	6.1 ± 0.1	5.6 ± 0.4
Second peak (3.25 Å)	4.1 ± 0.3	3.6 ± 0.3
Third peak (5.45 Å)	1.5 ± 0.1	1.3 ± 0.0

Figure IV.15. RDF measured for the first replica of the Azo-A/dT_{20} complex, for the *trans* (left) and *cis* (right) isomers. RDFs for the other simulations (second replica of Azo-A and simulations of Azo-T) follow a very similar trend (data not shown). The RDF value indicates the density of ions (Na^+ or Cl^-, in blue and red, respectively), normalized by the average density of ions in the full simulation box, as a function of the distance from the carbon atom of the carboxylate groups. For instance, a value of six means that the ion density is six times higher than in the bulk, indicating a strong local attraction. The RDF values for the first three peaks of the Na^+ distribution are indicated in the table (statistics calculated on all replicas).

DNA template, which may explain the vanishing of H-bonds for several molecules, unable to maintain both H-bonds and π-type interactions, and prioritizing the latter.

Globally, our simulations reveal similar trends for the Azo-A *trans*/dT_{20} and the Azo-T *trans*/dA_{20} complexes, without obvious differences in the assembly. The reason behind the apparition of ICD signals of opposite signs remains unclear. However, other ligands presenting a large aromatic region were shown to interact with an ssDNA strand without following its helicity.[48] Ligands maintaining strong stacking interactions when assembled along DNA templates can form complex chiral structures, whose interpretation is not straightforward. Further work is needed to get more precise information on the nature of these ICD signals.



C

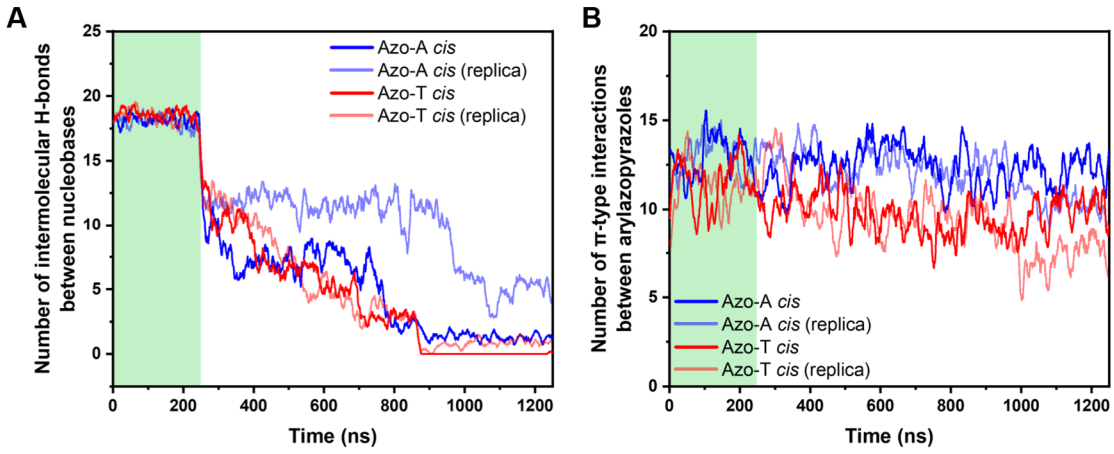
Rotation between stacked arylazopyrazoles (°)	
Average ± Std. dev.	**23.9 ± 17.6**
Median	**20.0**

Figure IV.16. Measurements of the rotation between pairs of stacked azo compounds. **(A)** Schematic representation of the methodology used to compute the rotation angle, using vectors defined along the conjugated region. **(B)** Distribution of the rotation angle values. The lines delimiting the box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside the box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. **(C)** Table summarizing the data.

IV-B.4. *Trans* to *cis* photoisomerization disorganizes ligands stacking and weakens the supramolecular assembly

Chiroptical spectra show a complete loss of the ICD signal when the molecules are switched in their *cis* configuration. To shine light on the effect of photoisomerization on the supramolecular assembly, MD simulations were also performed on the *cis* azo compounds organized along their complementary DNA template, following the same methodology than before. Similarly than for the *trans* compounds, the number of H-bonds quickly decreased after removal of the restraints **(Figure IV.17, A)**. However, here, the interactions are nearly completely lost at the end of the simulations, except in one case (replica of Azo-A, light blue curve). This outlier suggests that, although binding to the template seems weaker for the *cis* isomers, the photoisomerization does not always imply a complete dissociation. The number of π-type interactions between the arylazopyrazoles is also significantly lower for the *cis* isomers, with on average around 10 interactions per conformation at the end of the simulation (against around 19 for the *trans* compounds) **(Figure IV.17, B)**. The lack of planarity of the conjugated region for the *cis* isomers prevents the formation of well-organized stacks along the template. Instead, highly disordered and dense aggregates are formed

(Figure IV.18 A, B). Given the major role of π-type interactions in maintaining the assembly of the *trans* isomers, it is unsurprising that the molecules in their *cis* form quickly lose interactions with the oligonucleotide. Consequently, dissociation from the template is more likely for these isomers (see final MD snapshots of the other simulations in **Figure IV.S3**). This can be observed on the heatmap of H-bonds, with only one molecule still forming H-bonds with the oligonucleotide at the end of the simulation (Azo-A 2 with the nucleotide dT_9) **(Figure IV.18 C)**. The heatmap of π-type interactions also shows that interactions are much weaker for the *cis* compounds, in comparison to the same heatmap for their *trans* counterparts (see **Figure IV.14 D**). H-bonds and stacking heatmaps for the other simulations (replica of Azo-A *cis*/dT_{20} and the simulations of Azo-T *cis*/dA_{20}) are presented in the right column of **Figures IV.S1 and IV.S2**. Except for the replica of Azo-A *cis* mentioned earlier, where H-bonds remained surprisingly well-organized despite the apparent disorder of the ligands, these heatmaps confirm that the *trans* isomers bind more efficiently to the template and establish a stronger network of π-type interactions. Therefore, in



Average ± Std. dev. (last 500 ns)	Azo-A	Azo-A (replica)	Azo-T	Azo-T (replica)
H-bonds (nucleobases)	1.7 ± 1.2	7.7 ± 3.5	0.7 ± 1.3	1.3 ± 1.3
π-type interactions (arylazopyrazoles)	12.0 ± 1.9	11.2 ± 1.8	9.4 ± 1.9	8.3 ± 2.1

Figure IV.17. Interactions between the *cis* azo compounds and their complementary DNA template. **(A)** Evolution of the number of H-bonds between the complementary nucleobases of the arylazopyrazoles and the DNA template. **(B)** Evolution of the number of π-type interactions between the arylazopyrazoles. Blue and red curves represent the Azo-A/dT_{20} and Azo-T/dA_{20} systems, respectively. The running average including the five previous and five subsequent conformations is displayed, for ease of visualization. **(C)** Table summarizing the data during the last 500 ns of the simulation.

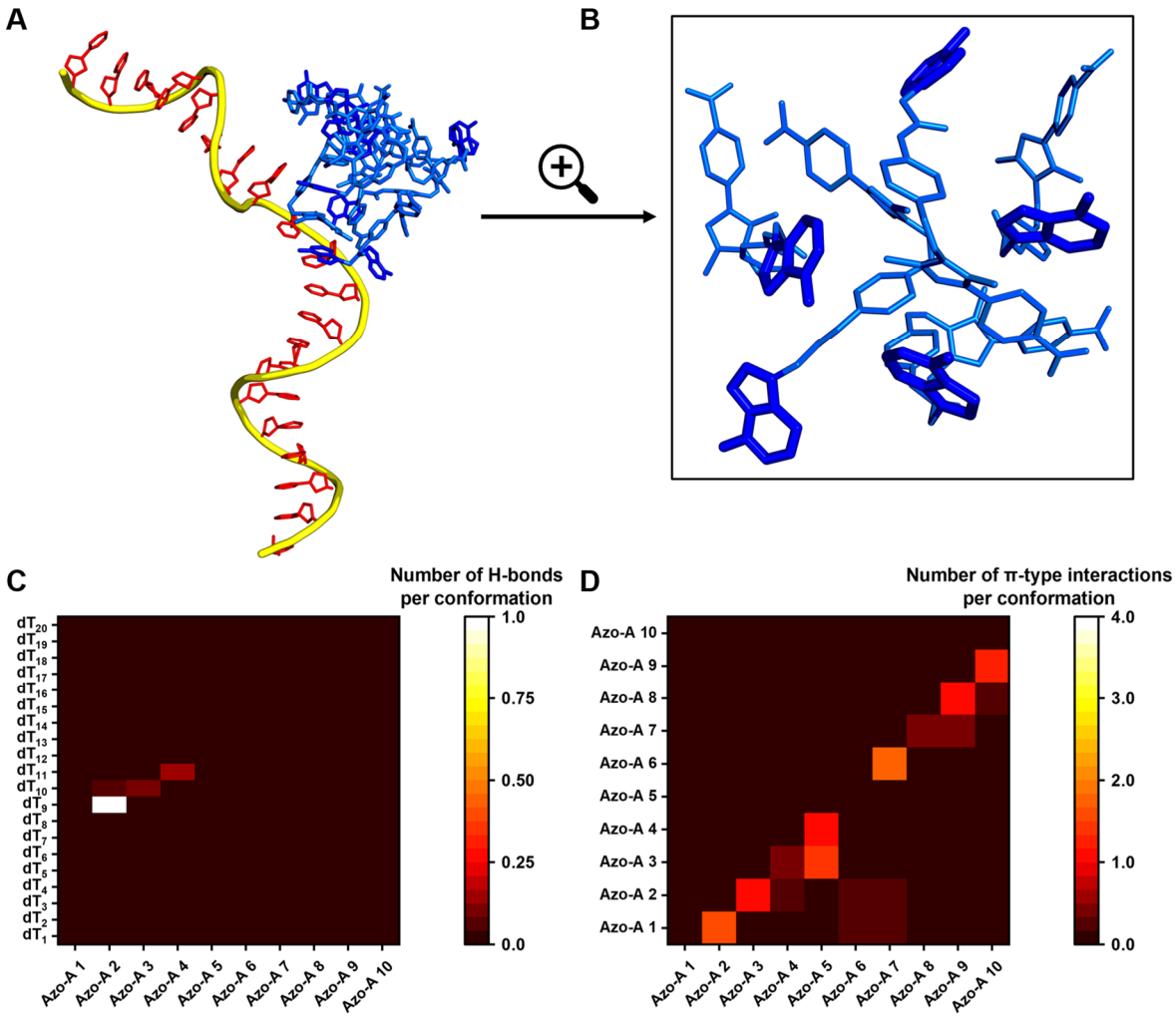


Figure IV.18. Overview of the simulation of the Azo-A *cis*/dT_{20} complex (first replica). **(A)** Final MD snapshot showing the 10 azo compounds (in blue) wrapped around the DNA strand (represented as a yellow tube with the nucleosides in red). **(B)** Zoom on five arylazopyrazoles forming a disordered aggregate. **(C)** Heatmap of H-bonds between the complementary nucleobases of the 10 azo compounds (represented on the x-axis, from Azo-A 1 to Azo-A 10) and of the DNA template (the 20 nucleotides are depicted in the y-axis, from dT_1 to dT_{20}). **(D)** Heatmap of π-type interactions between the 10 arylazopyrazole molecules (represented on the x- and y-axes).

addition to the partial disassembly from the oligonucleotide, the loss of the ICD signal observed for the *cis* isomers could also stem from the formation of highly disordered aggregates that do not strongly interact with the template, hence do not follow its helical structure.

IV-B.5. Conclusion

The binding modes of the azo compounds along their complementary DNA templates were shown to be strongly influenced by their configuration. In the *trans* form, a combination of H-bonds between complementary nucleobases and π-type interactions, with a predominance of the latter, is key to induce and maintain the supramolecular assembly. Although some molecules do not maintain their H-bonds with the template, they remain stacked with other ligands, ensuring that the chirality of DNA is transmitted to the azo compounds. These supramolecular stacks are stabilized by the presence of a high concentration of Na^+ ions. Inversely, the *cis* isomers, lacking planarity in their conjugated region, are not able to maintain ordered π-type interactions. This significantly weakens the binding to the DNA strand, and leads to partial disassembly from the template and the formation of disordered aggregates. These results are in line with the experimental CD spectra, which indicate the need of a high ionic strength to promote supramolecular assembly, and loss of the ICD signals upon photoisomerization into the *cis* configuration.
However, our simulations do not explain the unexpected nature of the ICD signals, which are of opposite signs for the two azo compounds. As mentioned previously, the sign of the ICD signal is not always dictated by the chirality of the DNA template in the case of strongly conjugated molecules.[48] Therefore, a deeper understanding of these chiroptical experiments would require theoretical calculations of CD spectra based on conformations extracted from MD simulations. This constitutes a perspective to this work, and new simulations have already been performed using the GROMACS package, with a more accurate reparametrization of the force field.[51] CD spectra have begun to be simulated with the *VeloxChem* software from these new simulations, in collaboration with the group of M. Linares and P. Norman at KTH Royal Institute of Technology in Stockholm. An example for each azo compound is shown in **Figure IV.S4**.[52] In brief, these first attempts suggest that very small conformational modifications can induce a completely opposite response. Our goal now is to understand which structural parameters determine the sign of the ICD signals, in order to correlate the experimental observations with an accurate atomic-scale picture of the assemblies.

IV-B.6. Simulation protocol

Concerning the reparametrization step, 540 conformers were generated for the fragment 1 to scan the Φ_1, Φ_2 and Φ_3 dihedral angles. New parameters were derived for all three angles based on this scan. For the fragment 2 and the reparametrization of Φ_4, 400 conformers were generated. Individual scans, presented in **Figure IV.11**, were

then realized to evaluate the quality of the new parameters. The QM calculations were performed with the *Gaussian 16* software, using the MP2 method (post-Hartree-Fock) and the cc-pvdz basis set.[53] Using fragments instead of the whole azo compounds has two interests: it reduces the number of atoms for the QM calculations, thus reducing computational cost, and allows to use the same set of new parameters for both Azo-A and Azo-T, which seems reasonable as their conjugated part is the same.

To build the azo compounds, the structure of the molecule was divided in 3 fragments, modeled with the *Avogadro 1.2.0* software.[54] The assembly of the fragments and all subsequent operations were carried out with the AMBER package.[46] The calculations of the atomic partial charges were performed with the *antechamber* module of AMBER, using the semi-empirical AM1-BCC method.[55] All force field parameters for the azo compounds were given by GAFF 2.11, except the reparametrized ones.[56] The fragments were assembled using the *LEaP* module of AMBER. The oligonucleotides were built with the *Nucleic Acid Builder* (NAB) tool implemented in AMBER and the DNA force field parameters were given by Parmbsc1.[57] The arylazopyrazoles/oligonucleotide supramolecular complexes were built within *LEaP*. The azo units were preorganized along the template (pairing of the complementary nucleobases) using *PyMol 2.5.4*.[58] The Azo-A/dT_{20} and Azo-T/dA_{20} complexes were simulated independently, in 2 replicas for each isomer (*trans* and *cis*), giving 8 independent simulations. All systems were solvated in truncated octahedral water boxes, with at least 25.0 Å between any solute atom and the edge of the box, in order to let enough space for the ligands to have the possibility to dissociate from the template. The 4-point OPC water model was used to describe the solvent and a NaCl concentration of 5 M was used to reproduce the experimental conditions, following the "SPLIT" method.[59,60] All MD simulations were performed with the GPU version of AMBER. They started with a geometry optimization performed by MM to get a stable starting point. 1,000 steps of steepest descent were followed by 9,000 steps of conjugated gradient on the solvent and salt residues. A second geometry optimization was done with the same protocol, on the whole system. Then, a heating step of 2 ns was performed in the NVT ensemble to bring the system to a temperature of 300 K, using positional restraints on the solute atoms with a force constant of 10 $kcal.mol^{-1}.Å^{-2}$. The temperature was maintained at 300 K with a Langevin thermostat, using a collision frequency of 1 ps^{-1}. The system was equilibrated during 10 ns in the NPT ensemble with a Monte Carlo barostat, with restraints to maintain the H-bonds between the complementary nucleobases of the arylazopyrazole units and the DNA template: a force constant of 40 $kcal.mol^{-1}.Å^{-2}$ was applied as soon as the distance between the donor and the acceptor of the H-bond exceeded 2.2 Å. These restraints were extended

for 250 ns, to let the system stabilize without disassembly of the ligands, followed by 1 µs of unrestrained simulation. A timestep of 2 fs was used and the SHAKE algorithm was applied to constrain bonds involving hydrogen atoms. To switch the azo units into their *cis* form, a constraint on the Φ_2 dihedral angle was imposed: a force constant of 100 kcal.mol^{-1}.Å^{-2} was applied as soon as the dihedral angle was going out of the range [-30.0 ; 30.0] degrees. In practice, this force constant helped the arylazopyrazoles to bypass the torsional barrier leading from the *trans* to *cis* configuration at the beginning of the simulation. Spontaneous back isomerization from *cis* to *trans* did not occur afterwards. A cut-off of 12.0 Å was selected for non-bonded interactions and the particle mesh Ewald (PME) scheme was used to treat electrostatic interactions. A snapshot was saved each ns and extracted for further analyses. *PyMol 2.5.4* was used to visualize the snapshots and to create images.[58]

To analyze the trajectories, the *cpptraj* module implemented in AMBER was used.[61] Hydrogen bonds were detected with geometric criteria: the distance between the acceptor and the donor heavy atoms must be ≤ 3.0 Å and the angle between the donor, the hydrogen atom and the acceptor must be ≥ 135°. H-bonds were measured between the atoms of the nucleobases of the azo compounds and of the oligonucleotide. π-type interactions between the aromatic cycles were detected with geometric criteria: two cycles are considered stacked if the distance between their centers of mass is ≤ 5 Å and if the angle between them is < 45° or > 135 °. The aromatic interactions were calculated for all pairs of aromatic cycles of the azo compounds, and were summed “by molecule” for the heatmaps (two molecules perfectly superimposed would form four interactions, as they possess four aromatic cycles). The heatmaps were computed over the last 500 ns of the simulations. Radial distribution functions were computed between the carbon atom of the carboxylate moiety of each arylazopyrazole and the Na^+ or Cl^- ions, over the last 500 ns, with a bin spacing of 0.1 Å. The density value used for normalization was calculated as the ratio between the number of ions in the box and the average volume of the simulation box. To measure the rotation between the conjugated parts of consecutive azo compounds, a vector was defined for each ligand, as represented by the purple arrows in **Figure IV.16 A**. The rotation angle between two stacked molecules was calculated from the dot-product of their vectors and was measured for each pair of consecutive azo compounds (Azo 1 – Azo 2; Azo 2 – Azo 3; Azo 3 – Azo 4; and so on). This angle was measured during the first 250 ns, when the H-bonds were constrained. As the stacks may present discontinuities, with two consecutive azo compounds not interacting, a criterion was added to remove these pairs from the calculation. We decided to exclude the pairs whose rotation angle was superior than or equal to 100° (which clearly indicates that the conjugated part are not superimposed) at least 10 % of the time.

IV-B.7. Additional data

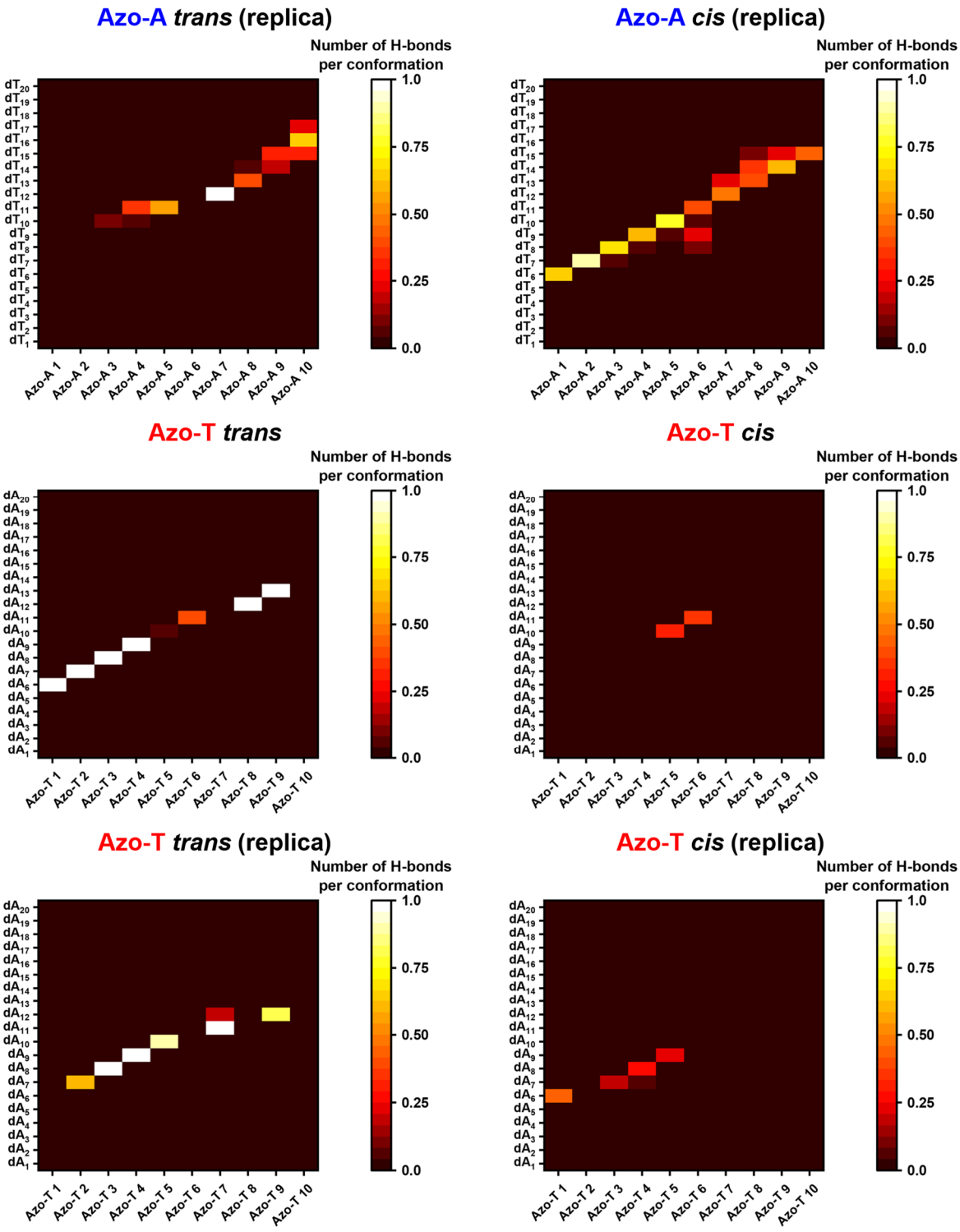


Figure IV.S1. Additional heatmaps of H-bonds, for the replicas of the *trans* and *cis* Azo-A/dT_{20} and both replicas of Azo-T/dA_{20}. Except for the replica of Azo-A in *cis*, the results generally indicate significantly more H-bonds for the *trans* isomers.

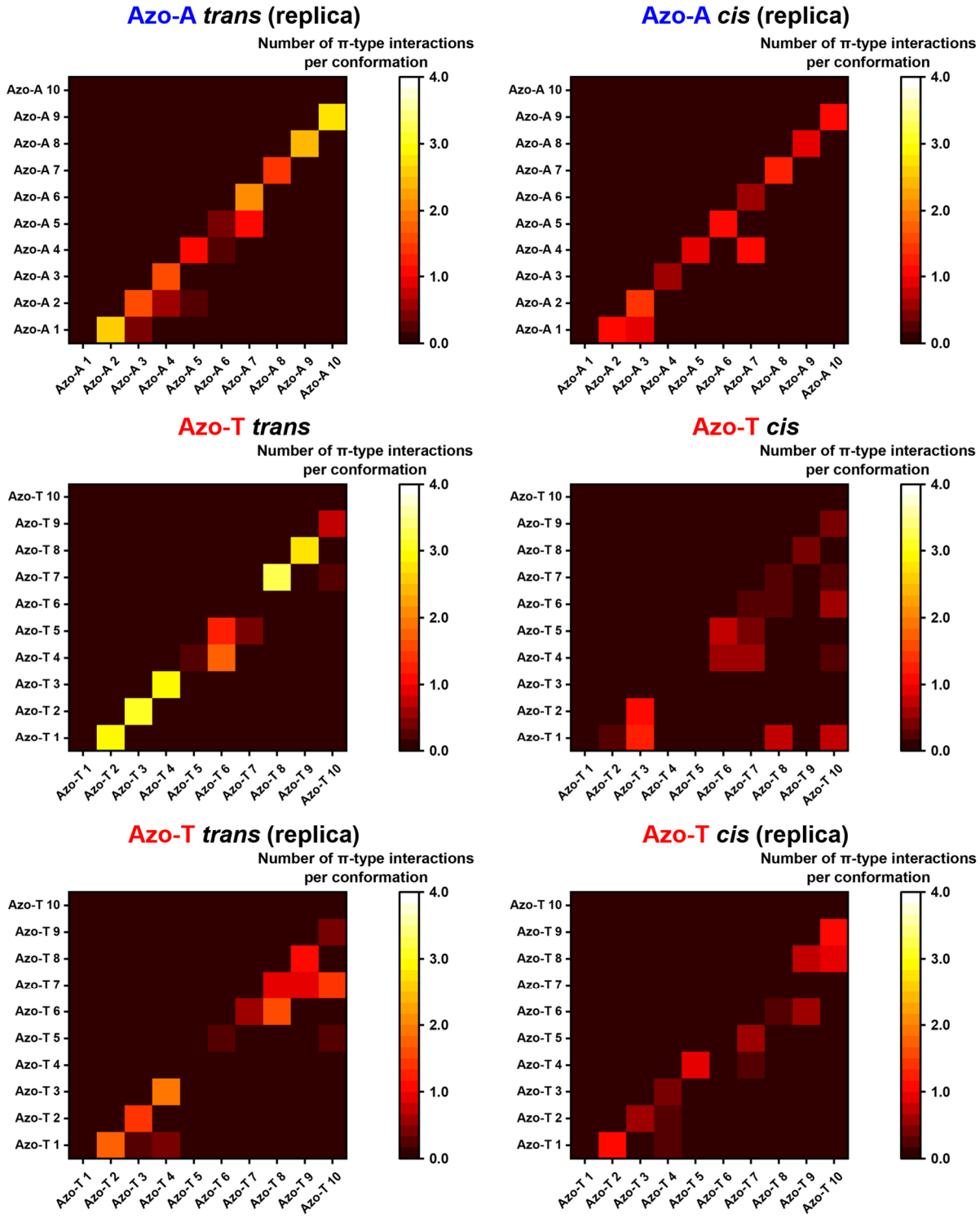


Figure IV.S2. Additional heatmaps of π-type interactions, for the replicas of the *trans* and *cis* Azo-A/dT_{20} and both replicas of Azo-T/dA_{20}. Except for the replica of Azo-T in *trans*, the results generally indicate significantly more stacking interactions for the *trans* isomers.

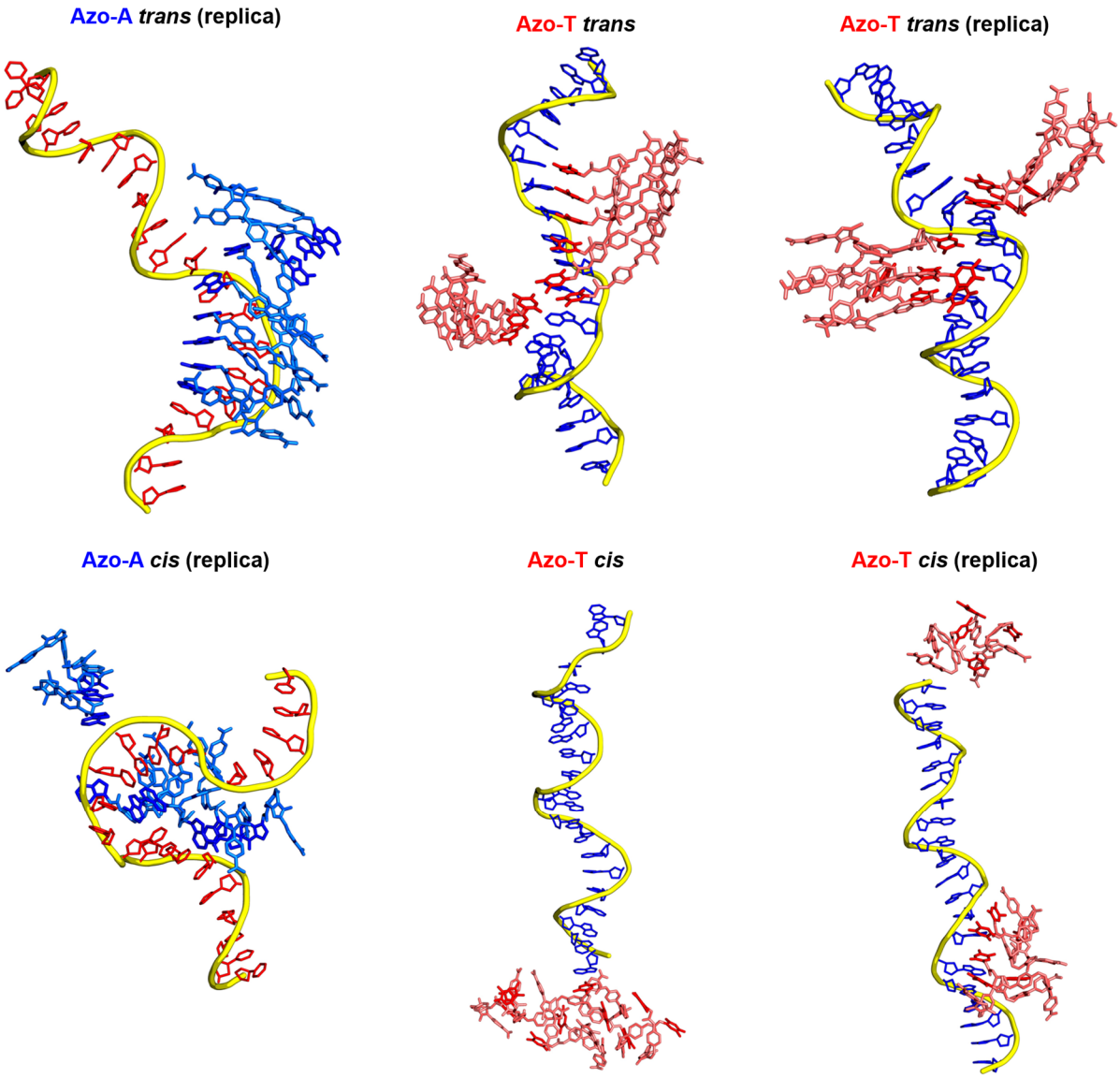


Figure IV.S3. Final MD snapshots of the six simulations not shown in the main text (replica for *trans* and *cis* Azo-A/dT_{20}, and simulations of Azo-T/dA_{20}).

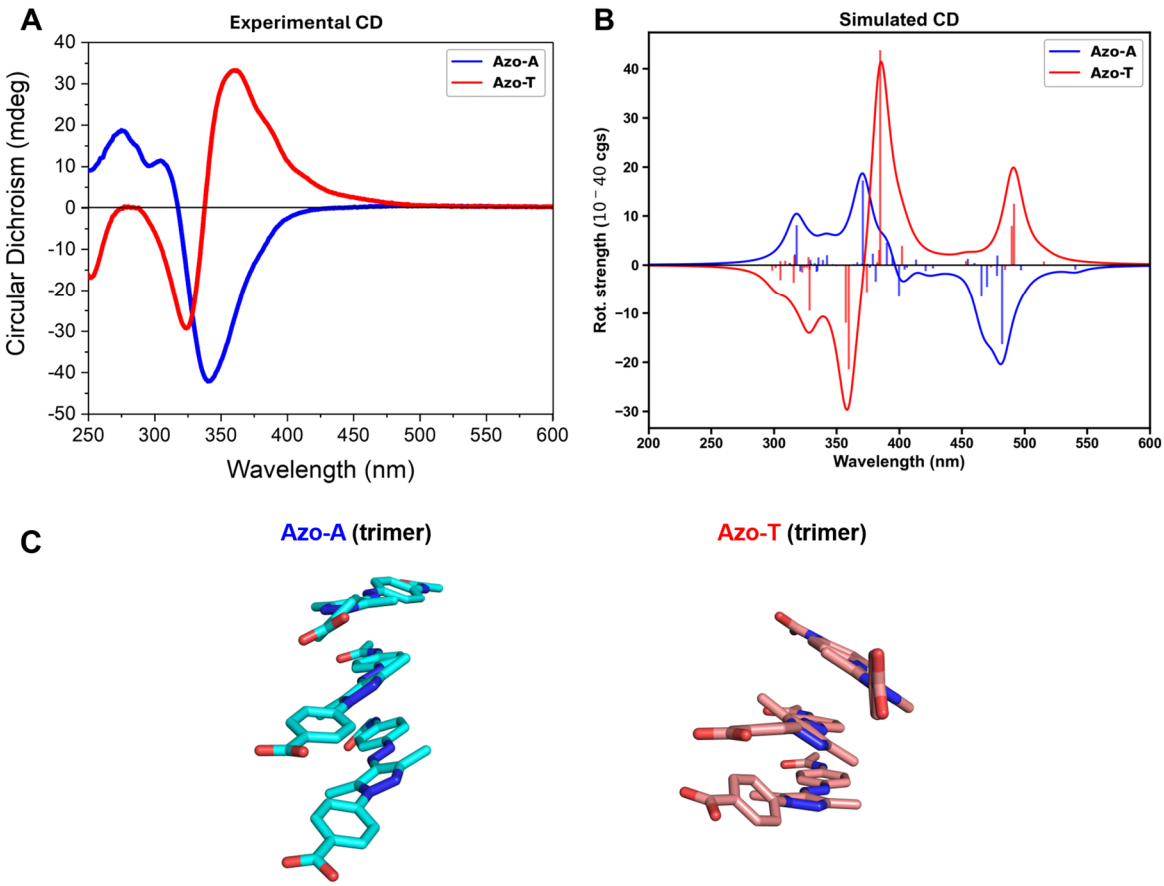


Figure IV.S4. Preliminary results of CD spectra simulation. **(A)** Experimental CD spectra measured at a NaCl concentration of 5 M. **(B)** Simulated CD spectra for both compounds measured on stacks extracted from one frame of the MD simulations (these curves were selected to show that it is possible to retrieve features similar to the experimental spectra, but are only issued from one frame; a more rigorous methodology will have to be implemented to get reliable results). The spectra were simulated by TD-DFT with the CAM-B3LYP functional and the def2-svpd basis set. **(C)** Structure of the stacks used to calculate the CD spectra. The nucleobases were stripped from the molecules and replaced by methyl groups to reduce computational cost, as the chromophore of interest is constituted by the conjugated region.

References

[1] F. A. Aldaye, A. L. Palmer, H. F. Sleiman. Assembling Materials with DNA as the Guide. *Science* **2008**, *321*, 1795–1799.

[2] N. C. Seeman. DNA in a material world. *Nature* **2003**, *421*, 427–431.

[3] C. Rossi-Gendron, F. El Fakih, L. Bourdon, K. Nakazawa, J. Finkel, N. Triomphe, L. Chocron, M. Endo, H. Sugiyama, G. Bellot, M. Morel, S. Rudiuk, D. Baigl. Isothermal self-assembly of multicomponent and evolutive DNA nanostructures. *Nat. Nanotechnol.* **2023**, *18*, 1311–1318.

[4] P. Zhan, A. Peil, Q. Jiang, D. Wang, S. Mousavi, Q. Xiong, Q. Shen, Y. Shang, B. Ding, C. Lin, Y. Ke, N. Liu. Recent Advances in DNA Origami-Engineered Nanomaterials and Applications. *Chem. Rev.* **2023**, *123*, 3976–4050.

[5] S. Samai, D. J. Bradley, T. L. Y. Choi, Y. Yan, D. S. Ginger. Temperature-Dependent Photoisomerization Quantum Yields for Azobenzene-Modified DNA. *J. Phys. Chem. C* **2017**, *121*, 6997–7004.

[6] T. J. Bandy, A. Brewer, J. R. Burns, G. Marth, T. Nguyen, E. Stulz. DNA as supramolecular scaffold for functional molecules: progress in DNA nanotechnology. *Chem. Soc. Rev.* **2011**, *40*, 138–148.

[7] W. Wang, W. Wan, H.-H. Zhou, S. Niu, A. D. Q. Li. Alternating DNA and π-Conjugated Sequences. Thermophilic Foldable Polymers. *J. Am. Chem. Soc.* **2003**, *125*, 5248–5249.

[8] H. Yang, F. Altvater, A. D. de Bruijn, C. K. McLaughlin, P. K. Lo, H. F. Sleiman. Chiral Metal–DNA Four-Arm Junctions and Metalated Nanotubular Structures. *Angew. Chem. Int. Ed.* **2011**, *50*, 4620–4623.

[9] Y. Vyborna, M. Vybornyi, R. Häner. Functional DNA-grafted supramolecular polymers – chirality, cargo binding and hierarchical organization. *Chem. Commun.* **2017**, *53*, 5179–5181.

[10] S. P. W. Wijnands, E. W. Meijer, M. Merkx. DNA-Functionalized Supramolecular Polymers: Dynamic Multicomponent Assemblies with Emergent Properties. *Bioconjugate Chem.* **2019**, *30*, 1905–1914.

[11] S. Rothenbühler, A. Gonzalez, I. Iacovache, S. M. Langenegger, B. Zuber, R. Häner. Tetraphenylethylene–DNA conjugates: influence of sticky ends and DNA sequence length on the supramolecular assembly of AIE-active vesicles. *Org. Biomol. Chem.* **2022**, *20*, 3703–3707.

[12] J. Thiede, T. Schneeberger, I. Iacovache, S. M. Langenegger, B. Zuber, R. Häner. Supramolecular assembly of phenanthrene–DNA conjugates into light-harvesting nanospheres. *New J. Chem.* **2024**, *48*, 15731–15734.

[13] C. K. McLaughlin, G. D. Hamblin, H. F. Sleiman. Supramolecular DNA assembly. *Chem. Soc. Rev.* **2011**, *40*, 5647.

[14] P. Ensslen, F. Brandl, S. Sezi, R. Varghese, R. Kutta, B. Dick, H. Wagenknecht. DNA-Based Oligochromophores as Light-Harvesting Systems. *Chem. Eur. J.* **2015**, *21*, 9349–9354.

[15] P. Ensslen, S. Gärtner, K. Glaser, A. Colsmann, H. Wagenknecht. A DNA–Fullerene Conjugate as a Template for Supramolecular Chromophore Assemblies: Towards DNA-Based Solar Cells. *Angew. Chem. Int. Ed.* **2016**, *55*, 1904–1908.

[16] R. Varghese, H. Wagenknecht. White-Light-Emitting DNA (WED). *Chem. Eur. J.* **2009**, *15*, 9307–9310.

[17] J. Kapuscinski. DAPI: a DNA-Specific Fluorescent Probe. *Biotech. Histochem.* **1995**, *70*, 220–233.

[18] M. Coste, C. Kotras, Y. Bessin, V. Gervais, D. Dellemme, M. Leclercq, M. Fossépré, S. Richeter, S. Clément, M. Surin, S. Ulrich. Synthesis, Self-Assembly, and Nucleic Acid Recognition of an

Acylhydrazone-Conjugated Cationic Tetraphenylethene Ligand. *Eur. J. Org. Chem.* **2021**, *2021*, 1123–1135.

[19] K. Zakrzewska, R. Lavery, B. Pullman. Theoretical studies of the selective binding to DNA of two non-intercalating ligands: netropsin and SN 18071. *Nucleic Acids Res.* **1983**, *11*, 8825–8839.

[20] J. A. Mondragón-Sánchez, R. Santamaria, R. Garduño-Juárez. Docking on the DNA G-quadruplex: A molecular electrostatic potential study. *Biopolymers* **2011**, *95*, 641–650.

[21] S. Chakrabarti, D. Bhattacharyya, D. Dasgupta. Structural basis of DNA recognition by anticancer antibiotics, chromomycin A3, and mithramycin: Roles of minor groove width and ligand flexibility. *Biopolymers* **2000**, *56*, 85–95.

[22] M. Fosséprés, I. Tuvi-Arad, D. Beljonne, S. Richeter, S. Clément, M. Surin. Binding Mode Multiplicity and Multiscale Chirality in the Supramolecular Assembly of DNA and a π-Conjugated Polymer. *ChemPhysChem* **2020**, *21*, 2543–2552.

[23] Y. Hu, D. Han, Q. Zhang, T. Wu, F. Li, L. Niu. Perylene ligand wrapping G-quadruplex DNA for label-free fluorescence potassium recognition. *Biosens. Bioelectron.* **2012**, *38*, 396–401.

[24] M. Surin. From nucleobase to DNA templates for precision supramolecular assemblies and synthetic polymers. *Polym. Chem.* **2016**, *7*, 4137–4150.

[25] M. Surin, S. Ulrich. From Interaction to Function in DNA-Templated Supramolecular Self-Assemblies. *ChemistryOpen* **2020**, *9*, 480–498.

[26] J. Schill, B. J. H. M. Rosier, B. Gumí Audenis, E. Magdalena Estirado, T. F. A. de Greef, L. Brunsveld. Assembly of Dynamic Supramolecular Polymers on a DNA Origami Platform. *Angew. Chem. Int. Ed.* **2021**, *60*, 7612–7616.

[27] S. Müller, Y. Fritz, H. Wagenknecht. Control of Energy Transfer Between Pyrene- and Perylene-Nucleosides by the Sequence of DNA-Templated Supramolecular Assemblies. *ChemistryOpen* **2020**, *9*, 389–392.

[28] H. Ucar, H.-A. Wagenknecht. Aggregation-induced emission by sequence-selective assembly of cyanolated distyrylbenzene in supramolecular DNA architectures. *Chem. Commun.* **2022**, *58*, 6437–6440.

[29] B. Tassignon, Z. Wang, A. Galanti, J. De Winter, P. Samorì, J. Cornil, K. Moth-Poulsen, P. Gerbaux. Site Selectivity of Peptoids as Azobenzene Scaffold for Molecular Solar Thermal Energy Storage. *Chem. Eur. J.* **2023**, *29*, e202303168.

[30] L. Osorio-Planes, C. Rodríguez-Escrich, M. A. Pericàs. Photoswitchable Thioureas for the External Manipulation of Catalytic Activity. *Org. Lett.* **2014**, *16*, 1704–1707.

[31] R. Liu, X. Zhang, F. Xia, Y. Dai. Azobenzene-based photoswitchable catalysts: State of the art and perspectives. *J. Catal.* **2022**, *409*, 33–40.

[32] M. Olesińska-Mönch, C. Deo. Small-molecule photoswitches for fluorescence bioimaging: engineering and applications. *Chem. Commun.* **2023**, *59*, 660–669.

[33] A. S. Lubbe, W. Szymanski, B. L. Feringa. Recent developments in reversible photoregulation of oligonucleotide structure and function. *Chem. Soc. Rev.* **2017**, *46*, 1052–1079.

[34] D. Y. Tam, X. Zhuang, S. W. Wong, P. K. Lo. Photoresponsive Self-Assembled DNA Nanomaterials: Design, Working Principles, and Applications. *Small* **2019**, *15*, 1805481.

[35] J. Rubio-Magnieto, T. Phan, M. Fossépré, V. Matot, J. Knoops, T. Jarrosson, P. Dumy, F. Serein-Spirau, C. Niebel, S. Ulrich, M. Surin. Photomodulation of DNA-Templated Supramolecular Assemblies. *Chem. Eur. J.* **2018**, *24*, 706–714.

[36] A. Diguet, N. K. Mani, M. Geoffroy, M. Sollogoub, D. Baigl. Photosensitive Surfactants with Various Hydrophobic Tail Lengths for the Photocontrol of Genomic DNA Conformation with Improved Efficiency. *Chem. Eur. J.* **2010**, *16*, 11890–11896.

[37] N. A. Simeth, S. Kobayashi, P. Kobauri, S. Crespi, W. Szymanski, K. Nakatani, C. Dohno, B. L. Feringa. Rational design of a photoswitchable DNA glue enabling high regulatory function and supramolecular chirality transfer. *Chem. Sci.* **2021**, *12*, 9207–9220.

[38] Y. Kamiya, T. Takagi, H. Ooi, H. Ito, X. Liang, H. Asanuma. Synthetic Gene Involving Azobenzene-Tethered T7 Promoter for the Photocontrol of Gene Expression by Visible Light. *ACS Synth. Biol.* **2015**, *4*, 365–370.

[39] N. A. Simeth, P. de Mendoza, V. R. A. Dubach, M. C. A. Stuart, J. W. Smith, T. Kudernac, W. R. Browne, B. L. Feringa. Photoswitchable architecture transformation of a DNA-hybrid assembly at the microscopic and macroscopic scale. *Chem. Sci.* **2022**, *13*, 3263–3272.

[40] L. Stricker, E.-C. Fritz, M. Peterlechner, N. L. Doltsinis, B. J. Ravoo. Arylazopyrazoles as Light-Responsive Molecular Switches in Cyclodextrin-Based Supramolecular Systems. *J. Am. Chem. Soc.* **2016**, *138*, 4547–4554.

[41] M. A. Gerkman, R. S. L. Gibson, J. Calbo, Y. Shi, M. J. Fuchter, G. G. D. Han. Arylazopyrazoles for Long-Term Thermal Energy Storage and Optically Triggered Heat Release below 0 °C. *J. Am. Chem. Soc.* **2020**, *142*, 8688–8695.

[42] R. Iwaura, Y. Kikkawa, M. Ohnishi-Kameyama, T. Shimizu. Effects of oligoDNA template length and sequence on binary self-assembly of a nucleotide bolaamphiphile. *Org. Biomol. Chem.* **2007**, *5*, 3450.

[43] P. G. A. Janssen, S. Jabbari-Farouji, M. Surin, X. Vila, J. C. Gielen, T. F. A. de Greef, M. R. J. Vos, P. H. H. Bomans, N. A. J. M. Sommerdijk, P. C. M. Christianen, P. Leclère, R. Lazzaroni, P. van der Schoot, E. W. Meijer, A. P. H. J. Schenning. Insights into Templated Supramolecular Polymerization: Binding of Naphthalene Derivatives to ssDNA Templates of Different Lengths. *J. Am. Chem. Soc.* **2009**, *131*, 1222–1231.

[44] F. J. Rizzuto, C. M. Platnich, X. Luo, Y. Shen, M. D. Dore, C. Lachance-Brais, A. Guarné, G. Cosa, H. F. Sleiman. A dissipative pathway for the structural evolution of DNA fibres. *Nat. Chem.* **2021**, *13*, 843–849.

[45] N. Nogal, S. Guisán, D. Dellemme, M. Surin, A. de la Escosura. Selectivity in the chiral self-assembly of nucleobase-arylazopyrazole photoswitches along DNA templates. *J. Mater. Chem. B* **2024**, *12*, 3703–3709.

[46] D. A. Case, T. E. Cheatham Iii, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[47] L. Settimo, K. Bellman, R. M. A. Knegtel. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm. Res.* **2014**, *31*, 1082–1095.

[48] R. Hofsäβ, P. Ensslen, H.-A. Wagenknecht. Control of helical chirality in supramolecular chromophore–DNA architectures. *Chem. Commun.* **2019**, *55*, 1330–1333.

[49] M. Surin, P. G. A. Janssen, R. Lazzaroni, P. Leclère, E. W. Meijer, A. P. H. J. Schenning. Supramolecular Organization of ssDNA-Templated π-Conjugated Oligomers via Hydrogen Bonding. *Adv. Mater.* **2009**, *21*, 1126–1130.

[50] D. Paolantoni, J. Rubio-Magnieto, S. Cantel, J. Martinez, P. Dumy, M. Surin, S. Ulrich. Probing the importance of π-stacking interactions in DNA-templated self-assembly of bisfunctionalized guanidinium compounds. *Chem. Commun.* **2014**, *50*, 14257–14260.

[51] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

[52] Z. Rinkevicius, X. Li, O. Vahtras, K. Ahmadzadeh, M. Brand, M. Ringholm, N. H. List, M. Scheurer, M. Scott, A. Dreuw, P. Norman. VeloxChem: A Python-driven density-functional theory program for spectroscopy simulations in high-performance computing environments. *WIREs Comput. Mol. Sci.* **2020**, *10*, e1457.

[53] Gaussian 16, Revision A.03, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, **2016**.

[54] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, G. R. Hutchison. Open Access Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Software* **2012**, *4*, 17.

[55] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[56] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case. Development and Testing of a General Amber Force Field. **2004**, *25*, 1157–1174.

[57] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case, M. Orozco. Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **2016**, *13*, 55–58.

[58] Schrödinger LLC, *The PyMOL Molecular Graphics System, Version 2.5.4*, **2015**.

[59] S. Izadi, R. Anandakrishnan, A. V. Onufriev. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **2014**, *5*, 3863–3871.

[60] M. R. Machado, S. Pantano. Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 1367–1372.

[61] D. R. Roe, T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

V. Dynamic self-assembly of supramolecular catalysts from precision macromolecules

The next results chapter of our thesis focuses on the supramolecular assembly of two SDMs forming a catalytic duplex. Taking a step further from natural systems towards artificial materials, we studied oligomers with a purely synthetic backbone, functionalized with nucleobases – inspired by the recognition mechanisms of nucleic acids – and catalytic units that must self-assemble in close spatial proximity, reminiscent of enzymatic active sites. MD simulations combined with network theory were used to characterize the 3D structure and dynamics of the supramolecular complex. This approach helped to rationalize experimental trends in catalytic activity and provided insights for improving molecular design.

Part of this work is reported in: *Dynamic self-assembly of supramolecular catalysts from precision macromolecules*.
Q. Qin, J. Li, D. Dellemme, M. Fosséprè, G. Barozzino-Consiglio, I. Nekkaa, A. Boborodea, A. E. Fernandes, K. Glinel, M. Surin, A. M. Jonas, *Chem. Sci.*, **2023**, 14, 9283-9292.

V.1. Introduction

Living systems rely on many complex metabolic pathways, such as the Krebs cycle, photosynthesis, or the urea cycle, enabling energy production and the degradation of harmful species.[1] These processes involve cascades of chemical reactions, which must be highly efficient and tightly regulated to ensure the proper functioning of living organisms. This role is fulfilled by enzymes, biocatalysts displaying well-defined catalytic pockets able to host specific substrates, which are stabilized by shape complementarity with the active site and interactions involving a series of ideally positioned amino acids. After binding of the substrate, its transformation is catalyzed by the cooperative action of several chemical groups, assembled in spatially close positions. This mechanism is very dynamic, the conformational flexibility of the enzyme ensuring accessibility to and from the catalytic pocket and the realization of the reaction transition state.[2,3] The remarkable efficiency and selectivity displayed by enzymes have long attracted the interest of researchers, but the translation of such self-assembled multifunctional catalysts into synthetic systems remains challenging. A key parameter consists in maximizing the probability of encounter of the different components forming the catalytic site. Various approaches have been investigated, including the use of natural SDMs, such as peptides or DNA, which can fold and self-assemble into programmable and well-defined nanostructures, and guide the

organization of catalytic components.[4–14] Preorganizing the catalytic units within synthetic polymer backbones able to form controlled structures in solution is another option, with the examples of single-chain polymeric nanoparticles (SCPNs), foldamers or even supramolecular polymers.[15–20] More recently, synthetic SDMs have been envisioned as a very promising avenue to approach the efficiency of biocatalysts.[21,22] Distributing the catalytic units at precise locations within a defined primary structure could help in building artificial systems mimicking the organization of enzymes, without reproducing their full size and complexity. Additionally, the order of monomers can be used to modulate the catalytic properties.[18,23–25] Impressive sequence effects were demonstrated for short catalytic trimers grafted on silica particles, the primary structure influencing interchain interactions and the spatial proximity of the catalytic units, thus the catalytic activity.[23] The role of the sequence was also demonstrated at the single-chain level for similar trimers, whose catalytic properties were rationalized by MD simulations and network analyses.[24]

Our work follows this trend, aiming at exploiting SDMs to optimize the spatial proximity and organization of the components of a multifunctional catalytic system. However, instead of relying on single-chain folding, our approach requires the supramolecular assembly of two complementary sequence-defined oligomers. The catalytic units are distributed among the two chains, which must therefore self-assemble into a supramolecular duplex to form the active center. To this end, our SDMs are also functionalized with complementary recognition units. Several examples have shown that synthetic SDMs can be advantageously used to precisely position recognition motifs and promote the formation of controlled assemblies.[26–30] Consequently, encoding a programmed recognition into synthetic SDMs to maximize the encounter of catalytic units, replicating the sequence control of biomacromolecules, constitutes an interesting strategy in view of approaching artificial enzymes. MD simulations were carried out on the supramolecular complex formed by the assembly of the two oligomers to better understand its 3D structure and dynamics in solution. Our results indicate the formation of a disordered globular duplex, stabilized by a myriad of interactions and inside which the individual oligomers retain a high flexibility **(Figure V.1)**. However, with the support of network representations, we were able to demonstrate the important role played by the recognition units in the supramolecular assembly, showing that persistent and specific interactions arise in the globule. Our results, combined with experiments, give precious insights into the role of each monomer unit on the properties of the complex, helping us to rationalize peculiar trends in catalytic activity. All the experimental results presented in this chapter were obtained by Qian Qin, Jie Li, Gabrielle Barozzino-Consiglio and Adrian Boborodea, in the frame of a collaboration with the group of Profs. K. Glinel and A. M.

Jonas at the Université Catholique de Louvain.[31] The compounds were synthesized by Qian Qin, Jie Li and Imane Nekkaa.

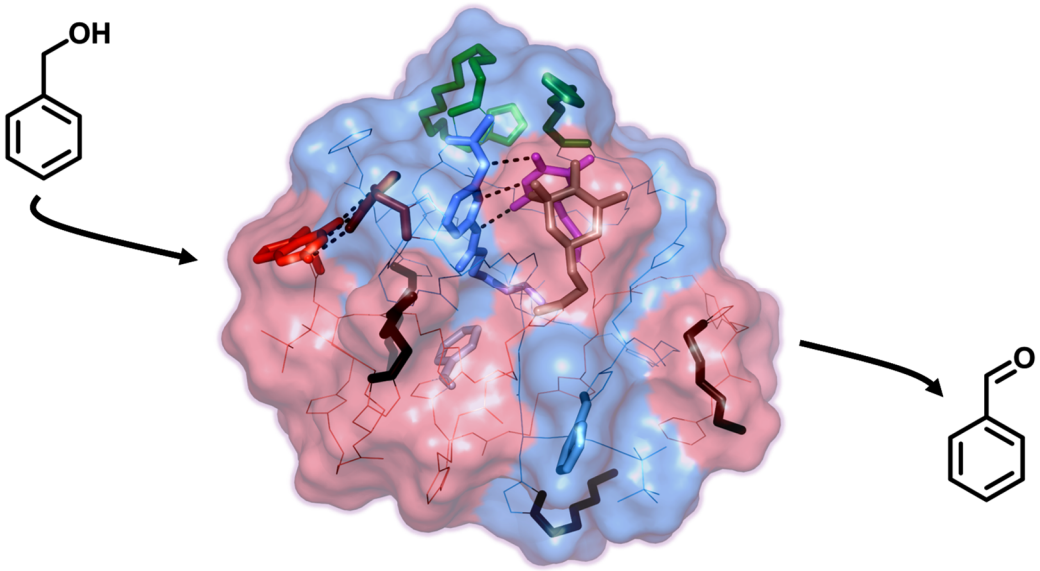


Figure V.1. Illustration of the globular supramolecular duplex formed by the assembly of two sequence-defined oligomers (one chain is represented in red, the other in blue), involved in the aerobic oxidation of alcohols. Each colored unit represents a side-chain of the oligomers (see their chemical structure in **Figure V.2**).

V.2. Design of the oligomers and supramolecular assemblies studied by MD simulations

To explore whether a supramolecular catalytic center can be encoded within self-assembled SDMs, we selected a multifunctional catalytic system developed for the aerobic oxidation of alcohols, based on 2,2,6,6-tetramethylpiperidine-1-oxyl (TEMPO), $Cu^{(I)}$-complexes (involving bidentate nitrogen ligands, such as bipyridine) and imidazoles.[32–35] The exact catalytic mechanism is still under discussion, but recent studies corroborate the formation of a four-membered intermediate composed of a dinuclear copper complex supported by two auxiliary ligands and interacting with a TEMPO moiety **(Figure V.2 A)**.[36–38] These five units were therefore attached as side-chains in two oligomeric strands, O_a and O_b, based on an oligo(urethane triazole) backbone **(Figure V.2 B)**. The first oligomer, O_a, contains the TEMPO radical (M) and a pyridyltriazole-copper complex (P), while O_b contains two imidazole co-ligands (I and I' having different spacer lengths for optimal accessibility) and the second P unit **(Figure V.2 C)**. For the system to be active, O_a and O_b must self-assemble; they were therefore both functionalized with two nucleobases, selected as the recognition units. These substituents, well-known for driving the secondary structures of nucleic acids, have been successfully used to control the supramolecular assembly of various synthetic systems.[39–42] O_a is decorated with a guanine (G) and a thymine (T), and O_b

with a cytosine (C) and a 2,6-diamidopyridine (D) unit, complementary through G---C and T---D hydrogen bonding interactions. The unnatural nucleobase D was preferred to the natural adenine because it can form three H-bonds with the thymine, instead of two for adenine. Finally, the oligomers carry hexyl side-chains (C_6) on both extremities, to improve solubility in the acetonitrile : dimethylsulfoxide 95 : 5 v/v solvent mixture and possibly contribute to stabilization of the self-assembled structure. In summary, the supramolecular catalyst is made by the combination of an hexamer, O_a, of sequence C_6-G-**M**-**P**-T-C_6 (catalytic units shown in bold), and an heptamer, O_b, of sequence C_6-C-**I'**-**I**-**P**-D-C_6. To demonstrate that all five units are required for catalytic activity, two alternative oligomers were designed as substitutes to O_b: O_{b2} and O_{b3}, lacking a P and an I group, respectively, yielding the incomplete catalytic centers O_a/O_{b2} and O_a/O_{b3}.

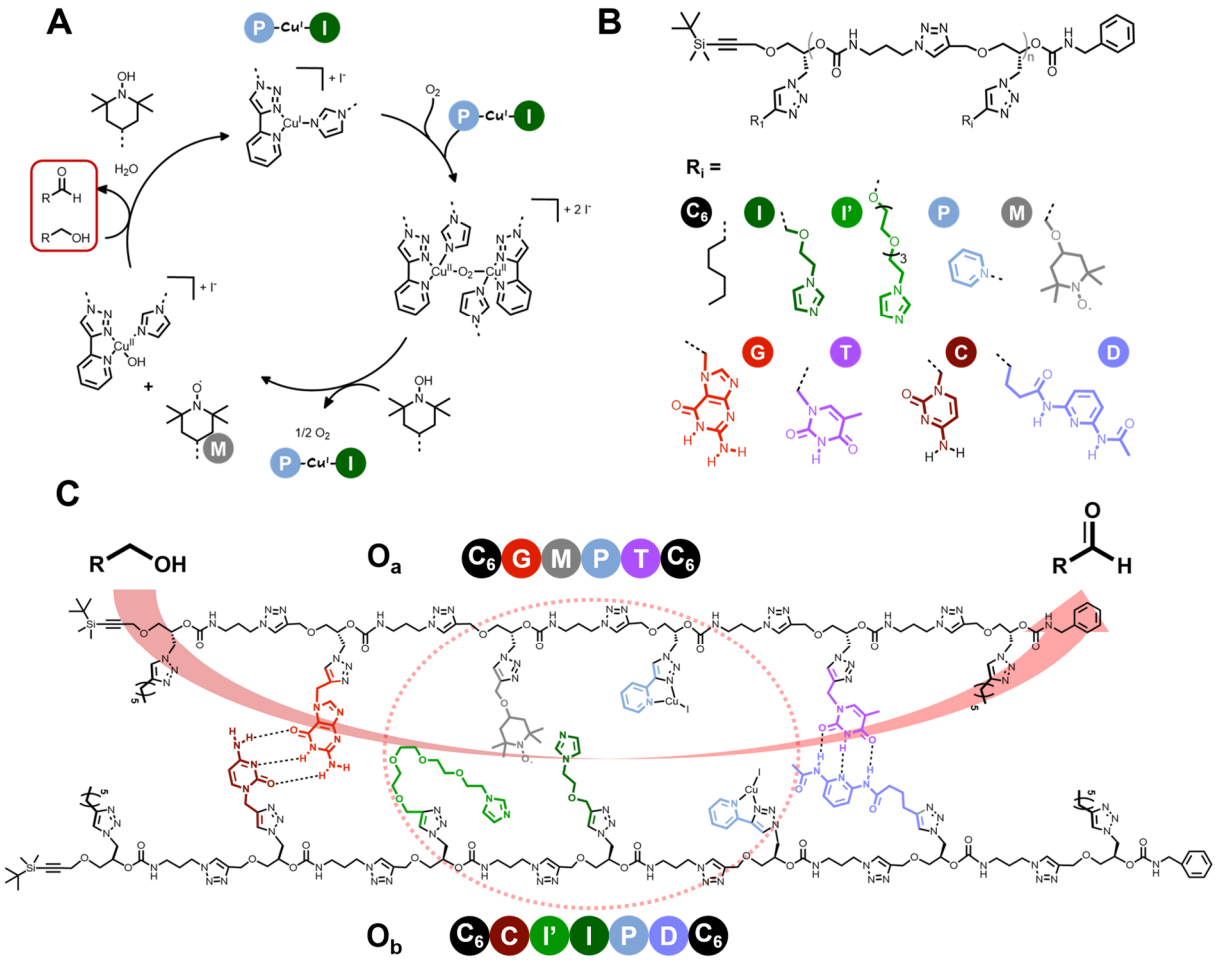
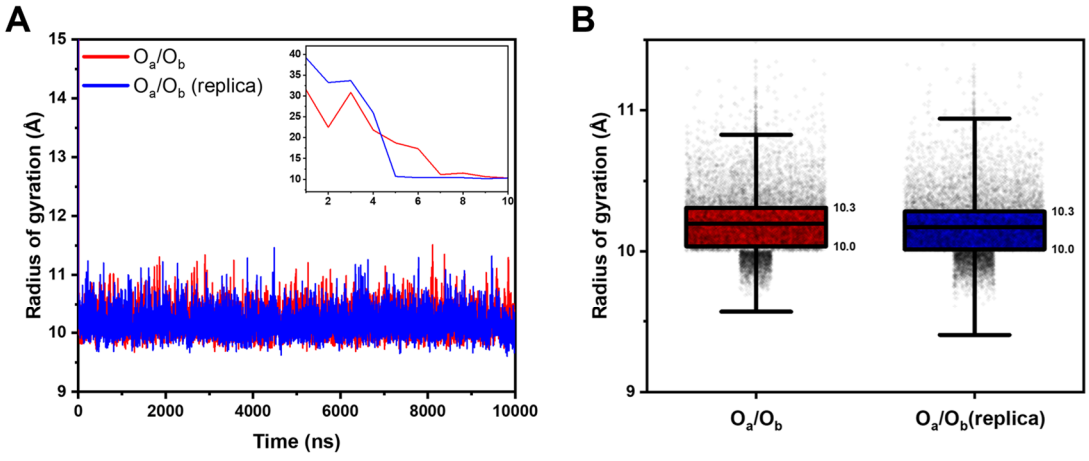


Figure V.2. Overview of the catalytic system studied. **(A)** Simplified catalytic mechanism of the aerobic oxidation of alcohols, showing the formation of the five-membered intermediate (shown in the right part of the cycle). The substrate transformation is highlighted in the red rectangle (shown in the left). **(B)** General chemical structure of the oligo(urethane triazole) backbone and side-chains library. **(C)** Chemical structure of the complete O_a/O_b catalytic duplex. The catalytic center is shown in the red dotted circle. Adapted from Ref. 31.

MD simulations were realized on the complete catalytic system, O_a/O_b, to elucidate the mechanisms of interaction between the two chains. Simulations were also carried out on an incomplete duplex, O_a/O_{b2}, which will be shortly discussed in comparison of O_a/O_b. The simulations were performed with the AMBER package, using an implicit solvent model with the dielectric constant of acetonitrile at 20 °C. Two independent replicas of 10 μs were generated for each duplex, with the O_a and O_b (or O_{b2}) chains being separated by more than 50 Å in the starting structure, to avoid initial contacts. Additional simulations were performed on the individual chains O_a and O_b for 5 μs, in two replicas (see **Section V.7** for full details of the simulation protocol).

V.3. The oligomers quickly fold and assemble into a highly dynamic globular duplex

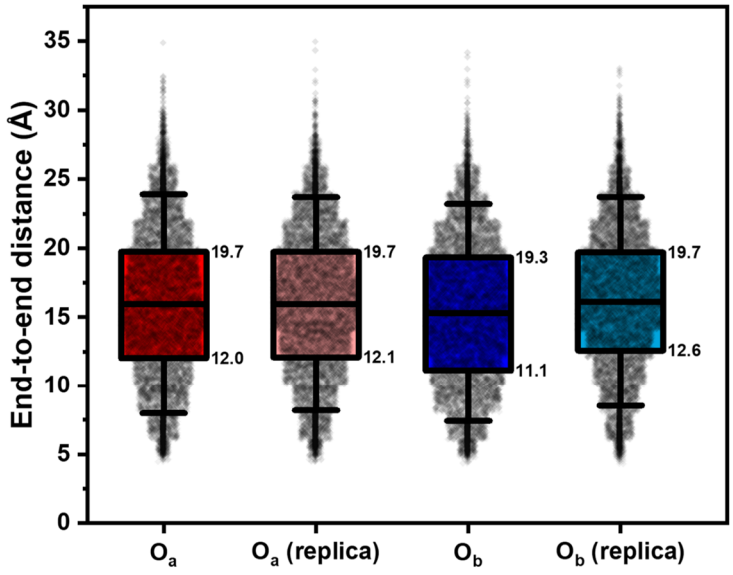
At the beginning of the simulations, the two strands (O_a and O_b) are separated, and the system is characterized by high radius of gyration (R_G) values (see inset in **Figure V.3 A**). The oligomers quickly assembled, within 10 ns, as indicated by the sharp decrease in the R_G for both replicas. The chains formed a compact and globular heteromolecular duplex, which remained stable during the whole simulations, as indicated by the R_G of the system oscillating around 10 Å. This value is in very good agreement with



Radius of gyration (Å)	Average ± Std. dev.	Initial	Final
O_a/O_b	10.2 ± 0.4	31.6	10.1
O_a/O_b (replica)	10.2 ± 0.5	39.2	9.9

Figure V.3. Data on the radius of gyration of the O_a/O_b duplex. **(A)** Evolution of the radius of gyration for both replicas of the O_a/O_b duplex over time, with an inset showing the fast decrease occurring in the first 10 ns. **(B)** Distribution of values for the whole simulations. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. **(C)** Table summarizing the data.

experimental measurements of the hydrodynamic radius (R_H), comprised between 8.6 and 14.6 Å for solutions of equimolar mixtures of O_a and O_b in dilute regimes (between 1 and 5 mM). These radii were calculated with the Stokes-Einstein equation, using diffusion coefficients measured by DOSY.[31] Inside the stable globule, the two strands undergo significant folding and remain highly flexible, allowing the different units to dynamically reorganize. This is shown by the end-to-end distance values measured for O_a and O_b, ranging from around 5 to 30 Å **(Figure V.4)**. It indicates that the chains can adopt very extended or very compact conformations in the duplex, without affecting its global globular shape.



End-to-end distance (Å)	O_a	O_a (replica)	O_b	O_b (replica)
Average ± Std. dev.	**15.9 ± 5.3**	**16.0 ± 5.2**	**15.3± 5.3**	**16.1 ± 5.1**

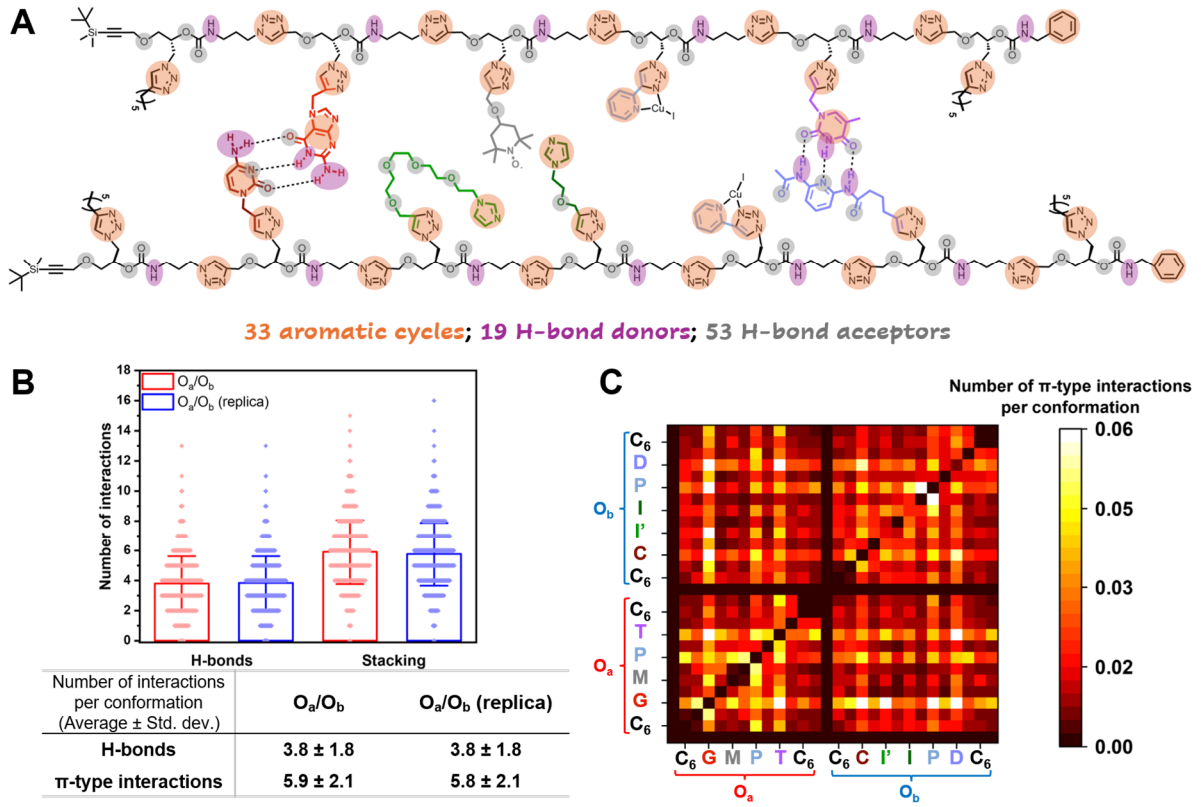
Figure V.4. Distribution of the end-to-end distance values for the O_a and O_b oligomers, measured for both replicas of the O_a/O_b duplex simulations. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. Statistics are given in the table below.

These observations reveal the formation of a disordered and highly dynamic supramolecular complex, where the side-chains appear randomly mixed within the globule, in marked contrast with the idealized 2D representation where each unit is precisely positioned along a linear backbone (see **Figure V.2 C**). During the simulations, the chains explore a large conformational space allowed by their intrinsic flexibility, arising from the presence of many freely rotatable bonds in their backbone

and side-chains. Globular and folded conformations are stabilized by a vast network of interactions, in particular π-type interactions[1] between the large number of triazole rings, but also with the nucleobases, imidazole moieties and pyridyl groups **(Figure V.5 A)**. The oligomers also present many H-bond donors and acceptors, such as the ether and urethane moieties in the backbone, as well as the nucleobases **(Figure V.5 A)**. Interestingly, we found very few hydrogen bonding interactions involving the triazole rings, which is in line with NMR measurements realized on monomer units.[43] We detected slightly more π-type interactions (around six per conformation) than H-bonds (around four per conformation) **(Figure V.5 B)**. The heatmap of π-type interactions highlights a homogeneous distribution of the stacking, involving all monomers and side-chains (although some units perform more interactions, in particular G) (**Figure V.5 C**, see **Figure V.11** in **Section V.7** for explanations on the per-residue decomposition adopted in the heatmap). While the total number of π-type interactions is significant, the contacts are weakly persistent, with only slightly more than one interaction every 20 conformations, on average, for the most frequent residue – residue interactions (see the white spots on the heatmap, corresponding to an average of 0.06 interaction per conformation). It shows that, within the globule, all residues dynamically interact together and can spatially regroup, even if they seem far from each other in the 2D representation.

Overall, these results indicate that a defined sequence of monomers does not necessarily translate into a well-defined 3D structure, particularly in the case of highly flexible chains displaying a large number of interaction sites. Such characteristics promote the formation of a disordered, globular system able to dynamically reorganize, thereby partially blurring the influence of the primary structure. The nucleobases, which contain H-bond donors and acceptors in addition to being aromatic structures, also contribute to many unspecific interactions, contrary to their organized behavior typically adopted within nucleic acids. However, in natural systems, these groups are combined with a rather rigid backbone, short side-chains, and a structure promoting well-defined stacking.[44] The chemistry of our system is very different, and the nucleobases are not tightly paired with their complementary partner. While the dynamic behavior of the assembly is not optimal to favor duplex formation over other poly(oligomeric) species, it could reveal advantageous for the catalytic process, which requires conformational flexibility for substrate binding and product release, as observed for enzymes.

[1] π-type interactions are counted between aromatic cycles following these geometric criteria: the distance between their centers of mass must be ≤ 5 Å, and the angle between their planes must be < 45 ° or > 135 °.



Number of interactions per conformation (Average ± Std. dev.)	O_a/O_b	O_a/O_b (replica)
H-bonds	3.8 ± 1.8	3.8 ± 1.8
π-type interactions	5.9 ± 2.1	5.8 ± 2.1

Figure V.5. Overview of the interactions stabilizing the O_a/O_b globule. **(A)** Chemical structure of the O_a/O_b duplex, highlighting the various interaction sites. The triazoles were not counted as H-bond acceptors, as they mainly perform aromatic interactions. **(B)** Number of H-bonds and π-type interactions per conformation measured during the whole simulations. Data is shown as mean ± standard deviation. The pale lines represent the distribution of measurements. Statistics are given in the table below. **(C)** Heatmap showing the decomposition of the π-type interactions by residue pairs (see **Figure V.11** for explanations on the per-residue decomposition). Each square, localized at the crossing of two residues, indicates the number of interactions per conformation detected between these residues. The heatmap is symmetrical with respect to the diagonal.

V.4. Specific interactions arise in the disordered duplex

Given the formation of a disordered globular duplex made of highly flexible chains, it seems that all monomers and side-chains contribute to a network of rather unspecific interactions. Interestingly, while this is true for the π-type interactions, a different behavior is observed for the H-bonds, as evidenced by a heatmap, used once more to localize the interactions **(Figure V.6 A)**. This heatmap displays three bright spots, indicating persistent interaction sites, while the other residue pairs contribute to a lesser extent to the network of H-bonds. The two most frequent H-bond sites (white squares) are located for the G---C and T---D pairs of residues, *i.e.* the complementary recognition units. In addition to these interactions, G---D pairing (yellow square)

provides a significant but less frequent mechanism of stabilization. This graph demonstrates that, despite the apparent disorder of the globule, specific hydrogen bonding interactions are able to emerge. A similar conclusion can be drawn from a second heatmap, showing the enthalpy of binding decomposed by residue pairs, highlighting the most stabilizing residue – residue interactions (darkest blue squares) in the duplex **(Figure V.6 B)**. These are localized for the pairs G---D, T---D and G---C, in excellent agreement with the heatmap of H-bonds. These analyses demonstrate that the recognition units play their role efficiently in the supramolecular assembly, although the duplex adopts compact conformations that allow all units to interact.

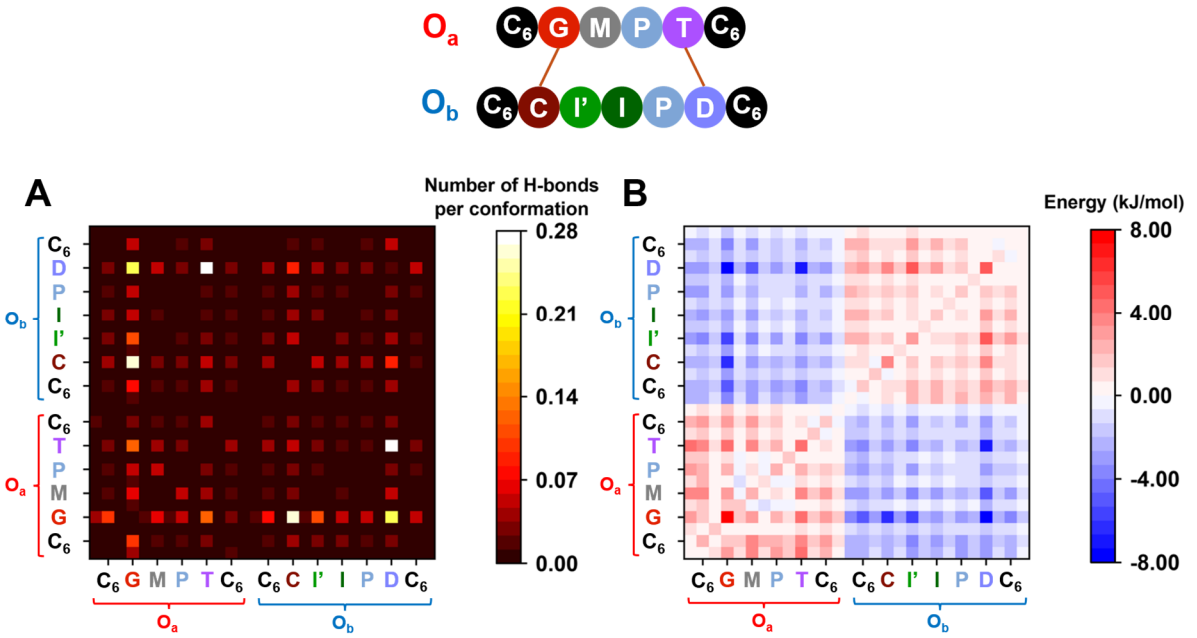


Figure V.6. Localization of the interactions stabilizing the supramolecular assembly. A cartoon representation of the primary structure of the O_a/O_b duplex is displayed above the heatmaps. **(A)** Heatmap showing the decomposition of H-bonds by residue pairs. **(B)** Heatmap showing the decomposition of the enthalpy of binding by residue pairs.

Another way to study the intermolecular contacts stabilizing the assembly is the use of network theory. The 3D conformations generated during the simulations can be converted into 2D networks, where each atom (except hydrogens) constitutes a node **(Figure V.7)**. Two nodes are connected by an edge if they are spatially close: here, the distance cutoff was set at 5 Å, to take into account contacts through H-bonds and π-type interactions. We computed an average network from the whole simulations to highlight the most persistent contacts, thus the strongest contributions to the assembly (see full details of the methodology in **Section V.7**). In line with the previous heatmaps, the network indicates that interchain connections arise dominantly from the nucleobases, especially through G---C, T---D and G---D pairings, whereas

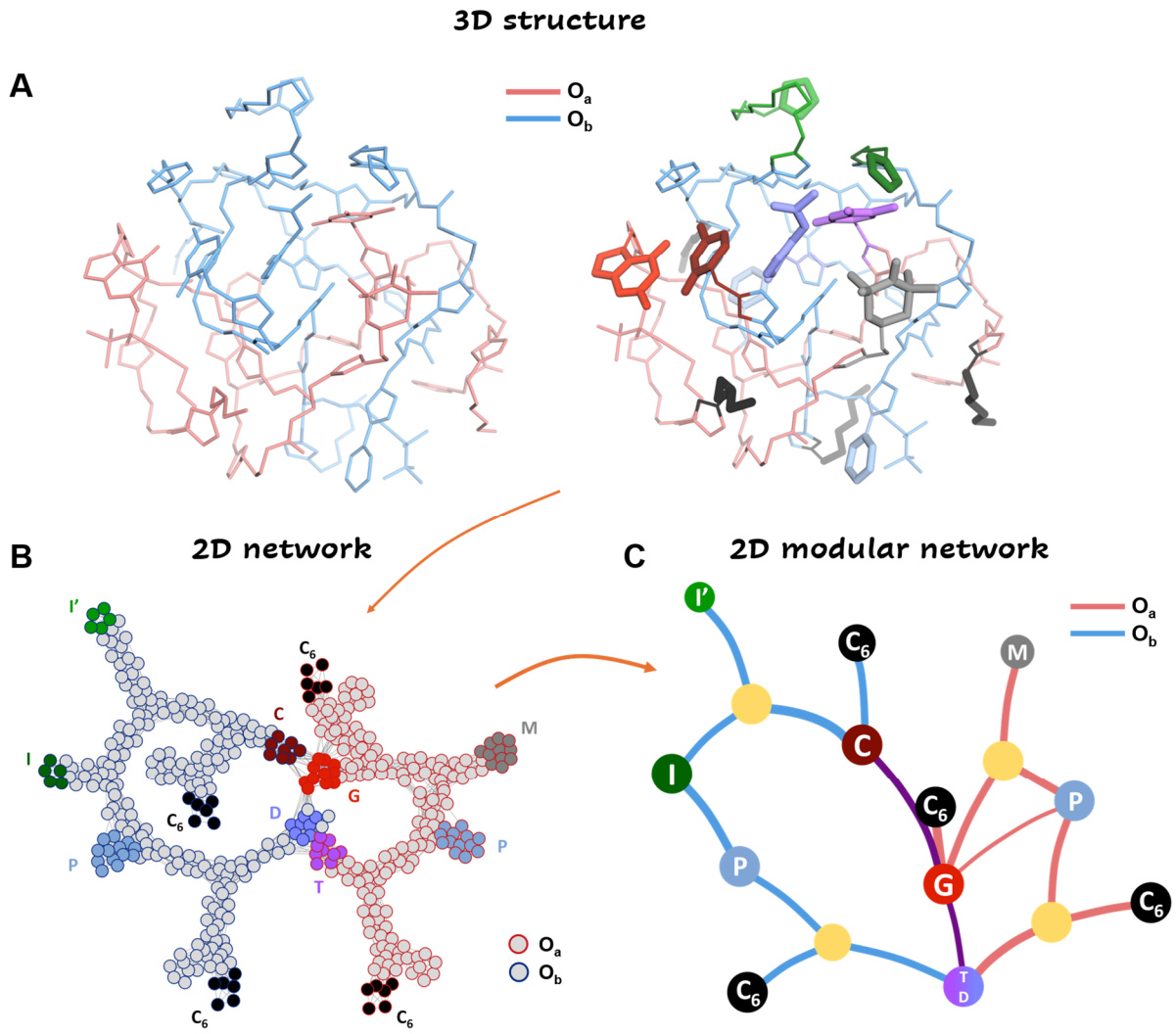
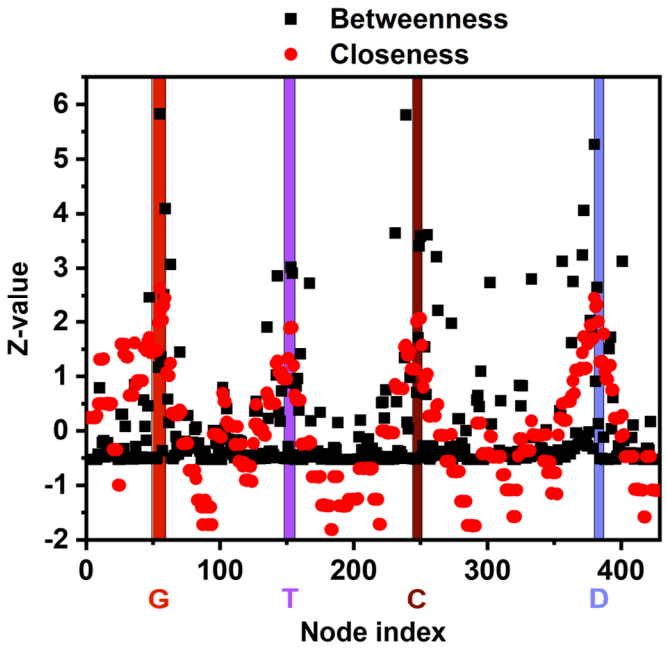


Figure V.7. Structure of the O_a/O_b duplex. **(A)** MD snapshot illustrating the globular structure of the assembly, with one color by oligomer chain (left) and with one color by substituent (right). The color code of the side-chains is the same as in the other figures. **(B)** Network representation of the system, highlighting the persistent contacts observed during the MD simulations. The nodes belonging to O_a and O_b are circled in red and blue, respectively, with the same color code for the functional groups. **(C)** Modular representation of the network. The modules in yellow contain backbone or chain-ends nodes. Intramolecular connections relying nodes belonging to O_a or O_b are represented by red and blue lines, respectively. Intermolecular connections between O_a and O_b are represented by purples lines. The nucleobases T and D are merged in the same module, indicating high connectivity.

intrachain connectivity replicates the primary structure of the oligomers **(Figure V.7 B)**. These visual observations are confirmed by several descriptors, such as the betweenness and closeness centralities, which are presented with their Z-values, *i.e.* the number of standard deviations by which the value is below or above the mean value **(Figure V.8)**. The betweenness describes the importance of a node to create

connections between the other nodes, while the closeness is related to the ability of a node to communicate with all the others (see more details in **Section V.7**). These two values are particularly high for the nodes belonging to the nucleobases, as expected given their strong contribution to the persistent contacts between the two chains.



	Hexyl tails (C_6)	Catalytic units (M, P, I', I)	Nucleobases (G, T, C, D)	Average (all nodes)
Closeness	0.12 ± 0.01	0.12 ± 0.01	0.20 ± 0.01	0.15 ± 0.03
Closeness (Z-value)	-0.46	-0.45	0.86	0
Betweenness (x10^2)	0.15 ± 0.12	0.20 ± 0.32	3.62 ± 4.80	1.37 ± 2.62
Betweenness (Z-value)	-1.18	-1.00	1.68	0

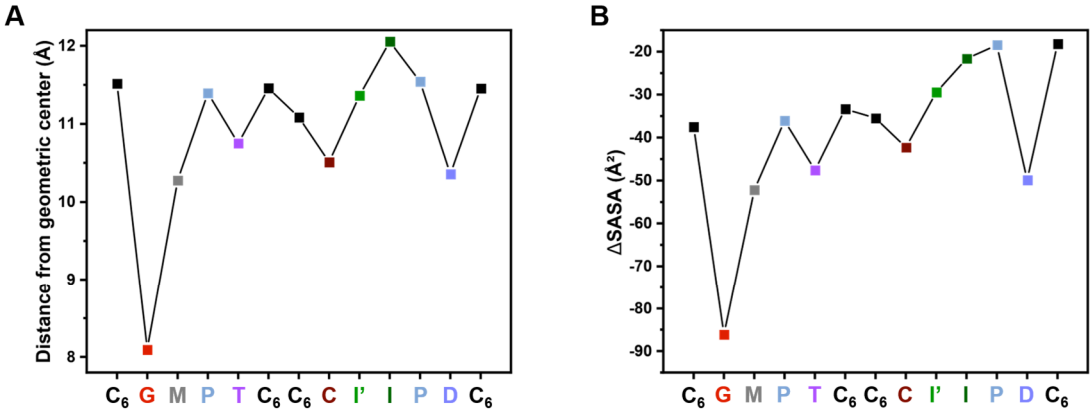
Figure V.8. Characterization of the connectivity inside the network with the betweenness and closeness centralities (see definitions in **Section V.7**), presented with their Z-value. The nodes belonging to the nucleobases (G, T, C and D) are highlighted in the graph, showing that higher Z-values are located for these nodes. Statistics are given in the table below.

A modular representation of the network gives a simpler, coarse-grained view of the system and shows again that the catalytic modules (M, I', I and P in **Figure V.7 C**) are connected through links involving the nucleobases. The network representations also suggest that the C_6 units do not contribute to the assembly, and do not tend to intertwin with each other, as could be thought when looking at the 2D structure. Viewing them on the 3D structures (see units in black in **Figures V.1 and V.7 A, right**), it appears likely that they are too short, compared to the size of the globule, to contribute to the binding. Overall, our simulations indicate that the duplex is stabilized by a vast network of dynamic interactions involving all residue pairs, no matter their position in the sequence, with more persistent contacts emerging dominantly between the recognition units, ensuring some specificity in the recognition.

Interestingly, the recognition pattern of the incomplete catalytic system O_a/O_{b2} follows the same trend **(Figure V.S1** in **Section V.8)**. This is not unexpected, O_{b2} having the same sequence as O_b except for one P monomer. Our simulations suggest that it is possible to bring small modifications in the composition of the catalytic center without affecting the mechanisms of assembly between the oligomers. It means that variations in the catalytic activity between the O_a/O_b and O_a/O_{b2} systems can be directly related to the contribution of individual catalytic units, making our SDMs particularly interesting to study the role of each monomer in the catalytic mechanism.

V.5. Rationalizing the trends in catalytic activities by combining MD simulations and experiments

For an efficient catalysis, the substrate must interact with the catalytic moieties when the O_a/O_b duplex is formed. We therefore investigated the accessibility of the different functional units to their surrounding environment **(Figure V.9)**. Two variables were used: the average distance from the geometric center of the globule, and the ΔSASA, *i.e.* the difference in accessibility of one residue in the duplex compared to its accessibility when the oligomer is alone (see **Section V.7** for details on the methodology). When looking at the distance from the geometric center, it appears that, on average, the nucleobases tend to be located closer to the center of the globule than



C

	Hexyl tails (C_6)	Catalytic units (M, P, I', I)	Nucleobases (G, T, C, D)
Distance from geometric center (Å) (Average ± Std. dev.)	11.4 ± 0.17	11.3 ± 0.6	9.9 ± 1.1
ΔSASA (Å²) (Average ± Std. Dev.)	-31.2 ± 7.6	-31.6 ± 12.0	-56.5 ± 17.3

Figure V.9. Data on the accessibility of the different residues in the O_a/O_b duplex to their environment. **(A)** Average distance of each functional unit from the geometric center of the assembly. **(B)** Average ΔSASA value of each functional unit, measured as the difference between the SASA of this unit in the O_a/O_b duplex and the SASA of the unit in the O_a or O_b chain alone. A high negative ΔSASA value indicates that the residue is significantly hindered by the formation of the assembly. **(C)** Table summarizing the data for each kind of substituents. For detailed statistics on the individual residues, see **Table V.S1** in **Section V.8**.

the hexyl tails or the catalytic units **(Figure V.9 A)**. The only exception to this trend is the TEMPO moiety (M), which is one of the five units with the shortest average distance to the geometric center. The measurements of ΔSASA confirm this trend **(Figure V.9 B)**. The nucleobases present the greater decrease of accessibility going from the isolated oligomers to the intermolecular duplex, as indicated by their higher negative ΔSASA values. Again, the hexyl tails and catalytic units display similar values, with a lower decrease of their SASA upon formation of the globule, at the exception of the TEMPO moiety. Although the differences between the different kinds of substituents are rather small, these two descriptors reveal that, on average, the hexyl tails and catalytic units spend more time at the periphery of the globule, indicating that substrates should be able to interact with the catalytic moieties. The nucleobases, on the other hand, tend to be more isolated from their environment, located more often at the interface of the two oligomers. This view is concordant with our network representations, and the fact that the nucleobases significantly contribute to intermolecular interactions.

With all this information in hand, we can now bring some insights into the experimental trends in catalytic activity. The systems were tested in the aerobic oxidation of benzyl alcohol into benzaldehyde, for different temperatures between 30 and 60 °C and at different molar concentrations of catalyst. $Cu^{(I)}$ was introduced in stoichiometric amount with respect to P groups, and catalyst concentration is expressed as the content in molar units relative to the molar concentration of introduced alcohol (0.2 M). The catalytic activity is represented by the turnover frequency (TOF) **(Figure V.10)**. The complete O_a/O_b system presents a peculiar bell-shaped curve, with a maximum TOF at around 1 mol % of catalyst concentration, and a decrease of activity at higher concentrations **(Figure V.10 A)**. We attribute this behavior to the formation of poly(oligomeric) aggregates at higher concentrations, where an important steric crowding would decrease accessibility to the catalytic center. These complexes would be stabilized by various unspecific interactions, involving the backbone H-bond donors and acceptors and the numerous aromatic rings. The maximum of catalytic activity for the O_a/O_b system is presumably reached when the heteromolecular duplex is predominant in solution. In comparison, the incomplete O_a/O_{b2} and O_a/O_{b3} complexes exhibit a significantly lower TOF **(Figure V.10 B)**. Additionally, their activity continues to increase well after 1 mol % of catalyst, meaning that, contrarily to O_a/O_b, they benefit from the formation of poly(oligomeric) species. This is not surprising if we consider that these systems require the assembly of at least three chains to regroup all five catalytic units, and it demonstrates that only one missing catalytic moiety in the chains is enough to significantly decrease the catalytic activity. Individual chains, which also contain an incomplete sequence of catalytic

units, equally display a very low TOF (diamonds in **Figure V.10 C**). A mixture of monomeric units with the same composition as the complete duplex displays a completely different trend, the catalytic activity increasing linearly with the concentration, as expected for a system relying on random encounters (crosses in **Figure V.10 C**). For the same reason, a mixture of MP dimers and II'P trimers, *i.e.* oligomers lacking the hydrogen binding units and the hexyl tails, also display a linear evolution of the TOF with respect to the catalyst concentration (pentagons in **Figure V.10 C**). The most striking feature of the O_a/O_b system is its resistance to dilution. It is the only system able to maintain a catalytic activity even at very low catalyst concentrations, a particularity that we ascribe to its self-assembling properties, leading to the formation of active heteromolecular duplexes in solution. Upon dilution, the TOF of the other systems quickly decreases, either because they lack recognition units or because catalytic units are missing when forming duplexes. Therefore, the unique catalytic properties of the O_a/O_b system must emerge from the formation of self-assembled duplexes, maintained through various interactions, with a particularly important role of the nucleobases and the presence of the five required catalytic moieties. However, the presence of a large number of interaction sites favors the formation of aggregates at higher concentrations, limiting the efficiency of the system above 1 mol % of catalyst.

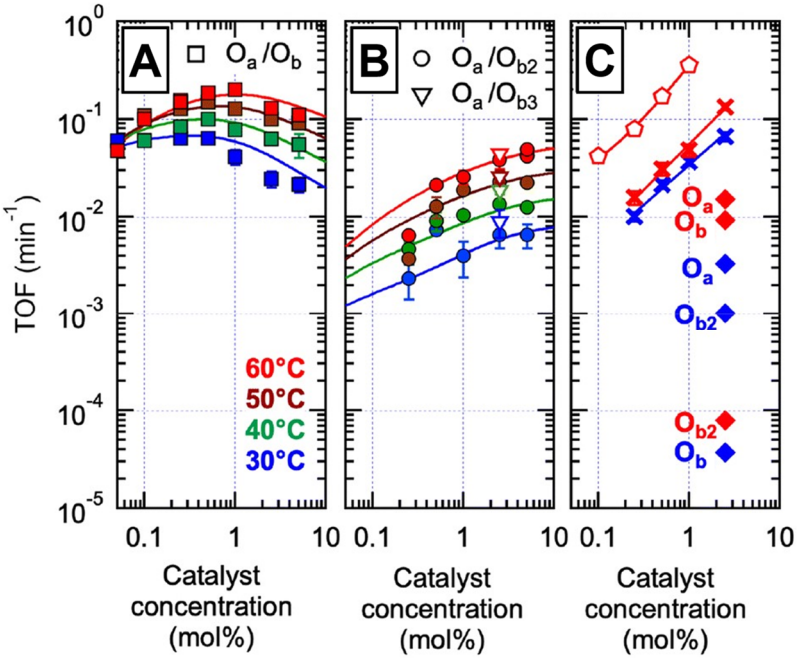


Figure V.10. Experimental measurements of the catalytic activity of the self-assembled oligomers. The TOF was calculated as the slope of the yield of oxidation versus time, divided by the molar content in M units in the catalyst. **(A)-(C)** TOF versus catalyst concentration at four temperatures (colors as indicated), for **(A)** the complete O_a/O_b system, **(B)** the incomplete O_a/O_{b2} (circles) and O_a/O_{b3} (open triangles) systems and **(C)** various control systems. Reproduced from Ref. 31.

V.6. Conclusion

We investigated the possibility to exploit precisely designed SDMs to build a complex multifunctional catalytic system, requiring the supramolecular assembly of two oligomers to be active. MD simulations revealed the formation of a globular duplex, inside which each chain is strongly folded and highly flexible. Our results indicate that 2D chemical structures can be misleading: here, this representation suggests the formation of a well-organized duplex, where each monomer unit faces a partner in the other chain in a precisely programmed manner. The 3D view given by the simulations provides important information on the supramolecular assembly of the two chains, stabilized by a myriad of interactions, and where the primary structure is partially blurred by the flexibility of the oligomers. Despite the apparent disorder in the duplex, heatmaps and network representations demonstrated that interactions involving the nucleobases contribute dominantly to the duplex stabilization, in particular through complementary H-bond pairings. Additionally, our simulations indicate that the catalytic units tend to remain accessible to their environment, at the periphery of the globule, while the nucleobases spend more time at the interface of the two chains. The simulations helped us to understand the peculiar trends in catalytic activity, where the O_a/O_b system revealed to be particularly efficient at high dilution. Such resistance to dilution was not observed on any other system, which either lacked the recognition units or only one catalytic moiety. These results indicate that the formation of the self-assembled duplex containing all five catalytic units is key to catalytic activity. Precisely controlling the monomeric composition of the system is therefore crucial, which is an important validation of our approach combining supramolecular assembly and catalytic activity inside SDMs. Additionally, such SDMs are very powerful in view of mechanistic studies, to probe the role of each substituent in the catalytic mechanism. However, it seems likely that the precise monomer ordering in our chains is not crucial for their activity, given the high flexibility of the system. Still, it allowed us to place the nucleobases far from each other inside the chains, which revealed to be very efficient to minimize intramolecular interactions, which would be detrimental to the intermolecular assembly. In comparison, tetramers with the same oligo(urethane triazole) backbone functionalized with four adjacent nucleobases displayed important intramolecular interactions.[43] MD simulations, combined with network theory, revealed to be a very precious tool to decipher the structure, dynamics and mechanisms of assembly of this complex supramolecular system, presenting a large number of interaction sites and an important flexibility.

Based on our simulations, several leads could be considered to improve the molecular design of the chains. Ideally, the equilibrium of species in solution should be even more

biased towards duplex formation, through supplementary interactions. It could be interesting to increase the number of recognition units, for example by replacing the hexyl tails, which seem too short to contribute to the stability of the duplex, by additional nucleobases. The nucleobases could also be substituted by other H-bonding units containing either only donor or only acceptor sites, thus promoting duplex formation over intramolecular folding, as demonstrated by others.[45,46] Modifying the chemical nature of the backbone may also be an option, to avoid the presence of numerous unspecific interaction sites, such as the carbamate units and the triazole rings, which promote aggregation at higher concentrations. Increasing the rigidity of the backbone to better retain the information encoded in the primary structure, thus increasing the specificity of the recognition, could also be envisioned. Of course, while these options are trivial from the point of view of a computational chemist, it may bring significant synthetic challenges; the triazole rings, for example, are unavoidable when using the click chemistry route exploited here. In conclusion, while there are plenty of alternative designs to test, it remains difficult to embark into lengthy syntheses without a clearer picture of the systems that are really worth the effort. Computational methods and a better fundamental understanding of sequence – structure relationships, leading to accurate predictive models, will be important tools to rationalize the design of such complex sequence-defined nanomaterials in the future.

V.7. Simulation protocol

The oligomeric chains were built as a series of fragments, or residues, with the *Avogadro* software **(Figure V.11)**.[47]

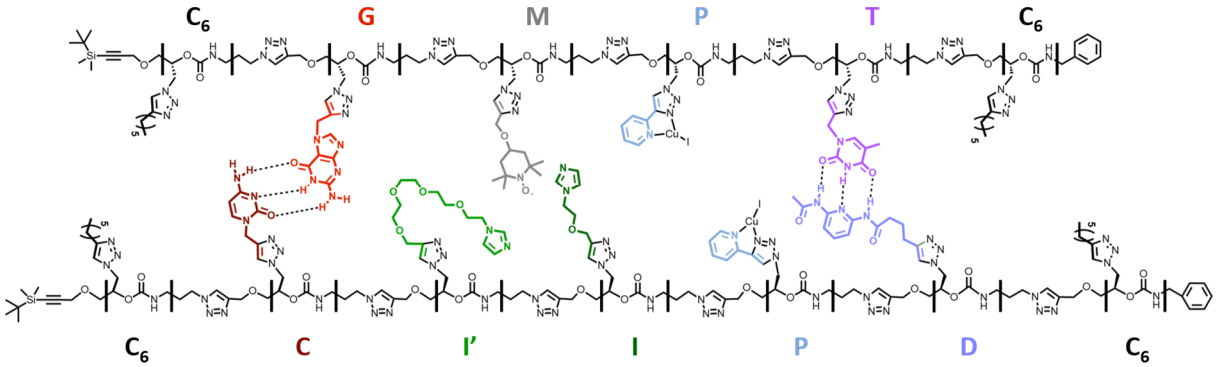


Figure V.11. Chemical structure of the complete O_a/O_b catalytic system. The two strands, O_a and O_b, are decomposed into a series of 28 residues, separated by black dots. The residues containing the functional side-chains are identified by letters.

These fragments were subsequently connected to one another to form the complete strands. The calculations were then performed with the AMBER simulation package, while free energy calculations were performed with the 2020 version of *ambertools*.[48]

The atomic partial charges were assigned to the fragments using the *antechamber* module of AMBER with the semi-empirical AM1-BCC method.[49] Structural parameters and partial charges of the TEMPO radical were used as reported by Stendardo *et al.* after an *ab initio* reparametrization.[50] Structural parameters and partial charges of the pyridyltriazole copper ligands were obtained by calculations at the quantum chemical level using density functional theory (DFT) with the B3LYP/6-31G** model and extra basis sets LanL2DZ and MIDI! for the copper and iodine atoms, respectively. All other force field parameters were given by GAFF 2.11.[51] The individual molecular fragments were connected in the desired sequence with the *LEaP* module of AMBER to constitute the complete oligomeric chains. When building the O_a/O_b catalytic system, O_b was translated by 40 Å in the *x*, *y* and *z* directions from O_a in order to avoid contacts between the two oligomers in the starting structure. A geometry optimization was then performed by MM, with a total of 10,000 steps distributed in 1,000 steps of steepest descent and 9,000 steps of conjugated gradient, in order to get a stable starting point for the subsequent MD simulations. These were carried out with an implicit solvent model, the Generalized Born (GB) model, to ensure a sufficient conformational sampling in a reasonable computational time.[52] The dielectric constant was set as the one of acetonitrile at 20 °C (ε = 37.5). For the catalytic system, constituted of the two strands O_a and O_b, two replicas of 10 µs were realized (same for the O_a/O_{b2} duplex). Each oligomer was also simulated alone in two replicas of 5 µs. The timestep was fixed to 1 fs and the temperature was maintained at 300 K with a Langevin thermostat, with a collision frequency of 1 ps^{-1}. A bond restraint was applied in the simulations of the two strands to avoid that they translate in opposite directions and never meet each other, as we are not in periodic conditions: when the distance between the chains exceeds about 75 Å, a force constant of 10 $kcal.mol^{-1}.Å^{-2}$ is activated to prevent the chains from moving further away. An infinite cut-off was selected for the non-bonded interactions. A snapshot was saved each ns during the MD simulations and extracted for further analyses, giving a total of 20,000 conformations for the catalytic duplexes and 10,000 for each oligomeric chain alone. The GPU version of AMBER was used for all minimizations and MD simulations. *PyMOL 2.5.4* was used to visualize the snapshots and to extract images.[53]

The analyses of the simulations were performed using the *cpptraj* module of AMBER.[54] Radii of gyration (R_G) were computed with respect to heavy atoms (all atoms except hydrogens). End-to-end distances were calculated as the distance between the carbon directly linked to the silicon in the tert-butyl moiety at one end, and the carbon in para position of the phenyl ring at the other end. Hydrogen bonds were detected with the default *cpptraj* parameters, *i.e.* a distance cutoff of 3.0 Å

between the acceptor and the donor heavy atom and an angle cutoff of 135 ° between the donor, the hydrogen atom and the acceptor. π-type interactions (parallel stacking) were detected using geometric criteria: two aromatic units are considered in interaction if the distance between their centers of mass is ≤ 5.0 Å and if the angle between the normal vectors of their planes is < 45° or > 135 °. The heatmaps of H-bonds and π-type interactions were built with in-house scripts. The residue decomposition follows the sequence order, and corresponds to the fragments used to build the chains, as is represented in **Figure V.11**. To compute the distance of each side-chain to the geometric center, only the heavy atoms located after the triazole ring were included in the calculation. For instance, for a C_6 unit, only the six carbon atoms of the hexyl tail were considered: the distance is measured between the geometric center of these six carbon atoms and the geometric center of the globule (same approach for the other side-chains). The first 100 ns were not included in this calculation, to let the duplex form and equilibrate. Solvent-accessible surface area (SASA) values were calculated with the LCPO model, as implemented within *cpptraj*.[55] The per-residue ΔSASA was computed as the difference between the SASA calculated for a residue in the simulation of the O_a/O_b duplex and the SASA of the same residue calculated in the simulation of the individual chains, O_a or O_b, using the per-residue scheme shown in **Figure V.11**. The ΔSASA can only be inferior or equal to zero: a residue that would be located far from the interface of the two strands, without contact with the second chain, would have a ΔSASA of zero, meaning that it is equally accessible with or without the second chain. Binding enthalpy calculations were performed with the Molecular mechanics Poisson-Boltzmann surface area (MMPBSA) method, using the parallelized version of the Python program *MMPBSA.py 14.0*, implemented in AMBER.[56] The binding enthalpy is given as the difference between the energy calculated for the complex (here, the complete catalytic system) and the sum of energies for the receptor (O_a) and ligand (O_b) alone. The multiple-trajectory approach was followed, which means that the conformations for the complex, receptor and ligand were obtained from independent simulations. *MMPBSA.py* post-processes the trajectories and calculates the energy of each frame with an implicit representation of the solvent. The energy is divided in two parts: an internal contribution and a solvation contribution. The internal contribution was given by the force field and can be seen as the energy of the system in vacuum (bonds, angles, dihedrals, van der Waals and electrostatics). The solvation contribution is further divided into a polar and a non-polar part. The polar part represents the electrostatic interactions between the solute and the solvent and was obtained by solving the PB equation with a finite difference method. The non-polar part was calculated as the sum of a favorable “dispersion term” and an unfavorable “cavity term”, representing the stabilizing solute – solvent

dispersion interactions and the cost of creating a cavity in the solvent, respectively. These two terms are proportional to the SASA. Several calculations were carried out by varying the internal dielectric constant, *i.e.* the dielectric constant of the solute, ranging from 1 to 4. The default value is set to 1, but in some cases, a better agreement with experimental results was obtained with higher values of the internal dielectric constant.[57,58] In our case, a better agreement was reached with an internal dielectric value of 4 (note that this parameter did not affect the qualitative trend observed in the residue – residue interactions). The first 100 ns of each replica were skipped for the calculations to let the systems reach equilibrium, giving a total of 19,800 conformations for the O_a/O_b duplex and 9,800 conformations for the oligomeric strands alone. The external dielectric constant was fixed at 37.5, as in the MD simulations. *MMPBSA.py* offers the possibility to decompose the energy by residue and by pairs of residues (using the same residue division as the one shown in **Figure V.11**). The residue pairwise decomposition scheme was used to highlight pairs of residues playing an important part in the binding of the two oligomeric chains.

The 2D networks were built from the 3D conformations generated during the simulations. In this representation, all heavy atoms constitute *nodes*, and two nodes are connected by an *edge* if their distance is inferior than or equal to 5 Å. This cutoff value allows to take into account hydrogen bonding interactions as well as π-type interactions. In practice, one network file was created for each conformation as soon as the duplex was formed (interchain distance < 14 Å) using in-house scripts, resulting in a total of 19,988 networks for O_a/O_b and 19,953 for O_a/O_{b2}. These files, representing one conformation each, have been used to build one global network for each system, following this procedure: two nodes are considered connected by an edge only if they have been in contact during at least 10 % of the MD time, *i.e.* if their contact was detected in at least 10 % of the 19,998 conformations for O_a/O_b and of the 19,953 conformations for O_a/O_{b2}. The resulting network is thus focusing on persistent contacts. All edges are undirected and unweighted. To analyze and visualize the network, the *Cytoscape 3.9.1* software was used with its included analyzer *NetworkAnalyzer 4.4.8*.[59,60] Two descriptors were chosen to characterize the nodes inside the network. The betweenness centrality C_b for one node is proportional to the number of shortest paths connecting two other nodes passing through this node, *i.e.* the importance of the node to put the other ones into communication. The closeness centrality C_c for one node reflects the reciprocal of the average shortest paths length connecting this node to all the other nodes in the network: the higher the closeness, the more "central" is the node in the network, the more easily it communicates with the other nodes. The network file was then submitted to the *Infomap* algorithm, in

order to detect *communities* or *modules*, which are defined as highly connected groups of nodes. The resulting modular representation can thus be considered as a coarse-grained view of the previous network. Various methods exist to decipher the modular topology of a network: the one that we used is called the *map equation*.[61,62] It is a flow-based method, which means that it focuses on "how is flowing, propagating the information from one node to another?". The propagation of information is materialized by a random walker that can move on the edges between nodes while an information cost, in bits, is associated with the movements of the walker. The main idea behind the map equation is that finding modules in the network can be seen as an encoding problem: to reduce at best the information cost associated to the random walk, it is necessary to efficiently partition, modularize the network. The map equation, based on Shannon's source coding theorem, gives the theoretical lower limit of the information cost associated with one step of the random walker, on average, on the network.[63] *Infomap* is the algorithm used to minimize the map equation. The principle is as follows: each node begins in its own module. Then, the nodes are moved into their neighboring module that reduces the most the map equation. This operation is repeated, the newly formed modules are merged with their neighbors, until no more minimization can be attained. The main goal of *Infomap* is thus to find the best partition of the network, *i.e.* the optimal organization of nodes inside the optimal number of modules, to reduce the most efficiently the information cost associated with the movements of a random walker. A two-level partition of the network was chosen, such as there is only one layer of modules containing the nodes (no possibility to have "modules inside a module"). Visualization and analysis of the network were done with the *2.6.0* version of the web server utility of *Infomap*.

V.8. Additional data

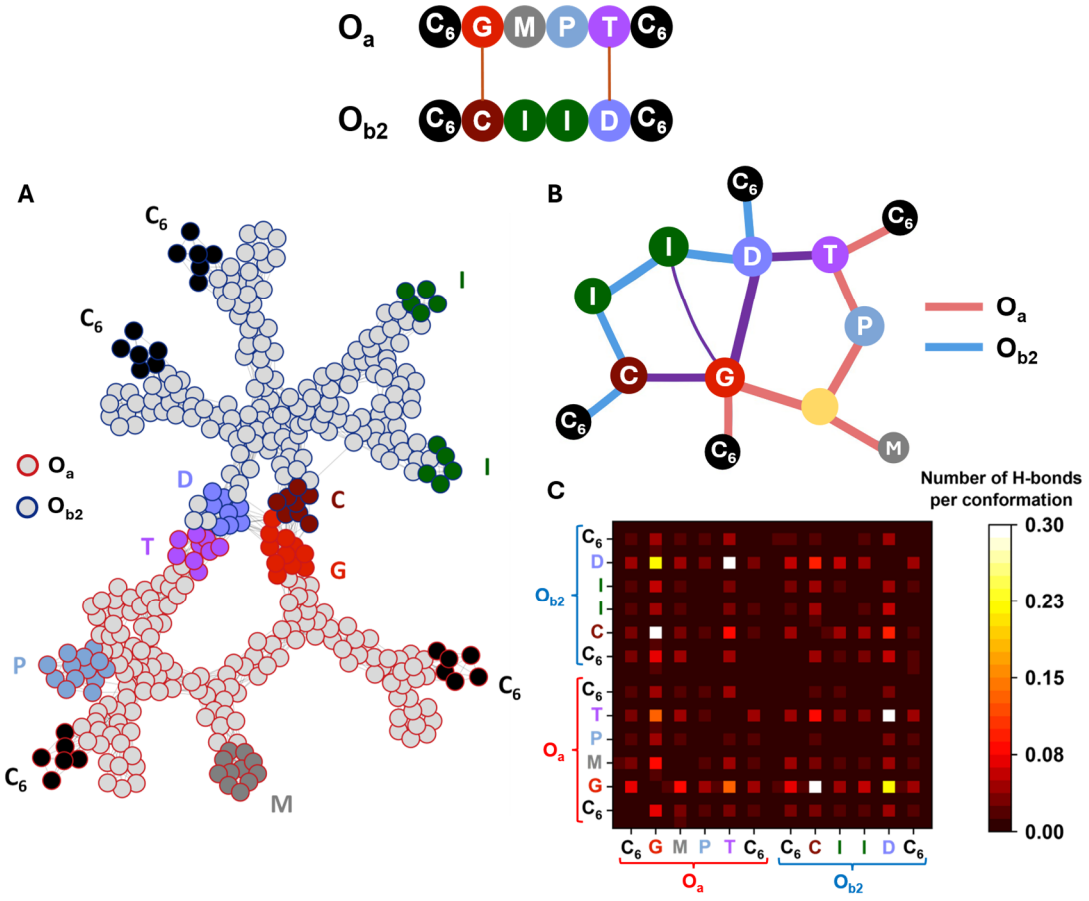


Figure V.S1. Overview of the mechanisms of assembly of the incomplete O_a/O_{b2} duplex. A cartoon representation of the primary structure of the incomplete O_a/O_{b2} duplex is displayed above. **(A)** Network representation of the system, highlighting the persistent contacts observed during the MD simulations. The nodes belonging to O_a and O_{b2} are circled in red and blue, respectively, with the same color code as in the other figures for the functional groups. **(B)** Modular representation of the network. The module in yellow contains backbone or chain-ends nodes. Intramolecular connections relying nodes belonging to O_a or O_{b2} are represented by red and blue lines, respectively. Intermolecular connections between O_a and O_{b2} are represented by purples lines. **(C)** Heatmap showing the decomposition of H-bonds by residue pairs.

	C_6	G	M	P	T	C_6	C_6	C	I'	I	P	D	C_6
Distance from geometric center (Å)	11.5 ± 2.2	8.1 ± 3.1	10.3 ± 2.3	11.4 ± 2.2	10.8 ± 2.6	11.5 ± 2.2	11.1 ± 2.3	10.5 ± 3.3	11.4 ± 2.0	12.1 ± 2.3	11.5 ± 2.1	10.4 ± 2.2	11.5 ± 2.2
Delta SASA (Å²)	-37.5 ± 0.8	-86.1 ± 0.8	-52.2 ± 0.9	-36.1 ± 0.6	-47.6 ± 0.9	-33.4 ± 0.7	-35.5 ± 0.9	-42.3 ± 0.9	-29.5 ± 1.1	-21.7 ± 0.8	-18.5 ± 0.6	-49.9 ± 1.0	-18.2 ± 0.8

Table V.S1. Statistics on the accessibility of the O_a/O_b functional units to their environment. The distance from the geometric center is given as mean ± standard deviation. The ΔSASA is given as difference between means ($mean_{complex} - mean_{chain\ alone}$, as explained in **Section V.7**) ± standard error of the difference between the two means.

References

[1] A. Judge, M. S. Dodd. Metabolism. *Essays Biochem.* **2020**, *64*, 607–647.

[2] A. Kohen. Role of Dynamics in Enzyme Catalysis: Substantial versus Semantic Controversies. *Acc. Chem. Res.* **2015**, *48*, 466–473.

[3] P. K. Agarwal, D. N. Bernard, K. Bafna, N. Doucet. Enzyme Dynamics: Looking Beyond a Single Structure. *ChemCatChem* **2020**, *12*, 4704–4720.

[4] C. Zhang, X. Xue, Q. Luo, Y. Li, K. Yang, X. Zhuang, Y. Jiang, J. Zhang, J. Liu, G. Zou, X.-J. Liang. Self-Assembled Peptide Nanofibers Designed as Biological Enzymes for Catalyzing Ester Hydrolysis. *ACS Nano* **2014**, *8*, 11715–11723.

[5] S. Liu, P. Du, H. Sun, H.-Y. Yu, Z.-G. Wang. Bioinspired Supramolecular Catalysts from Designed Self-Assembly of DNA or Peptides. *ACS Catal.* **2020**, *10*, 14937–14958.

[6] S. A. Green, H. R. Montgomery, T. R. Benton, N. J. Chan, H. M. Nelson. Regulating Transition-Metal Catalysis through Interference by Short RNAs. *Angew. Chem. Int. Ed.* **2019**, *58*, 16400–16404.

[7] J. J. Marek, R. P. Singh, A. Heuer, U. Hennecke. Enantioselective Catalysis by Using Short, Structurally Defined DNA Hairpins as Scaffold for Hybrid Catalysts. *Chem. Eur. J.* **2017**, *23*, 6004–6008.

[8] M. A. Aleman Garcia, Y. Hu, I. Willner. Switchable supramolecular catalysis using DNA-templated scaffolds. *Chem. Commun.* **2016**, *52*, 2153–2156.

[9] I. Drienovská, G. Roelfes. Artificial Metalloenzymes for Asymmetric Catalysis by Creation of Novel Active Sites in Protein and DNA Scaffolds. *Isr. J. Chem.* **2015**, *55*, 21–31.

[10] E. B. Pimentel, T. M. Peters-Clarke, J. J. Coon, J. D. Martell. DNA-Scaffolded Synergistic Catalysis. *J. Am. Chem. Soc.* **2021**, *143*, 21402–21409.

[11] M. Kurbasic, A. M. Garcia, S. Viada, S. Marchesan. Heterochiral tetrapeptide self-assembly into hydrogel biomaterials for hydrolase mimicry. *J. Pep. Sci.* **2022**, *28*, e3304.

[12] C. Zhang, R. Shafi, A. Lampel, D. MacPherson, C. G. Pappas, V. Narang, T. Wang, C. Maldarelli, R. V. Ulijn. Switchable Hydrolase Based on Reversible Formation of Supramolecular Catalytic Site Using a Self-Assembling Peptide. *Angew. Chem. Int. Ed.* **2017**, *56*, 14511–14515.

[13] M. Tena-Solsona, J. Nanda, S. Díaz-Oltra, A. Chotera, G. Ashkenasy, B. Escuder. Emergent Catalytic Behavior of Self-Assembled Low Molecular Weight Peptide-Based Aggregates and Hydrogels. *Chem. Eur. J.* **2016**, *22*, 6687–6694.

[14] G. Gulseren, M. A. Khalily, A. B. Tekinay, M. O. Guler. Catalytic supramolecular self-assembled peptide nanostructures for ester hydrolysis. *J. Mater. Chem. B* **2016**, *4*, 4605–4611.

[15] T. Terashima, T. Mes, T. F. A. De Greef, M. A. J. Gillissen, P. Besenius, A. R. A. Palmans, E. W. Meijer. Single-Chain Folding of Polymers for Catalytic Systems in Water. *J. Am. Chem. Soc.* **2011**, *133*, 4742–4745.

[16] H. Rothfuss, N. D. Knöfel, P. W. Roesky, C. Barner-Kowollik. Single-Chain Nanoparticles as Catalytic Nanoreactors. *J. Am. Chem. Soc.* **2018**, *140*, 5875–5881.

[17] Y. Liu, S. Pujals, P. J. M. Stals, T. Paulöhrl, S. I. Presolski, E. W. Meijer, L. Albertazzi, A. R. A. Palmans. Catalytically Active Single-Chain Polymeric Nanoparticles: Exploring Their Functions in Complex Biological Media. *J. Am. Chem. Soc.* **2018**, *140*, 3423–3433.

[18] K. J. Prathap, G. Maayan. Metallopeptoids as efficient biomimetic catalysts. *Chem. Commun.* **2015**, *51*, 11096–11099.

[19] M. Raynal, F. Portier, P. W. N. M. van Leeuwen, L. Bouteiller. Tunable Asymmetric Catalysis through Ligand Stacking in Chiral Rigid Rods. *J. Am. Chem. Soc.* **2013**, *135*, 17687–17690.

[20] Z. C. Girvin, S. H. Gellman. Exploration of Diverse Reactive Diad Geometries for Bifunctional Catalysis via Foldamer Backbone Variation. *J. Am. Chem. Soc.* **2018**, *140*, 12476–12483.

[21] J.-F. Lutz. The future of sequence-defined polymers. *Eur. Polym. J.* **2023**, *199*, 112465.

[22] R. Szweda. Sequence- and stereo-defined macromolecules: Properties and emerging functionalities. *Prog. Polym. Sci.* **2023**, *145*, 101737.

[23] P. Chandra, A. M. Jonas, A. E. Fernandes. Sequence and Surface Confinement Direct Cooperativity in Catalytic Precision Oligomers. *J. Am. Chem. Soc.* **2018**, *140*, 5179–5184.

[24] J. Li, Q. Qin, S. Kardas, M. Fossépré, M. Surin, A. E. Fernandes, K. Glinel, A. M. Jonas. Sequence Rules the Functional Connections and Efficiency of Catalytic Precision Oligomers. *ACS Catal.* **2022**, *12*, 2126–2131.

[25] S. Kardas, M. Fossépré, V. Lemaur, A. E. Fernandes, K. Glinel, A. M. Jonas, M. Surin. Revealing the Organization of Catalytic Sequence-Defined Oligomers via Combined Molecular Dynamics Simulations and Network Analysis. *J. Chem. Inf. Model.* **2022**, *62*, 2761–2770.

[26] B. Gong. Molecular Duplexes with Encoded Sequences and Stabilities. *Acc. Chem. Res.* **2012**, *45*, 2077–2087.

[27] S. C. Leguizamon, T. F. Scott. Sequence-selective dynamic covalent assembly of information-bearing oligomers. *Nat. Commun.* **2020**, *11*, 784.

[28] K. R. Strom, J. W. Szostak. Folding and Duplex Formation in Sequence-Defined Aniline Benzaldehyde Oligoarylacetylenes. *J. Am. Chem. Soc.* **2022**, *144*, 18350–18358.

[29] G. Iadevaia, C. A. Hunter. Recognition-Encoded Synthetic Information Molecules. *Acc. Chem. Res.* **2023**, *56*, 712–727.

[30] M. Dhiman, J. T. Smith, C. A. Hunter. Supramolecular assembly properties of a mixed-sequence recognition-encoded melamine oligomer. *Org. Biomol. Chem.* **2025**, *23*, 6948–6956.

[31] Q. Qin, J. Li, D. Dellemme, M. Fossépré, G. Barozzino-Consiglio, I. Nekkaa, A. Boborodea, A. E. Fernandes, K. Glinel, M. Surin, A. M. Jonas. Dynamic self-assembly of supramolecular catalysts from precision macromolecules. *Chem. Sci.* **2023**, *14*, 9283–9292.

[32] J. M. Hoover, S. S. Stahl. Highly Practical Copper(I)/TEMPO Catalyst System for Chemoselective Aerobic Oxidation of Primary Alcohols. *J. Am. Chem. Soc.* **2011**, *133*, 16901–16910.

[33] J. M. Hoover, B. L. Ryland, S. S. Stahl. Mechanism of Copper(I)/TEMPO-Catalyzed Aerobic Alcohol Oxidation. *J. Am. Chem. Soc.* **2013**, *135*, 2357–2367.

[34] J. M. Hoover, B. L. Ryland, S. S. Stahl. Copper/TEMPO-Catalyzed Aerobic Alcohol Oxidation: Mechanistic Assessment of Different Catalyst Systems. *ACS Catal.* **2013**, *3*, 2599–2605.

[35] B. L. Ryland, S. D. McCann, T. C. Brunold, S. S. Stahl. Mechanism of Alcohol Oxidation Mediated by Copper(II) and Nitroxyl Radicals. *J. Am. Chem. Soc.* **2014**, *136*, 12166–12173.

[36] J. Rabeah, U. Bentrup, R. Stößer, A. Brückner. Selective Alcohol Oxidation by a Copper TEMPO Catalyst: Mechanistic Insights by Simultaneously Coupled Operando EPR/UV-Vis/ATR-IR Spectroscopy. *Angew. Chem. Int. Ed.* **2015**, *127*, 11957–11960.

[37] K. Warm, G. Tripodi, E. Andris, S. Mebs, U. Kuhlmann, H. Dau, P. Hildebrandt, J. Roithová, K. Ray. Spectroscopic Characterization of a Reactive [Cu 2 (μ-OH) 2] 2+ Intermediate in Cu/TEMPO Catalyzed Aerobic Alcohol Oxidation Reaction. *Angew. Chem. Int. Ed.* **2021**, *60*, 23018–23024.

[38] W. Zhong, J. Luo, Z. Liu, G. Zhan, L. Zhu, C. Lu, Z. Shen, X. Li, X. Liu. The mechanistic diversity of the selective aerobic oxidation of alcohols catalyzed by systems derived from CuI and a diamine ligand. *Inorg. Chem. Front.* **2023**, *10*, 3139–3150.

[39] A. Sikder, C. Esen, R. K. O'Reilly. Nucleobase-Interaction-Directed Biomimetic Supramolecular Self-Assembly. *Acc. Chem. Res.* **2022**, *55*, 1609–1619.

[40] A. del Prado, D. González-Rodríguez, Y. Wu. Functional Systems Derived from Nucleobase Self-assembly. *ChemistryOpen* **2020**, *9*, 409–430.

[41] S. Cheng, M. Zhang, N. Dixit, R. B. Moore, T. E. Long. Nucleobase Self-Assembly in Supramolecular Adhesives. *Macromolecules* **2012**, *45*, 805–812.

[42] M. Surin. From nucleobase to DNA templates for precision supramolecular assemblies and synthetic polymers. *Polym. Chem.* **2016**, *7*, 4137–4150.

[43] K. Grafskaia, Q. Qin, J. Li, D. Magnin, D. Dellemme, M. Surin, K. Glinel, A. M. Jonas. Chain stretching in brushes favors sequence recognition for nucleobase-functionalized flexible precise oligomers. *Soft Matter* **2024**, *20*, 8303–8311.

[44] E. T. Kool. Hydrogen Bonding, Base Stacking, and Steric Effects in DNA Replication. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 1–22.

[45] D. Núñez-Villanueva, G. Iadevaia, A. E. Stross, M. A. Jinks, J. A. Swain, C. A. Hunter. H-Bond Self-Assembly: Folding versus Duplex Formation. *J. Am. Chem. Soc.* **2017**, *139*, 6654–6662.

[46] M. Dhiman, J. T. Smith, C. A. Hunter. Supramolecular assembly properties of a mixed-sequence recognition-encoded melamine oligomer. *Org. Biomol. Chem.* **2025**, *23*, 6948–6956.

[47] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, G. R. Hutchison. Open Access Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Software* **2012**, *4*, 17.

[48] D. A. Case, T. E. Cheatham Iii, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[49] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[50] E. Stendardo, A. Pedone, P. Cimino, M. Cristina Menziani, O. Crescenzi, V. Barone. Extension of the AMBER force-field for the study of large nitroxides in condensed phases: an ab initio parameterization. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11697.

[51] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case. Development and Testing of a General Amber Force Field. **2004**, *25*, 1157–1174.

[52] A. V. Onufriev, D. A. Case. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **2019**, *48*, 275–296.

[53] Schrödinger LLC, *The PyMOL Molecular Graphics System, Version 2.5.4*, **2015**.

[54] D. R. Roe, T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

[55] J. Weiser, P. S. Shenkin, W. Clark Still. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217230.

[56] B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, A. E. Roitberg. *MMPBSA.py* : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314–3321.

[57] H. Sun, Y. Li, S. Tian, L. Xu, T. Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys. Chem. Chem. Phys.* **2014**, *16*, 16719–16729.

[58] S. Genheden, U. Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery* **2015**, *10*, 449–461.

[59] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.

[60] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, M. Albrecht. Computing topological parameters of biological networks. *Bioinformatics* **2008**, *24*, 282–284.

[61] M. Rosvall, D. Axelsson, C. T. Bergstrom. The map equation. *Eur. Phys. J.: Spec. Top.* **2009**, *178*, 13–23.

[62] M. Rosvall, C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1118–1123.

[63] C. E. Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

VI. Revealing the folding of single-chain polymeric nanoparticles at the atomistic scale by combining computational modeling and X-ray scattering

The last results chapter of our thesis is dedicated to the study of purely synthetic polymers, whose folding in water was studied by a combination of MD simulations and experiments, in particular small-angle X-ray scattering (SAXS). Despite their artificial nature, these systems aim to reproduce the controlled folding properties of natural SDMs. We applied our methodology to elucidate the 3D structure and folding dynamics of two families of polymers functionalized with different types of hydrophilic side-chains. Unlike the other systems studied during our thesis, these macromolecules are not characterized by a perfect control over the sequence of monomers. However, sequence effects were investigated *in silico*, in order to understand whether controlling the primary structure would be beneficial for these chains.

Part of this work is reported in: *Revealing the folding of single-chain polymeric nanoparticles at the atomistic scale by combining computational modeling and X-ray scattering*.
S. Wijker, D. Dellemme, L. Deng, B. Fehér, I. K. Voets, M. Surin, A. R. A. Palmans, *ACS Macro Lett.*, **2025**, 14, 426-433.

VI.1. Introduction

As a core concept in this thesis, functional biomacromolecules display exquisite properties, acquired through their well-defined 3D structure and programmed folding process. Enzymes and many biological receptors need to fold to acquire their function, and their activity can be modulated through small conformational changes.[1,2] The field of single-chain polymeric nanoparticles (SCPNs) aims to reproduce this ability in synthetic materials, enabling them to acquire a function through controlled folding in solution.[3–5] To control the conformations of the polymer chains in solution, it is tempting to take inspiration from protein folding. Solvophilic/solvophobic effects and non-covalent interactions, such as hydrogen bonding, metal coordination or host-guest complexation have been extensively used to design SCPNs.[6–18] Introducing covalent intramolecular crosslinks is also an efficient way to stabilize single-chain systems, and this approach is sometimes combined with supramolecular interactions and amphiphilic effects to get a better control on the folded structures.[19–21] Recently, an SCPN was designed to collapse in water due to hydrophobic effects and supramolecular interactions, while remaining in a random coil state in tetrahydrofuran (THF).[20] Using photoinduced covalent crosslinking, the compact and extended conformations

in water could be “locked” and retained when introducing the SCPN in THF. SCPNs are envisioned for various applications, notably for catalysis in water or in complex media.[22–29] Biomedical applications are also strongly investigated, with examples in drug delivery, cellular targeting, bioimaging, or biosensing.[30–33] For such applications, which operate in complex biological media, it is crucial to assess the conformational stability of SCPNs.[19,34] More generally, it is of utmost importance to resolve the 3D structure of SCPNs, as their morphology is key to their function. However, getting a precise picture of the 3D structure of such nanoparticles in solution, in particular concerning their internal structure, remains challenging.[35,36] Typically, techniques such as dynamic light scattering (DLS), size exclusion chromatography (SEC), or nuclear magnetic resonance (NMR) are used to obtain information on the size of the nanoparticles. These methods can detect size variations, enabling a distinction between presumably extended and folded chains. A decrease in the measured size is generally attributed to the efficient folding of the system, without providing any information on its conformation in solution. As discussed in this chapter, it is very difficult to distinguish pure single-chains from small aggregates based on size measurements alone. Fluorescent probes were also used to investigate the formation of hydrophobic compartments in surface-immobilized SCPNs, allowing single-chain resolution.[37] While being impressive, this example does not provide clear information on the 3D structure of the chains, and does not unequivocally demonstrate SCPN formation. The introduction of scattering techniques such as small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS) were important to get finer insights on the 3D structure of SCPNs in solution. These methods challenged the naive view that ‘any’ copolymer structure composed of the correct ratio of hydrophilic and hydrophobic grafts would form a globular core-shell structure.[9,10,38] It is now well established that the conformational landscape of SCPNs more closely resembles that of intrinsically disordered proteins than that of globular proteins.[39] The improvements in computational power have also enabled the use of MD simulations, with the unique ability to provide a direct picture of the 3D structure of SCPNs in solution, to investigate their internal structure, and to reveal their folding dynamics.[40–42] Recently, MD simulations based on a very simple physical model were combined with machine learning to provide a complete mapping of the conformational landscape of SCPNs based on the position of their cross-linking units.[43]

Our work aims at combining the structural information brought by MD simulations and SAXS experiments. While the theoretical investigation provides a picture of unmatched atomistic resolution on the structure of SCPNs, it is necessary to compare our models to experiments to validate their robustness. We applied our MD protocol to two different kinds of polymers, distinguished by the nature of their hydrophilic

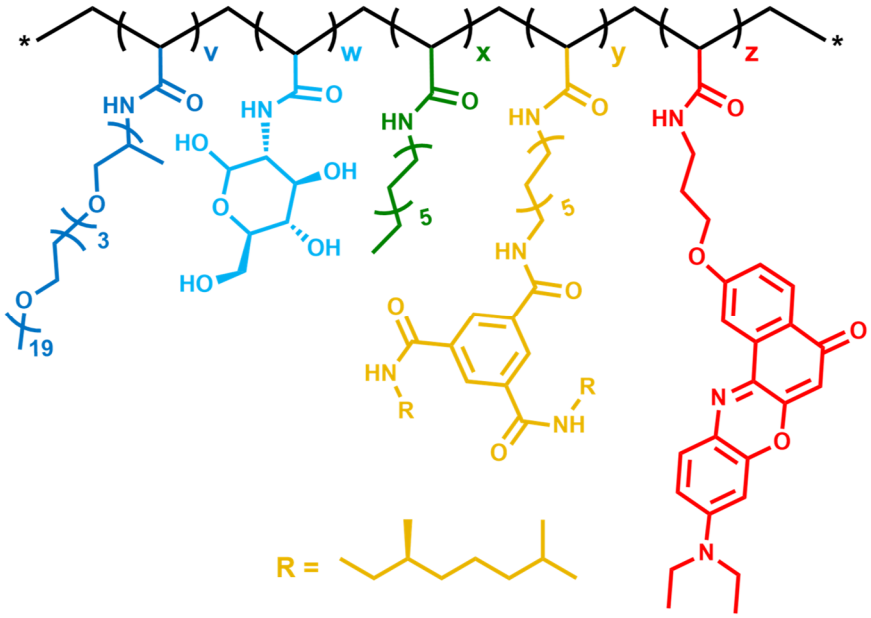
grafts. Our results indicate that very different morphologies are obtained in water. Furthermore, the combination of MD simulations and SAXS allowed us to reveal the formation of small aggregates having a very similar size and shape than the single-chain systems for one polymer family, demonstrating the interest of our approach. This work has been carried out through a collaboration with the group of Prof. A. Palmans at Technische Universiteit Eindhoven. All the experimental results were obtained by Stefan Wijker and Bence Fehér.[44] The polymers were synthesized by Stefan Wijker and Linlin Deng.

VI.2. Design of the two polymer families studied by MD simulations in water

Two families of polymers and random copolymers were designed through sequential amine postfunctionalization of poly(pentafluorophenyl acrylate), and functionalized with five different kinds of grafts **(Figure VI.1)**. The two families are mainly distinguished by the nature of their hydrophilic grafts, added to impart water solubility, which are either JeffamineM1000 (J, v in **Figure VI.1**) or glucosamine (G, w), giving rise to p(J) and p(G), respectively. JeffamineM1000 is an oligoether with a molecular weight of around 1000 g/mol and an average degree of polymerization (DP) of 22 (~19 ethylene oxide and ~3 propylene oxide). In comparison, glucosamine is a much smaller graft of high hydrophilicity, owing to its many hydroxyl groups. In addition to these fully hydrophilic polymers, random copolymers were designed. They incorporate hydrophobic and/or supramolecular side-chains, which are dodecylamine (D, x in **Figure VI.1**) and a chiral benzenetricarboxamide (BTA) derivative (B, y), respectively. Both units induce the formation of hydrophobic domains, and the chiral BTAs are also able to self-assemble into cylindrical helical stacks with preferred handedness via 3-fold hydrogen bonding. The Jeffamine-based copolymers, p(J-BD), incorporate both substituents, while the glucose-based copolymers contain either dodecyl or BTA, in p(G-D) and p(G-B), respectively. Additionally, the glucose-based polymers contain one Nile Red substituent (z in **Figure VI.1**), a fluorescent dye whose emission wavelength depends on the polarity of its environment.[45] A last difference between the Jeffamine- and glucose-based (co)polymers is their length, with an average DP of 186 and 103, respectively.

The five polymers were studied by MD simulations at the atomistic scale. While a coarse-grained (CG) modeling approach has been successfully applied elsewhere and is computationally faster, it misses the information at the atomic level, such as hydrogen bonds between different units or with the water solvent, which is important for the systems investigated here.[41,46] Also, applying CG models to synthetic polymers would require a challenging parametrization, particularly when dealing with a variety

of complex side-chains. [47–50] All-atom approaches tend to rely on more transferable force fields, offer access to finer atomistic details and information on the dynamics of the system. With the current computational power at hand, an all-atom approach seems more appropriate to study single-chain systems. Thus, MD simulations were performed for each polymer as an isolated single chain in explicit water boxes, starting from fully extended conformations and simulated on the 2 μs time scale. Three independent simulation replicas were run for each system with the AMBER package, using parameters coming from GAFF 2.11 to describe the polymers (see details of the protocol in **Section VI.8**).[51,52] To explore sequence effects, p(J-BD) was simulated as random (*r*), block (*b*), and multiblock (*mb*) polymer chains. These structures are denominated as p(J-*r*-BD), p(J-*b*-BD) and p(J-*mb*-BD), respectively. The p(G-D) and p(G-B) copolymers were only simulated as random sequences.



Polymer	v (%)	w (%)	x (%)	y (%)	z (%)	DP (-)
p(J)	100	0	0	0	0	186
p(J-BD)	81	0	15	4	0	186
p(G)	0	99	0	0	1	103
p(G-D)	0	84	15	0	1	103
p(G-B)	0	94	0	5	1	103

Figure VI.1. General chemical structure of the (random co)polymers studied and details of their monomeric composition. Adapted from Ref. 44.

All systems were properly equilibrated after 2 μs of simulation, as indicated by the convergence of their root mean square deviation (RMSD) values **(Figure VI.2)**. Larger fluctuations are observed for the p(G) system, showing that this macromolecule remains flexible and does not stabilize into one specific conformation.

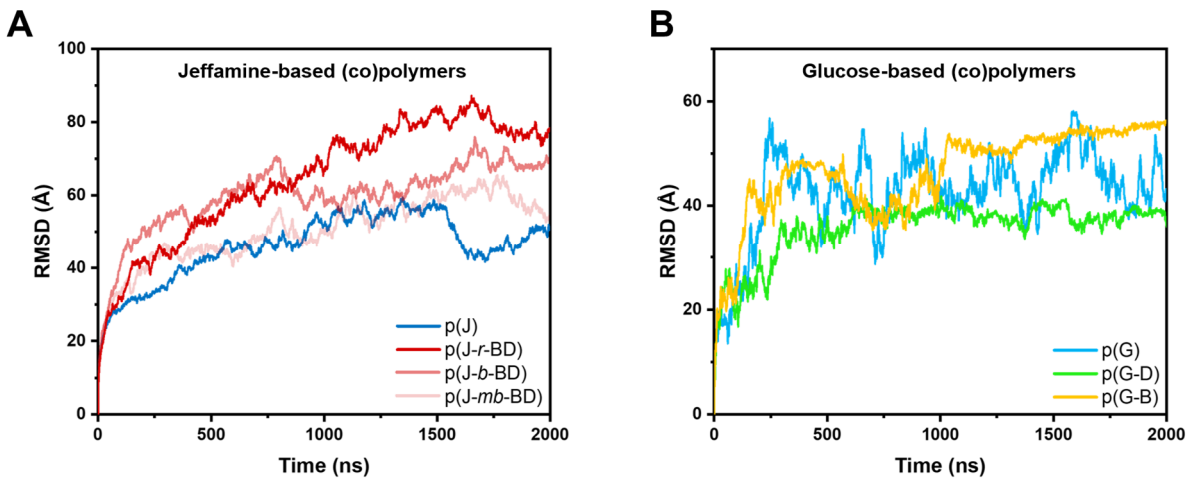
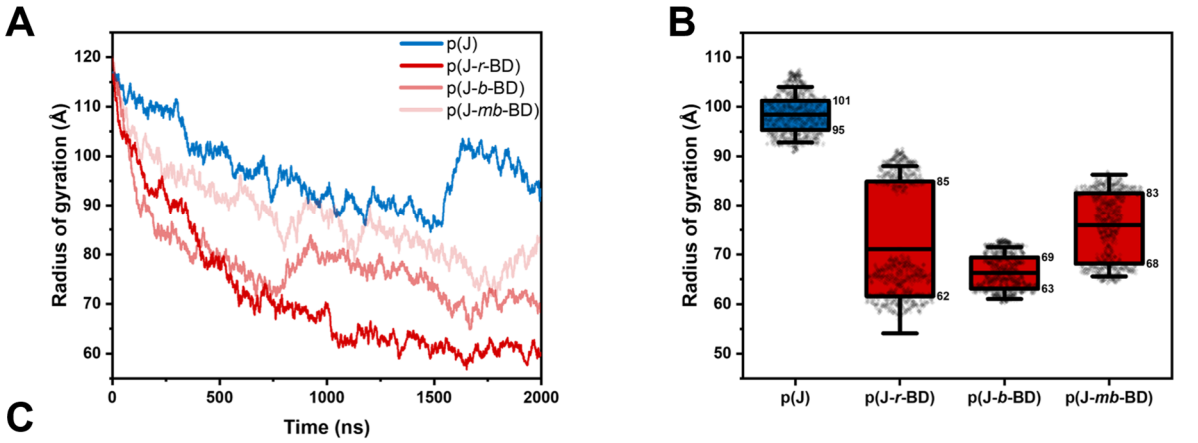


Figure VI.2. Evolution of the RMSD values over time for the **(A)** Jeffamine-based (co)polymers and for the **(B)** glucose-based (co)polymers. Only the first replica is displayed, for the sake of clarity; see **Figure VI.S1** in **Section VI.9** for all replicas.

VI.3. Jeffamine-based polymers form worm-like structures in water

The Jeffamine-based (co)polymers, starting from a fully elongated structure, tend to rapidly coil in water. However, the fully hydrophilic systems remain quite extended, with a radius of gyration (R_G) of around 10 nm **(Figure VI.3)**. The introduction of hydrophobic grafts in the copolymers leads to more compact conformations at the end of the simulations ($R_G \approx 6.5$ to 8 nm), no matter their microstructure (random, block or multiblock). This trend is consistent with the experimentally derived R_G values from SAXS where p(J) displays higher values than p(J-BD) (R_G = 11.1 nm for p(J) and R_G = 9.3 nm for p(J-BD)). Given the fact that experimental samples have a molar mass dispersity (both polymers) as well as heterogeneity in microstructures (for p(J-BD)), the simulated R_G values are well in line with the experimentally derived ones. The worm-like structure of the Jeffamine-based polymers can be observed in the snapshots presented in **Figure VI.4** (see **Figure VI.S4** for the final snapshots of all systems). Although the copolymers display more compaction, they do not completely fold, but rather form "kinked tube" with local folding around the hydrophobic groups, as shown in **Figure VI.4 B**. The information encoded in the primary structure is retained: units that are far in the sequence remain far in the 3D structures. For the random copolymer, p(J-*r*-BD), hydrophobic moieties close to each other in the sequence are able to merge into the same cluster, but do not meet units at the other end of the copolymer (see red and pink circles in **Figure VI.4 B**). Consequently, multiple local hydrophobic pockets form along the chain. This is also observed in p(J-*mb*-BD), where the hydrophobic groups were preorganized into three different clusters, which never merge during the

simulations. P(J-*b*-BD) contains a single central hydrophobic core, which does not split into smaller clusters. The simulations reveal that controlling the sequence of the Jeffamine-based copolymers, in particular the distribution of hydrophobic groups, could be exploited to control their morphology in water.



C

Radius of gyration (Å)	Initial	Final	Average (last 400 ns)	Median (last 400 ns)
p(J)	118 ± 1	97 ± 5	98 ± 4	98
p(J-*r*-BD)	115 ± 1	70 ± 11	71 ± 11	66
p(J-*b*-BD)	119 ± 1	65 ± 3	66 ± 4	66
p(J-*mb*-BD)	118 ± 2	78 ± 7	76 ± 7	77

Figure VI.3. Data on the R_G for the Jeffamine-based systems. **(A)** Development of the R_G for each system over time. Only one replica is shown, for the sake of clarity (see **Figure VI.S2** in **Section VI.9** for all replicas). **(B)** Distribution of the R_G values over the last 400 ns of the simulations for each system, averaged over the three replicas. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. **(C)** Table summarizing the data.

To validate our theoretical model, the 3D structures obtained from the MD simulations were used to simulate SAXS curves, which were compared to experimental SAXS measurements **(Figure VI.5)**. Basically, in SAXS, the intensity of the light scattered by a sample, *I(q)*, is measured as a function of the scattering vector, *q*, which is directly related to the scattering angle *θ*. Information about the size, shape and internal structure of the nanoparticles can be extracted from the evolution of the scattering curve. Different scales are probed depending on the *q* value: large distances at small *q* (the whole particle contributes to the scattering) and small distances at high *q* (as contributions coming from atoms separated by a large distance disappear, due to destructive interference). Therefore, the scattering intensity reaches its maximum value at small *q*, where it generally forms a horizontal plateau, and decreases with

increasing q. At some point, the signal may become too low to emerge from the background, which explains the important noise observed at high q in **Figure VI.5**. SAXS curves were simulated at different times (520 – 700 ns and 1820 – 2000 ns), and from additional accelerated MD (aMD) simulations, to ensure sufficient sampling.[53] This method applies a boost to the dihedral and potential energy of the system, helping it to escape local minima, thus improving sampling efficiency (full details on the aMD protocol is given in **Section VI.8**). Although MD simulations and SAXS experiments scan the matter at a different scale (with ideal systems with no molar mass dispersity for MD and disperse, heterogeneous mixture of chains with different microstructures for SAXS), the agreement between the experimental and simulated data is remarkable. The quality of the fit can be assessed

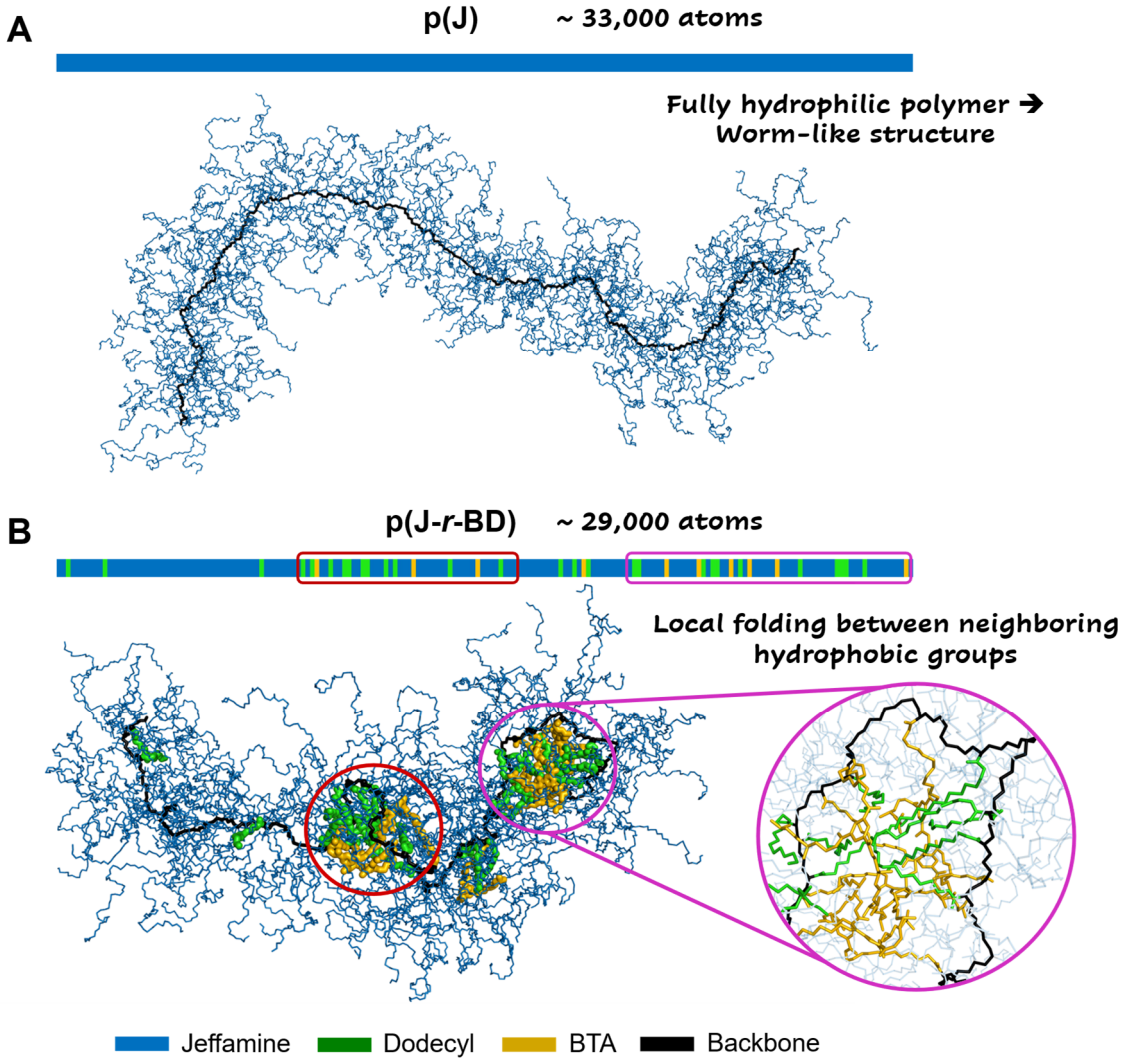
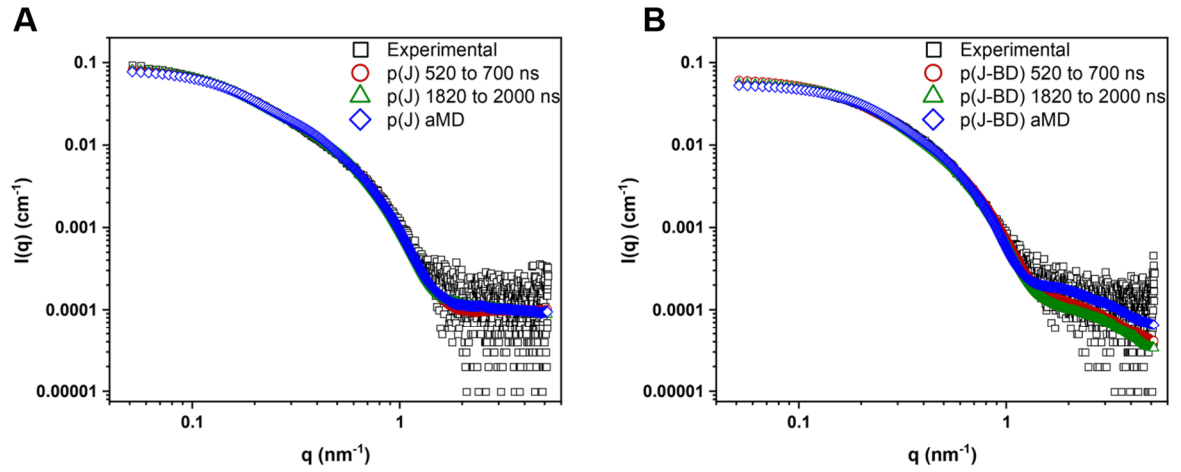


Figure VI.4. Final MD snapshot of the Jeffamine-based (co)polymers. The number of atoms is given for each system, and the sequence of monomers is represented as a colored bar (see bottom legend). **(A)** Fully hydrophilic polymer, p(J). **(B)** Copolymer with the random sequence, p(J-*r*-BD). Clusters of hydrophobic groups are highlighted in the sequence and in the 3D structure, with the same color.

through the χ^2 values, displayed in the tables in **Figure VI.5** (the higher the χ^2 value, the less the curves overlap; see **Section VI.8** for details). The global shape of the experimental scattering curves is in good agreement with that expected for graft polymers with extended conformations forming worm-like chains (see Ref. 44 for in-depth analysis of the scattering curves). Notably, two different power law regimes are detected, *i.e.* the curves display different slopes, at intermediate q ($0.1 < q < 0.6$ nm^{-1}) and high q ($0.6 < q < 1.5$ nm^{-1}), in agreement with the curves expected for semi-flexible polymer chains. The scattering curve of p(J-BD) lacks a clear oscillation around $q = 1$ nm^{-1}, which indicates that p(J-BD) does not form a defined, single hydrophobic interior as expected in core-shell structures.[8] Overall, the comparison with experimental SAXS data corroborates that MD simulations reflect the nature of the formed structures well, namely as extended worm-like chains for p(J) and p(J-BD), and the formation of local hydrophobic pockets in p(J-BD). A careful validation against experimental measurements turned out to be extremely important. We found that the model used to compute the partial charges of our polymers, thus the representation of their electrostatic properties, strongly influenced the resulting conformations. Initially, we computed the partial charges with the AM1-BCC model, which is very common and often used by our group.[54,55] This method led to underestimated absolute values for the charges on the oxygen and carbon atoms of the Jeffamine grafts, giving them a low polarity **(Figure VI.6 A)**. It resulted in the



Simulated curve	χ^2 (-)
p(J) - 520 to 700 ns	2.7 ± 0.7
p(J) - 1820 to 2000 ns	2.5 ± 0.8
p(J) - aMD	3.5 ± 1.0

Simulated curves	χ^2 (-)
p(J-BD) - 520 to 700 ns	4.8 ± 1.6
p(J-BD) - 1820 to 2000 ns	5.3 ± 2.8
p(J-BD) - aMD	3.0 ± 0.6

Figure VI.5. Experimental (black squares) and simulated (colored shapes) SAXS curves in water for the **(A)** p(J) polymer and **(B)** p(J-BD) copolymers. Simulated curves were generated from two time intervals during the conventional MD simulations, and from aMD simulations. The experimental polymer concentration is 1.5 mg.mL^{-1}. The χ^2 values, assessing the accuracy of fit, are given in the tables below.

formation of a fully folded, compact globule, which was contradictory to experimental results. Turning to the RESP methodology to compute the partial charges, we obtained much better results, with the formation of the worm-like chains discussed above.[56] Simulations on simple PEG chains indicated the same trend, with much more compact chains when the charges are computed using the AM1-BCC model **(Figure VI.6 B)**. The atomic charges were recently shown to strongly influence the interactions of polyethers with water, explaining the sensitivity of the system to partial charges.[57] This example demonstrates to which extent small inaccuracies on charge description can lead to wrong predictions on the shape and size of macromolecular structures. However, precious information can be extracted from this error. First, it gives us an idea of the minimal size that a p(J) particle could reach if fully collapsed, the globule having an R_G of about 3.5 nm. Then, it demonstrates that folding could occur at the microsecond timescale if the Jeffamine chains were less polar, thus not interacting with water. In other words, the Jeffamine grafts limit the flexibility of the backbone not only because they are long and generate steric hindrance, but also because they are polar, thus stay extended in the solvent and attract a lot of water molecules.

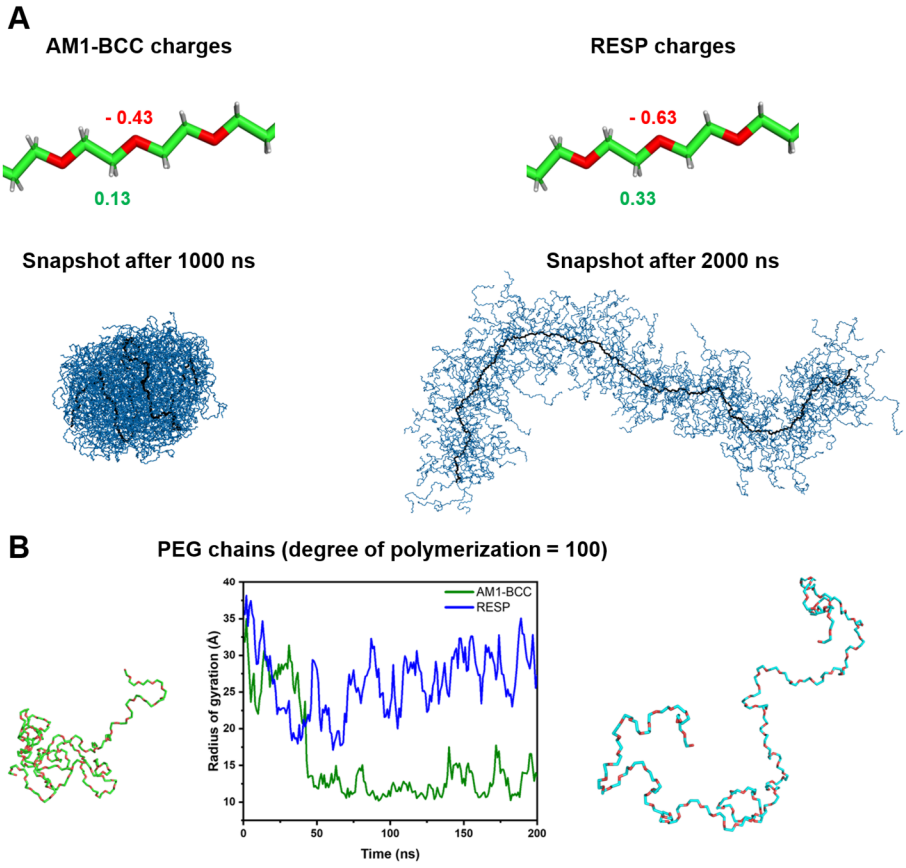
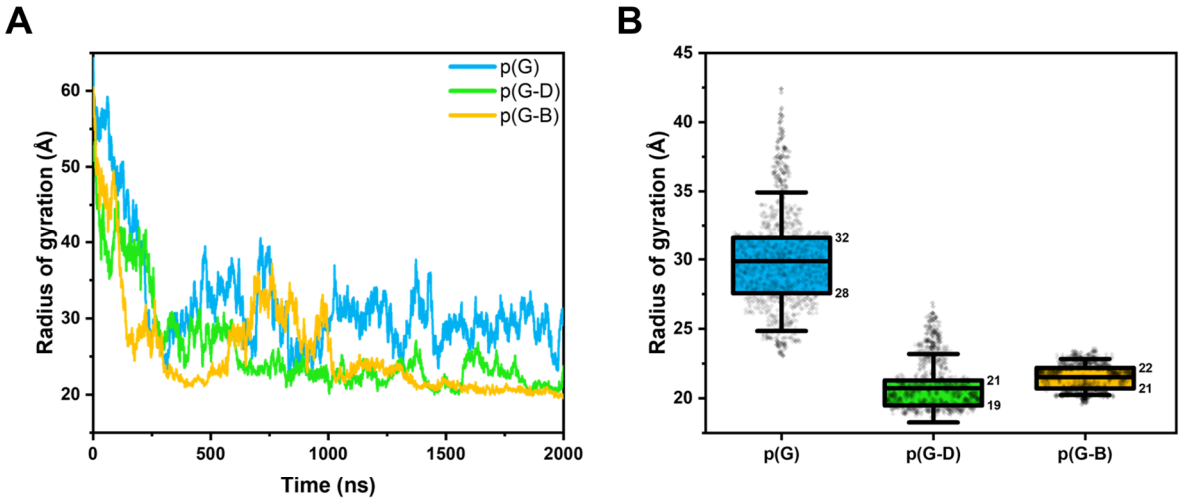


Figure VI.6. Investigation of the effect of the set of partial charges (AM1-BCC or RESP) on the conformations adopted by **(A)** the p(J) polymers and **(B)** simple PEG chains. The typical value of the charges computed on the atoms in the middle of a Jeffamine graft are represented in red and green for the oxygen and carbon atoms, respectively.

VI.4. Glucose-based polymers form core-shell structures in water

The fully hydrophilic p(G) polymers, as their Jeffamine-based counterpart, quickly coil in water but do not fold, and their R_G oscillates around 3 nm **(Figure VI.7)**. The glucose-based copolymers, however, display a completely different behavior: p(G-D) and p(G-B) both fold into core-shell nanoparticles ($R_G \approx 2$ nm). The small glucose residues form a shell around a hydrophobic core comprising the dodecyl or BTA grafts and the Nile Red moiety, as can be seen in the snapshots in **Figure VI.8**.



C

Radius of gyration (Å)	Initial	Final	Average (last 400 ns)	Median (last 400 ns)
p(G)	61 ± 3	32 ± 4	30 ± 3	30
p(G-D)	52 ± 2	21 ± 2	21 ± 2	20
p(G-B)	63 ± 2	21 ± 1	22 ± 1	22

Figure VI.7. Data on the R_G for the glucose-based systems. **(A)** Development of the R_G for each system over time. Only one replica is shown, for the sake of clarity (see **Figure VI.S3** in **Section VI.9** for all replicas). **(B)** Distribution of the R_G values over the last 400 ns of the simulations for each system, averaged over the three replicas. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. **(C)** Table summarizing the data.

This can also be inferred from the significant decrease in solvent-accessible surface area (SASA) of Nile Red during the simulations, indicating a reduction of Nile Red exposure to its environment during chain folding **(Figure VI.9)**. The SASA of Nile Red decreases even in the fully hydrophilic p(G) polymers, which tend to coil around it to shield it from water. It shows that the presence of a single hydrophobic unit can induce local structuration of the chain. However, the SASA values are lower in the

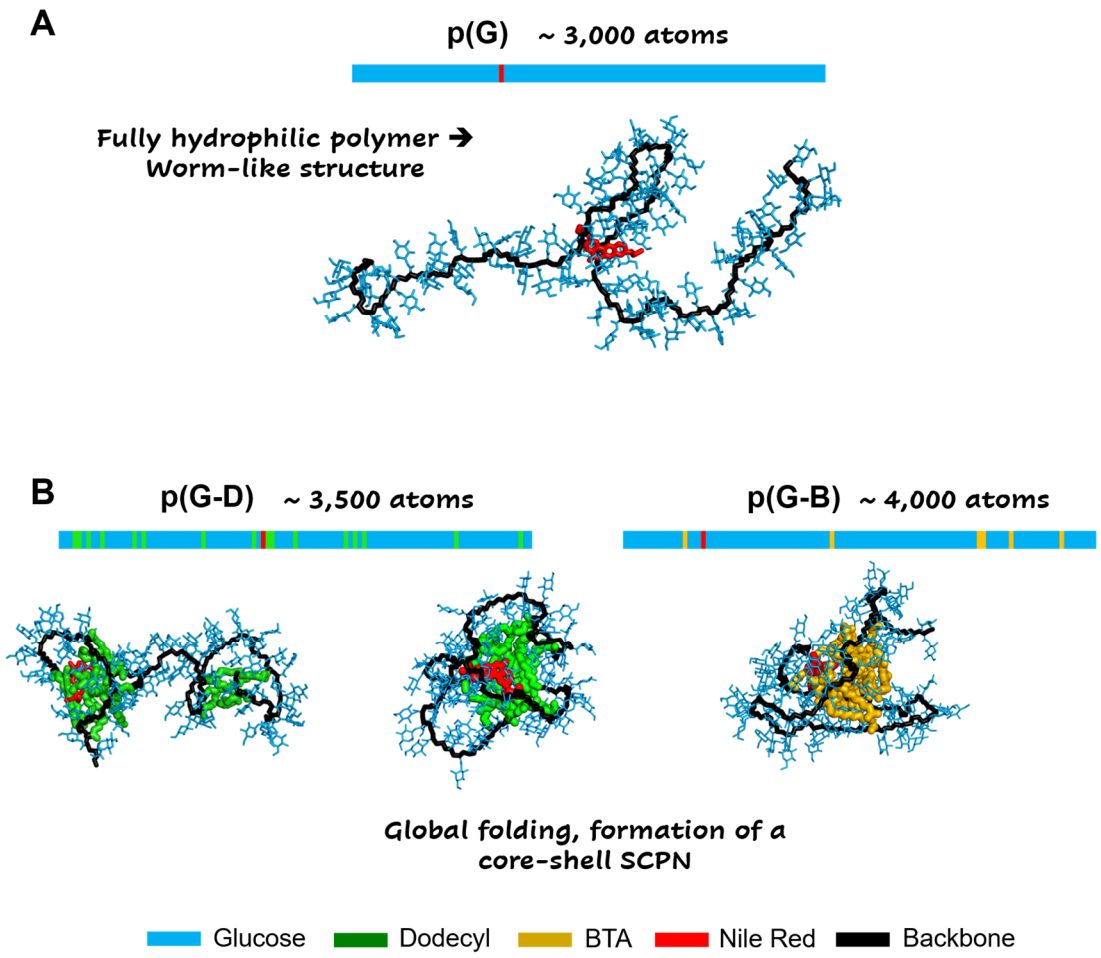
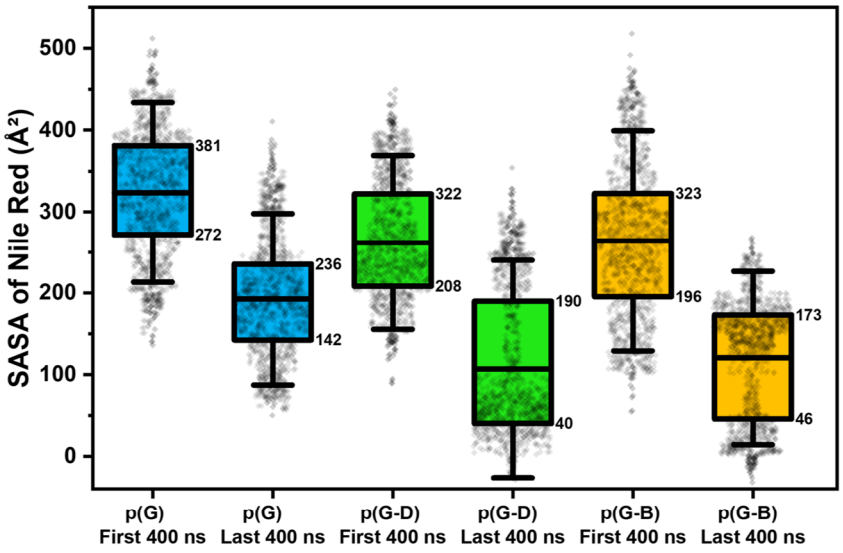


Figure VI.8. Final MD snapshots of the glucose-based (co)polymers. The number of atoms is given for each system, and the sequence of monomers is represented as a colored bar (see bottom legend). **(A)** Fully hydrophilic polymer, p(G). **(B)** From left to right: p(G-D) simulated by conventional MD, p(G-D) simulated by aMD, p(G-B) simulated by conventional MD.

copolymers, where Nile Red is incorporated into hydrophobic domains and more efficiently shielded. The presence of Nile Red in hydrophobic compartments in the p(G-D) and p(G-B) systems was also detected experimentally, in agreement with the simulations.[44] The folding of the backbone in p(G-D) and p(G-B) allows hydrophobic units that are far in the sequence to become spatially close in the 3D structure, as can be seen in **Figure VI.8**. Hydrophobic groups quickly merge with their neighbors, and the formed clusters merge together until forming a single hydrophobic core. Inside these globules, the backbone dynamics are strongly reduced, as indicated by the decrease of the dihedral angles' fluctuations along the backbone, showing that the conformational space is reduced as the polymer folds into a compact core-shell structure **(Figure VI.10)**. A similar trend was observed in other folded amphiphilic copolymers.[40] The compact p(G-D) and p(G-B) conformations are further stabilized by intramolecular hydrogen bonds that increase in number over time, while these

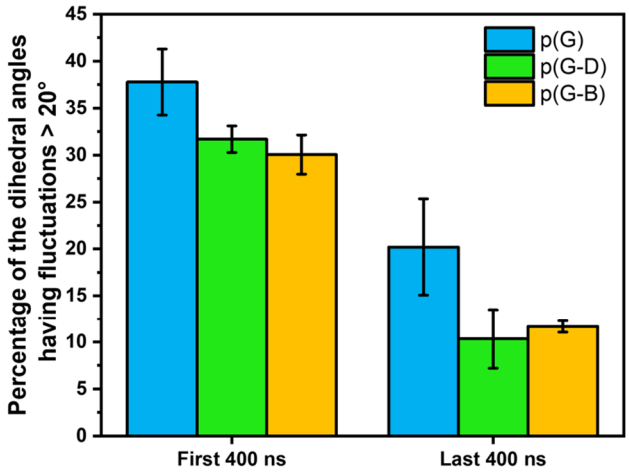


SASA of Nile Red (Å²)	p(G) First 400 ns	p(G) Last 400 ns	p(G-D) First 400 ns	p(G-D) Last 400 ns	p(G-B) First 400 ns	p(G-B) Last 400 ns
Average ± Std. dev.	324 ± 73	193 ± 70	262 ± 71	107 ± 89	264 ± 90	121 ± 71
Median	331	188	249	69	266	144

Figure VI.9. Distribution of the SASA values measured for the Nile Red moiety in the different glucose-based polymers, averaged over the first and last 400 ns of the simulation. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. Statistics are summarized in the table below.

remain constant for p(G) **(Figure VI.11)**. The folding of p(G-D) and p(G-B) is reminiscent of the early stages of protein folding and the formation of "molten globules", characterized by nonspecific and local interactions between side-chains promoted by hydrophobic effects, and increasing backbone rigidity.[58–60]

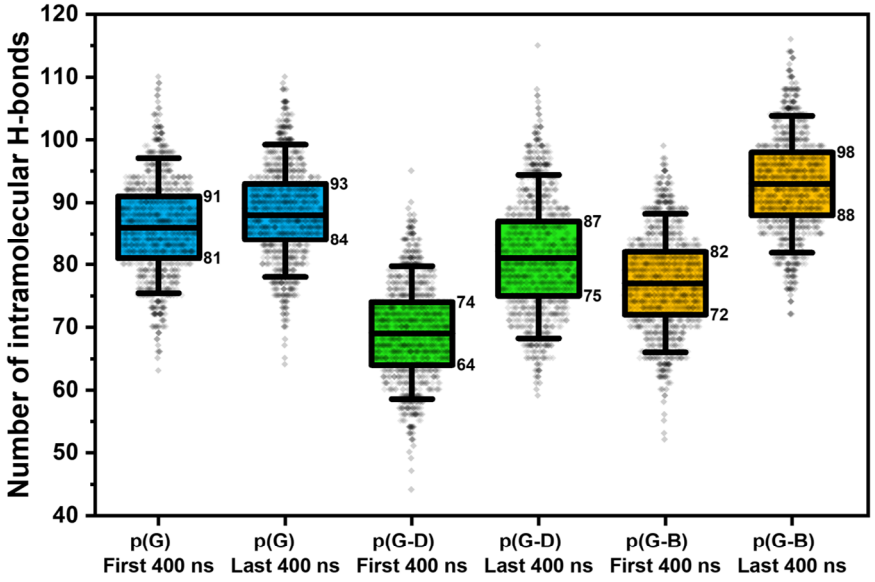
During their folding, the copolymers may remain trapped for several hundreds of nanoseconds in partly folded states, when the hydrophobic units are grouped in two or more clusters (see **Figure VI.8 B**, structure on the left). As mentioned above, the formation of these stable clusters decreases the flexibility of the backbone. To avoid spending too much time in these metastable states, aMD simulations are particularly useful, and this protocol was successfully applied to the p(G-D) systems (see **Figure VI.8 B**, structure on the middle). The fully folded structure comprising a single hydrophobic core was formed after 300 ns, compared to 2 µs (or more) with conventional MD simulations. In contrast, the p(G-B) systems folded without the need for the accelerated protocol despite containing only 5 % of hydrophobic monomers, compared to 15 % for the p(G-D) systems. However, the presence of more hydrophobic units means that more encounters between the hydrophobic groups are required to



% of the dihedral angles with fluctuations > 20°	p(G)	p(G-D)	p(G-B)
First 400 ns - Average ± Std. dev.	37.8 ± 3.5	31.7 ± 1.4	30.0 ± 2.1
Last 400 ns - Average ± Std. dev.	20.2 ± 5.1	10.3 ± 3.1	11.7 ± 0.6

Figure VI.10. Percentage of the backbone dihedral angles undergoing fluctuations superior than 20° in the glucose-based systems, over the first and last 400 ns of the simulations. Statistics are summarized in the table below.

form the complete hydrophobic core. Additionally, in terms of carbon content, one BTA unit, which bears three alkyl chains, is roughly equivalent to three dodecyl groups. The larger size of the BTA grafts probably increases the likelihood of hydrophobic contacts along the chain, which overall leads to a more efficient folding. The 3D structures of the p(G-D) copolymer generated from the MD simulations were used to simulate SAXS curves, in order to compare them to the experimental SAXS measurements **(Figure VI.12)**. The p(G-D) particles being much smaller than the p(J) and p(J-BD) polymers, the horizontal plateau extends to higher q values, up to around q = 0.4 nm^{-1}. The experimental curve indicates the formation of core-shell nanoparticles of small size (R_G = 3.8 nm), owing to the oscillation in the scattering curve around q = 1 - 2 nm^{-1}. Surprisingly, the simulated SAXS curve for a single-chain of p(G-D) did not match the experimental curve well, although the 3D structures from MD show the formation of core-shell structures. We attributed this discrepancy to the presence of aggregates in solution, as suggested by the upturn below q = 0.1 nm^{-1}. To support this hypothesis, mixtures of two and three chains were simulated for the p(G-D) copolymer, starting from extended chains. As observed in **Figure VI.12 A**, the overlap between the experimental and simulated curves is significantly improved when considering multichain aggregates (see also the χ^2 values in the table). The maximum of the peak around q = 1 – 2 nm^{-1} appears at a smaller q for the aggregate comprising three chains compared to the single-chain system, which is consistent with the formation of a larger hydrophobic core. An equilibrium of species, comprising SCPNs but also small

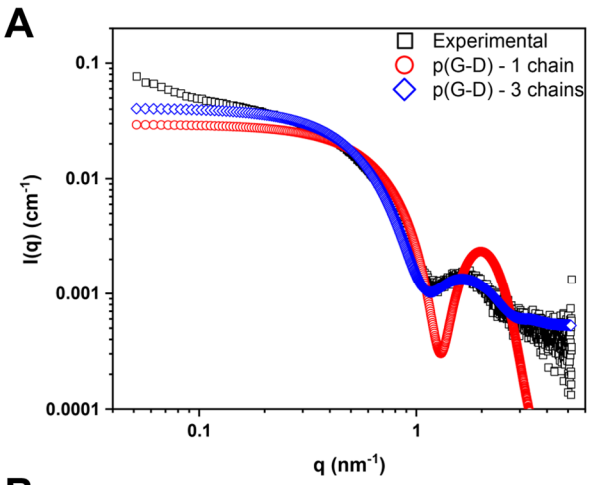


Number of intramolecular H-bonds	p(G) First 400 ns	p(G) Last 400 ns	p(G-D) First 400 ns	p(G-D) Last 400 ns	p(G-B) First 400 ns	p(G-B) Last 400 ns
Average ± Std. dev.	86 ± 7	89 ± 7	69 ± 7	81 ± 9	77 ± 7	93 ± 7
Median	86	88	69	81	77	93
Delta (median)	+ 2		+ 12		+ 16	

Figure VI.11. Distribution of the number of intramolecular H-bonds in the different glucose-based polymers, averaged over the first and last 400 ns of the simulations. The lines delimiting a box represent the first and third quartiles, whose values are annotated at the edges of the box. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. Statistics are summarized in the table below.

aggregates, probably coexist in solution. This would be in line with DLS measurements, which show a wide distribution of sizes for the p(G-D) particles, from about 2 to 10 nm.[44] Aggregation likely occurs in the early steps of folding, before the complete shielding of the hydrophobic moieties, similarly to what happens to misfolded proteins that expose hydrophobic groups and thus tend to aggregate. This example demonstrates the robustness of our approach combining MD simulations and SAXS, as it allows to distinguish particles of similar shape, core-shell structures, but of slightly different size (see the similarity between structures comprising one, two or three chains, in the snapshots in **Figure VI.12 B**). Such resolution is difficult to attain using experimental means alone. It also demonstrates that aggregate formation is not necessarily accompanied by a strong increase in R_G, making it very difficult to infer SCPN formation by measuring the size of the particles in solution. It seems likely that some experimental results describing the formation of SCPNs were in fact characterizing mixtures of single-chain systems and small aggregates, which could

explain, in part, the surprisingly wide range of sizes reported in the literature for SCPNs of similar molecular weights.[36] Much larger aggregates were detected experimentally for the p(G-B) copolymers. This is probably due to the larger size of the hydrophobic BTA grafts, which facilitates interparticle contacts, and to their ability to make H-bonds with BTA units of other chains. The formation of large aggregates was demonstrated to be dependent on the percentage of BTA units in the chain, for similar polymers.[11] Given the higher computational cost associated with the simulation of such multichain systems and the uncertainty concerning the number of chains comprised in the p(G-B) aggregates, we did not investigate them by MD simulations.



Simulated curves	χ^2 (-)	χ^2 excluding 50 first points (-)
p(G-D) - 1 chain	116.1 ± 6.6	51.6 ± 3.8
p(G-D) - 2 chains	17.4 ± 2.8	3.6 ± 0.7
p(G-D) - 3 chains	11.3 ± 1.0	3.0 ± 0.4

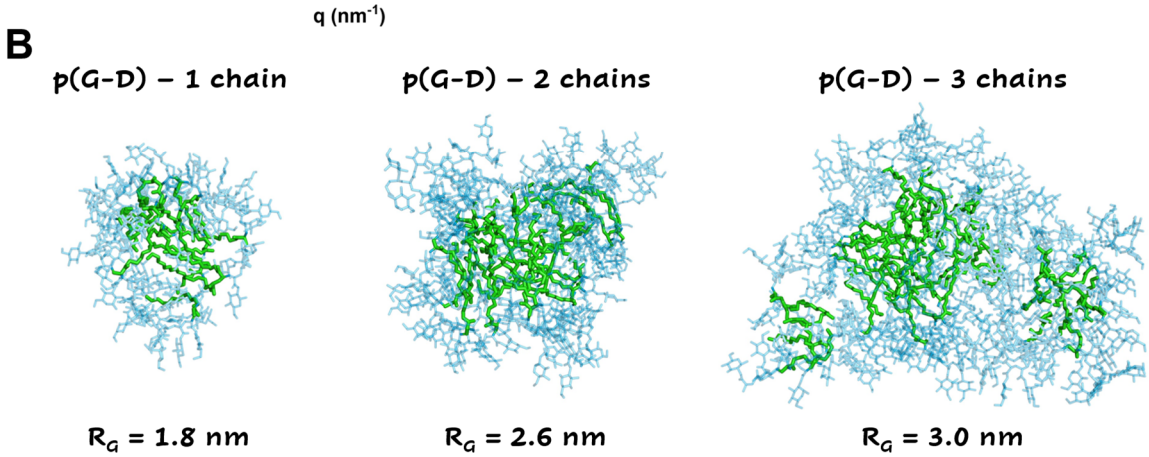
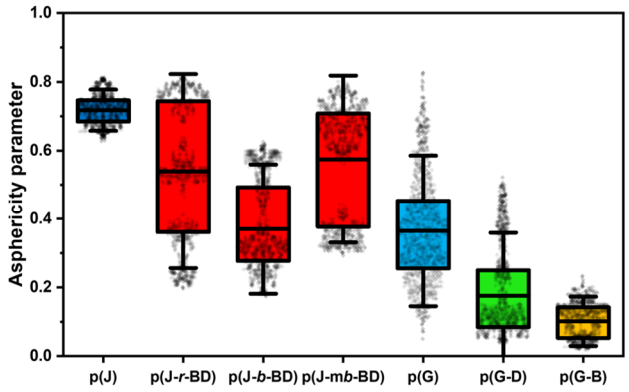


Figure VI.12. Comparison between simulated and experimental SAXS curves for the p(G-D) copolymer. **(A)** Experimental (black squares) and simulated SAXS curves in water for the p(G-D) systems comprising one (red circles) or three (blue diamonds) chains. The curve generated from two chains is not shown, for the sake of clarity. The experimental polymer concentration is 2.5 mg.mL^{-1}. The χ^2 values, assessing the accuracy of fit, are given in the table next to the graph. They were also calculated without the first 50 points, because the upturn detected experimentally in the low-q region of the spectrum (q < 0.1 nm^{-1}) is caused by a population of larger aggregates, that our simulations cannot reproduce. **(B)** Final snapshot of p(G-D) comprising one, two or three chains generated from the aMD simulations and used to simulate the SAXS curves. The glucose and dodecyl units are colored in blue and green, respectively.

VI.5. Comparison between Jeffamine- and glucose-based (co)polymers

As described in the two preceding sections, the nature of the hydrophilic grafts has a strong impact on the morphology of the chains in water. While the Jeffamine-based systems remain quite extended and form worm-like structures, the glucose-based copolymers are able to fold into core-shell SCPNs (with the competition of intermolecular aggregation). This is well reflected by measurements of the asphericity parameter, which show that the p(G-D) and p(G-B) systems are the more spherical systems, although they do not form perfect spheres, in agreement with the ellipsoidal core-shell structures detected experimentally **(Figure VI.13)**.

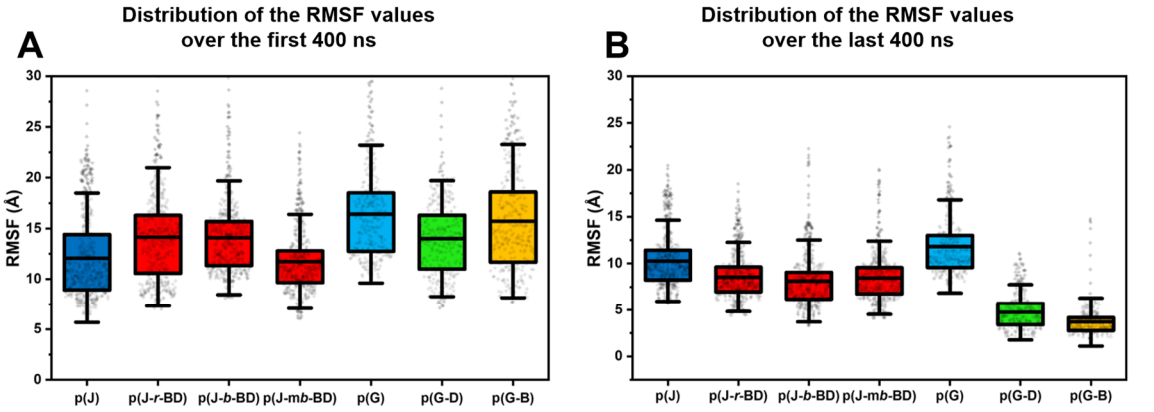


Asphericity	Average ± Std. dev.
p(J)	0.72 ± 0.04
p(J-*r*-BD)	0.54 ± 0.19
p(J-*b*-BD)	0.37 ± 0.13
p(J-m*b*-BD)	0.57 ± 0.16
p(G)	0.37 ± 0.15
p(G-D)	0.18 ± 0.13
p(G-B)	0.10 ± 0.05

Figure VI.13. Distribution of the asphericity parameter for each system, computed over the last 400 ns of the simulations. The lines delimiting a box represent the first and third quartiles. The line inside a box indicates the mean value. The error bar is given as mean ± 1.5 x standard deviation. Statistics are summarized in the table.

The Jeffamine-based polymers exhibit higher values, as expected given their extended, rod-like character. The p(J-BD) copolymers, more compact due to the formation of the hydrophobic domains, display lower values than p(J). Among them, p(J-*b*-BD) is significantly more spherical, although it remains extended (see snapshots in **Figure VI.S4**). We attribute this behavior to the formation of the dense hydrophobic core located at the center of the chain, as preorganized in the primary structure, demonstrating that the sequence of monomers influences the morphology of the system. In both polymer families, the introduction of hydrophobic grafts induces compaction of the main chain, with a stabilization around local or global hydrophobic domains. The formation of these domains is associated with reduced backbone mobility, as indicated by a larger decrease in root mean square fluctuation (RMSF) values for the copolymers compared to their fully hydrophilic counterparts throughout the simulations **(Figure VI.14)**. This effect is particularly pronounced in the p(G-D) and p(G-B) systems, whose conformational flexibility significantly drops upon folding

into compact globules, as already evidenced by the reduced fluctuations of their dihedral angles **(Figure VI.10)**.



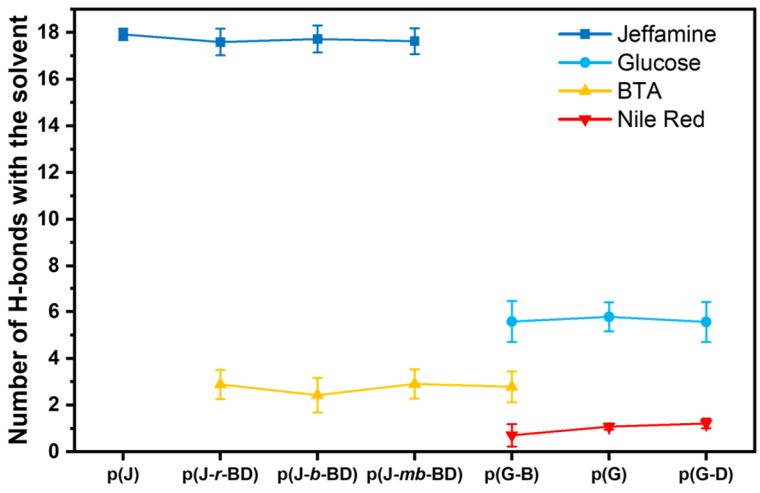
C

RMSF (Å)	p(J)	p(J-r-BD)	p(J-b-BD)	p(J-mb-BD)	p(G)	p(G-D)	p(G-B)
First 400 ns Average ± Std. dev. *(median)*	12 ± 4 *(11)*	14 ± 5 *(14)*	14 ± 4 *(13)*	12 ± 3 *(11)*	16 ± 4 *(16)*	14 ± 4 *(14)*	16 ± 5 *(15)*
Last 400 ns Average ± Std. dev. *(median)*	10 ± 3 *(10)*	9 ± 2 *(8)*	8 ± 3 *(8)*	8 ± 3 *(8)*	12 ± 3 *(11)*	5 ± 2 *(4)*	4 ± 2 *(3)*
Delta median	*- 1*	*- 6*	*- 5*	*- 3*	*- 5*	*- 10*	*- 12*

Figure VI.14. Distribution of the RMSF values over the **(A)** first and **(B)** last 400 ns of the simulations. **(C)** Table summarizing the data.

Simulations and experiments both demonstrated that the nature of the hydrophilic grafts strongly impacts the morphology of the chains in water. Global folding of the Jeffamine-based copolymers is prevented by the highly polar Jeffamine grafts, which remain extended and interact with many water molecules. This behavior is well reflected by the average number of H-bonds performed by each kind of side-chain with the solvent **(Figure VI.15)**. A Jeffamine chain performs, on average, 18 H-bonds with water molecules per conformation, which is significantly more than the glucose units (6 H-bonds per conformation). Jeffamine – water interactions significantly contribute to the limited flexibility of these polymers, preventing complete folding. This graph also indicates that the number of interactions with the solvent for a given side-chain is independent of the system to which the graft belongs. For instance, a Jeffamine graft performs 18 H-bonds whether it is in the fully hydrophilic p(J) or in a p(J-BD) copolymer. The local folding of the Jeffamine-based copolymers is also related to the weaker hydrophobic driving force, compared to the glucose-based analogues. The long Jeffamine grafts efficiently shield the dodecyl and BTA units, as indicated by measurements of their SASA values **(Figure VI.16)**. The dodecyl and BTA grafts are slightly more exposed to the solvent in the p(G-D) and p(G-B) systems than in the p(J-BD) copolymers, suggesting that the latter do not require global compaction to shield their hydrophobic groups. In contrast, the glucose moieties being much shorter, the glucose-based copolymers must fold into compact structures around a single hydrophobic core to minimize the exposure of their hydrophobic units to water.

A

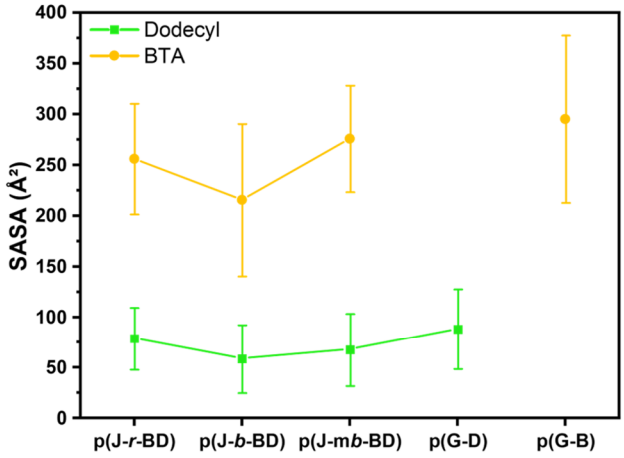


B

Number of H-bonds with the solvent per conformation	Jeffamine	Glucose	BTA	Nile Red
p(J)	18 ± 0	/	/	/
p(J-*r*-BD)	18 ± 1	/	3 ± 1	/
p(J-*b*-BD)	18 ± 1	/	2 ± 1	/
p(J-m*b*-BD)	18 ± 1	/	3 ± 1	/
p(G-B)	/	6 ± 1	3 ± 1	1 ± 0
p(G)	/	6 ± 1	/	1 ± 0
p(G-D)	/	6 ± 1	/	1 ± 0

Figure VI.15. Number of H-bonds with the solvent per conformation, for each kind of side-chain in each system, averaged over the last 400 ns of the simulations. Statistics are summarized in the table below.

A



B

SASA (Å²)		p(J-*r*-BD)	p(J-*b*-BD)	p(J-m*b*-BD)	p(G-D)	p(G-B)
Dodecyl	Average ± Std. dev.	79 ± 31	59 ± 34	68 ± 36	88 ± 39	/
	Median	72	51	68	91	/
BTA	Average ± Std. dev.	256 ± 54	216 ± 75	276 ± 52	/	295 ± 82
	Median	253	204	262	/	303

Figure VI.16. SASA value for each kind of hydrophobic moiety in each system, averaged over the last 400 ns of the simulations. Statistics are summarized in the table below.

VI.6. Role of the BTA units in the folding process

Finally, we investigated more specifically the role played by the BTA grafts, which are often introduced with the aim of forming "structured" hydrophobic domains through their helical stacking.[9,10] Interactions between BTA units have been supported by circular dichroism (CD) experiments in various systems[9,11,20]; however, such measurements do not distinguish between intra and interchain interactions, nor do they provide a direct visualization of the spatial organization of the BTAs. All-atom MD simulations offer a powerful approach to probe their role in the folding process of SCPNs.

Overall, the BTA units appear to contribute to the folding primarily through their hydrophobic nature. In the p(J-BD) systems, their alkyl chains merge with the dodecyl units in the hydrophobic domains (see snapshot and zoom in **Figure VI.4 B**), while in the p(G-B) copolymers, they constitute the central hydrophobic core of the core-shell structures (see snapshot in **Figure VI.8 B**). However, unlike the dodecyl grafts, the BTAs are amphiphilic, with three amide moieties surrounding their aromatic core.

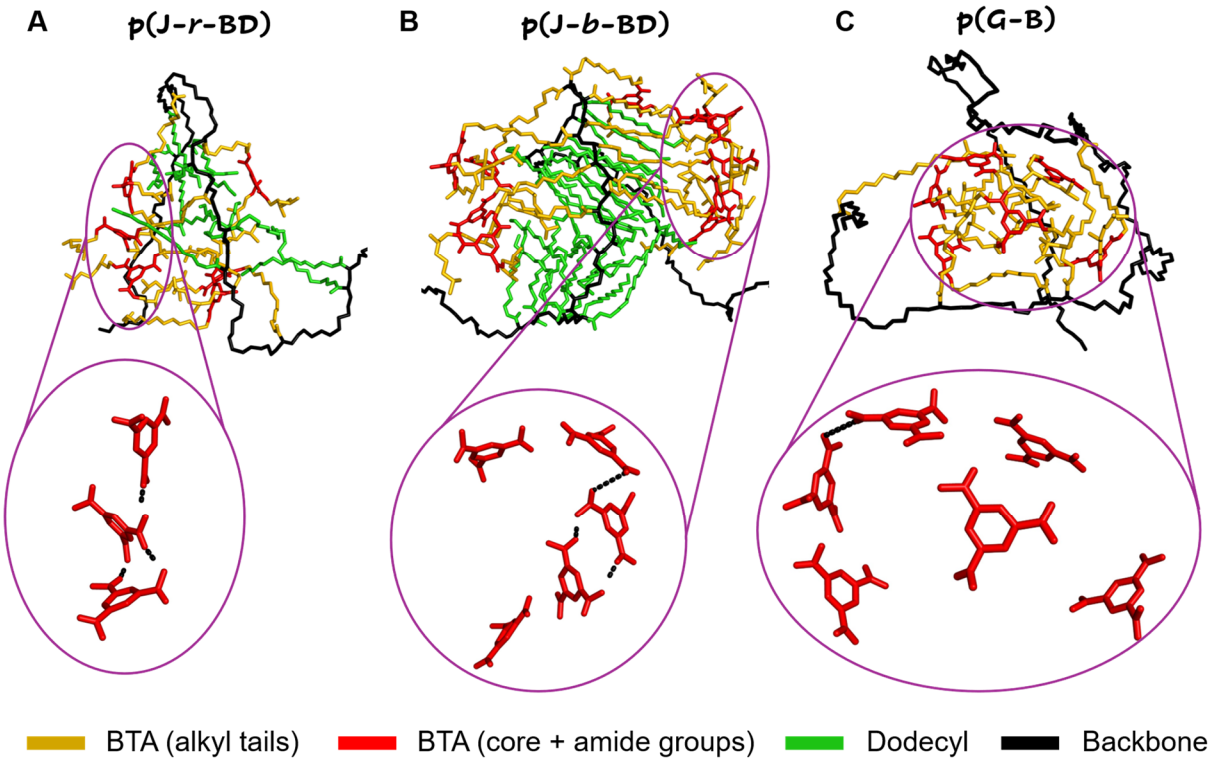
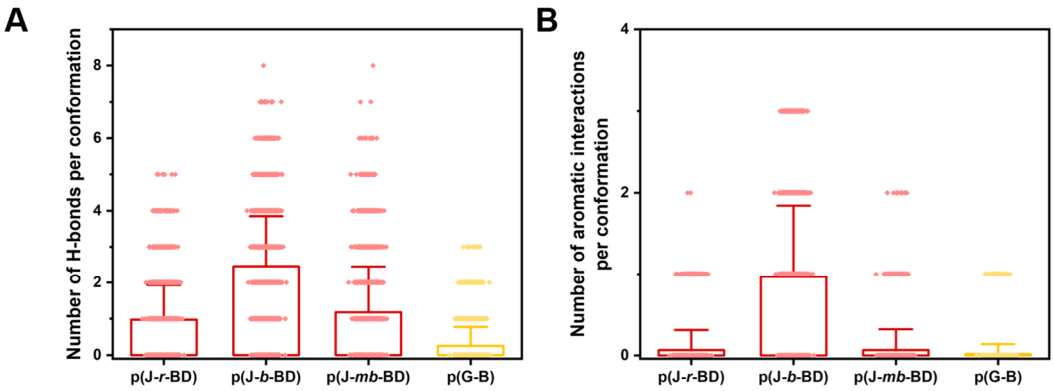


Figure VI.17. Final MD snapshots zooming on hydrophobic clusters belonging to the **(A)** p(J-*r*-BD), **(B)** p(J-*b*-BD) and **(C)** p(G-B) systems. For each system, a focus is made on the BTA cores (represented in red), where H-bonds are shown as black dots. The snapshots show that the BTA cores are mainly found at the periphery of the hydrophobic clusters and do not form well-organized helices, although they perform some H-bonds. Hydrophilic grafts are not shown, for the sake of clarity.

These polar parts are essentially found at the periphery of the hydrophobic domains, where they contribute to the shielding of the hydrophobic groups **(Figure VI.17)**. This is reflected by the number of H-bonds performed by the BTA units with the solvent, around three per conformation (see **Figure VI.15**). These numerous BTA – water interactions are in competition with the establishment of persistent contacts between the BTA units, as indicated by their weak number of H-bonds and π-type interactions[1] **(Figure VI.18)**. The p(J-*b*-BD) system displays slightly more interactions: in this copolymer, all the BTA units are preorganized into the same hydrophobic core, increasing their chances of encounter in comparison to the other Jeffamine-based microstructures, where BTAs are dispersed into several hydrophobic clusters. Interestingly, in the p(G-B) systems, which also regroup all their BTA units within the same hydrophobic core, the number of interactions between BTAs is almost zero, on average. This could be due to the strong decrease in backbone flexibility accompanying the formation of the globule, which limits the reorganization of the BTA cores after folding. Additionally, the BTAs are slightly more exposed to their solvent in the p(G-B) systems (see SASA measurements, **Figure VI.16**), which could further favor BTA – water over BTA – BTA interactions. The more efficient screening performed by the long Jeffamine grafts seems to increase the contacts between BTAs, although the interactions remain limited. Our simulations provide a very different picture than the one commonly used to describe the role of BTAs in such SCPNs, often representing them in well-organized supramolecular helices within the hydrophobic domains.[9,10] This should make us reconsider the "structuring" effect of these units, and their ability to transmit their chirality; for example, it was shown in similar polymers bearing catalytic units that, despite efficient catalytic properties and the presence of chiral BTA groups, the reactions were performed without enantioselectivity.[61] The measured CD signals, which undoubtedly confirm the presence of BTA interactions, could originate from transient contacts between BTAs located within the same hydrophobic domains, rather than from the formation of well-organized supramolecular helices. Intermolecular interactions within larger aggregates could also contribute to the CD response. Of course, we cannot rule out the possibility that our simulations misbalance BTA – water and BTA – BTA interactions, although similar simulation protocols have previously been successfully applied to study BTA-based supramolecular assemblies.[62,63]

[1] π-type interactions are counted between aromatic cycles following these geometric criteria: the distance between their centers of mass must be ≤ 5 Å, and the angle between their planes must be < 45 ° or > 135 °.



C

Average number of interactions per conformation	p(J-*r*-BD)	p(J-*b*-BD)	p(J-*mb*-BD)	p(G-B)
H-bonds	1.0 ± 1.0	2.5 ± 1.4	1.2 ± 1.3	0.3 ± 0.5
π-type	0.1 ± 0.2	1.0 ± 0.9	0.1 ± 0.3	0.0 ± 0.1

Figure VI.18. Data on the interactions between BTA units. **(A)** Average number of H-bonds and **(B)** π-type interactions per conformation, measured during the whole simulations. Data is shown as mean ± standard deviation. The pale lines represent the distribution of measurements. **(C)** Table summarizing the data.

VI.7. Conclusion

Throughout this chapter, we have demonstrated that atomistic scale MD simulations constitute a promising tool to gain insight into the folding behavior of amphiphilic heterograft polymers. The current computational power using GPUs is now adapted to treat these very large systems (more than 30,000 atoms for p(J) in a box of ~ 2,500,000 atoms of solvent) at the atomistic scale. A combination of MD simulations and SAXS experiments revealed that Jeffamine-based copolymers adopt globally extended structures capable of forming local hydrophobic domains. Copolymers functionalized with hydrophilic glucose grafts are instead capable of global folding into core-shell structures comprising a single central hydrophobic core. Importantly, our combined MD and SAXS approach allowed us to elucidate, at the atomistic level, the formation of small aggregates for the p(G-D) particles. These aggregates, which are very similar in size and shape to the SCPNs formed by the same system, could not have been detected by experiments or simulations alone. Their presence could only be confirmed by confronting the simulated SAXS curves to the experimental ones. Additionally, comparing our theoretical results to experiments was crucial, as we have seen that MD simulations done with inaccurate parameters can lead to completely wrong predictions. This demonstrates the robustness of our approach, and the complementarity between MD simulations and SAXS experiments to elucidate the 3D structure of SCPNs. The atomistic view provided by the simulations allowed us to investigate the influence of the primary structure on the folded conformations and

internal organization of the Jeffamine-based copolymers, revealing differences between random, blocky or multiblocky microstructures. These systems are particularly interesting to study sequence – structure relationships, as their limited folding permits to retain the information encoded in the sequence into the 3D structures. This property could be valuable for various applications: for instance, a recent example showed that the distribution of catalytic units within the sequence of an SCPN influenced its catalytic activity.[64] Atomistic MD simulations, especially accelerated methodologies, are now at a timely stage to be used as a predictive tool to guide the design of SCPNs for targeted applications, before embarking on lengthy synthesis procedures.

VI.8. Simulation protocol

MD simulations were carried out using the AMBER package.[51] The polymer chains were assembled in several steps. First, the monomeric units and chain-ends were built individually with the *Avogadro 1.2.0* software.[65] Each of these residues was then assigned atomic partial charges following the AM1-BCC[54] or the RESP[56] methodology (as discussed in **Section VI.3**), using the *antechamber* module of AMBER. The QM calculations were done with the *Gaussian 16* software.[66] The polymer chains were then built by assembling the monomers in the desired sequence, with randomized chirality, using the *sequence* command of the *LEaP* module of AMBER. The ratio between the α and β anomers of the glucosamine monomers was set as 60 % α and 40 % β, as measured experimentally in aqueous solution.[67] Three different sequences were investigated for the p(J-BD) copolymers, to study the effect of the primary structure on the 3D structures. The first one is a random copolymer, denominated as p(J-*r*-BD). The second one is a bloc copolymer, denominated as p(J-*b*-BD), in which all dodecyl and BTA side-chains are placed consecutively in the center of the chain. The third one is a multiblock copolymer, p(J-*mb*-BD). The dodecyl and BTA grafts are distributed in three clusters, at the beginning, the middle and the end of the copolymer. The glucose-based copolymers, p(G-D) and p(G-B), were only studied as random copolymers. Each polymer was simulated in three replicas, identified by the Roman numerals I, II and III. For the Jeffamine-based systems, the same sequence was used for all three replicas, *i.e.* p(J-*r*-BD) I, II and III all have the same sequence of monomers. For the glucose-based systems, a new (random) sequence was inputted for each replica. All force field parameters for the polymers and their side-chains (Jeffamine, glucose, dodecyl, BTA and Nile Red) were given by GAFF 2.11.[52] The starting structure of the polymer chains were reworked by hand to remove most of the steric clashes using the *PyMOL* software, which was also used to produce all the MD snapshots.[68] This step was followed by a geometry optimization in implicit

solvation, with 1,000 steps of steepest descent followed by 9,000 steps of conjugated gradient. The stable molecules were then solvated in rectangular water boxes, ensuring a minimal distance between any solute atom and the edge of the box of 25 and 40 Å for the Jeffamine- and glucose-based systems, respectively. One Na^+ ion was added to bring the system to electroneutrality. The OPC3 water model was used to describe the solvent.[69] The hydrogen mass repartitioning (HMR) scheme was applied on all solute atoms, enabling the use of a timestep of 4 fs.[70] All subsequent simulations were performed with the GPU version of AMBER. The MD protocol followed five steps. First, a 10,000 steps minimization (1,000 steps of steepest descent and 9,000 steps of conjugated gradient) was carried out on the solvent molecules and ion only, using positional restraints on the solute with a force constant of 25 $kcal.mol^{-1}.Å^{-2}$. A second minimization step was carried out without restraints, with the same methodology. Then the system was heated in 1 ns from 10 to 300 K in the NVT ensemble, with 1 more ns of equilibration under these conditions. During heating, positional restraints were applied on the solute atoms with a force-constant of 10 $kcal.mol^{-1}.Å^{-2}$. The temperature was maintained at 300 K with a Langevin thermostat, using a collision frequency of 1 ps^{-1}. The system was then equilibrated for 10 ns in the NPT ensemble. The pressure was maintained at 1 bar with a Monte Carlo barostat, and the pressure relaxation time was set at 2 ps. Finally, the production phase of the simulation was launched in the same conditions for 2 μs. This portion of the simulation was analyzed, saving a snapshot each ns. For all these steps, the cutoff for non-bonded interactions was fixed at 8.0 Å and the long-range electrostatic interactions were treated by the particle mesh Ewald method. The SHAKE algorithm was applied to constrain bonds involving hydrogen atoms. Note that the simulations on the Jeffamine-based (co)polymers were restarted after 1.2 μs: the last snapshot was extracted and re-solvated in a smaller solvent box, and the simulation was extended until 2 μs, such as to save computational time.

Accelerated MD (aMD) simulations were performed for 400 ns on the p(J) and p(J-*r*-BD) systems (starting from the snapshot extracted after 1.2 μs) and for 300 ns on the p(G-D) copolymer (starting from the initial extended conformation).[53] The building of the system and the first four steps of the simulations, before the production phase, followed the protocol described above. However, the HMR scheme was not applied and the timestep was set to 2 fs. The p(G-D) aggregates of two or three chains were simulated with the same aMD protocol, for more than 1 μs. The macromolecules started as fully extended chains, with initial intermolecular contacts between some dodecyl moieties, such as to promote intermolecular assembly instead of single-chain folding. The basic principle of aMD is to provide a boost on the energy when the system

reaches stable states, to facilitate transitions between local minima separated by high energy barriers. Here, two boosts were applied: one on the dihedral energy, and one on the potential energy. They depend on two boost parameters, E and α, which were determined as follows for the dihedral energy:

$$E_D = (4\ x\ N_{residues}) + E_{dihed,avg} \quad \textbf{(VI.I)}$$

$$\alpha_D = (0.8\ x\ N_{residues}) \quad \textbf{(VI.II)}$$

With E_D and α_D, the dihedral boost parameters, $N_{residues}$, the number of solute residues and $E_{\text{dihed,avg}}$, the average dihedral energy, measured during the 10 ns of equilibration in the NPT ensemble. Similarly, the boost parameters for the potential energy, E_P and α_P:

$$E_P = (0.2\ x\ N_{atoms}) + E_{pot,avg} \quad \textbf{(VI.III)}$$

$$\alpha_P = (0.2\ x\ N_{atoms}) \quad \textbf{(VI.IV)}$$

With N_{atoms}, the total number of atoms in the system (including solvent) and $E_{pot,avg}$, the average potential energy, measured during the 10 ns of equilibration in the NPT ensemble. Note that the boost parameters may be adapted for a higher or lower acceleration.

After the simulations, all analyses were done with the *cpptraj* module of Amber.[71] The root mean square deviation (RMSD) values were computed after removal of the translational and rotational movements, taking the first snapshot of the production phase as the reference structure. The radius of gyration (R_G) is a measure of compactness and gives the average distance of an atom to the geometric center of the system. The R_G was measured on all atoms except hydrogens. The solvent-accessible surface area (SASA) measures the exposure of a group of atoms to its surrounding environment. The higher the SASA, the more the moiety is exposed. SASA values were calculated with the LCPO algorithm, using a van der Waals radius of 1.4 Å for the solvent probe.[72] Root mean square fluctuations (RMSF) are an indicator of the mobility of an atom or group of atoms. The higher the RMSF, the greater the positional fluctuations. First, translational and rotational movements were suppressed by aligning all structures to a reference, generally the first conformation of the production phase (for the RMSF calculated in the last 400 ns, the reference structure was the first of this time interval). Then, RMSF values were computed for each monomer on the backbone carbon atom bearing the side-chain. The fluctuations of the dihedral angles were measured for all bonds in the backbone as their average deviation to their mean value. The fluctuation of one angle θ_A around its mean value θ_{mean}, calculated for the N conformations sampled, was computed as follows:

$$\frac{\sum_i^N | [(\theta_{A,i} - \theta_{mean} + 180) \, modulo \, 360] - 180 |}{N} \quad \textbf{(VI.V)}$$

To avoid the problem of working with a periodic variable, the mean dihedral angle was computed in the cartesian space.[73] Each individual angle, expressed in degrees in the range [-180 ° ; 180 °], is converted in (x,y) coordinates. The average values over the N conformations of the x and y coordinates define the mean dihedral angle in the cartesian space. This angle is then converted back to polar coordinates, in degrees, as θ_{mean}. In the formula, the addition of 180 ° and the application of modulo 360 are done to ensure that $\theta_i - \theta_{mean}$ values are expressed in the range [0 ° ; 360 °]. Then, 180 ° are subtracted to measure the difference in the desired [-180 ° ; 180 °] interval, and the absolute value is taken, as we are only interested in the absolute difference. The hydrogen bonds were detected with the *hbond* command of *cpptraj*, with distance and angle cutoffs of 3.0 Å and 135°, respectively. π-type interactions (parallel stacking) were detected using geometric criteria: two aromatic units are considered in interaction if the distance between their centers of mass is less than or equal to 5.0 Å and if the angle between the normal vectors of their planes is < 45 or > 135°. The asphericity parameter, whose value ranges between 0 for a perfect sphere and 1 for rod-like conformations, was computed based on the gyration tensor values, as described elsewhere.[74] The simulated SAXS curves were generated using the *CRYSOL 3.2.1* software.[75] The average displaced solvent volume per atomic group, the contrast of the hydration shell and the relative background used to generate the simulated SAXS curves were optimized against the experimental SAXS curves (experimental concentration of 1.5 mg.mL^{-1} and 2.5 mg.mL^{-1} for the Jeffamine- and glucose-based (co)polymers, respectively). The discrepancy between the simulated and experimental curves is quantified by *CRYSOL* with a χ^2 value, which compares, for each data point (each *q* value), the simulated intensity and the experimental intensity. The higher the χ^2 value, the less the curves overlap (see Ref. 75 for mathematical details). For the p(J) system, three average curves were obtained, at different times: in the range 520 – 700 ns, in the range 1820 – 2000 ns and in the last 100 ns of the accelerated simulation, to ensure that the conformations probed during the simulations remain in agreement with the experimental SAXS spectra over time. 10 conformations of each replicate (p(J) I, p(J) II and p(J) III), one conformation each 20 ns (or each 10 ns for the aMD simulations) were extracted to compute the average curves. Similarly, the scattering curves of p(J-BD) were obtained by averaging over the three replicas of the three sequences, p(J-*r*-BD), p(J-*b*-BD), and p(J-*mb*-BD). The scattering curves of p(G-D) for one, two or three chains were obtained by averaging the spectra obtained for 10 conformations, generated through aMD simulations.

VI.9. Additional data

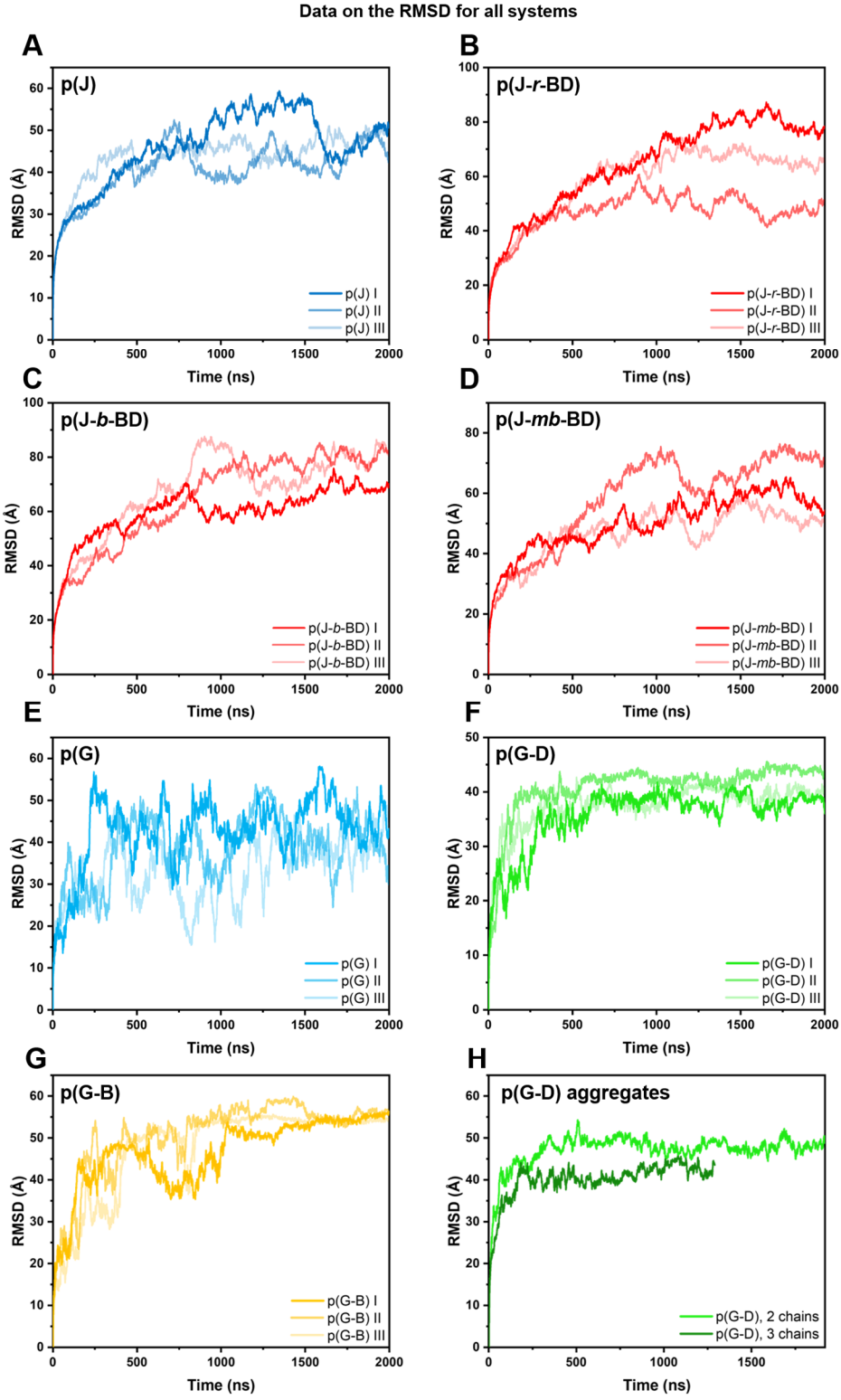


Figure VI.S1. Evolution of the RMSD values over time for the **(A)** p(J), **(B)** p(J-*r*-BD), **(C)** p(J-*b*-BD), **(D)** p(J-*mb*-BD), **(E)** p(G), **(F)** p(G-D), **(G)** p(G-B) systems and **(H)** p(G-D) aggregates of two and three chains. In some cases, all replicas of the same microstructure do not converge to the same RMSD value, reflecting that different conformations may be obtained from a given primary structure.

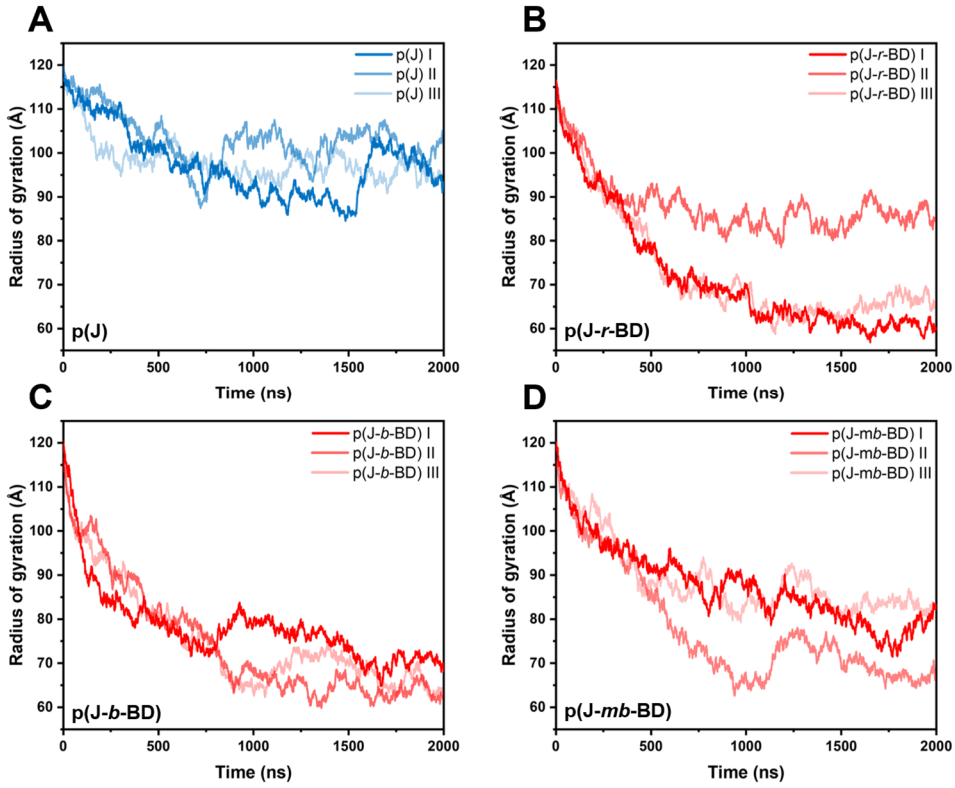


Figure VI.S2. Development of the R_G for the Jeffamine-based (co)polymers during the 2000 ns simulations. **(A)** p(J), **(B)** p(J-*r*-BD), **(C)** p(J-*b*-BD) and **(D)** p(J-*mb*-BD).

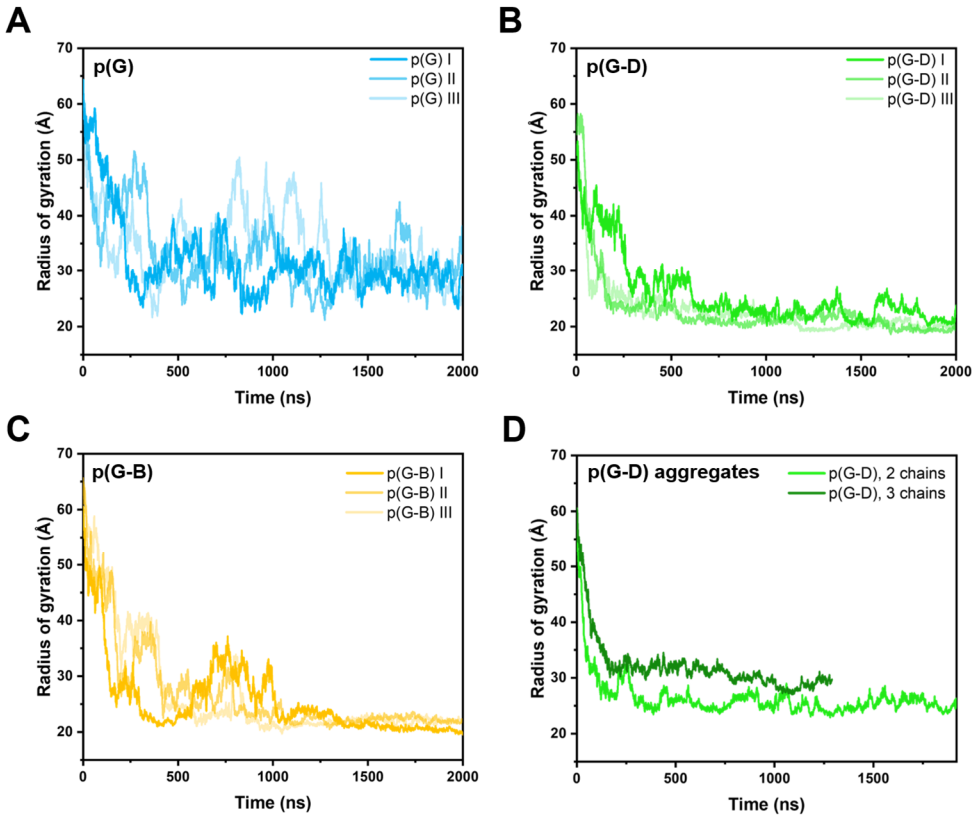
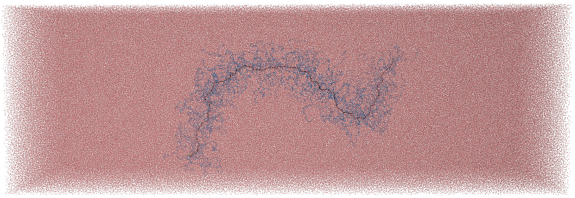
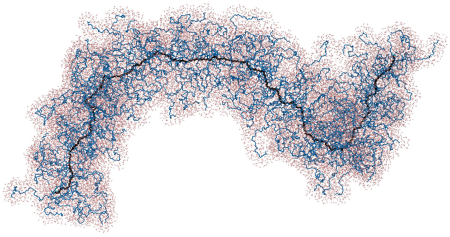


Figure VI.S3. Development of the R_G for the glucose-based (co)polymers during the 2000 ns simulations. **(A)** p(G), **(B)** p(G-D), **(C)** p(G-B) and **(D)** p(G-D) aggregates.

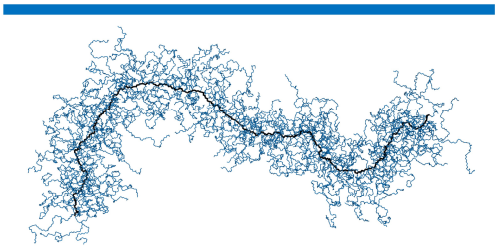
p(J) I – Full water box



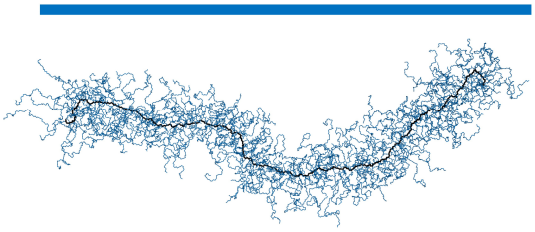
p(J) I – 35,000 closest water molecules



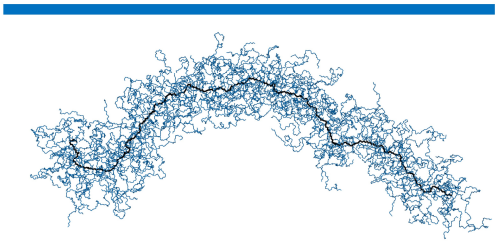
p(J) I



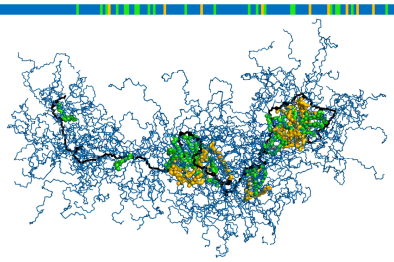
p(J) II



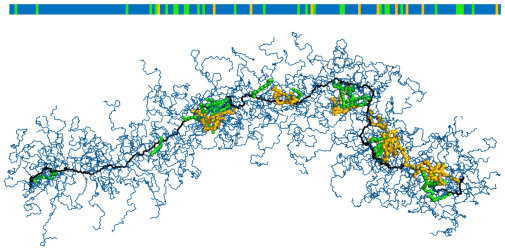
p(J) III



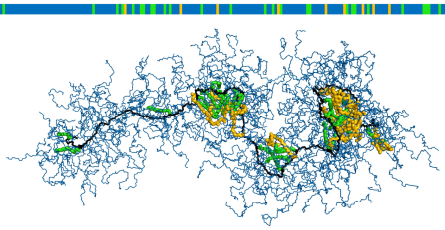
p(J-*r*-BD) I



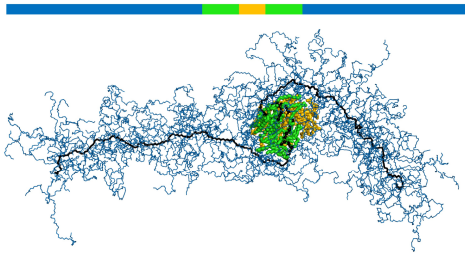
p(J-*r*-BD) II



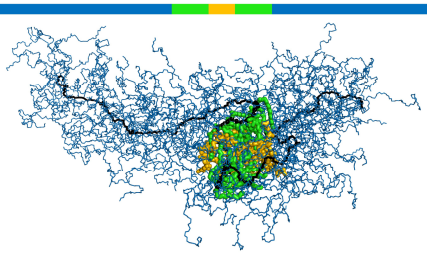
p(J-*r*-BD) III



p(J-*b*-BD) I



p(J-*b*-BD) II



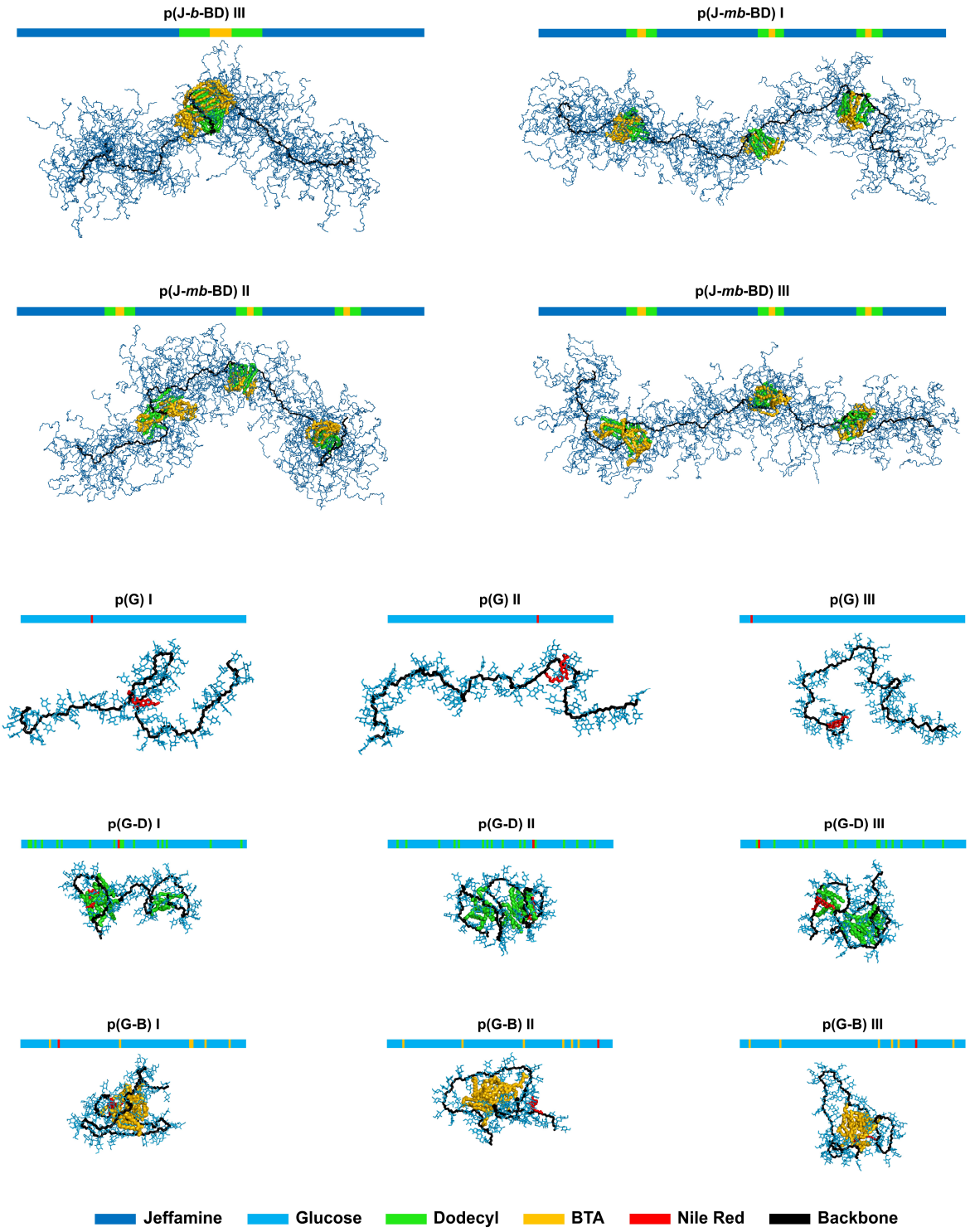


Figure VI.S4. MD snapshots of the three replicas of all systems.

References

[1] P. Ojeda-May, A. U. Mushtaq, P. Rogne, A. Verma, V. Ovchinnikov, C. Grundström, B. Dulko-Smith, U. H. Sauer, M. Wolf-Watz, K. Nam. Dynamic Connection between Enzymatic Catalysis and Collective Protein Motions. *Biochemistry* **2021**, *60*, 2246–2258.

[2] J.-P. Changeux, S. J. Edelstein. Allosteric Mechanisms of Signal Transduction. *Science* **2005**, *308*, 1424–1428.

[3] R. Chen, E. B. Berda. 100th Anniversary of Macromolecular Science Viewpoint: Re-examining Single-Chain Nanoparticles. *ACS Macro Lett.* **2020**, *9*, 1836–1843.

[4] A. Nitti, R. Carfora, G. Assanelli, M. Notari, D. Pasini. Single-Chain Polymer Nanoparticles for Addressing Morphologies and Functions at the Nanoscale: A Review. *ACS Appl. Nano Mater.* **2022**, *5*, 13985–13997.

[5] A. Latorre-Sánchez, J. A. Pomposo. Recent bioinspired applications of single-chain nanoparticles. *Polym. Int.* **2016**, *65*, 855–860.

[6] G. Hattori, Y. Hirai, M. Sawamoto, T. Terashima. Self-assembly of PEG/dodecyl-graft amphiphilic copolymers in water: consequences of the monomer sequence and chain flexibility on uniform micelles. *Polym. Chem.* **2017**, *8*, 7248–7259.

[7] T. Terashima, T. Sugita, K. Fukae, M. Sawamoto. Synthesis and Single-Chain Folding of Amphiphilic Random Copolymers in Water. *Macromolecules* **2014**, *47*, 589–600.

[8] Y. Hirai, T. Terashima, M. Takenaka, M. Sawamoto. Precision Self-Assembly of Amphiphilic Random Copolymers into Uniform and Self-Sorting Nanocompartments in Water. *Macromolecules* **2016**, *49*, 5084–5091.

[9] P. J. M. Stals, M. A. J. Gillissen, T. F. E. Paffen, T. F. A. de Greef, P. Lindner, E. W. Meijer, A. R. A. Palmans, I. K. Voets. Folding Polymers with Pendant Hydrogen Bonding Motifs in Water: The Effect of Polymer Length and Concentration on the Shape and Size of Single-Chain Polymeric Nanoparticles. *Macromolecules* **2014**, *47*, 2947–2954.

[10] M. A. J. Gillissen, T. Terashima, E. W. Meijer, A. R. A. Palmans, I. K. Voets. Sticky Supramolecular Grafts Stretch Single Polymer Chains. *Macromolecules* **2013**, *46*, 4120–4125.

[11] G. M. ter Huurne, L. N. J. de Windt, Y. Liu, E. W. Meijer, I. K. Voets, A. R. A. Palmans. Improving the Folding of Supramolecular Copolymers by Controlling the Assembly Pathway Complexity. *Macromolecules* **2017**, *50*, 8562–8569.

[12] N. Hosono, M. A. J. Gillissen, Y. Li, S. S. Sheiko, A. R. A. Palmans, E. W. Meijer. Orthogonal Self-Assembly in Folding Block Copolymers. *J. Am. Chem. Soc.* **2013**, *135*, 501–510.

[13] K. Matsumoto, T. Terashima, T. Sugita, M. Takenaka, M. Sawamoto. Amphiphilic Random Copolymers with Hydrophobic/Hydrogen-Bonding Urea Pendants: Self-Folding Polymers in Aqueous and Organic Media. *Macromolecules* **2016**, *49*, 7917–7927.

[14] Z. Cui, H. Cao, Y. Ding, P. Gao, X. Lu, Y. Cai. Compartmentalization of an ABC triblock copolymer single-chain nanoparticle via coordination-driven orthogonal self-assembly. *Polym. Chem.* **2017**, *8*, 3755–3763.

[15] Z. Zhu, N. Xu, Q. Yu, L. Guo, H. Cao, X. Lu, Y. Cai. Construction and Self-Assembly of Single-Chain Polymer Nanoparticles via Coordination Association and Electrostatic Repulsion in Water. *Macromol. Rapid Commun.* **2015**, *36*, 1521–1527.

[16] F. Wang, H. Pu, M. Jin, D. Wan. Supramolecular Nanoparticles via Single-Chain Folding Driven by Ferrous Ions. *Macromol. Rapid Commun.* **2016**, *37*, 330–336.

[17] E. A. Appel, J. Dyson, J. del Barrio, Z. Walsh, O. A. Scherman. Formation of Single-Chain Polymer Nanoparticles in Water through Host–Guest Interactions. *Angew. Chem. Int. Ed.* **2012**, *51*, 4185–4189.

[18] F. Huang, J. Liu, M. Li, Y. Liu. Nanoconstruction on Living Cell Surfaces with Cucurbit[7]uril-Based Supramolecular Polymer Chemistry: Toward Cell-Based Delivery of Bio-Orthogonal Catalytic Systems. *J. Am. Chem. Soc.* **2023**, *145*, 26983–26992.

[19] S. Wijker, R. Monnink, L. Rijnders, L. Deng, A. R. A. Palmans. Simultaneously controlling conformational and operational stability of single-chain polymeric nanoparticles in complex media. *Chem. Commun.* **2023**, *59*, 5407–5410.

[20] S. Wijker, L. Deng, F. Eisenreich, I. K. Voets, A. R. A. Palmans. En Route to Stabilized Compact Conformations of Single-Chain Polymeric Nanoparticles in Complex Media. *Macromolecules* **2022**, *55*, 6220–6230.

[21] M. Matsumoto, T. Terashima, K. Matsumoto, M. Takenaka, M. Sawamoto. Compartmentalization Technologies via Self-Assembly and Cross-Linking of Amphiphilic Random Block Copolymers in Water. *J. Am. Chem. Soc.* **2017**, *139*, 7164–7167.

[22] J. Chen, J. Wang, Y. Bai, K. Li, E. S. Garcia, A. L. Ferguson, S. C. Zimmerman. Enzyme-like Click Catalysis by a Copper-Containing Single-Chain Nanoparticle. *J. Am. Chem. Soc.* **2018**, *140*, 13695–13702.

[23] D. Arena, E. Verde-Sesto, I. Rivilla, J. A. Pomposo. Artificial Photosynthases: Single-Chain Nanoparticles with Manifold Visible-Light Photocatalytic Activity for Challenging 'in Water' Organic Reactions. *J. Am. Chem. Soc.* **2024**, *146*, 14397–14403.

[24] Y. Liu, T. Pauloehrl, S. I. Presolski, L. Albertazzi, A. R. A. Palmans, E. W. Meijer. Modular Synthetic Platform for the Construction of Functional Single-Chain Polymeric Nanoparticles: From Aqueous Catalysis to Photosensitization. *J. Am. Chem. Soc.* **2015**, *137*, 13096–13105.

[25] F. Eisenreich, A. R. A. Palmans. *"Supramolecular Catalysis"*, Wiley, **2022**, pp. 489–506.

[26] Y. Liu, S. Pujals, P. J. M. Stals, T. Paulöhrl, S. I. Presolski, E. W. Meijer, L. Albertazzi, A. R. A. Palmans. Catalytically Active Single-Chain Polymeric Nanoparticles: Exploring Their Functions in Complex Biological Media. *J. Am. Chem. Soc.* **2018**, *140*, 3423–3433.

[27] A. Sathyan, S. Croke, A. M. Pérez-López, B. F. M. de Waal, A. Unciti-Broceta, A. R. A. Palmans. Developing Pd(ii) based amphiphilic polymeric nanoparticles for pro-drug activation in complex media. *Mol. Syst. Des. Eng.* **2022**, *7*, 1736–1748.

[28] L. Deng, A. Sathyan, C. Adam, A. Unciti-Broceta, V. Sebastian, A. R. A. Palmans. Enhanced Efficiency of Pd(0)-Based Single Chain Polymeric Nanoparticles for *in Vitro* Prodrug Activation by Modulating the Polymer's Microstructure. *Nano Lett*. **2024**, *24*, 2242–2249.

[29] K. Mundsinger, A. Izuagbe, B. T. Tuten, P. W. Roesky, C. Barner-Kowollik. Single Chain Nanoparticles in Catalysis. *Angew. Chem. Int. Ed.* **2024**, *63*, e202311734.

[30] C.-C. Cheng, D.-J. Lee, Z.-S. Liao, J.-J. Huang. Stimuli-responsive single-chain polymeric nanoparticles towards the development of efficient drug delivery systems. *Polym. Chem.* **2016**, *7*, 6164–6169.

[31] A. P. P. Kröger, J. M. J. Paulusse. Single-chain polymer nanoparticles in controlled drug delivery and targeted imaging. *J. Controlled Release* **2018**, *286*, 326–347.

[32] J. Chen, K. Li, J. S. "Lucy. Shon, S. C. Zimmerman. Single-Chain Nanoparticle Delivers a Partner Enzyme for Concurrent and Tandem Catalysis in Cells. *J. Am. Chem. Soc.* **2020**, *142*, 4565–4569.

[33] D. N. F. Bajj, M. V. Tran, H.-Y. Tsai, H. Kim, N. R. Paisley, W. R. Algar, Z. M. Hudson. Fluorescent Heterotelechelic Single-Chain Polymer Nanoparticles: Synthesis, Spectroscopy, and Cellular Imaging. *ACS Appl. Nano Mater.* **2019**, *2*, 898–909.

[34] L. Deng, L. Albertazzi, A. R. A. Palmans. Elucidating the Stability of Single-Chain Polymeric Nanoparticles in Biological Media and Living Cells. *Biomacromolecules* **2022**, *23*, 326–338.

[35] M. Artar, E. Huerta, E. W. Meijer, A. R. A. Palmans. *"Sequence-Controlled Polymers: Synthesis, Self-Assembly, and Properties"*, ACS, **2014**, pp. 313–325.

[36] E. Blasco, B. T. Tuten, H. Frisch, A. Lederer, C. Barner-Kowollik. Characterizing single chain nanoparticles (SCNPs): a critical survey. *Polym. Chem.* **2017**, *8*, 5845–5851.

[37] E. Archontakis, L. Deng, P. Zijlstra, A. R. A. Palmans, L. Albertazzi. Spectrally PAINTing a Single Chain Polymeric Nanoparticle at Super-Resolution. *J. Am. Chem. Soc.* **2022**, *144*, 23698–23707.

[38] G. M. ter Huurne, M. A. J. Gillissen, A. R. A. Palmans, I. K. Voets, E. W. Meijer. The Coil-to-Globule Transition of Single-Chain Polymeric Nanoparticles with a Chiral Internal Secondary Structure. *Macromolecules* **2015**, *48*, 3949–3956.

[39] J. A. Pomposo, I. Perez-Baena, F. Lo Verso, A. J. Moreno, A. Arbe, J. Colmenero. How Far Are Single-Chain Polymer Nanoparticles in Solution from the Globular State?. *ACS Macro Lett.* **2014**, *3*, 767–772.

[40] S. L. Hilburg, Z. Ruan, T. Xu, A. Alexander-Katz. Behavior of Protein-Inspired Synthetic Random Heteropolymers. *Macromolecules* **2020**, *53*, 9187–9199.

[41] A. Arbe, J. A. Pomposo, A. J. Moreno, F. LoVerso, M. González-Burgos, I. Asenjo-Sanz, A. Iturrospe, A. Radulescu, O. Ivanova, J. Colmenero. Structure and dynamics of single-chain nanoparticles in solution. *Polymer* **2016**, *105*, 532–544.

[42] Y. Zhang, X. Jia, R. Shi, S. Li, H. Zhao, H. Qian, Z. Lu. Synthesis of Polymer Single-Chain Nanoparticle with High Compactness in Cosolvent Condition: A Computer Simulation Study. *Macromol. Rapid Commun.* **2020**, *41*, 1900655.

[43] R. A. Patel, S. Colmenares, M. A. Webb. Sequence Patterning, Morphology, and Dispersity in Single-Chain Nanoparticles: Insights from Simulation and Machine Learning. *ACS Polym. Au* **2023**, *3*, 284–294.

[44] S. Wijker, D. Dellemme, L. Deng, B. Fehér, I. K. Voets, M. Surin, A. R. A. Palmans. Revealing the Folding of Single-Chain Polymeric Nanoparticles at the Atomistic Scale by Combining Computational Modeling and X-ray Scattering. *ACS Macro Lett.* **2025**, *14*, 428–433.

[45] D. L. Sackett, J. Wolff. Nile red as a polarity-sensitive fluorescent probe of hydrophobic protein surfaces. *Anal. Biochem.* **1987**, *167*, 228–234.

[46] H. Zhang, L. Zhang, J. You, N. Zhang, L. Yu, H. Zhao, H.-J. Qian, Z.-Y. Lu. Controlling the Chain Folding for the Synthesis of Single-Chain Polymer Nanoparticles Using Thermoresponsive Polymers. *CCS Chemistry* **2021**, *3*, 2143–2154.

[47] W. G. Noid. Perspective: Advances, Challenges, and Insight for Predictive Coarse-Grained Models. *J. Phys. Chem. B* **2023**, *127*, 4174–4207.

[48] S. Dhamankar, M. A. Webb. Chemically specific coarse-graining of polymers: Methods and prospects. *J. Polym. Sci.* **2021**, *59*, 2613–2643.

[49] J. Jin, A. J. Pak, A. E. P. Durumeric, T. D. Loose, G. A. Voth. Bottom-up Coarse-Graining: Principles and Perspectives. *J. Chem. Theory. Comput.* **2022**, *18*, 5759–5791.

[50] F. Schmid. Understanding and Modeling Polymers: The Challenge of Multiple Scales. *ACS Polym. Au* **2023**, *3*, 28–58.

[51] D. A. Case, T. E. Cheatham Iii, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[52] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[53] D. Hamelberg, J. Mongan, J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

[54] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[55] D. Dellemme, S. Kardas, C. Tonneaux, J. Lernould, M. Fossépré, M. Surin. From Sequence Definition to Structure-Property Relationships in Discrete Synthetic Macromolecules: Insights from Molecular Modeling. *Angew. Chem. Int. Ed.* **2025**, *64*, e202420179.

[56] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[57] B. Ensing, A. Tiwari, M. Tros, J. Hunger, S. R. Domingos, C. Pérez, G. Smits, M. Bonn, D. Bonn, S. Woutersen. On the origin of the extremely different solubilities of polyethers in water. *Nat. Commun.* **2019**, *10*, 2893.

[58] V. E. Bychkova, G. V. Semisotnov, V. A. Balobanov, A. V. Finkelstein. The Molten Globule Concept: 45 Years Later. *Biochemistry (Moscow)* **2018**, *83*, S33–S47.

[59] C. M. Quezada, B. A. Schulman, J. J. Froggatt, C. M. Dobson, C. Redfield. Local and Global Cooperativity in the Human α-Lactalbumin Molten Globule. *J. Mol. Biol.* **2004**, *338*, 149–158.

[60] R. Pancsa, D. Raimondi, E. Cilia, W. F. Vranken. Early Folding Events, Local Interactions, and Conservation of Protein Backbone Rigidity. *Biophys. J.* **2016**, *110*, 572–583.

[61] M. Artar, E. R. J. Souren, T. Terashima, E. W. Meijer, A. R. A. Palmans. Single Chain Polymeric Nanoparticles as Selective Hydrophobic Reaction Spaces in Water. *ACS Macro Lett.* **2015**, *4*, 1099–1103.

[62] M. B. Baker, L. Albertazzi, I. K. Voets, C. M. A. Leenders, A. R. A. Palmans, G. M. Pavan, E. W. Meijer. Consequences of chirality on the dynamics of a water-soluble supramolecular polymer. *Nat. Commun.* **2015**, *6*, 6234.

[63] M. F. J. Mabesoone, S. Kardas, H. Soria-Carrera, J. Barberá, J. M. de la Fuente, A. R. A. Palmans, M. Fossépré, M. Surin, R. Martín-Rapún. Competitive hydrogen bonding in supramolecular polymerizations of tribenzylbenzene-1,3,5-tricarboxamides. *Mol. Syst. Des. Eng.* **2020**, *5*, 820–828.

[64] I. Asenjo-Sanz, T. Claros, E. González, J. Pinacho-Olaciregui, E. Verde-Sesto, J. A. Pomposo. Significant effect of intra-chain distribution of catalytic sites on catalytic activity in 'clickase' single-chain nanoparticles. *Mater. Lett.* **2021**, *304*, 130622.

[65] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, G. R. Hutchison. Open Access Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Software* **2012**, *4*, 17.

[66] Gaussian 16, Revision A.03, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, **2016**.

[67] D. L. Bertuzzi, T. B. Becher, N. M. R. Capreti, J. Amorim, I. D. Jurberg, J. D. Megiatto, C. Ornelas. General Protocol to Obtain D-Glucosamine from Biomass Residues: Shrimp Shells, Cicada Sloughs and Cockroaches. *Global Challenges* **2018**, *2*, 1800046.

[68] Schrödinger LLC, *The PyMOL Molecular Graphics System, Version 2.5.4*, **2015**.

[69] S. Izadi, A. V. Onufriev. Accuracy limit of rigid 3-point water models. *J. Chem. Phys.* **2016**, *145*, 074501.

[70] C. W. Hopkins, S. Le Grand, R. C. Walker, A. E. Roitberg. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874.

[71] D. R. Roe, T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

[72] J. Weiser, P. S. Shenkin, W. Clark Still. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217230.

[73] Bishop, C. M. "*Pattern Recognition and Machine Learning*", Springer New York, **2006**.

[74] Blavatska, von Ferber, Holovatch. Shapes of macromolecules in good solvents: field theoretical renormalization group approach. *Condens. Matter Phys.* **2011**, *14*, 33701.

[75] D. Svergun, C. Barberato, M. H. J. Koch. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **1995**, *28*, 768–773.

VII. Conclusion and perspectives

Throughout our thesis, we attempted to get a glimpse into sequence – structure – function relationships for a variety of SCMs and SDMs, from the impressive properties displayed by the biomacromolecules of life to purely artificial systems, designed following chemical intuition. After reviewing the most important results gathered from our work, we will regroup the general lessons we could learn, and provide a general opinion on the current and future roles that we envision for human-made SCMs and SDMs.

The defined sequence and folding properties of natural SDMs were harnessed for biorecognition applications in **Chapter IV**. In **Chapter IV-A**, we investigated the interaction between collagen-mimetic peptides and a collagen-binding receptor. MD simulations showed that very subtle changes, here concerning the stereochemistry of a small number of AAs in the peptide, could strongly affect its interactions with the receptor. This molecular-level information could be connected to experimental observations showing a reduced ability of this peptide to support cell adhesion and migration. In **Chapter IV-B**, MD simulations helped us understand how light-induced *trans* to *cis* isomerization could affect the binding of photoswitchable ligands to a complementary DNA template. Our results indicated that hydrogen bonding interactions between the ligands in their *trans* configuration and the template were reinforced and even dominated by a vast network of π-type interactions between the ligands. These interactions were weaker for the molecules in their *cis* configuration, leading to more disordered assemblies and weaker H-bonds with the template. These examples demonstrate that natural SDMs are very sensitive to small changes in the 3D structure of their ligands.

In **Chapter V**, we investigated the possibility to precisely preorganize catalytic and recognition units within two synthetic SDMs, which need to self-assemble to form an active supramolecular catalyst. Our results indicated that the molecules were highly flexible and adopted folded conformations, leading to the generation of a disordered and globular duplex, where all functional groups could interact. This shows that precisely controlling both sequence and stereochemistry (as the SDMs are enantiopure) does not necessarily imply the formation of well-defined 3D structures, unlike what is observed for natural SDMs. However, controlling the monomeric composition was essential for achieving high catalytic activity, and specific interactions were detected between the complementary recognition units despite the apparent disorder within the supramolecular duplex.

In **Chapter VI**, we studied the folding and 3D structures of purely synthetic amphiphilic SCPNs in water. Our results revealed that, depending on the nature of the hydrophilic grafts, very different morphologies were obtained. In particular, polymers functionalized with oligo(ethylene oxide) grafts adopted extended, worm-like structures, with local folding around the hydrophobic moieties. Sequence effects were studied for these polymers *in silico*, and it was shown that their restricted folding allowed them to retain the information encoded in their sequence within their 3D structure. This work demonstrates that MD simulations are reaching a stage where they can be reliably used as a predictive tool to guide the design of SCPNs. Recently, a strategy combining MD simulations and machine learning (ML) has been applied to predict the conformational landscape of SCPNs.[1] While this study relied on a simplified physical model to describe the polymers, we now have the computational resources to investigate more complex systems with all-atom MD simulations, accounting for realistic chemical structures that incorporate diverse functional groups.

Our results revealed that sequence and chirality are of crucial importance when interactions with natural SDMs are involved. In **Chapter IV-A**, the presence of a glutamate moiety inside an AA recognition motif GXX'GEX" was crucial for the binding to the receptor. Sequence alone, however, is not enough: two peptides distinguished only by their chirality displayed different behaviors. In view of creating synthetic SDMs replicating the properties of biomacromolecules, a particular attention should be directed to stereochemistry, as indicated by others.[2] This contrasts with traditional polymers, for which tacticity is generally neglected. Mismatches between R- and S-monomers along the chain could be detrimental to the formation of controlled structures. The importance of sequence was also demonstrated for synthetic SDMs in **Chapter V**, as only one missing catalytic unit could significantly decrease the catalytic activity of the duplex. This work also revealed that synthetic SDMs could display very different properties than their natural counterparts. While proteins and nucleic acids rely on a rather rigid backbone, the oligomers presented here probably incorporate too many rotatable bonds between their functional units to retain the information encoded in their primary structure. This led to very flexible chains, where the precise monomer ordering does not seem to be crucial for the catalytic activity. Other SDMs, based on different chemistries, also exhibited an important flexibility.[3] While sequence effects were clearly demonstrated even within flexible and folded systems, they were sometimes counter-intuitive, difficult to predict, and challenging to rationalize.[4,5] Therefore, while the ability to undergo conformational changes is required for some applications, it seems important to find an optimized balance between flexibility and rigidity within synthetic SDMs. Currently, they generally do not tend to adopt well-

defined 3D structures. We mentioned in **Chapter I** the idea of constituting databases of MD-generated structures, to train ML algorithms aimed at developing accurate predictive models for artificial SCMs. However, for such models to be efficient, very large datasets of well-resolved structures are required, as illustrated with the Protein Data Bank for proteins. We saw in **Chapter II** that ML algorithms already struggle with single-stranded nucleic acids, which can adopt a variety of folded structures and for which experimental data remains sparse. If these natural SDMs based on a well-known backbone and functionalized with only four different monomers are already challenging, it seems very difficult to envision accurate predictive models for synthetic SDMs. To make progress, general design principles may need to be established in order to reduce the conformational space. One direction could be to restrict the number of rotatable bonds between functional units, favoring short backbones and side-chains, as observed in natural SDMs. This strategy would at least limit folding for short chains, giving more weight to the encoded primary structure, and facilitating the establishment of sequence – structure relationships, before gradually increasing the complexity of the studied systems.

Based on these lessons, what can we expect for the future? The field of SCMs and SDMs has attracted a considerable interest, driven by the possibility to design "artificial proteins" with completely novel chemistries. Several examples, in the literature and in our thesis, have demonstrated the promises of synthetic SCMs. However, although significant progress has been made in the past 15 years, the synthesis of SDMs still essentially rely on tedious step-by-step approaches. Without innovations and the discovery of new synthetic methodologies, an absolute control over the sequence of long polymer chains (more than ~20 units) still seems far of reach.[6] In the same time, we have seen that characterizing the 3D structures of such molecules remains extremely challenging, even for short chains. In the medium term, the practical applications of SDMs are likely to remain limited. Among them, **information storage** stands out. The impossibility to reach high DP for synthetic SDMs is mitigated by the possibility of incorporating large monomer libraries, which easily outperform the four nucleobases of DNA, also considered for such applications. Another advantage of SDMs over DNA is the possibility of adapting the backbone chemistry to the detection method, for instance by designing backbones with predictable fragmentation patterns suitable for MS/MS. Moreover, the polymer can be tailored for stability under any desired conditions, whereas DNA imposes constraints for long-term storage. Beyond applications, SDMs provide an ideal platform for fundamental studies, enabling precise investigation of the role of specific monomer units. This was illustrated in our thesis in **Chapter IV-A**, with the role of glutamate, and in **Chapter**

V, where all five catalytic units were required to achieve efficient catalysis. Therefore, there remains a lot of efforts to engage in fundamental research before any practical application, in particular on investigating the role of individual functional units within complex processes. This is extremely valuable to study catalytic mechanisms or (bio)recognition phenomena. For applications where the control over the 3D structure is required, SDMs do not appear to be the best option. While there is something beautiful in trying to reproduce the absolute sequence and structure definition of proteins or nucleic acids into synthetic materials, we are still far from being able to rationally implement these design principles into functional systems. To mimic natural biomacromolecules, SCPNs incorporating a limited control over their primary structure appear more promising. This is particularly true for systems with restricted folding, such as the Jeffamine-based polymers studied in **Chapter VI**. SCMs based on this design, with block or multiblock architectures, could be ideal targets to study sequence – structure – function relationships, as the information encoded in their sequence is retained in their 3D structure. Polydispersity and the lack of absolute sequence definition would not be problematic, provided that the global morphology can be tuned by adjusting the ratio and nature of solvophilic and solvophobic units. Furthermore, the design of such single-chain systems can now be rationalized using all-atom MD simulations, enabling the investigation of sequence – structure relationships *in silico* to guide the development of efficient sequence-controlled SCPNs. For these systems, where clearer links between sequence and structure appear, the emergence of predictive ML algorithms seems more realistic. In this sense, sequence-controlled SCPNs may represent a major step towards synthetic materials with protein-like levels of control, opening the door to a new generation of functional artificial macromolecules.

The modeling strategy used in our thesis will also undoubtedly benefit from further advances in computational resources and the development of new methodologies. The simulation of natural SDMs is steadily improving, in particular concerning proteins, with the help of ML predictive tools such as AlphaFold[7] and the emergence of ML force fields approaching QM accuracy within reasonable timescales.[8] Another particularly exciting perspective is the ability to simulate more realistic biological environments. For instance, in **Chapter IV-A**, the interactions between peptides and the binding site of an integrin were investigated in isolation, neglecting the influence of the surrounding cellular context. In the future, we can envision the simulation of increasingly complex biological environments, potentially up to whole-cell models, through multiscale approaches exploiting coarse-grained (CG) representations and ML tools.[9] This would allow us to study protein-ligand complexes in realistic cellular

environments, and to explore how these complexes influence and are influenced by the myriad of other cellular components with which they dynamically interact. Advances in this direction would make MD simulations an even more powerful computational microscope, capable of observing dynamic cellular processes with atomistic resolution.

References

[1] R. A. Patel, S. Colmenares, M. A. Webb. Sequence Patterning, Morphology, and Dispersity in Single-Chain Nanoparticles: Insights from Simulation and Machine Learning. *ACS Polymers Au* **2023**, *3*, 284–294.

[2] R. Szweda. Sequence- and stereo-defined macromolecules: Properties and emerging functionalities. *Prog. Polym. Sci.* **2023**, *145*, 101737.

[3] D. Dellemme, S. Kardas, C. Tonneaux, J. Lernould, M. Fossepre, M. Surin. From sequence definition to structure-property relationships in discrete synthetic macromolecules: insights from molecular modeling. *Angew. Chem. Int. Ed.* **2025**, *64*, e202420179.

[4] J. Li, Q. Qin, S. Kardas, M. Fossépré, M. Surin, A. E. Fernandes, K. Glinel, A. M. Jonas. Sequence Rules the Functional Connections and Efficiency of Catalytic Precision Oligomers. *ACS Catal.* **2022**, *12*, 2126–2131.

[5] R. Aksakal, C. Tonneaux, A. Uvyn, M. Fossépré, H. Turgut, N. Badi, M. Surin, B. G. De Geest, Filip. E. Du Prez. Sequence-defined antibody-recruiting macromolecules. *Chem. Sci.* **2023**, *14*, 6572–6578.

[6] J.-F. Lutz. The future of sequence-defined polymers. *Eur. Polym. J.* **2023**, *199*, 112465.

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[8] T. Wang, X. He, M. Li, Y. Li, R. Bi, Y. Wang, C. Cheng, X. Shen, J. Meng, H. Zhang, H. Liu, Z. Wang, S. Li, B. Shao, T.-Y. Liu. Ab initio characterization of protein molecular dynamics with AI2BMD. *Nature* **2024**, *635*, 1019–1027.

[9] G. Bonollo, G. Trèves, D. Komarov, S. Mansoor, E. Moroni, G. Colombo. Advancing Molecular Simulations: Merging Physical Models, Experiments, and AI to Tackle Multiscale Complexity. *J. Phys. Chem. Lett.* **2025**, *16*, 3606–3615.

Scientific activities

Peer-reviewed publications

A. Remson, **D. Dellemme**, M. Luciano, M. Surin, S. Gabriele. Chiral mismatch in collagen-mimetic peptides modulates cell migration through integrin-mediated molecular recognition. *In preparation*.

L. Groignet, **D. Dellemme**, Q. Duez, A. Fizazi, J. M. Colet, P. Brocorens, M. Surin, P. Gerbaux, J. De Winter. Impact of post-translational succination on small ubiquitin-like modifier 1 structure: a dual approach combining gas phase and solution studies. *ACS Pharmacol. Transl. Sci.* **2025**, 8, 2683-2693.

D. Dellemme, S. Kardas, C. Tonneaux, J. Lernould, M. Fosséprés, M. Surin. From sequence definition to structure-property relationships in discrete macromolecules: insights from molecular modeling. *Angew. Chem. Int. Ed.* **2025**, 64, e202420179.

S. Wijker, **D. Dellemme**, L. Deng, B. Fehér, I. K. Voets, M. Surin, A. R. A. Palmans. Revealing the folding of single-chain polymeric nanoparticles at the atomistic scale by combining computational modeling and X-ray scattering. *ACS Macro Lett.* **2025**, 14, 428-433.

K. Grafskaia, Q. Qin, J. Li, D. Magnin, **D. Dellemme**, M. Surin, K. Glinel, A. M. Jonas. Chain stretching in brushes favors sequence recognition for nucleobase-functionalized flexible precise oligomers. *Soft Matter*, **2024**, 20, 8303-8311.

N. Nogal, S. Guisán, **D. Dellemme**, M. Surin, A. de la Escosura. Selectivity in the chiral self-assembly of nucleobase-arylazopyrazole photoswitches along DNA templates. *J. Mater. Chem. B*. **2024**, 12, 3703-3709.

Q. Qin, J. Li, **D. Dellemme**, M. Fosséprés, G. Barrozino-Consiglio, I. Nekkaa, A. Boborodea, A. E. Fernandes, K. Glinel, M. Surin, A. M. Jonas. Dynamic self-assembly of supramolecular catalysts from precision macromolecules. *Chem. Sci.* **2023**, 14, 9283-9292.

M. Coste, C. Kotras, Y. Bessin, V. Gervais, **D. Dellemme**, M. Leclercq, M. Fosséprés, S. Richeter, S. Clément, M. Surin, S. Ulrich. Synthesis, self-assembly, and nucleic acid recognition of an acylhydrazone-conjugated cationic tetraphenylethene ligand. *Eur. J. Org. Chem.* **2021**, 2021, 1123-1135.

D. Dellemme, M. Leclercq, M. Fossépré, J. Li, Q. Qin, A. M. Jonas, K. Glinel, M. Surin. Folding and self-assembly of sequence-defined oligomers containing nucleobases. *Chimie Nouvelle*, **2021**, 138, 7-12.

Oral presentations

Elucidating the 3D structure and dynamics of single-chain polymeric nanoparticles – Annual Meeting of the Belgian Polymer Group 2025 (BPG 2025), May 2025, Houffalize, Belgium – Best oral presentation prize.

Molecular dynamics simulations of multimillion-atom systems using the power of LUCIA and LUMI GPUs – 13th CECI Users Meeting, April 2024, Louvain-la-Neuve, Belgium.

Molecular modeling of single-chain nanoparticles as enzyme mimetics – EDT-CHIM PhD Scientific Day 2024, January 2024, Liège, Belgium.

Self-assembly of supramolecular catalysts from precision macromolecules: insights from molecular dynamics simulations – Journées d'études des Polymères 2023 (JEPO 2023), October 2023, Eppe-Sauvage, France.

Understanding the self-assembly of precision catalytic oligomers through molecular dynamics simulations – SRC Young Chemists' Day 2023, May 2023, Mons, Belgium.

Sequence-defined oligomers bearing nucleobases: targeting specificity in folding and supramolecular assembly – Annual Meeting of the Belgian Polymer Group 2022 (BPG 2022), November 2022, Blankenberge, Belgium – Master's Thesis Award French Speaking Community.

Poster presentations

Chiral mismatch in collagen-mimetic peptides modulates cell migration through integrin-mediated molecular recognition – Journées André Collet de la Chiralité 2025 (JACC 2025), June 2025, Rouen, France.

Influence of the size of hydrophilic grafts on the structure of single-chain polymeric nanoparticles: insights from molecular dynamics simulations – IUPAC MACRO 2024, July 2024, Warwick, United Kingdom.

Self-assembly of precision macromolecules forming a supramolecular catalyst: insights from molecular dynamics simulations – Supr@Paris 2024, May 2024, Paris, France.

Understanding the self-assembly of precision catalytic oligomers through molecular dynamics simulations – 2nde Journée de Chimie Supramoléculaire 2023 (JCS 2023), June 2023, Montpellier, France – Prize for best poster presentation.

Sequence-defined stereocontrolled oligomers bearing nucleobases: targeting specificity in supramolecular assembly – SRC 2022 Scientific Day, October 2022, Liège, Belgium.

Sequence-defined stereocontrolled oligomers bearing nucleobases: targeting specificity in supramolecular assembly – Journées André Collet de la Chiralité 2022 (JACC 2022), October 2022, Biarritz, France.

This work has been supported by the F.R.S.-FNRS through a FRIA grant funded by the French Community of Belgium.