MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
Vincent Conticello	Chemistry	Emory University	vcontic@emory.edu
James Kindt	Chemistry	Emory University	

Project Title

Data Driven Discovery of Structurally Defined 2D Organic Materials

Project Synopsis (approx. 100 words)

Two-dimensional organic assemblies (i.e., nanosheets) represent a promising structural platform to arrange molecular and supramolecular substrates with precision in ordered arrays. In addition, these types of nanomaterials may be integrated more straightforwardly into devices using conventional nanofabrication methods. General principles for the design of structurally ordered 2D assemblies from synthetic protomers are not readily available, especially for non-peptidic oligomers. The latter materials are particularly appealing from the perspective of potential applications as they often display interesting optical, electronic, and mechanical properties. A broad-based materials informatics approach could assist in mining available structural databases to identify organic scaffolds that promote the formation of 2D assemblies of defined plane symmetry and internal structure. Initial efforts would focus on peptidic assemblies in order to define paradigms that would be applicable to more conventional, non-peptidic systems.

Identified Data-Science Collaborative Need (approx. 100 words)

1. Effective methods to identify well-packed 2D layers within 3D structures within large structural databases.
2. Definition of criteria to distinguish stable interfaces from weak crystallization artifacts.
3. Development of machine learning or alternative classification approaches to glean information on the relationship between local protomer structure and 2D lattice symmetry
4. Correlated database development to enable identification of novel sequences that would be compatible with the formation of ordered 2D lattices.

Data Origin and Access (*data must be available and sharable with data science teams* – please address: data source/origin, access privileges, sharing privileges)

Protein Data Bank: freely available and accessible

Cambridge Structural Database: subscription service through which we have access through an annual institutional subscription

Polymer Database (National Institute of Materials Science, Japan): free access, requires registration.

Scientific Literature: it is possible that not all of the polymer crystal structure data has been effectively compiled or is readily available via conventional literature search engines (see below). Would require access to scientific literature that may be behind a paywall. Emory has many institutional subscriptions, but may not have complete coverage.

Project Description

Rationale and Research Goals: Research on 2D materials is currently undergoing expansive growth due to the convergence of a variety of factors: (1) demonstration of compelling optical, electronic, and magnetic properties, (2) potential ease of integration into conventional device architectures, and (3) continuously improving methods for controlled fabrication on the nano- to meso-scale.[1-2] However, the rational and predictable design of these materials remains a significant challenge. Sequence-specific polymers, such as peptides, proteins and structurally related synthetic polyamide foldamers, have been examined as substrates for the construction of structurally ordered two-dimensional assemblies (nanosheets) and represent promising structural platforms to arrange molecular and supramolecular substrates with precision in ordered arrays.[3] The well-defined correlations between sequence and structure that are observed for these types of materials offer the opportunity for rational design. Structurally defined nanosheets have been created from self-assembly of a variety of appropriately designed structural motifs including collagen-mimetic peptides, β-sheet peptides, straight α-helices, α-helical coiled-coils, peptoids, and globular proteins.[3]

Despite these recent successes, most of the aforementioned nanosheets arose fortuitously, i.e., with relatively limited insight from rational sequence design. Computational methods have been employed to facilitate the process of sequence selection, however the large size of these assemblies and structural complexity of the protomers currently rule out atomistic modeling and, consequently, *ab initio* design. Successful efforts in computational design have focused on structurally well-characterized protomers of specific, crystallographically defined rotational symmetry, which are subsequently arranged on 2D lattices with compatible plane symmetry.[4,5] This approach has met thus far with some degree of success, although the observed 2D lattice symmetry is not always in agreement with the computational design. The challenges observed for design of 2D peptide assemblies are even more daunting in the case of conventional polymers based on non-peptidic backbone chemistries. Sequence-structure correlations are not as well developed for synthetic polymers as for peptides and proteins, particularly as far less high resolution structural data are available. Nonetheless, it is well known that polymers often crystallize from dilute solution into 2D layered crystals through adjacent reentry chain folding.[6,7] The polymer backbone is oriented perpendicular to the surface of the sheet and defines the thickness of the lamellar crystals. The development of methods to rationally design 2D polymer/oligomer crystals would open up a wide range of opportunities to engineer nanosheets from conventional materials that display significantly different (but complementary) physico-chemical properties in comparison to peptide-based materials.

Big Data Driven Approaches to 2D Materials Design: Rich troves of structural information are available for proteins and small molecules (peptide oligomers, etc.) in the Protein Data Bank (137K+ structures) and the Cambridge Structural Database (800K+ structures), respectively. Big Data methods could be potentially employed to leverage this structural information to guide the design of novel types of 2D assemblies. Effective structural search methods are needed that would identify well-packed 2D lattices within the crystallographic data and classify them on the basis of plane symmetry. Machine learning (or other informatics based methods) would be desirable to analyze the structures and identify relationships embedded within the structures that might specify different packing symmetries of the respective 2D lattices. Rubrics need to be defined that would enable the identification of "stable" 2D layers within the 3D crystal structures of peptides and proteins within the databases. This process would entail the identification of physico-chemical criteria that could be employed as descriptors for structural stability.

The most commonly observed 2D peptide nanostructures are based on appropriately designed, sequence-specific chiral rods (corresponding to helical or super-helical subunits or protomers) that laterally associate with structural specificity to afford ordered two-dimensional lattices in which sheet thickness, internal structure, and surface chemistry can be controlled through molecular-level interactions (electrostatic attraction, hydrogen-bonding, and complementary side-chain packing). Helical and super-helical structural motifs are amenable to parametric design approaches using computational design engines once the helical fold is identified and parameterized.[8] The immediate objective in this materials informatics strategy would be to identify protein folds compatible with specific 2D packing symmetries such that a parametric description can be generated for the individual structural elements (protomers). Polar sequence patterning often plays an influential role in determination of the local peptide chain conformation and, additionally, may influence lateral packing arrangements within the 2D lattice (Figure 1).[9] The limited structural evidence from well-characterized 2D peptide assemblies suggests that electrostatic interactions play a significant structural role in determining the packing symmetry and local peptide orientation. These observations suggest that residue-based coarse-grained modeling might be a useful approach to simulate the 2D packing if data classification methods could identify relationships that would restrict the pool of polar sequence patterns compatible with specific local helical symmetries and extended plane symmetries.

If successful, this inverse strategy, i.e., the identification of sequence patterns and general design principles from peptide-derived structural correlations, could be applied in the forward direction to select and design the sequences of synthetic polymer chain architectures that are compatible with the formation of structurally defined 2D crystals of controllable symmetry and internal structure. Polymer single crystals arise from chain folding in which turn sequences are localized at the surface of the crystal while helical segments define the lamellar thickness. In theory, these helical polymer segments can be parameterized in a similar manner to peptide chain conformations. Helical and supra-helical patterned architectures that promote specific 2D lattice formation in peptide structures (*see above*) can be screened to identify synthetic polymer structures that would be compatible with the different lattice sub-types. The most desirable outcome would be to identify synthetic sequence-complex oligomers based on conventional polymer backbones that would form structurally well-defined 2D materials. The need for materials informatics in this aspect of the research would be the generation of a searchable database in which polymer/oligomer structural data were compiled and cross-correlated with respect to backbone chemistry, helical symmetry, unit cell information, and structure-related physical properties (e.g., polydispersity, molecular weight, melting point, solvent interaction parameter, etc.). This information can be drawn from structural databases, such as the Cambridge Structural Database, the Polymer Database at NIMS (20K+ polymers and co-polymers), along with other web-based data resources on polymers. However, consideration should be given in the analysis to the reliability of the archived data and completeness of coverage of the polymer literature.[10] The optimal target materials would be discrete-length block-like oligomers based on non-peptidic backbones. The blocks would comprise distinct polymer compositions in which the two domains displayed different solvent preferences. This architecture would selectively segregate the crystalline block while the solvatophilicity of the other block promoted the formation of a stable colloid dispersion. *The proposed study would serve to identify best-practice methods based on data analysis to create structurally homogeneous 2D assemblies that would be broadly generalizable to the field of soft materials. The ultimate goal would be to provide a set of structural prototypes that could be developed as functional platforms for devices in the future.*

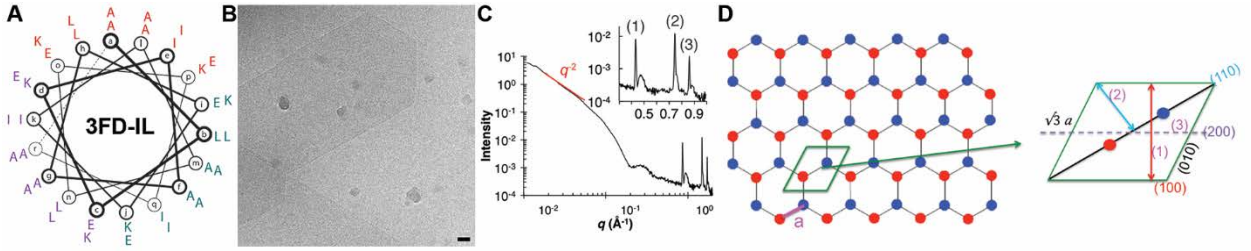

Figure 1. **A**. Helical wheel diagram of the sequence of **3FD-IL** in which the three-fold faces are depicted in different colors. **B**. Cryo-TEM image of a **3FD-IL** nanosheet. **C**. SAXS scattering profile for **3FD-IL** in which the Bragg diffraction peaks associated with the hexagonal lattice are depicted in the inset. **D**. Hexagonal honeycomb lattice model of **3FD-IL** in which the 2D unit cell is depicted with lattice planes.

References Cited:

1. Govindaraju T, Avinash MB. (2012) Two-dimensional nanoarchitectonics: organic and hybrid materials. *Nanoscale.* **4**, 6102-17. (doi: 10.1039/c2nr31167d)
2. Aono M, Ariga K. (2016) Nanoarchitectonics for dynamic functional materials from atomic-/molecular-level manipulation to macroscopic action. *Adv Mater.* **28**, 1251-86. (doi: 10.1002/adma.201502545)
3. Magnotti E, Conticello V. (2016) Two-dimensional peptide and protein assemblies. *Adv Exp Med Biol.* **940**, 29-60. (doi: 10.1007/978-3-319-39196-0_3)
4. Zhang HV, Polzer F, Haider MJ, Tian Y, Villegas JA, Kiick KL, Pochan DJ, Saven JG. (2016) Computationally designed peptides for self-assembly of nanostructured lattices. *Sci Adv.* **2**, e1600307. (doi: 10.1126/sciadv.1600307)
5. Gonen S, DiMaio F, Gonen T, Baker D. (2015) Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science*. **348**, 1365-8. doi: 10.1126/science.aaa9897
6. Keller, A. (1962) Polymer single crystals. *Polymer*. **3**, 393-421. (doi: 10.1016/0032-3861(62)90094-0)
7. Xu J, Ma Y, Hu W, Rehahn M, Reiter G. (2009) Cloning polymer single crystals through self-seeding. *Nat Mater.* **8**, 348-53. (doi: 10.1038/nmat2405)
8. Huang PS, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D. (2014) High thermodynamic stability of parametrically designed helical bundles. *Science*. 346, 481-5. (doi: 10.1126/science.1257481)
9. Magnotti EL, Hughes SA, Dillard RS, Wang S, Hough L, Karumbamkandathil A, Lian T, Wall JS, Zuo X, Wright ER, Conticello V. (2016) Self-assembly of an α-helical peptide into a crystalline two-dimensional nanoporous framework. *J Am Chem Soc.* **138**, 16274-82. (doi: 10.1021/jacs.6b06592)
10. Seebach D, Zass E, Schweizer WB, Thompson AJ, French A, Davis BG, Kyd G, Bruno IJ. (2009) Polymer backbone conformation--a challenging task for database information retrieval. *Angew Chem Int Ed Engl*. **48**, 9596-8. (doi: 10.1002/anie.200904422)