OPEN-SOURCE WORKFLOWS FOR REPRODUCIBLE MOLECULAR SIMULATION

by

Jenny W. Fothergill

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Materials Science and Engineering

Boise State University

August 2022

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Jenny W. Fothergill

Dissertation Title: Open-source Workflows for Reproducible Molecular Simulation

Date of Final Oral Examination: 23 May 2022

The following individuals read and discussed the dissertation submitted by student Jenny W. Fothergill, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Eric Jankowski Ph.D.	Chair, Supervisory Committee
Matthew D. King Ph.D.	Co-Chair, Supervisory Committee
Christopher R. Iacovella Ph.D.	Member, Supervisory Committee
Matthew L. Ferguson Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Eric Jankowski Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

ACKNOWLEDGMENT

I am exceptionally grateful to my advisors, Dr. Eric Jankowski and Dr. Matt King, for their support, patience, and advice. I could not have asked for better advisors; they have motivated me to grow as a person while helping me tailor this project to achieve my goals. My time in Dr. Jankowski's lab has been a joy; he has worked to cultivate a welcoming and collaborative environment that prioritizes teamwork and kindness, and I consider myself very lucky to have been a part of it. I want to sincerely thank my labmates for their encouragement which helped me develop confidence as a molecular scientist, their willingness to take the time to work through a problem together, and their kindness and humor. I also want to thank my collaborators in the open-source software community, in particular the Molecular Simulation and Design Framework (MoSDeF), HOOMD, and Signac framework. Their kindness in welcoming my first contributions has helped develop my passion to promote inclusivity for others, and for that I am very grateful. Last but not least, I want to thank my family and friends for their understanding and support throughout my PhD and my life in general.

ABSTRACT

We apply molecular simulation to predict the equilibrium structure of organic molecular aggregates and how these structures determine material properties, with a focus on software engineering practices for ensuring correctness. Because simulations are implemented in software, there is potential for authentic scientific reproducibility in such work: An entire experimental apparatus (codebase) can be given to another investigator who should be able to use the same processes to find the same answers. Yet in practice, there are many barriers which stand in the way of reproducible molecular simulations that we address through automation, generalization, and software packaging. Collaboration on and application of the Molecular Simulation and Design Framework (MoSDeF) features prominently.

We present structural investigations of organic molecule aggregates and the development of infrastructure and workflows that help manage, initialize, and analyze molecular simulation results through the following scientific applications (1) A screening study wherein we validate self-assembled poly-3-hexylthiophene (P3HT) morphologies show the same state dependency as in prior work, and (2) A multi-university collaborative reproducibility study where we examine modeling choices that give rise to differences between simulation engines. In aggregate, we reinforce the need for pipelines and practices emphasizing transferability, repro-

ducibility, useability, and extensibility in molecular simulation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:

INTRODUCTION

Molecular simulations give us the ability to predict the structures of molecules and materials through models of atomic interactions. Predicting structure is invaluable to engineers because atomic structure determines the mechanical, electrical, chemical, and optical properties of a molecule or material, and therefore how it may be applied to applications in energy generation, sensing, quantum computing, or medicine. Realizing the potential of molecular simulations to inform and transform materials science and engineering is held back by theoretical and practical limitations to the library of simulation techniques. At the smallest length scales, using density functional theory (DFT), it is possible to obtain optimized ground state structures of collections of atoms and predict induced shifts in the electron density in the form of charges. The poor computational scaling of DFT, however, places theoretical limits on the number of atoms whose structure can be predicted with today's common computer architectures. Molecular dynamics (MD) simulations that sample equilibrium structures of thermodynamic ensembles expand the length scales and time scales that are accessible to molecular simulations, with the trade-off of simplifying assumptions about how atoms interact. The $O(N^2)$ scaling of force calculations in MD simulations is better than the $O(N_e^3)$ scaling of some

DFT techniques (doubling the number of atoms N makes the computational effort of advancing one "step" 4x harder in MD, versus a factor 8x harder when doubling the number of electrons N_e in DFT), but this polynomial scaling still places practical limits of a few million simulation elements (atoms, or bodies representing multiple atoms) using modern supercomputers. In addition to these theoretical limits, the practical difficulties of coding simulation software, creating input files, running simulations, and performing analysis are barriers to correctness and ultimately the utility of molecular simulations. In this work we study and improve practical aspects of performing molecular simulations with the aim of making simulations more accessible, reproducible, and ultimately more impactful to society.

The circular scientific process, consisting of observation, hypothesis, experiment, analysis, and conclusion, often neglects to emphasize a vital extra step which is review and reproduction of the results. A Nature study from 2016 found that more than 70% of researchers had tried and failed to reproduce another scientist's experiments—over half had failed to reproduce their own past work [1]. The experience of trying to replicate a seemingly simple method and failing, is common and discouraging. Pushing the boundaries of knowledge with new discoveries is important, but it is equally important to validate that these new discoveries are robust and verifiable.

Computer simulation, in which all parameters can be controlled, should be the most straightforward to reproduce and get identical results. This is in contrast to a wet-laboratory setting where many unforeseen factors can influence the result. For example, in the case of measuring fullerene solubility in water, exposure of the experiment to sunlight or ozone in the atmosphere may affect the result [2]. In a

computer simulation, however, no particle or force exists unless it is specified by the user, so theoretically the experiment should be exactly reproducible. A computational scientist should be able to read a paper in their field, repeat the experiment with the information they gained, and achieve statistically identical results. Unfortunately, although this goal is simple in theory, many barriers stand between computational scientists and reproducibility. These struggles can be broken into two categories: those using established code bases and those using bespoke scripts. When reporting experiments conducted using established codebases, which are stable and distributed open or closed source, the following reproducibility issues can occur:

- Lack of documentation of the version of the software or its dependencies. (e.g. BLAS libraries, compiler, etc.)
- Potential errors when moving data between software (e.g. improperly handling file type or unit conversions)
- Manual editing of files leading to typographical errors
- User unfamiliarity with which simulation parameters should be reported
- In the case of closed-source software, the simulation details may be obscured; users cannot view the source code and must rely on the documentation being accurate

When reporting experiments conducted using custom code, in addition to the above hurdles, the following barriers stand in the way of reproducibility:

- Researchers may not provide the source code, whether due to fears of others

using the code to publish before them or unfamiliarity with software distribution

- If the code is available, it may be non-functioning due to lack of unit-tests and continuous integration
- As user developed code is often under revision, inadequate version reporting or complete absence of version control
- Poor documentation and examples may make the code base unusable to all but its developers
- In the case of a workflow consisting of multiple separate scripts, poor documentation of the process of moving data between scripts or reliance on users manually editing the data may make the logic hard to follow

So how can we help computational scientists to make their work more reproducible? The solution can be divided into three categories: training, tools, and community.

Many scientists are subject matter experts first and learn code development on an as-needed basis, or never have any formal training in software development best practices. By investing in initial training, ultimately researchers can save time by working more efficiently. For example, a new scientist may need to learn bash, git, a programming language, and the specifics of a library or engine that they commonly work with. Many engines may provide their own tutorials [3, 4, 5]. Collating some relevant tutorials within a lab (and perhaps creating some additional training materials) can help to standardize the training process and make sure that

new researchers are not missing material [6]. There are many short workshops and references to help newcomers get started, such as Software Carpentry [7]. Software Carpentry is an open-source, community-driven organization which hosts a collection of lessons and conducts in-person and remote workshops to teach basic lab skills for research computing [8]. Using tools like git can help computational scientists keep track of how each result was produced, like a lab notebook [9]. Automation, or using scripts to avoid any manual data manipulation, will help reduce error and boost efficiency and these scripts should be hosted in open version-controlled repositories [9]. For example, if a public version controlled repository is set up from the beginning, when it comes time to publish, no additional work is required to present the code. Although participation in software-specific training may add to a scientist's upfront workload, the long-term efficiency gain results in saving time.

This training can also help scientists use available tools to increase their productivity while helping to clearly communicate their process to others. Version control software, like git, is a great benefit for tracking changes to a code repository, and many repository hosting services, like GitHub or GitLab, also provide issue tracking and project boards which can be used to manage and organize collaboration and development [10]. To help users understand the purpose behind a code and how it should be run, there are many tools for creating and hosting documentation such as `ReadtheDocs` and `sphinx`. Providing examples and tutorials can help make the code more usable. `Jupyter` notebooks, which contain cells of formatted text and images alongside runnable code, can be a great way for scientists to show and tell the story of their work [11]. In addition to documentation to help users get the

code running, developers can package their code to help users more easily build the software stack necessary to run it. Providing the exact names and versions of software aids reproducibility; however, going a step further to provide containers or virtual machines which contain the exact software stack is even better. This saves users the hassle of installation and prevents opportunities for version mismatching [12, 13]. These tools may add to the cognitive load of a computational scientist, but ultimately make the developed code more useable by the community.

Providing adequate training upfront and introduction to commonly used tools can help new scientists feel more prepared to be part of the open source science community. Training, inclusivity, and collaboration benefit the software development community, which can be demonstrated by the recent influx of contributors to open source molecular simulation codes[14]. This increase in eyes on the code helps to find bugs, discover new use cases, and make the code more usable to people from diverse backgrounds, but it all relies on scientists being willing to share their code. One reason that scientists may not share their code is because they view their work as unfinished or messy and don't have time to prepare it for publication or worry about being judged [15]. Some scientists may worry that by sharing their work openly, their discoveries will be scooped, but using version control can not only help to track their work—as it provides a signed and timestamped commit—but also these commit signatures can be used to prove priority [16]. Participation in the open source software development helps scientists find communities which share the same struggles and can help provide solutions. Often this feeling of community, not prestige, is what drives open-source developers to contribute [17]. By providing newcomers with public contributing guides, codes of conduct, prompt

respectful responses to their questions, and acknowledgement of their contributions [18], we can help to make this community more welcoming and accessible. Collaboration and participation in a community is vital to helping make sure that computational science is reproducible.

Additionally, reproducibility can be aided by not only sharing the source code but also the unedited raw data and analysis methods [19]. By having the complete process—from data collection, to analysis, and even figure creation—be completely automated, transparent, and open, computational results can be the most reproducible and useful to other researchers and society as a whole [20, 21]. These principles help make computational research transparent, reproducible, useable by others, and extensible (TRUE) [22]. TRUE principles don't necessarily guarantee that the results are correct, but instead ensure results are reported in a way that facilitates replication and testing by readers and reviews. Reporting work according to the TRUE principles supports correctness by allowing subsequent works to more easy scrutinize the results obtained using the exact code used to obtain them. When code is provided, many corrections can be found and addressed which ultimately adds to the credibility and promise of molecular simulation [23]. An area where transparency, reproducibility, useability, and extensibility is really brought to the forefront is in my experience developing scientific software with the Molecular Simulation and Design Framework (MoSDeF) team. I have contributed to the development of open-source packages including `mBuild`, `Foyer`, `Signac`, `Fresnel`, and `HOOMD` [24, 25, 26, 27, 28]. These contributions have given me ample opportunity to exercise best software development practices such as writing documentation, using version control, developing unit tests, employing continuous integra-

tion (CI), using a fork & pull workflow, etc. This experience will help me to be more efficient in my future work. I've also gained valuable experience working as part of a team. As my work has been intertwined with existing projects, good collaborative skills were necessary. The experience has been empowering: submitting changes to reputable code bases has helped me to place myself in the computational science community. Helping others to realize this feeling for themselves is part of the reason I so strongly believe in the principles of TRUE science.

Throughout my journey as a graduate student, I have grown as a computational scientist. My research focus shifted part way through, but I would say all of my struggle and efforts have taught me valuable lessons about how to present my work in a way that is most helpful to the next generation of computational scientists. Broadly, my research has used molecular simulation to determine the morphology of the bulk structure from thermodynamic self-assembly of its constituent parts. Learning the intricacies and struggles and what matters to each simulation and how to disseminate and visualize these results. I then have used various analysis methods to validate these simulated morphologies with those observed in experiment. Along the way, I have found my passion in helping to make computational sciences more accessible and reproducible. There is so much for a new molecular simulator to learn: When entering the field, these scientists must train on the underlying theories (e.g., statistical mechanics, thermodynamics, quantum mechanics, etc.) and the specifics of their materials system (e.g., polymers, organic photovoltaics, etc.) all while navigating the computational ecosystem (e.g., git, compilers, paths, the terminal, clusters, python, etc.). With all this cognitive overhead to consider, it makes sense that scientists cobble together what they can

in order to run their experiment; they may not feel they have the time to learn best practices or they may not be aware these best practices even exist! My work has focused on developing and using computational tools and workflows to predict and understand self assembly of electronically active molecules with a focus on making these tools usable by others and demonstrating best practices following TRUE principles. The resulting four journal publications, two in-preparation manuscripts, two posters, and one conference presentation span all scales from first-principles calculation of electronically active dimers to the creation of new tools for analyzing the long-range periodicities of large scale atomistic and coarse-grain (CG) simulations.

In this dissertation I describe projects in which I was both the lead code developer as well as projects where I needed to learn existing code to run experiments. An additional focus of this dissertation is the work that I have done to make the codebases I have contributed to TRUE in the hopes that they will serve others after I leave.

In my first publication on dimer formation, an investigation of the excitonic splitting observed in cyanine dye dimers was done using density functional theory (DFT) [29]. By comparing calculated with experimental absorption spectra it was found that the solvent may play a large role in the dimer formation and peak shifting. I performed all simulations under the guidance of Dr. Li and Dr. Yurke, created all figures, and wrote all stages of draft this paper with review feedback from Dr. Li, Dr. Knowlton, and Dr. Yurke.

In my second publication, an investigation of the supramolecular interactions

which guide crystal self-assembly in diphenylurea transition metal complexes was done using DFT [30]. I contributed DFT energies and optimized structures under the guidance of Dr King.

My third publication was a thorough investigation of the effect of para-substituents to the crystal structure of diphenylurea molecules. I contributed DFT calculations, data analysis, writing in the initial draft, and images made in Figures 3, 4, 5, 6, and 8 [31].

My contribution to the lab's perspective paper was mostly about my on-boarding experience trying to reproduce an existing model; a task which should be very straightforward but, which any computational scientist can attest, is far from simple [14].

These published works are included in the appendices for readers with interest in these topics. The two core chapters here focus on the development of TRUE workflows, highlighting software development efforts.

Chapter 2 details validation of an updated workflow to explore the state-space dependence of the ordering of poly-3-hexylthiophene (P3HT) polymer. This workflow was highlighted in a talk given at the 2021 American Institute of Chemical Engineers conference. One of the tools used in the workflow, GIXStapose, was also presented in my first poster contribution as Boise State University's 2020 Research Computing days and at the 2020 SciPy conference [32, 33]. The original workflow being validated was transparent in that it freely distributed all it's source material, from raw data to all scripts used for processing. It was also reproducible in that the software stack was documented and if necessary it could be recreated and re-

run. I added improvements to this workflow by making it easier to install by using and distributing a Docker container with the full software stack, more usable by others by creating modular packages with full documentation and tutorials, and more extensible by generalizing the workflow to accept any molecule as input. This chapter demonstrates what a TRUE workflow can look like, and how tools can be designed according to TRUE principles.

Chapter 3 summarizes my effort in a multi-university collaborative study to use MoSDeF tools reproducibly which is in preparation. This study follows in the footsteps of other studies which compare the result of analogous methods across different engines in hopes of validating method implementation[34, 35, 36]. This study demonstrates the hurdles to achieve reproducible results between different people using different engines, and the power of collaboration and community to tackle large problems. My contribution to this study include creation of all scripts needed to run MD simulations using HOOMD including implementation of a new method for tail correction and fixing all bugs found along the way.

Through my work spanning all scales of molecular simulation—from implementing details in MD engines, to creating analysis software, to performing DFT calculations of single molecules, simulations of collections of molecules, and performing thousands of simulations of collections of molecules across thermodynamic state space—the need and utility of pipelines and practices emphasizing transferability, reproducibility, useability, and extensibility has only been reinforced. The result of my work has been the validation of the general amber forcefield (GAFF) and an updated workflow for use with P3HT, the creation of extensible tools for mapping between coarse and atomistic representations, the development of tools

for reproducibly analyzing structure and diffraction patterns, and the creation and comparison of methods within the HOOMD engine.

CHAPTER 2:

VALIDATING STATE SPACE SCREENING CALCULATIONS, A P3HT CASE STUDY

The following chapter contains yet to be published work written by me with guidance from Dr. Jankowski.

A Jupyter notebook demonstrating the tools mentioned in this chapter can be found in Appendix A, and a Zenodo dataset containing all data used in this chapter can be found at Ref. Fothergill [37].

2.1 Introduction

An important facet of applying molecular simulations to solve real-world engineering problems is ensuring that the simulations are correct. While there exist general guidelines for improving simulation correctness (e.g., the TRUE principles [22]), specific instances will vary widely between disciplines because of individual workflows that govern scientific simulation pipelines. In this work we detail the application of TRUE principles for a common problem that arises in molecular simulations: performing molecular simulations across a set of thermodynamic state points to screen these conditions for structures of interest. Specifically, we

consider a case study of validating structural predictions of the organic photovoltaic polymer P3HT.

Organic photovoltaics (OPVs) are a focus of this research because they represent the best opportunity for cost-effective solar power. The theoretical efficiency limit (Shockley-Queisser limit) for a p-n single junction solar cell is about 30% [38]. Multijunction cells may achieve efficiencies that surpass this limit, but they often do not have lower energy payback times as their production is more costly. Figure 2.1 shows the efficiency gains made by single junction silicon and organic photovoltaic devices over the last 30 years. These are merely a selection among the many categories of PV devices. The trend of the silicon cells shows that the efficiency gain is leveling off as the devices approach the Shockley-Queisser limit. OPV devices, however, still have a lot of room for improvement, and recent increases in efficiency reflect this. The potential of OPVs to achieve higher power conversion efficiencies depends on the morphology of the active layer. Molecular simulation can help to predict the combinations of donor and acceptor which will robustly self-assemble into a morphology best able to transport charge.

A complication of the simulation of OPV polymers is that the properties which predict a device's efficiency, for example charge transport, span multiple length scales. In order to model charge movement through a device, we need to know the position of individual atoms in order to discern the electronic environment and thus the likelihood of a charge hop, but we also need a bulk structure large enough that we can observe morphological features like the interdigitation of polymer lamellae to compare these morphologies to experiment. But atomic resolution at length scales of hundreds of nanometers becomes computationally expensive, so simulat-

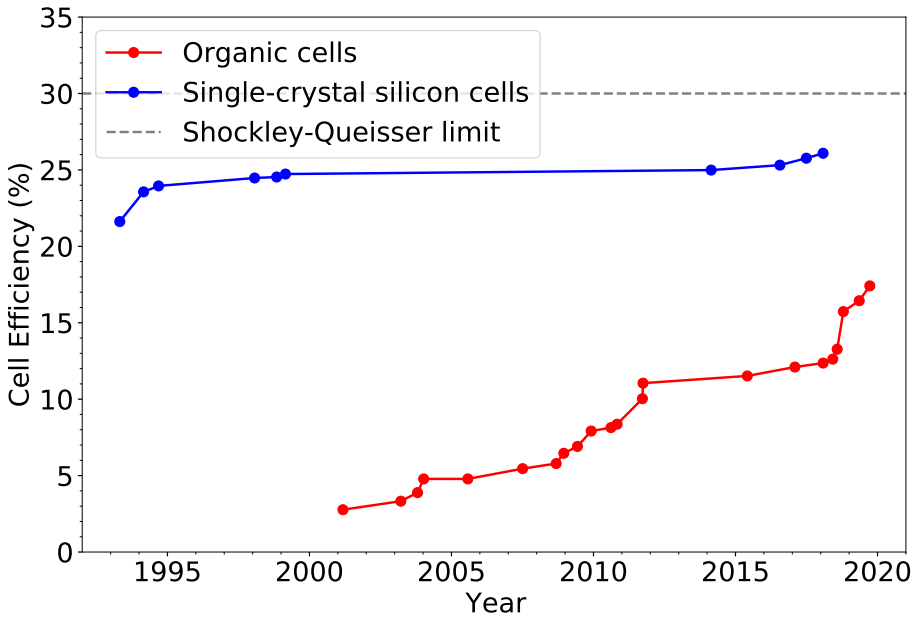


Figure 2.1: Comparison of efficiencies in organic and silicon photovoltaic technologies from 1990 to present. The current state-of-the-art OPV, a mixed polymer, small molecule, and fullerene device, achieves 18.2% efficiency [39]. Data taken from Nationational Renewable Energy Laboratory. U.S. Department of Energy [40].

ing larger morphologies necessitates a simplified model. By simplifying our model using coarse-graining techniques where multiple atoms are represented as a single bead we can more efficiently equilibrate larger length scales. In addition to using a simplified model, the statepoint variables (including temperature, density, and solvent) must be carefully tuned to find the conditions under which the OPV morphology has the best self-assembly. Choosing a model and statepoint are some of the important choices a simulator must make when simulating OPV polymers. These choices should be informed by data, which increases the scope of this problem. This chapter will discuss a previous work which laid the groundwork for making these choices using poly-3-hexylthiophene (P3HT), and the current work

which aims to reproduce this work with improvements to the underlying tools which make them more transparent, reproducible, usable, and extensible (TRUE) [22].

2.2 Background

In previous work from our lab, Ref. Miller et al. [41], hundreds of molecular dynamics simulations of P3HT polymer are performed across different temperatures, densities, and solvent qualities in order to ascertain at which statepoints the morphology would self-assemble into the most ordered structure. These hundreds of simulations at different statepoints were organized using custom python scripts which relied on creating and navigating a directory structure. Figure 2.2a helps

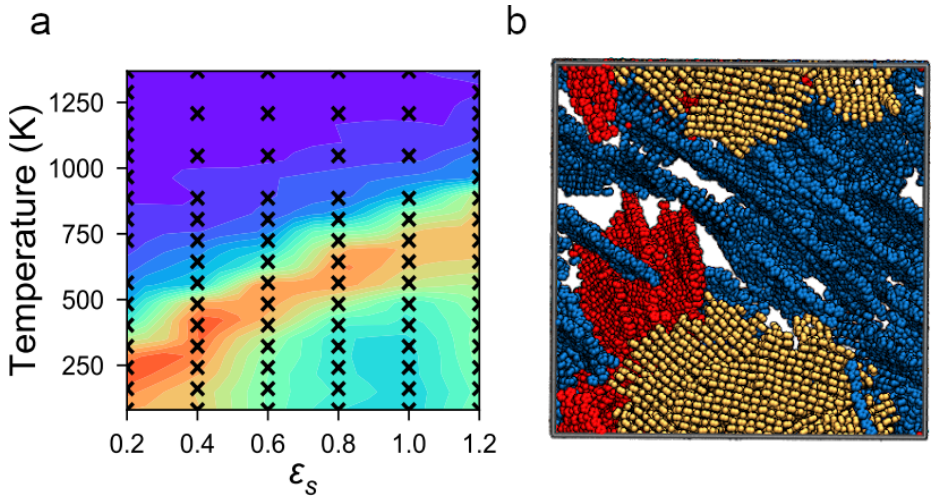


Figure 2.2: (a) The degree of ordering, Ψ, at various temperatures and solvent parameters taken from Ref. Miller et al. [41]. (b) The three largest clusters (colored blue, red, and yellow in order of descending size) in a P3HT morphology taken from Ref. Miller et al. [41].

us to visualize the temperature and solvent parameter state space. In order to aid the analysis of this state space, the order parameter (Ψ) was defined as the ratio

of thiophene moeities in "large" clusters. The clustering criteria takes into account the distance between thiophene centers and the angle between the planes of the thiophenes. In Miller et al. [41] the distance cutoff was 6 Å, the angle cutoff was 20 °, and a "large" cluster was defined as having 6 or more thiophenes. Figure 2.2b is a visualization of these clusters in a morphology. The order parameter analysis required selecting specific atoms from the trajectory file based on their index. The atom indices in this analysis were hard-coded for P3HT. All the input files were made programmatically, so having hard-coded index values worked; however, if this analysis was to be used on polydisperse polymer lengths or a different molecule, it would have to be changed. Simulated GIXS diffraction was used to compare the most highly ordered morphologies to experiment and they were found to show good agreement. Miller et al. [41] accomplished what they set out to do (validate a simplified model of P3HT) and performed work which was impressive in scope (hundreds of simulations!) and all the code is freely available to all, but how can we redesign these tools to be easier to use for the next molecule and the next user?

To reproduce the work of Miller et al. [41], this workflow must be designed to handle a broad computational scope. It must be able to create and manage a large parameter space, move data between multiple different pieces of code, translate units, and perform analysis on selected particles from simulation trajectories. Much of my work in the lab has been to design tools which makes these tasks more transferable, reproducible, usable, and extensible (TRUE) [22]. Transferable means the code is general enough to be used in different ways or applied to different problems. Reproducible means enough information is provided that the work can

be redone and no important details regarding the process are obscured. Useable by others means that other people—especially people outside your circle—can actually use this tool. This entails that not only is the source code freely available and easy to install but also that users can understand how to use it without too much cognitive overhead. This can be helped with permissive, open licensing, dependency documentation and packaging (including containers), and thorough documentation and examples. Extensible means that others can build upon the work you have done. If this project is going to have the desired impact (i.e., using molecular simulation to identify novel OPVs), every step of the process and the resulting data must be straightforward to reproduce and validate. The nature of a multiscale simulation requires translation between formats and units and accurate handling of large data, and thus requires infrastructure to reduce error. Although the molecular and computational scope of this project is broad, by using and building upon existing architectures and adhering to recent guidelines in the computational sciences community we can manage the multiscale nature of this project and contribute to reproducible science. This chapter will demonstrate that we can reproduce and validate our P3HT morphologies while using MoSDeF tools to help our methods to be more TRUE.

2.3 Statement of Need

In order to sample the parameter space, we need to be able to spin up a large number of simulations (i.e., initializing our simulation volume, implementing the model, and running the simulation) and accurately access each simulation output to perform analysis. The order parameter analysis involves selecting the atoms that are part of a chemical moiety and calculating the distance and angle between each

potential neighbor moiety. If the potential neighbor meets the clustering criteria, it is added to the cluster. The order parameter can then be calculated as the ratio of the number of the moiety in large clusters vs the total number in the morphology. And the GIXS analysis involves converting the simulation trajectory to the format required by the diffractometer package including unit conversion.

Before we address challenges in the implementation, it is important to note how we solved issues related to installing the software stack. Often this step goes unmentioned, but installation of software packages and managing dependencies can be a huge hurdle for computational scientists. To prevent this, we have built pipelines for all our code repositories (using GitHub Actions) which automatically build docker containers from each tagged version of the repository and the latest master branch. These containers can be used with the docker or singularity applications and contain the exact code state present in the repository along with static versions of its dependencies. This allows our software to be used without ever having to manually install it or manage its dependencies. Not only does this make our software stack easier to use, but it makes it more portable—we can have the exact same environment on our school cluster or XSEDE. And if someone wants to reproduce our work, they can access the exact container we used.

The first challenge, managing a large dataspace, is handled using the `signac` framework. Although designing code to work within a framework like `signac` does add some cognitive load, after this initial lift the project becomes more extensible and flexible regrading changes in the parameter space. We will cover the tool designed to manage and submit OPV MD simulations within this framework, `PlanckTon`, later in the chapter.

Although the application of the united-atom model to our P3HT system was not overly complicated as it only contained 3 types, it required manual atom typing, which presented another hurdle to applying this method to a new and perhaps more complicated compound. By using the `foyer` forcefield dissemination and atom-type engine [25] our atom types can be automatically assigned based on the chemical connectivity. This allows us to more easily extend to new compounds.

Calculating the order parameter originally depended on a workflow that was hard-coded for P3HT and selected specific atom indices—this depended on the particle order in the simulation being the same every time, which it was because routine initialization methods were used, but it did not allow for the method to be applied to other compounds.

2.4 Tools

Next we will discuss the various tools used in this project. Figure 2.3 gives an overview of how these tools work together to perform the structural analysis. This workflow was designed such that the output from each code can be used as the input for the next with little to no modification. Many of these code bases are collaborative works, so I will give credit to the contributors and creators of the underlying building blocks.

2.4.1 PlanckTon

In order to initialize simulations in a reproducible way, we have designed `PlanckTon`: a wrapper for the initialization, management, and submission of organic photovoltaic molecular dynamics simulations using the HOOMD-Blue engine [42, 14, 43]. The `PlanckTon` software includes some OPV compounds com-

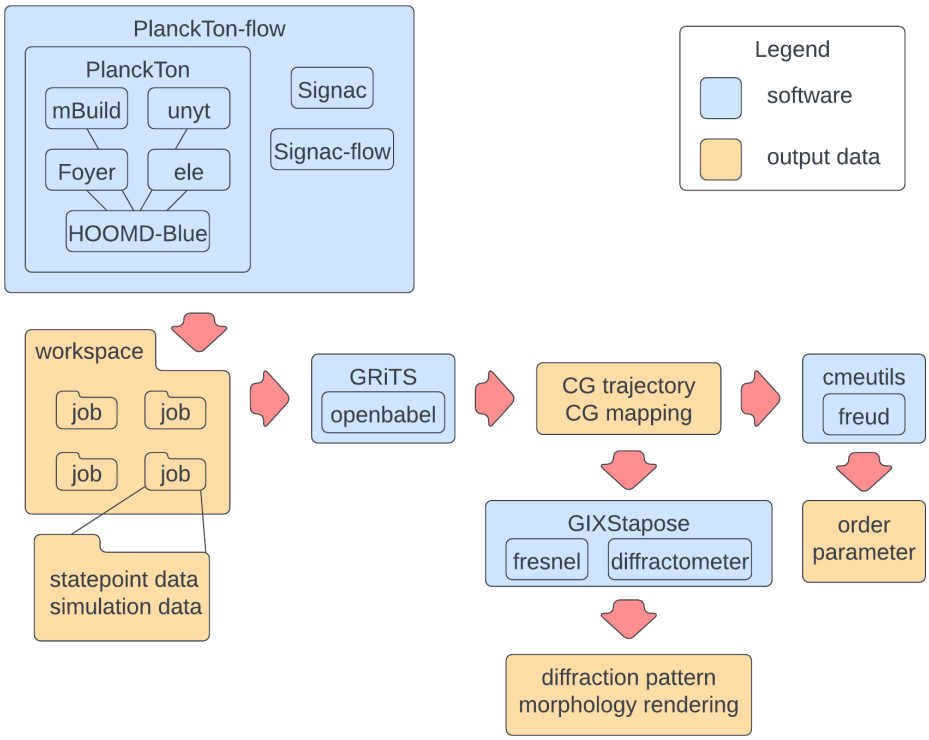


Figure 2.3: Overview of the structural analysis workflow. First, the workspace is created in the `PlanckTon-flow` framework and all simulations are run. Then GRiTS is used on the output simulation data to find the thiophene centers. Finally, the order parameter function in cmeutils is used to calculate the order parameter of each simulation trajectory and GIXStapose is used to perform GIXS analysis.

monly used in our lab to guarantee that all simulations are initialized with the exact same starting compound; however, the workflow is designed to easily accept any MoSDeF compatible chemical input file format including SMILES strings. The code base also ships with custom XMLs for OPLS-UA and GAFF and a separately packaged GAFF XML [44], but is compatible with any `foyer` XML. The initialization procedure of `PlanckTon` is as follows: First the user selects a compound or mixture of compounds, the number of said compounds, the temperature, the density, and the solvent parameter. As in Ref. Miller et al. [41], the solvent

in `PlanckTon` is modelled implicitly, so the solvent quality, ϵ_s, refers to a scaling factor on the non bonded forces. In order to more robustly achieve high density morphologies, the simulation volume is initialized at a lower density and then shrunk down to the desired density at high temperature[45, 46, 47, 48, 41]. The creation of our simulation volume, the forcefield atom-type assignment, and the creation of the `HOOMD` force objects is completely handled by `mBuild` and `foyer` in the MoSDeF framework[24, 25]. By using the modular, general system initialization of the `mBuild` and `foyer`, we get two benefits: (1) we draw on the knowledge base of the MoSDeF community–more users means we can find and resolve errors more quickly, and (2) we can more easily incorporate the system initialization in our other projects. Once the system and forces are initialized, `PlanckTon` uses the NVT ensemble via the Nosé-Hoover thermostat [49, 50] as implemented in the HOOMD-Blue molecular dynamics engine. The initial velocities and angular momenta are randomly assigned from the Maxwell-Boltzmann distribution. By having our entire process scripted, we can have everything about our process documented and available for scrutiny and reduce the potential human error inherent when switching between codes or transferring files. Through `PlanckTon-flow`, this tool is supported by the `signac` framework, which handles the parameter space initialization and management and submission to various computing clusters. Because this workspace is created using `signac`, `signac` can be used to navigate it. This allows to reference our data via the statepoint (e.g., what temperature and ϵ_s it was run at) without ever having to create or manage a complicated directory structure or naming scheme.

`PlanckTon` was started by my labmates Mike Henry, Evan Miller, and Matty

Jones[14, 42], and I have assumed responsibility for its development including: Updating to `HOOMD` version 2 allowed us to use the `HOOMD` simulation initialization functions in `mBuild`. This allows the package to be more modular and useable by others; any improvements or bugfixes needed in the `create_hoomd_simulation` function can be contributed back to `mBuild`, which allows the whole community to benefit. Adding automated docker container builds. Generalizing the initialization procedure to allow starting compounds to be loaded from SMILES strings or any input file and these untyped compounds can be typed on the fly using any `foyer` compatible forcefield. Adding support for all `HOOMD` neighborlists, which allows better sampling of sparse systems. And adding support for user specified temperature ramps, allowing users to perform temperature annealing, a technique used in OPV active layer synthesis. In order to be more clear about the units used in `PlanckTon`, the unyt package is used to handle all unit conversions. This package allows for values to be tagged with their unit and then all conversions can be handled by the package.

`PlanckTon` is unique, but there exist other tools to manage molecular dynamics simulations. There is a web-based application that facilitates MD simulation using cloud computing services and automates related tasks [51] and a command line tool that automates many common MD tasks [52], but I have not seen a pure python implementation which can also manage a large dataspace.

2.4.2 GRiTS

I developed GRiTS as a tool to assist in applying a coarse-grain mapping and to backmap a coarse-grain system to a fine one. Although the fine-graining capabilities are still under development, the coarse-graining is robust enough to apply

a mapping to an entire system trajectory. This mapping can also be used to pick out specific chemical moieties from a morphology. GRiTS works using SMARTS matching—implemented in OpenBabel—to map atomic indexes to coarse grain beads. Once all the atomic indices are assigned to a bead, then bonds are inferred between coarse-grain beads based on the atomistic bonding scheme. For example, if an atom in bead A is bonded to an atom in bead B, beads A and B are assumed to be bonded. When looking at an entire simulation trajectory, the SMARTS matching algorithm can be prohibitively expensive. So GRiTS uses the following simplifying assumptions: Chemical bonds are not formed or broken over the course of the simulation—this allows only the first frame of the trajectory to be used for mapping. Using only the first frame of the trajectory is also useful because this way the molecules in the system can be guaranteed to be chemically reasonable—i.e., the bonds and angles are not distorted, aromatic systems should be planar, etc. It is also assumed that if molecules have the same number of atoms, they have the same chemical structure. This allows us to use the clustering algorithms of implemented in the `freud` analysis library to break the snapshot into molecules and only perform SMARTS matching on each molecule type, then extrapolate the mapping out to other molecules of the same type.

When working with a UA system, GRiTS can also infer hydrogen positions and bonding. This allows for SMARTS strings to find matches in UA systems. The hydrogen inferring method uses OpenBabel and requires that the first frame of the trajectory be chemically reasonable.

There are other tools which handle creating coarse-grain mappings such as VOTCA [53]. This tool will crate a coarse grain trajectory from an atomistic one

given an XML file mapping which requires the user to specify everything from the atom types involved in the bead to the bond and angle information. VOTCA also has many other functions that GRiTS does not such as coarse-graining (including potential development), charge-transport, and excitation transport. VOTCA requires the user to edit an XML file in which formatting and whitespace have meaning, and define not only the mapping but also the bonding and angles for the coarse grain mapping, which could easily be inferred from the mapping scheme. Creating a mapping with GRiTS is more intuitive because the only parameter needed is a SMILES string, there are no potentially difficult to parse files for the user to edit, and all bonding and angles within the coarse grain structure are inferred.

2.4.3 GIXStapose

GIXStapose is a package which allows users to connect the real space view of a chemical structure to its simulated grazing incidence X-ray scattering (GIXS) pattern. (The name "GIXStapose" is a portmanteau of "GIXS" and "juxtapose.") I developed GIXStapose as wrapper for the diffractometer package and the fresnel ray tracing library [54, 47, 27] GIXStapose works by translating the structural information in many chemical input files into the formats needed by these two packages, and in the process can link rotations of the real space image to its diffraction pattern. The camera information used to generate the real space rendering and the diffraction pattern can be accessed and saved, allowing users to recreate the exact figure. The package has an interactive graphical user interface (GUI), but is completely modular and can be run in a scripted fashion to generate reproducible figures and diffraction patterns along with translation of units. This work was

presented at Scipy2020 [33].

2.5 Methods

2.5.1 Generate simulation data

A `PlanckTon-flow` workspace of 40 simulations defined by their temperature and solvent parameter (ϵ_s) was initialized. Each system comprised of 100 united-atom (UA) P3HT 16-mers parameterized with GAFF. All nonbonded forces used a cutoff value of 8.91 Å (2.5 times the largest sigma value). The target density for all was 0.56 $\frac{g}{cm^3}$. All simulations used a timestep of 2 fs and started by shrinking the box from 125 times the target volume to the target volume (789 nm^3) at 629 K, with a thermostat coupling value (tau) of 2 ps, run over 0.2 ns. Once the shrink step was finished, each simulation was run at its own combination of temperature and ϵ_s ranging from 125, 188 to 629 K and 0.2 to 1.0, respectively. These runs used a thermostat coupling value (tau) of 0.6 ps and were run for 200 ns. All were run in the `cmelab/planckton_gpu:v0.6.1` docker container.

2.5.2 Create coarse grain mapping

The GRiTS `CG_System` class was used to find indices of atoms in thiophenes using SMARTS string "c1cscc1" and inferred hydrogens. This created a new coarse-grain trajectory file with beads at thiophene centers and a json file which contains the SMARTS string which generates the mapping and the atom indices in the atomistic trajectory as a key value pair.

2.5.3 Calculate order parameter

The `order_parameter` function in cmeutils was designed to use the coarse grain trajectory and the mapping to calculate the average order parameter of the last ten

frames of the trajectory. An angle cutoff 10 ° and a distance cutoff of 6 Å was used as the clustering criteria.

2.6 Results and Discussion

With the updates made to this workflow, first let's check that we get quantitatively similar results.

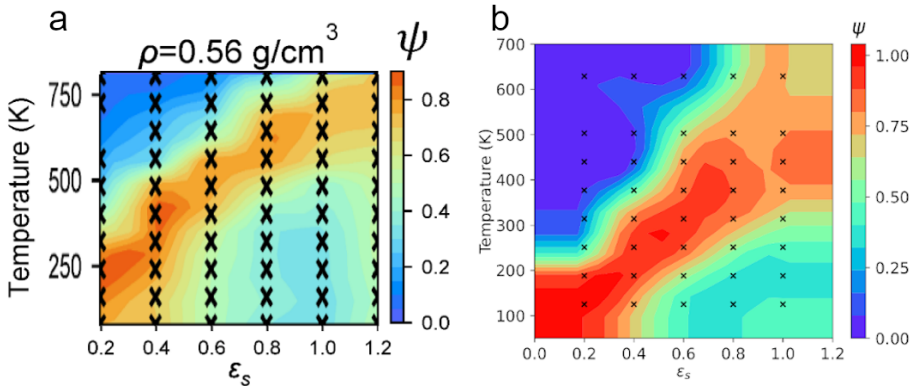


Figure 2.4: The degree of ordering, ψ, of P3HT at various solvent qualities, ϵ_s, and temperatures. Red regions denote high order, blue regions are disordered. Each black "x" indicates a measurement, values between are interpolated. Part a is taken from Miller et al. [41, Figure 3a] and has been cropped and scaled to focus on the region of interest and make the plot bounds closer in value. Part b was created using the order parameter value from each simulation trajectory and the RectBivariateSpline interpolation function from the Scipy library. The final z values were adjusted such that no order parameter was greater than 1 or less than 0.

Figure 2.4 shows good qualitative agreement between the previous findings of Miller et al. [41] and the order parameter trends found in this work. The absolute value of the order parameters seen in this work are, in general, higher than those observed by Miller et al. [41]. This may be due to the initial high temperature shrink period used in this work which is similar to Protocol 2 in Miller et al. [41] which was shown to achieve faster equilibration and more robust ordering.

Next the high order regions were further examined using GIXS analysis (see Figure 2.5).

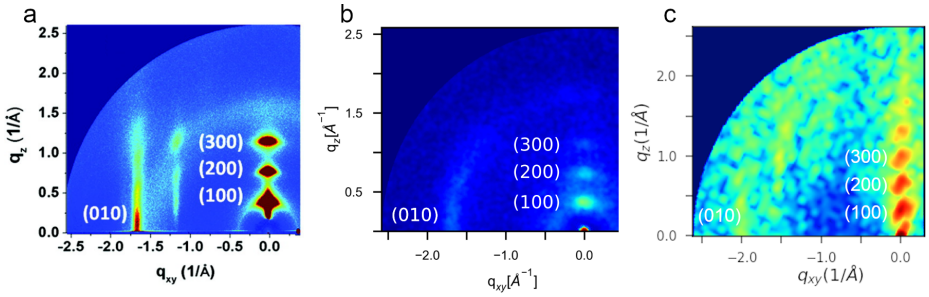


Figure 2.5: Diffraction patterns of P3HT from (a) experimental GIXS of neat P3HT [55], (b) simulated GIXS [41], and (c) this work. Part c was generated in GIXStapose using thiophene centers found using GRiTS from the simulation trajectory ran at ϵ_s 0.4 and T 251K.

Figure 2.5 shows good quantitative agreement between this work, experiment, and previous work. This also shows that high order parameter is a good predictor of clear GIXS peaks. The GIXS data in Figure 2.5c is noisier because was generated using a very small system (1,600 points vs 15,000 points in part b). However, even in spite of this, it is clear that qualitatively the same structural features seen in previous works and experiment are present. And we can check that the real space distances these peaks correspond to are reasonable using the following equation

$$d_{real} = \frac{2\pi}{d_{peak}} \tag{2.1}$$

where d_{peak} is the distance of the bright spot from the origin. The (100) peak corresponds to a periodic distance of 16.4 Å and and the (010) peak to a distance of 3.50 Å; Duong et al. [56] report a lamellar spacing of 16.5 Å and a pi-stacking spacing of 3.83 Å for neat P3HT. There is room for users to choose multiple points within

a large, smeary peak such as shown in Figure 2.5c, so these values are reasonable but could vary depending on the user's choice.

Using all particles in the united-atom thiophene may result in easier to see spots in the diffraction pattern (see Figure 2.6).

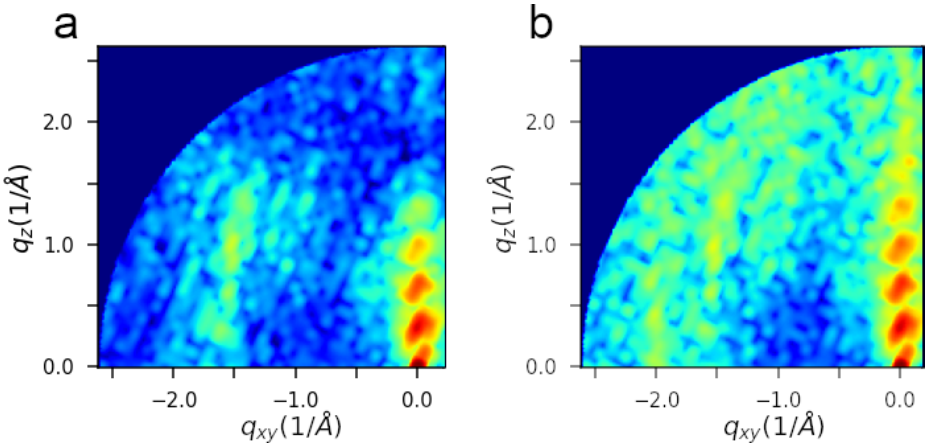


Figure 2.6: Diffraction patterns of P3HT using (a) all thiophene particles in the UA trajectory (b) thiophene centers found using GRiTS. Both patterns were generated using GIXStapose at the same camera angle on the simulation trajectory run at ϵ_s 0.4 and T 251K.

Although coarse-graining may provide clarity with much larger or less ordered structures, in this very small ordered system, the peaks are more easily visible when all the thiophene particles are used in diffraction.

As Figure 2.5c was generated using GIXStapose, we can examine the real space structure from the exact viewing angle the diffraction pattern was generated from.

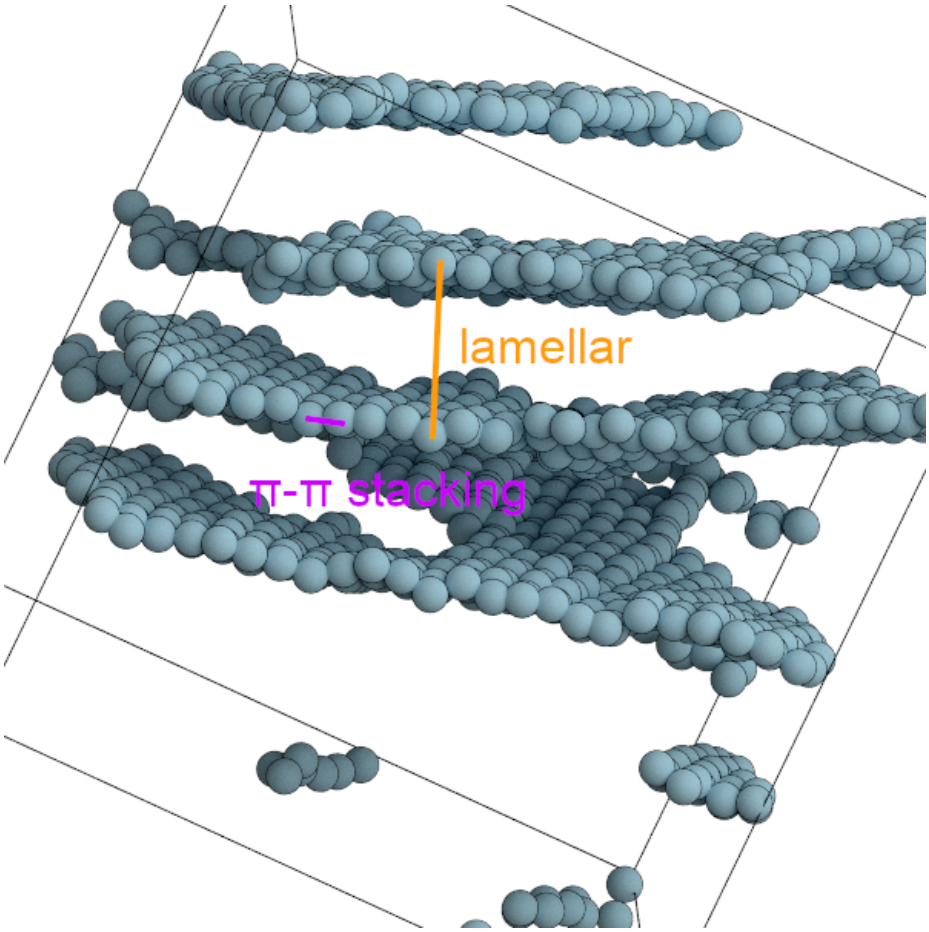


Figure 2.7: Real space structure of thiophene centers found using GRiTS from the simulation trajectory run at ϵ_s 0.4 and T 251K. This figure is generated using fresnel via GIXStapose with the same "camera angle" as Figure 2.5c.

Figure 2.7 shows the thiophene centers at a high-order statepoint. We can clearly see the lamellar spacing in the vertical direction, and less clearly the π-π stacking in the horizontal direction. By viewing the real space structure from the exact angle as the diffraction pattern, it is easier to correlate periodic features with their diffraction peaks.

To confirm that the thiophene center beads are being placed correctly, we can view

an overlay of the thiophene centers with a frame of the unaltered UA simulation trajectory.

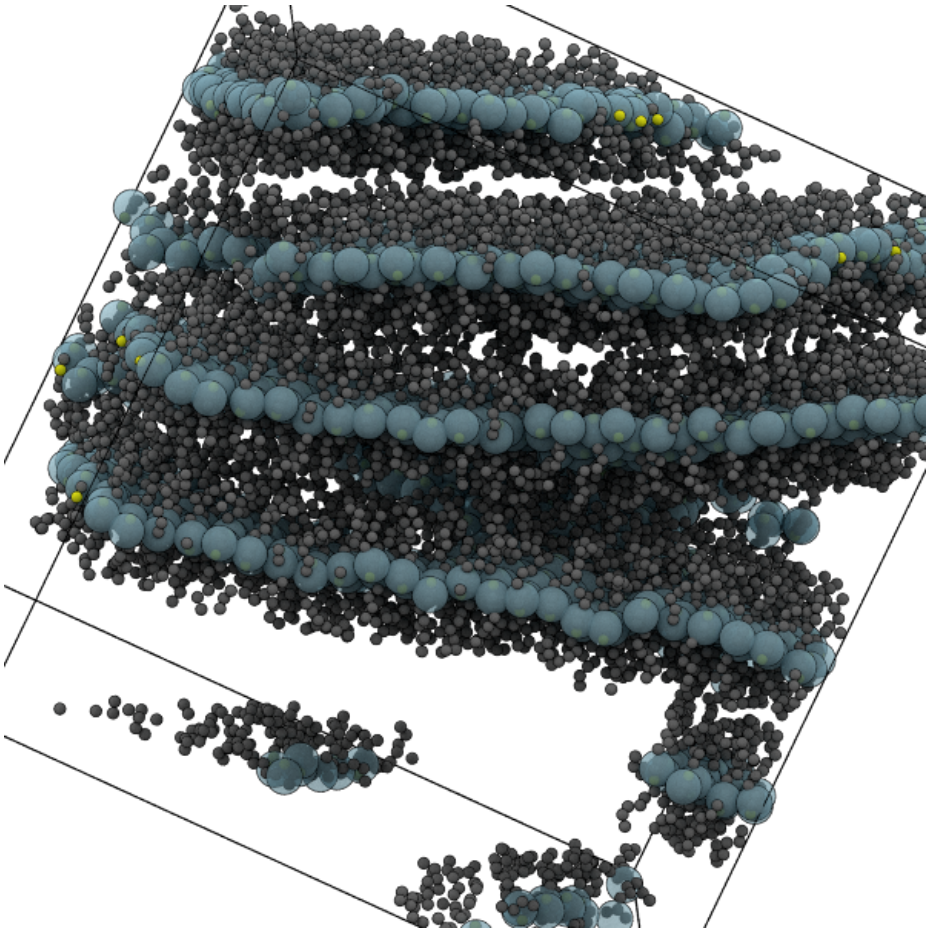


Figure 2.8: Overlay of thiophene bead centers (translucent blue) found using GRiTS with united atom carbon (grey) and sulfur (yellow) in P3HT simulation trajectory run at ϵ_s 0.4 and T 251K. This figure is generated using fresnel via GIXStapose with the same "camera angle" as Figure 2.5c and Figure 2.7

Figure 2.8 shows that the thiophene beads are correctly capturing the geometric centers of the thiophene moieties in the trajectory given that the sulfur atoms are found almost entirely within the thiophene beads. This view also gives us a better perspective on what forces may be causing the lamellar separation—namely, side-

chain interactions.

We can also show correlation between low order parameter and disordered systems with no GIXS peaks (see Figure 2.9 and Figure 2.10).

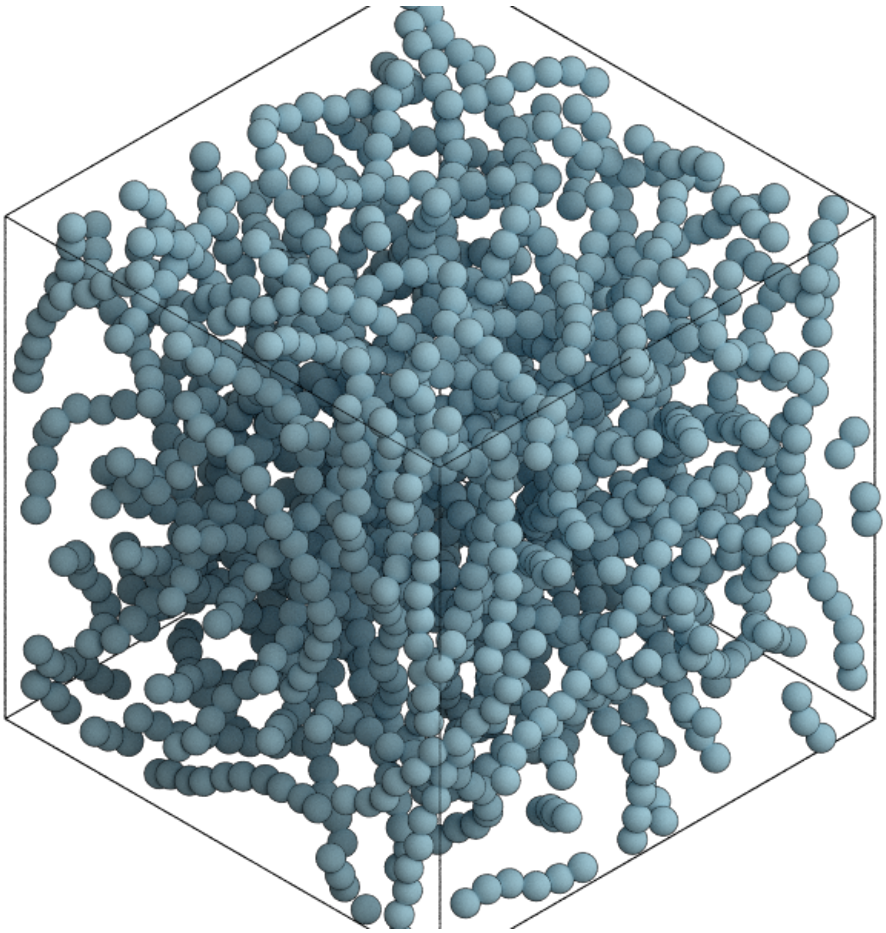


Figure 2.9: Real space structure of thiophene centers found using GRiTS from the simulation trajectory run at ϵ_s 0.2 and T 629K. This figure is generated using fresnel via GIXStapose with the same "camera angle" as Figure 2.10c

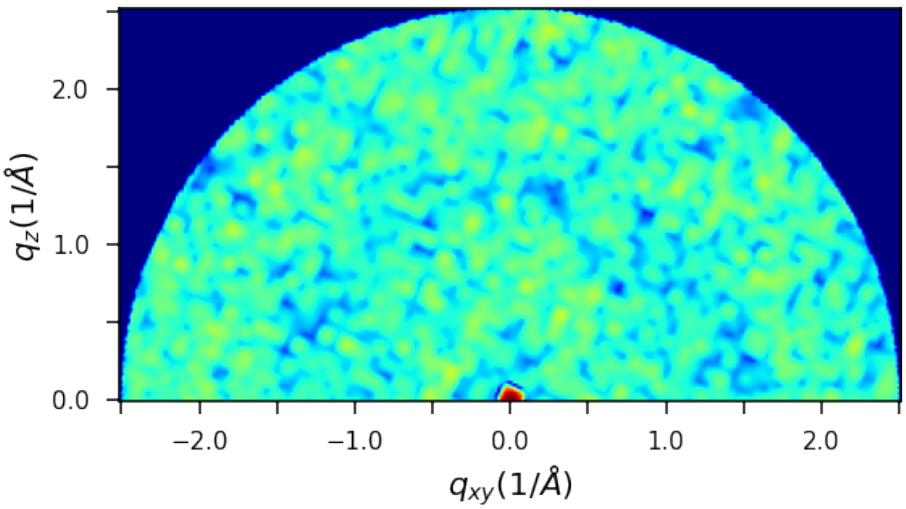


Figure 2.10: GIXS pattern generated in GIXStapose using thiophene centers found using GRiTS from the simulation trajectory run at ϵ_s 0.2 and T 629K.

Figure 2.9 shows the thiophene centers in a basically random configuration and Figure 2.10 shows a general lack of peaks.

So, we have demonstrated that our ordering trends observed in P3HT are robust to changes in the forcefield (OPLS-UA to GAFF-UA). And that the changes to simulation initialization (more incorporation of MoSDeF tools) and analysis (using GRiTS to detect thiophenes) have not skewed our results. Next let's discuss the potential effect of our TRUE changes.

2.7 Conclusions

By using `mBuild` with `foyer` in `PlanckTon` to initialize our simulations, we can more easily use different input file formats including smiles strings and any `foyer` forcefield. Using GRiTS to select the desired part of the molecule removes any need for manual indexing, which allows us to more easily extend the order parameter

calculation to other planar, conjugated molecules like perylene or ITIC. Working within the `signac` framework allows us to quickly and easily sample the necessary parameter space without needing to manually create or manage directories. We've also implemented semantic versioning in all our lab's codebases with version tagged docker containers which helps users to keep track of when changes are made and ensure we use the same code state. By using and building on existing code, we work with a community of open-source molecular simulators.

The ultimate success of this project to me is if someone else can use it. And in order to achieve that goal, I have tried to develop code with the next user in mind. I know it can be daunting to be handed someone else's code. If the only way to learn how to use it is to read through the code comments (or just the code itself) and try to figure it out, users may prefer to simply write their own. Even when good user documentation is provided, often these bespoke codebases have specific dependencies which may not be continuously maintained so they break in updated installations. We can consider the effort put in to providing containers with the complete functioning software stack and writing thoughtful user documentation and examples as a force multiplier. If the goal of a scientific codebase is to accomplish work, then the more people who can use this code, the better. Also writing our code to be more modular and general can help it to apply to more situations. For example the GRiTS code was designed to create coarse-grain mappings, but the mappings can also be used to find instances of a SMILES pattern with a trajectory as was done in this study.

It is hard to measure the success of that goal, but I submit the following examples: In the summer of 2021, our lab hosted two students from the NSF Research

Experience for Undergrads (REU) program and a high school teacher from the Research Experience for Teachers (RET) program. None of these students had any prior knowledge of shell scripting, python, or molecular simulation. After a week or so of introductory training, the three researchers were introduced to `PlanckTon`. Each researcher chose a molecule, and together with three undergrad researchers in the lab, were able to run 24K SUs worth of MD simulations on XSEDE using `PlanckTon`. A new PhD student in our lab with a strong background in software engineering but new to molecular simulation and chemistry was able to use GRiTS to create a coarse grain mapping in a matter of days. She did this without any one-on-one training beyond the documentation and examples.

By developing code with TRUE principles, other scientists can more easily use and extend this project. Hopefully this software ecosystem will continue to be dynamic and evolving. Big problems like solving the climate crisis with photovoltaic technology require robust solutions and writing usable code allows us to draw on the strength of the community.

CHAPTER 3:

REPRODUCIBILITY STUDY

The following chapter describes a yet to be published collaboration with manuscript in preparation. My contribution to this project was preparing the scripts to run the HOOMD simulations, writing documentation for the project, and involvement in numerous group discussions. This study brought to light many many issues which instigated many updates and improvements to the various codes used in the study. Scientific software improvements I am responsible for as a result of this study include:

1. Allow HOOMD functions to add to an existing snapshot to facilitate rigid bodies. (mBuild pull request #808)

2. Ported the rigid body constraint class from the HOOMD v2 to v3 API. (HOOMD pull request #888)

3. Allow `mBuild`'s xyz writer to write coordinates with greater precision, so MC-CCS could use it for this study. (mBuild pull request #948)

4. Added the required neighborlist buffer argument to the `mBuild` `create_hoomd_forcefield` function to address breaking change in the

HOOMD neighborlist API. (mBuild pull request #988)

5. Fixed a bug in the conversion of Ryckaert-Belleman (RB) to OPLS-style dihedrals. (mBuild pull request #996)

6. Corrected Coulomb's constant used in the charge conversion for HOOMD writers. (mBuild pull request #1011)

7. Add optional calculation of energy and pressure tail correction to the Lennard-Jones pair force. (HOOMD pull request #1138)

3.1 Introduction

Reproducibility in science means that we are transparent about our methods in a way that allows others to understand what was done. If the goal of science is to expand the knowledge of humankind, making sure experiments are reproducible helps to lay a strong foundation for the next discovery. In computational molecular simulation, where all parameters can be controlled, it would seem an easy task to reproduce an existing work. Molecular dynamics (MD) and Monte Carlo (MC) simulation are both driven by the core principles of statistical thermodynamics. Many codes for performing these methods exist each with their own algorithms and implementations, but if each code is correctly using the core principles, the end result should be the same.

This study aims to learn whether different codes can get the same result using the same model. Within the umbrella of model is the thermodynamic ensemble (e.g., NVT, NPT), the representation of the system (e.g., the force field, constraints, cutoffs), and the statepoint (i.e., the pressure, temperature). In this framework, a

"force field" is defined to supply the parameters for a given model. Other details related to the model are rigid representation and long-range and charged interactions.

In computational molecular simulation, already many hurdles to achieving simulation reproducibility have been reported. Using the same method, ensemble, and forcefield may not ensure reproducibility between engines, as different engines lack consistency in the functional forms, implementations, and options for handling long-range and charged interactions [57]. The energy calculated with certain codes has been found to depend on how the charges are calculated (whether the charge method uses the whole molecule or fragment) and which charge calculation method is used is often not reported—even differences in Coulomb's constant may cause different results [34]. Due to differences in available implementations in the difference engines there is no one size fits all protocol [35]. Systematic error was found between engines in as simple a task as calculating the potential energy and density of uncharged molecules in the liquid phase, even between groups using the same engine [36]. Although these hurdles may seem discouraging and lead some to disregard simulation entirely, I would argue that instead this is a challenge that be met by the community using TRUE principles.

This study, a multi-university collaboration, aims to determine what information is necessary to achieve statistically same results across engines. Researchers from eight universities (see Table 3.1) contributed to scripts to initialize and run each of these systems. Six different engines (see Table 3.2) were used to conduct molecular dynamics (MD) and Monte Carlo (MC) simulations. These scripts will use tools from the Molecular Simulation and Design Framework (MoSDeF) to help with re-

Table 3.1: Universities participating in the study and their abbreviation.

BSU	Boise State University
UD	University of Delaware
UH	University of Houston
UM	University of Michigan
UMN	University of Minnesota
UND	University of Notre Dame
VU	Vanderbilt University*
WSU	Wayne State University

Table 3.2: Simulation engines used in the study along with the publications in which the engine is described, simulation type, and the research group responsible for that engine.

Code	Type	Ref	Group
Cassandra	MC	[58]	UND
GOMC	MC	[58]	WSU
Gromacs	MD	[59, 60, 61, 62, 63, 64, 65]	VU
HOOMD	MD	[43, 66, 67, 68]	BSU, UM
LAMMPS	MD	[69]	UD, VU
MCCCS	MC	[70, 71]	UMN

producible simulation initialization. In order to help manage this large data space and submission to various clusters, the `signac` framework will be used to programmatically organize the statepoint directories and handle the idiosyncrasies of different cluster schedulers. This study is still ongoing, so some of the data shown in this chapter may be unfinished or incomplete. I will focus most on the simulations run using HOOMD, as this is the part of the study I am responsible for. I will also try to distill what we have learned, through the numerous issues we have discovered and overcome. This project involved a great deal of collaboration and discussion. Although the systems studied are very simple, we ran into many challenges: Keeping consistent units, sharing data, and overcoming the many unexpected hurdles to get consistency between engines due to the different available

functions in each engine. In addition, choosing what information to store and how to format that information so it is readable by one analysis script was challenging between multiple engines because not all engines were able to report the same values. The benefit of this study is that it provides comprehensive information: the results obtained and the exact code ran to obtain them. These data can be used by other simulators to validate their work. These workflows can also serve as valuable examples. At the time this study started, HOOMD v3 was still in beta, so the HOOMD part of this project can serve as a valuable example of how to (1) run atomistic simulations consistently with other engines, (2) initialize rigid bodies and bond constraints using the MoSDeF framework (this workflow is currently the only available example of how to initialize rigid bodies on an existing system), and (4) log individual energies such that they can be compared to other engines. We have learned that even when you control every step of the initialization process, there are still so many additional user-tuned parameters in each engine that guaranteeing reproducibility requires careful thought.

3.2 Models

The systems which are studied are small, simple, and chosen to demonstrate simulation in systems with varied degrees of complexity: The TraPPE united-atom (UA) forcefield was applied to UA methane (see Figure 3.1a) and pentane (see Figure 3.1b) models to investigate the simplest case of only Lennard-Jones interactions and a linear molecule with bonds, angles and dihedrals [72]. TraPPE-UA was also applied to a UA benzene model (see Figure 3.1c) to investigate rigid ring structures. (We initially tried a flexible model based on TraPPE-UA, but some engines were not able to get this model to run [73].) The SPC/E water model (see

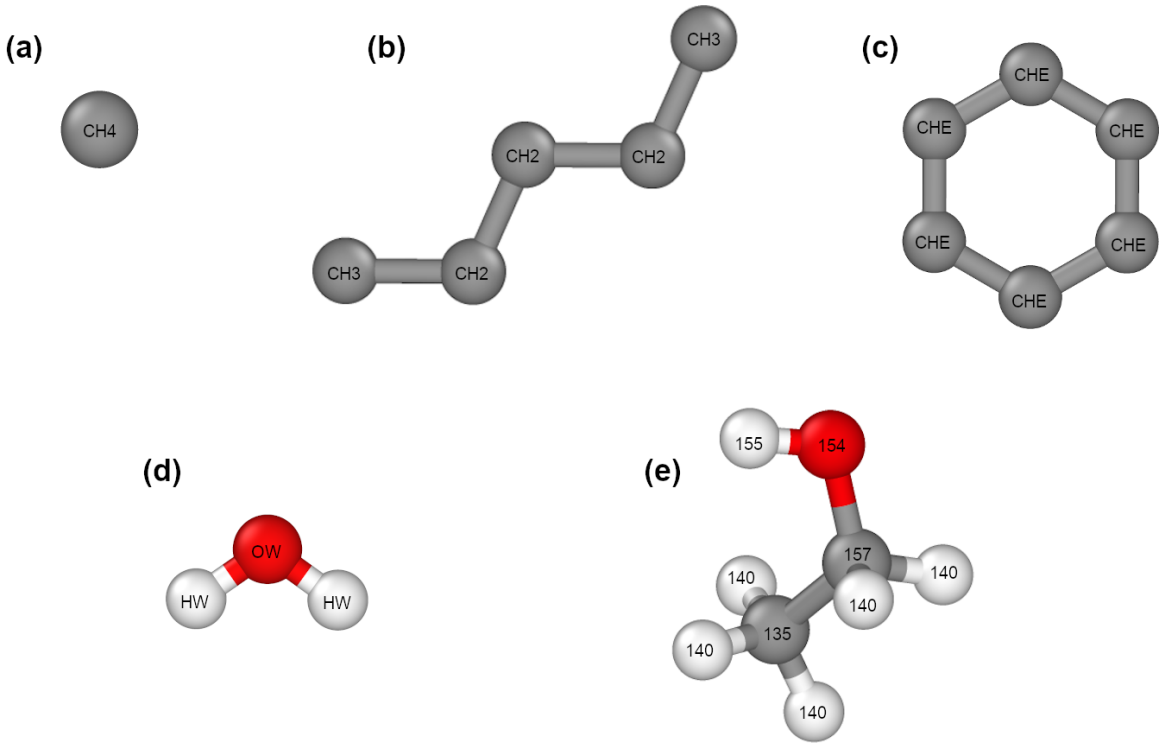


Figure 3.1: Structures of the five models with the atoms types labelled: (a) TraPPE-UA methane, (b) TraPPE-UA pentane, (c) TraPPE-UA benzene, (d) SPC/E water, (e) OPLS-AA ethanol

Figure 3.1d) was used to investigate a rigid molecule with electrostatics [74]. Finally, the most complex system, the OPLS all-atom (AA) forcefield was applied to ethanol (see Figure 3.1e) to investigate a fully atomistic, charged molecule [75]. The step-wise, increasing complexity in these systems was later very useful for pinpointing the source of a discrepancy between engines.

The nonbonded interactions were modelled as Lennard-Jones (LJ) interactions with long-range correction to energy and pressure. (This correction is discussed further in subsection 3.5.1.) And charged interactions are computed via the particle particle particle mesh (PPPM) method as described in Darden et al. [76] and

Lebard et al. [68]. The bonded interactions were modelled as harmonic bonds, harmonic angles, and OPLS style dihedrals. The forcefield parameters for TraPPE-UA and OPLS-AA are provided with `foyer` [25]. The non-bonded potential parameters for SPC/E water can be found in Table B.1. The non-bonded TraPPE-UA potential parameters used for benzene can be found in Table B.2. No bonded potentials were used for either water or benzene as both molecules were modelled as a rigid bodies [66, 67].

3.3 Methods

The scope of this project meant that it required thoughtful organization: A total of sixty-nine statepoints each in sixteen replicates were distributed to eight different universities. Devising a custom organization scheme would've been time-consuming and potentially unreliable, so the parameter space was handled by the `signac` framework[77, 78, 79, 80]. `Signac` is a tool designed for managing dynamic data spaces in a well-defined, indexable way. The study is formatted as a `signac` project, so all code—from system initialization to project analysis for all engines—is contained in a single repository, and access to the data, or "job", can be done using the statepoint parameters.

The workflow for the HOOMD simulations in shown in Figure 3.2; each engine will have similar although perhaps not identical workflows. First the statepoint

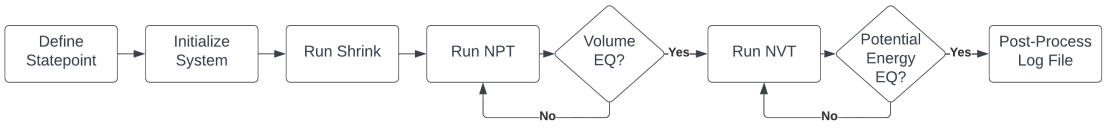


Figure 3.2: Workflow for HOOMD simulations in the reproducibility study.

data is used to define the job: The statepoint definition consists of the molecule being simulated, the temperature and pressure at which the simulation is run, the simulation ensemble, the number of molecules (N), the side length of the cubic box, the target density, the molecule mass, the forcefield, including the cutoff style and cutoff distance (r_{cut}), and the engine used to simulate that statepoint. To provide adequate data for statistical analysis, 16 replicates were run at each statepoint. Then using the statepoint data, the system (including the starting structure, box, etc.) is initialized programmatically.

To ensure that all engines are starting with the same system, this process is generalized: All engines use the `construct_system` function, which wraps `mBuild`'s `fill_box` to creates an `mBuild` `Compound` object containing the simulation box at the specified density populated with the specified number of molecules, and `load_ff`, which loads the correct `foyer` `Forcefield` object from the forcefield name in the statepoint. Then the forcefield is applied to the system box compound to generate a ParmEd structure[34], where the atoms are typed according to the forcefield and the relevant force parameters are also stored within the structure object. From there each engine takes the parameterized ParmEd structure and converts it to the necessary input format. For HOOMD this is done using the `create_hoomd_forcefield` function in `mBuild`, which handles the unit scaling, initial snapshot creation, and neighborlist and force creation using the information stored in the ParmEd structure. Some adjustments are made to the snapshot and forcefield after this to handle rigid bodies, constrained bonds, neighborlist exclusions, and long range correction. Then the simulation is initialized, the integrators, methods, and logging are set and the simulation can be run.

The first run step in this process is to run a brief shrink step to bring the system to the desired density. The shrink step starts with the volume expanded by 8 times (each length of the target box is 2 times larger) then the box is shrunk in the NVT ensemble over 100 ps with a thermostat coupling value of 1 ps at the simulation temperature. The NPT run is initialized from the final frame of the shrink trajectory and is run at high thermostat (1 ps) and barostat (5 ps) coupling values initially (for the first 1 ns) to prevent large fluctuations in the pressure and kinetic temperature and then they are lowered (to 0.1 ps and 1 ps, respectively) for the duration of the run. The NPT ensemble is run for a minimum of 6 ns, but after each NPT run, pymbar's `detectEquilibration` and `subsampleCorrelatedData` functions are used to check whether the volume has equilibrated and the desired number of decorrelated samples have been run [81, 82, 13]. A minimum of 100 independent samples at equilibrium were collected for each run, and this equilibration detection and subsampling is applied across all engines. By using a quantitative metric to achieve the same number of equilibrated, decorrelated samples, as opposed to deciding whether a simulation is equilibrated by eye or running for an arbitrary time, we aim to reduce user error in our sampling. Plots of the evolution of potential energy over time for the simplest (UA methane) and the most complex system (AA ethanol) can be found in Figure B.1 and Figure B.2. If the volume has not equilibrated or the minimum number of independent samples have not been run, then the NPT simulation continues at the lower thermostat and barostat coupling values for an additional 5 ns. If the volume has equilibrated, then the statistically independent samples of the equilibrated volume are averaged and this average is logged to the job document and the job moves to the NVT run. The NVT run is

initialized from the final frame of the NPT trajectory, but because this final frame may not be at the average volume, a short (20 ps) shrink period is done for the initial NVT run and then the simulation is run for 5 ns. After the NVT run, the same equilibration detection is used—this time looking at the potential energy values. If the potential energy is equilibrated, the NVT run is finished, otherwise it is run again for an additional 5 ns from the last frame of the proceeding NVT run.

Finally the log file containing thermodynamic information written out by HOOMD is processed to remove any extra headers within the data from restarted jobs, the pressure and temperature are converted from simulation units into kPa and K, and the density values are calculated from the volume and logged. For each engine, the final step is to convert its log and trajectory files to a standard format and unit to allow analysis to run over all jobs.

All scripts used to initialize, simulate, and analyze this project are publicly available on GitHub [83]. To share the large data workspace between universities, the workspace folder is uploaded to a shared Dropbox using Rclone. Once the study is finalized the complete data will be distributed using Zenodo. Along the way to achieving consensus between engines in this main project, many stumbling blocks were encountered. The following sections will detail the issues we encountered and their resolution.

3.4 Single point energy comparisons

In general when trying to achieve "correctness" in a complex system, it is advisable to first validate that correctness can be achieved in a simple system, then incrementally add complexity. In this study, single point energy calculations were done after

discrepancies were found in the simulation results. With the benefit of hindsight, it is clear that starting with a single point energy comparison would've been much more efficient—allowing discovery of many issues early in the study before numerous compute hours were consumed. It is a good rule of thumb for simulators to provide a snapshot along with its energy breakdown when publishing if they want to aid reproduction of their work.

The single point energy evaluation involved distributing the complete information for the starting system: atom coordinates and bonding and box information. The reason this starting structure needed to be distributed is because it was necessary to ensure all starting structures were identical: The packing functions in `mBuild`, including `fill_box` which was used to create the initial simulation inputs for this study, are wrappers for PACKMOL [84, 85]. PACKMOL uses an intrinsic random number generation method which is compiler dependent. The result is that PACKMOL creates different systems based on the operating system on which it was compiled. In general, simulators are interested in comparing a sampling of the equilibrium ensemble, which should not depend on the initial position. However, this discrepancy is important to note if one is trying to compare the energies of a single frame! For this reason, we could not rely on programmatically generated structures to compare the single point energies and it is good practice to distribute the input structure along with the code used to generate it. This starting system was then initialized following the same procedure as the NPT/NVT ensemble simulations, but the simulation was not allowed to progress in time allowing direct comparison of the energies between engines.

Through the single point energy comparisons, it was determined that information

about the 1-4 scaling and the combining rule were getting lost in the translation from `foyer` forcefield to ParmEd structure to engine input for Cassandra and Gromacs. These errors were fixed by `mBuild` PR 1004 and 1010. Initial comparison of the single point energies also showed large disagreement between the electrostatic energies of water in HOOMD and other engines. In HOOMD water was modelled as a rigid body. When the "rigid" neighborlist exclusion was added on to the default exclusions set by `create_hoomd_forcefield` ("bond" and "1-3"), it was found that the exclusions were double counted by the PPPM energy calculation resulting in a large discrepancy in the electrostatic energy. This is now noted in HOOMD's rigid body documentation. Comparison of the single point energy calculations also helped to tune the grid size needed for HOOMD's PPPM charge calculation.

The single point energy values after applying the fixes discussed are included in Table B.3-Table B.7, and in general these values were found to agree with some noted discrepancies: Some engines are not able to report as thorough of a breakdown for energies or they may partition the energies differently. For example, the long range and short range electrostatic energies are not the same between any engines; however, the total electrostatic energy is comparable. By first validating our method using comparison of the single point energies, we were able to move into more complex simulations with greater confidence.

3.5 Sensitivity of model to timestep

Designing a general simulation workflow to work with the range of systems in this study required some model specific adjustments. For example, the atomistic ethanol system required a smaller timestep. When a timestep which was too large

but small enough that the system didn't explode (i.e., the position update didn't push the particle outside of the simulation box) was used, the kinetic temperature was found to equilibrate to a lower value than was set in the NVT/NPT thermostat. Unlike the other systems studied, ethanol has explicit hydrogens with harmonic bonds. Because hydrogen is such a light particle, the oscillations of bonds involving hydrogen are very fast. (The SPC/E water system also has explicit hydrogens, but because it is rigid, the hydrogen bonds do not have these fast oscillations.)

The temperature could not equilibrate to the set temperature because the timestep did not allow adequate sampling of the harmonic bond. Consider the following example: The period of the harmonic bond is given by

$$T = \frac{2\pi}{\sqrt{\frac{k}{m}}} \tag{3.1}$$

where k is the potential constant for the bond given in the forcefield and m is the mass of the particle. Initially, I was using a timestep of 0.001 (in simulation time units, approximately 1 fs when converted to real units) while the period of the harmonic bond (see Equation 3.1) was 0.0118 (in simulation time units, approximately 11 fs). Figure 3.3 illustrates the sensitivity of the harmonic bond sampled at different timesteps. By using a smaller timestep (0.0005 in simulation time units, approximately 0.5 fs), the temperature equilibrated to the correct value. Figure 3.4 and Figure 3.5 show the sensitivity of the equilibrium temperature to timestep based on system complexity. Both plots are very noisy because they show only the initial 100 ps post-shrink equilibration in the NVT ensemble, however the methane

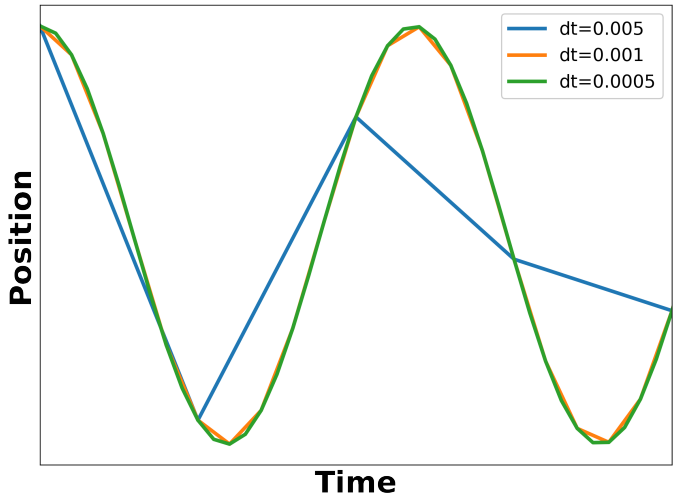


Figure 3.3: Demonstration of how choosing too large a timestep can lead to poor sampling of the position of atoms in a harmonic bond. Note the divergence from sinusoidal as dt increases. The dt values and the harmonic bond equation are accurate with the values given for the hydrogen bond.

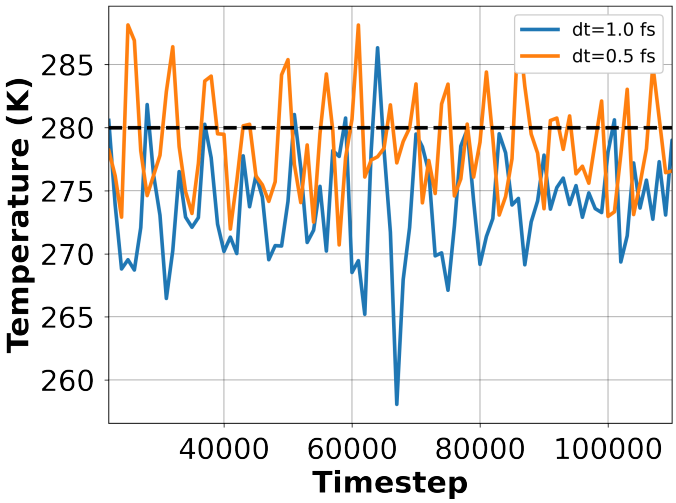


Figure 3.4: The evolution of temperature with time with the larger (1 fs) and a more reasonable (0.5 fs) timestep for the ethanol-AA system. The set temperature is shown as a dashed line.

temperature fluctuates around the set value and these fluctuations are small (only a couple of degrees K) and smooth while the ethanol fluctuations are much larger,

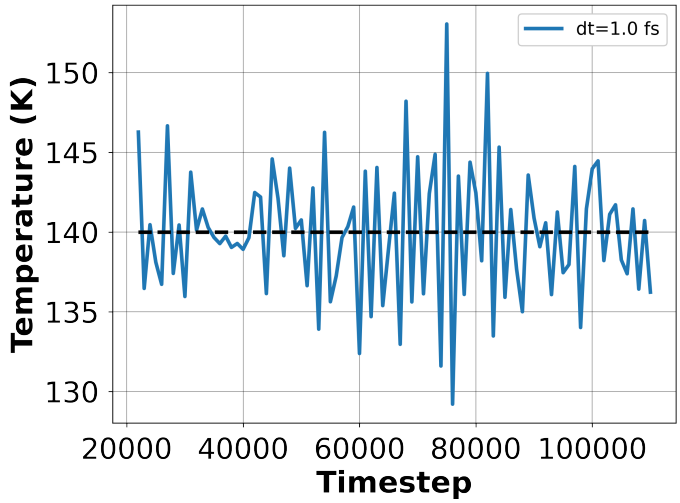


Figure 3.5: The evolution of temperature with time with the larger (1 fs) timestep for the methane-UA system. The set temperature is shown as a dashed line.

and with the larger timestep, the fluctuations are not about the set equilibrium temperature. This is due to the increased complexity of the ethanol system requiring finer sampling. When using general simulation workflows, it is important to check whether set defaults make sense for a particular system.

3.5.1 Cutoff schemes

As is common in molecular simulation, the Lennard-Jones equation (Equation (3.2)) was used to model the non-bonded potentials between particles. To save computational resources, it is common to truncate the potential at a certain distance, and, as a discontinuity in the potential energy can issues in energy conservation, there exist various smoothing schemes for handling values beyond the

cutoff.

$$U_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]; r < r_{cut} \tag{3.2}$$
$$= 0; \qquad r >= r_{cut}$$

Generally studies of molecular simulation will report the cutoff handling scheme used without much elaboration into why the particular scheme was chosen. Perhaps the engine used does not offer many alternatives or perhaps the forcefield used is parameterized with a specific cutoff in mind. This is by no means a comprehensive study of all cutoff schemes, but let us define three cases for how the potential beyond the cutoff is handled: hard, shifted, and long-range correction (LRC). In the "hard" cutoff scheme, the potential is simply set to zero beyond r_{cut} with no smoothing regardless of the potential's value at r_{cut}. The "shifted" cutoff scheme shifts the entire potential by the potential value at r_{cut} (a constant) such that the potential at r_{cut} is zero. Finally, the "LRC" scheme applies isotropic, integrated corrections to the energy and pressure based on the particle number densities beyond r_{cut}. The energy and pressure corrections ΔE and ΔP are given by

$$\Delta E = 2\pi \sum_{i=1}^{n} N_i \sum_{j=1}^{n} \rho_j \int_{r_{\text{cut}}}^{\infty} V_{ij}(r) r^2 \, \mathrm{d}r, \tag{3.3}$$

and

$$\Delta P = \frac{-2\pi}{3} \sum_{i=1}^{n} \rho_i \sum_{j=1}^{n} \rho_j \int_{r_{\text{cut}}}^{\infty} \left(r \frac{\mathrm{d}V_{ij}(r)}{\mathrm{d}r}\right) r^2 \, \mathrm{d}r \tag{3.4}$$

where n is the number of unique particle types in the system, ρ_i is the number density of particles of type i in the system, $V_{ij}(r)$ is the pair potential between

particles of type i and j, and N_i is the number of particles of type i in the system [86, 87]. These expressions assume that the radial pair distribution functions, $g_{ij}(r)$, are unity at the cutoff and beyond.

At the time this study was initiated, the only options for handling the cutoff shared between all engines were the "shifted" and "hard" schemes. It was found that neither cutoff scheme was adequate for equilibrating to the correct density or reaching agreement between MD and MC. So in order to use the "LRC" scheme in HOOMD, I contributed the tail correction calculations for the LJ pair object which adds a correction to energy and pressure according to Equation 3.3 and Equation 3.4. The tail correction code was included in the HOOMD v3.0.0 release.

In order to assess the impact of the cutoff scheme used to handle potential values beyond r_{cut}, the effect of cutoff scheme on the simplest system, united-atom methane, was examined. When a hard cutoff is used (see Figure 3.6), there is a large discrepancy between the MD engines (GROMACS, LAMMPS, and HOOMD) and the MC engines (GOMC, Cassandra, and MCCCS). The variation in densities between MC and MD when using the hard cutoff scheme is most likely due to innate differences in the two methods: MD calculates the interparticle forces—the derivative of the potential with respect to r—and uses these forces to update the particle velocities and, in turn, positions. MC, however, calculates the potential values outright and uses the change in potential to choose whether to accept or reject a particular move. Using a hard cutoff would cause a discontinuity in the potential, which results in how the force is handled near the cutoff to be ill-defined.

When a shifted cutoff is used (see Figure 3.7), all engines equilibrate to the same

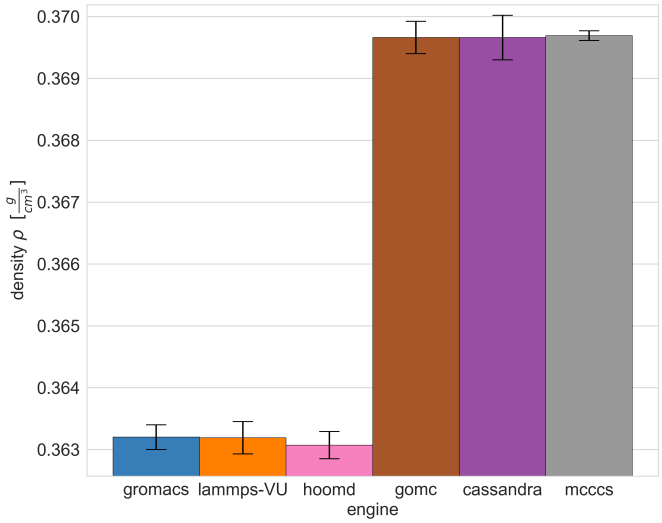


Figure 3.6: Average density from NPT simulation of methane using "hard" cutoff by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

density within error, but the density is much lower than the predicted value—the density of methane at 140K and 1318kPa is predicted to be 0.37808 g/cm^3[88]. This deviation from the correct density is due to the absolute values of the potential being shifted. Therefore the potential is different than it was originally created and parameterized, so it may yield different results. (The TraPPE-UA forcefield, for example, was designed to be used with analytical tail corrections as described in Equations (3.3) and (3.4).)

In order to get consensus between engines and closer to the correct density value, the correction to the energy and pressure was needed (see Figure 3.8). By using the energy-pressure correction, the energies and densities of both MD and MC are within the error tolerance of two standard deviations.

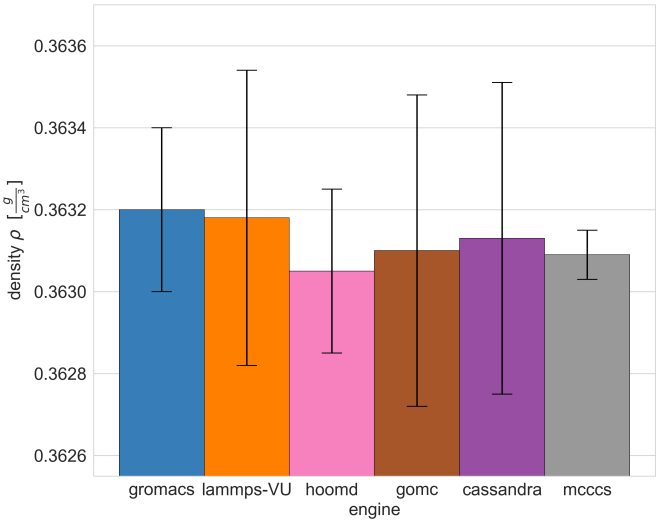


Figure 3.7: Average density from NPT simulation of methane using "shifted" cutoff by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

It is worth noting that differences in the computed pressure exist between engines regardless of cutoff scheme, see Figure 3.9. These discrepancies in pressure may be due to the ideal gas contribution to pressure. The pressure in molecular simulation can be computed by the following equation

$$P = \frac{Nk_BT}{V} + \frac{\sum_i^{N'} r_i \cdot f_i}{dV} \tag{3.5}$$

where N is the number of atoms, k_B is Boltzmann's constant, T is the temperature, d is the dimensionality of the system, V is the system volume, and r_i and f_i are the position and force vectors of atom i [89]. The first term in Equation 3.5 is the ideal gas contribution to pressure. Not including this ideal gas contribution is akin to calculating the pressure of the system at 0K. HOOMD calculates the pressure with

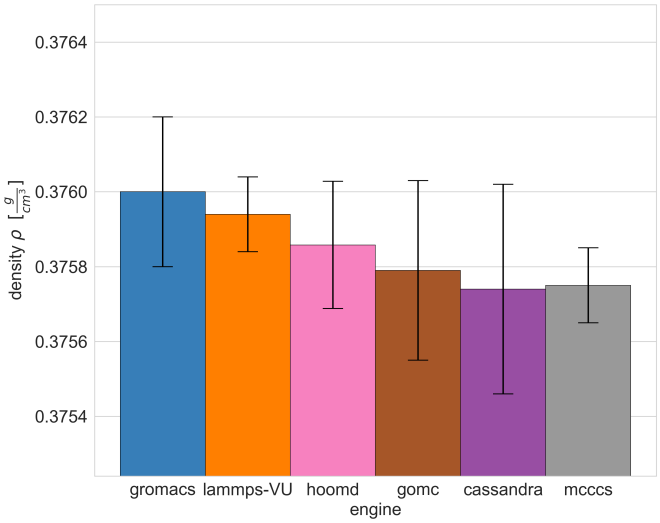


Figure 3.8: Average density from NPT simulation of methane using energy and pressure long range correction by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

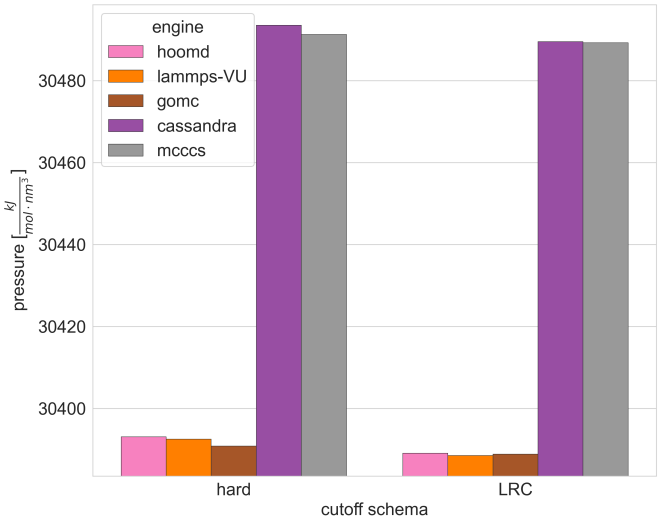


Figure 3.9: Instantaneous pressure from first frame of methane by engine comparing "hard" and "LRC" cutoff schemes.

the virial equation which includes accounting for the translational kinetic energy, essentially the kinetic temperature. However, calculating the pressure difference due to the ideal gas contribution from 900 particles at 140 K in a 63.9 nm^3 box accounts for a difference of about 16 $\frac{kJ}{mol \cdot nm^3}$ while Figure 3.9 shows the discrepancy is closer to 100 $\frac{kJ}{mol \cdot nm^3}$, so further investigation into the source of this discrepancy is required.

Through examination of the energy values of this first frame we also found a discrepancy in the energies reported in newer versions of GROMACS, see Figure 3.10. This should not have an effect on the end result but is worth noting for those who

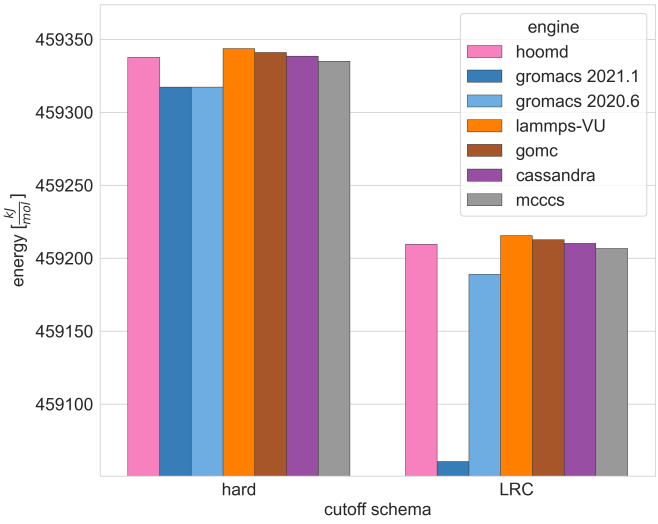


Figure 3.10: Instantaneous energies from first frame of methane by engine comparing "hard" and "LRC" cutoff schemes.

may want to compare instantaneous energies: Newer versions of GROMACS do not include the pressure correction to energy in the total potential energy when running an energy minimization step. This is documented in GROMACS Issue 4229.

Although it may seem of little significance, the choice of where the cutoff is set and how the cutoff is handled can greatly affect the simulation outcome.

3.5.2 Rigid constraints

Benzene and water are modelled as rigid bodies in this study. In HOOMD, rigid bodies consist of a central body particle and its constituent particles. All of these particles can have forces acting upon them, but the mass and moment of inertia of the body particle is set to be the full mass and moment of inertia of the body plus its constituents. Further description of how the rigid constraint forces are implemented can be found in Nguyen et al. [66] and Glaser et al. [67], and I contributed to updating the rigid body constraint to the v3 api. HOOMD provides functions to simplify the initialization of rigid bodies; however, using these rigid body functions with MoSDeF tools introduced additional hurdles. The difficulties which arise at the intersection between codebases are those with which many simulators will be familiar.

Initializing rigid bodies in HOOMD using MoSDeF tools required additional effort, as there are some idiosyncrasies that conflict with the typical workflow. In order to initialize a rigid body in HOOMD, all the body particles must be first in the snapshot and the constituent particles must all have the same relative orientation to the body particle. To reduce the cognitive load for users, HOOMD recommends creating the body constituent particles using the `create_bodies` function, but this precludes using the MoSDeF initialization functions which fill the simulation volume, apply the forcefield, and initialize the HOOMD forces. Therefore, we have developed a generalized workflow based on the assumption that each molecule is the same and will be its own rigid body. First, an initial snapshot is created with

the same number of body particle as molecules. Using a functionality which I contributed to the `create_hoomd_forcefield`, this initial snapshot can be passed in and the normal workflow—filling the box with all molecules in the same orientation and applying the forces—can add on to this initial snapshot. Next, the number of molecules and the number of particles in each molecule is used index through the snapshot to set the body ids for each constituent particle and to calculate and set the position, mass, and moment of inertia tensor for each body particle. Finally, the body constraint is defined based on the types, charges, diameters, and relative positions and orientations, of the first molecule.

Before using rigid bodies for water and benzene, the HOOMD simulations were taking an excessive amount of time to equilibrate and were equilibrating to vastly different results as other engines. There was some hesitation to use rigid bodies because it was a function that was not supported or implemented differently in other engines. However Figure 3.15 and Figure 3.17 suggest that the rigid models achieve agreement.

Currently the example for initializing rigid bodies in the HOOMD v3 API is still under development. This workflow can serve as an example of how run simulations of rigid bodies in HOOMD using MoSDeF tools.

3.6 Results

This study is still underway and data is still being collected, however we will examine the current results of this study up to this point, namely the densities and potential energies. First let's start with the simplest system, UA methane. The density of methane between engines agrees within error (see Figure 3.11). The MC

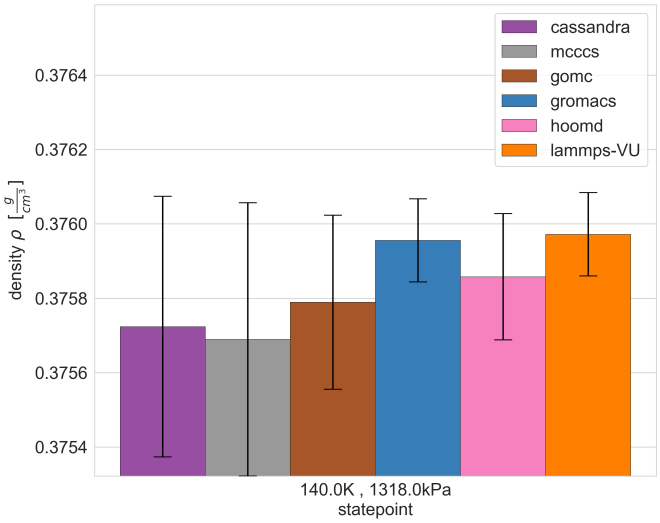


Figure 3.11: Average density of methane by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

density and potential energy data shows higher variation than the MD results. It appears that the MC engines in general equilibrate to a lower density, perhaps because they can more easily sample high density configurations using unphysical moves (e.g., moving through objects, etc). In contrast, MD may be more restricted by dynamics and must follow a physical path to reach a configuration, which could make getting to these high density configurations more difficult. The predicted density of methane at 140 K and 1318 kPa should be around 0.37808 g/cm^3 [88]. We can see that the MD engines, which get closer to the predicted density, have a lower potential energy Figure 3.12. Although there are some differences we can see here between MD and MC, the potential energies of methane also agree between engines within error.

Next, let's examine UA pentane, which adds bond, angle, and dihedral forces.

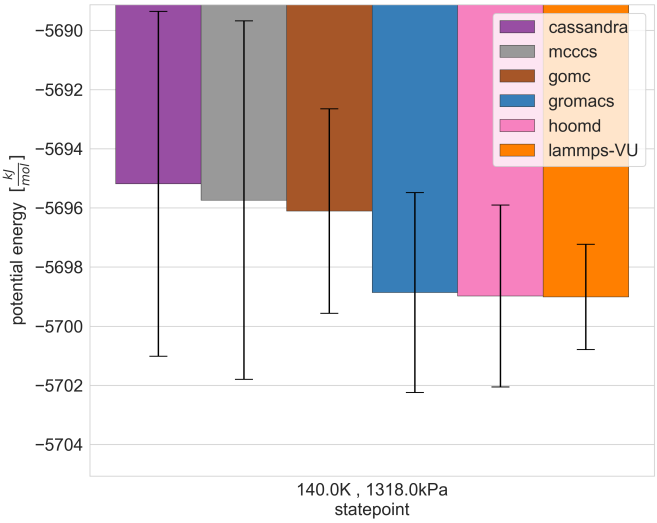


Figure 3.12: Average potential energy of methane by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

Within the pentane model, two bond types were used: flexible and fixed-length (also called constrained) bonds. This choice was made because the MC engines only have the capability to model constrained bonds. Therefore, by modelling a single system in each way (for the engines which are able) we can isolate and observe the effect of bond type. Figure 3.13 shows the densities of pentane with flexible and fixed bonds by engine. I do not yet have data for constrained bond from LAMMPS, and so the only engines we can compare at this point are GROMACS and HOOMD. Both engines seem to have greater variability in the density when using flexible bonds. This seems reasonable: allowing the bond lengths to fluctuate essentially allows the volume the molecule occupies to also fluctuate, so in order to keep the intermolecular forces consistent, the box volume must also fluctuate. The HOOMD flexible and constrained bond pentane densities appear

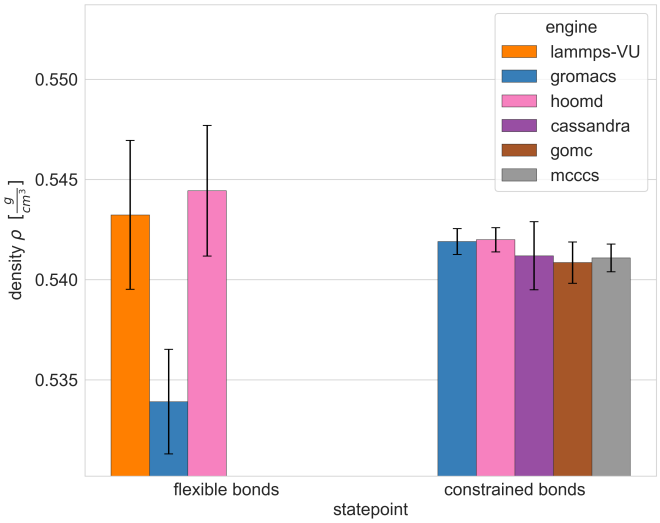


Figure 3.13: Average density of pentane with flexible or constrained bonds by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

to agree within error although the density of the flexible bond pentane is higher. For the GROMACS, however, the flexible pentane is much lower density and not within error of the constrained bond model. Further investigation into the cause of this will be required. The differences between the potential energy of the flexible and constrained bond pentane (see Figure 3.14) are much more clear. The potential energies for the constrained bond pentane are much lower because the bond lengths for all systems are fixed at their equilibrium length, which is the minima of the harmonic bond potential. In the flexible bond model, the bond is allowed to deviate from its equilibrium length, resulting in higher positive contribution to potential energy from the bond. This difference will be seen in the potential energies for other systems. When the bond lengths are fixed (as in the water and benzene models), the potential energy values observed by MD are closer to those observed

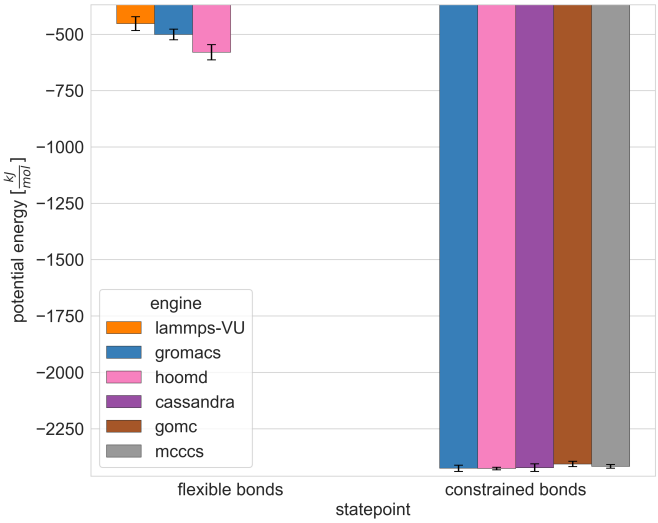


Figure 3.14: Average potential energy of pentane by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

by MC. (A direct comparison of flexible vs fixed bonds is shown in Figure 3.14, but Figure 3.18 and Figure 3.16 also are rigid models or have fixed bonds.)

Next, let's examine UA benzene, which is a ring structure modelled as a rigid body. I will use the term "rigid body", although the actual implementation may vary depending on the options available in each engine. Some engines may be fixing bonds, angles, and dihedrals, while others (like HOOMD) may support true rigid bodies. Although the implementations of rigidness may differ between engines, the average density of benzene between engines agrees within error (see Figure 3.15). Again we see slightly lower density with greater variation in MC, the reasons for this are likely the same as with methane. Figure 3.16 shows the potential energies of benzene by engine. With the exception of MCCCS, the potential energies agree within error. Further investigation will be required into the

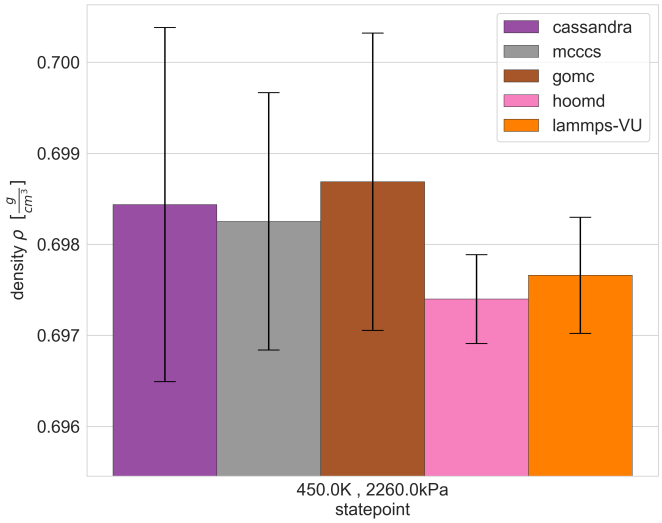


Figure 3.15: Average density of benzene by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

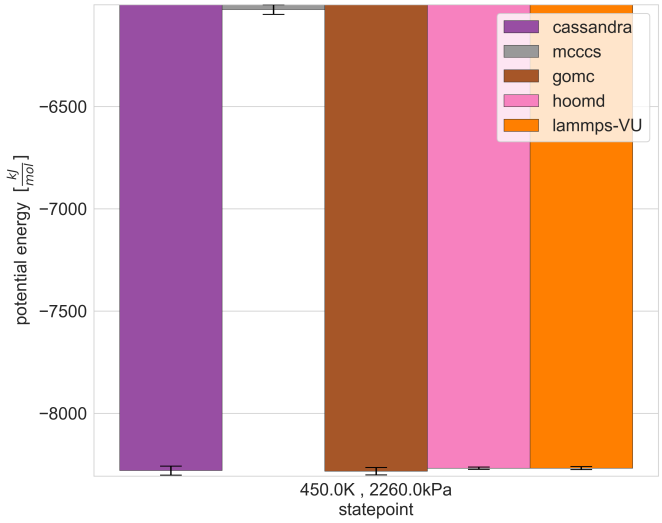


Figure 3.16: Average potential energy of benzene by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

incongruous energy value reported by MCCCS.

Next, let's examine SPC/E water, which is modelled as a rigid three-site charged molecule. Figure 3.17 shows the density of water by engines at three statepoints. The density of water at each statepoint appears to agree within error between all

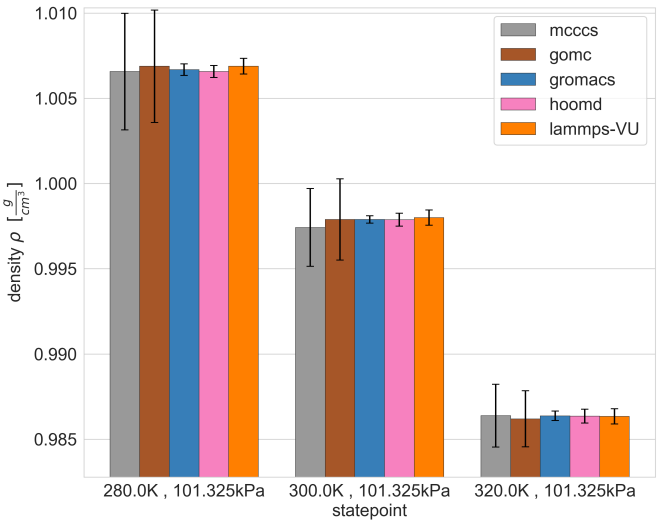


Figure 3.17: Average density of water at three statepoints by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

engines The average densities are also relatively close to the predicted values: 0.99991, 0.99656, and 0.98953 g/cm^3, respectively [88]. The potential energies of water at these three statepoints Figure 3.18 show similar agreement with the exception of LAMMPS. It seems that this discrepancy is related to the PPPM implementation in LAMMPS because the potential energy is systematically high in both charged systems (see Figure 3.18 and Figure 3.20). Further investigation into this discrepancy will be required.

Finally, let's look at the most complex system, all-atom ethanol, which is a fully-

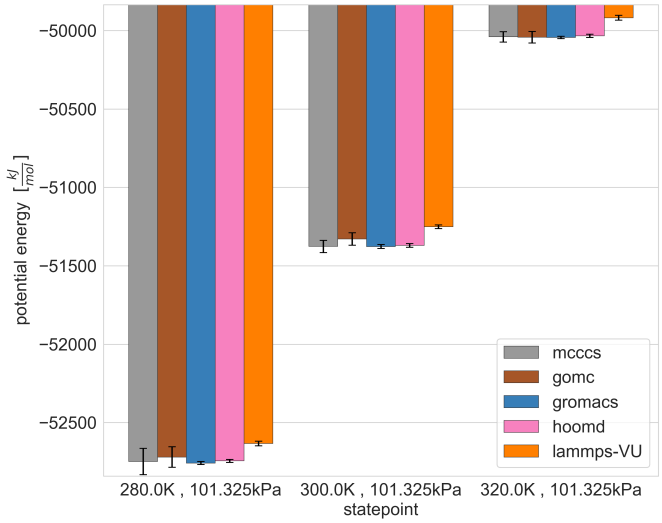


Figure 3.18: Average potential energy of water by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

flexible molecule with charges. (Full flexibility, of course, depends on the engine's capabilities, so MD engines modelled this system with flexible bonds and MC engines used fixed bonds.) The densities of ethanol by engines at three statepoints is shown in Figure 3.19. The densities agree within error for each method (MC or MD), but this time the density observed in MD is higher than that in MC. It is hard to make a clear statement for what is causing this discrepancy in densities between the two methods with the data available, but comparison with Figure 3.13 could suggest that the higher density observed in MD is due to the flexible bonds. The potential energies of ethanol (see Figure 3.20) show the greatest variability between engines; however, this can be attributed to two issues which have already been discussed. The first is that the MD engines are using flexible bonds while the MC engines are using fixed bonds. Figure 3.14 shows that when the only difference

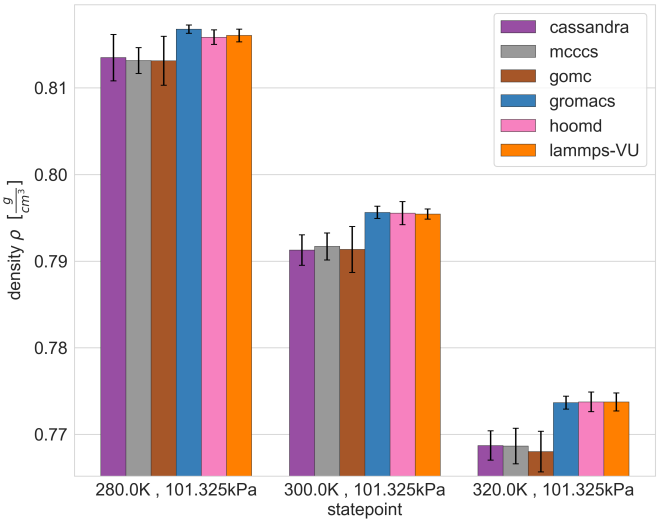


Figure 3.19: Average density of ethanol by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

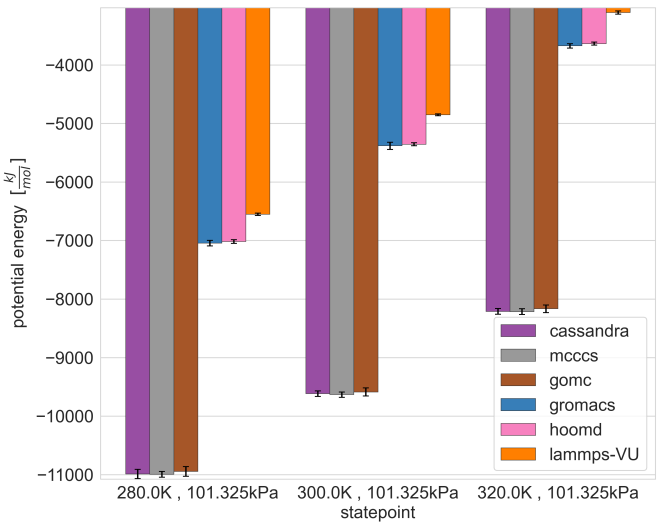


Figure 3.20: Average potential energy of ethanol by engine. The average is taken from independent samples of the equilibrated regions of 16 replicates. The error bars represent two standard deviations in each direction.

is whether bonds are flexible or fixed, the potential energy of the fixed bonds will be much lower, which explains the lower potential energy seen in the MC engines in Figure 3.20. The second is the yet to be understood discrepancy with LAMMPS in the PPPM model.

3.7 Conclusions

Achieving equivalent results in molecular simulations can be tricky because it's not always possible to implement models identically in different engines. For simple models like UA methane, sampling differences arise, but the engines largely agree. As our models increase in complexity so do the deviations in our results between engines and methods. There's moderately good consensus among the UA pentane constrained-bond models, but the flexible bond model has greater variation and highlights the discrepancy when comparing against the constrained model. The potential energy and density calculations of rigid UA benzene appear to demonstrate consensus, aside from MCCCS, but this deviation gives us a direction to investigate. In the SPC/E water and AA ethanol simulations, we see much more variance between MD and MC engines but consistency among each simulation method. Performing these consistency checks helps find bugs and identify modeling choices that give rise to these differences.

This study was a community-driven effort to show MoSDeF tools can help achieve reproducible results across engines. Beyond just an opportunity to validate that our results were "correct", working on this project together with other simulators was an invaluable opportunity to learn. Initially, this project seemed to be an easy task. However, it quickly became apparent that even seemingly minor choices rele-

vant to the simulation parameters required discussion. These discussions forced us to dig into each engines code in a way that helped us better understand the simulations we were running. Each group brought a different background and expertise which was of great benefit when trying to pinpoint the source of a discrepancy. Already, this collaboration has prompted many improvements to MoSDeF tools and the engines they support. Ultimately the goal of molecular simulation is to probe the thermodynamic equilibrium of materials in ways which may give insight that experiment cannot. In order to draw meaningful conclusions from simulation we must understand how the choices made in our model influence the results.

CHAPTER 4:

CONCLUSION AND OUTLOOK

In this work we have aimed to show what reproducibility can look like in molecular simulation in particular and in computational sciences in general. As someone who came into my PhD with almost no coding experience, I know the struggles of trying to reproduce an unfamiliar work—especially scientific scripts which may be done to achieve a purpose without really considering the next user. Although designing with the next user in mind may take more work upfront, with practice it becomes habit and helps promote efficiency and reproducibility. The effort toward clear documentation and transparency in methods is a benefit for anyone trying to reproduce the method, including the original author. Collaborative development through all stages of the project helps to ensure that our methods and documentation make sense from a different perspective. Being part of a broader collaboration where I was encouraged to build on and change some of the existing framework helped me build confidence and to learn how to make my own projects. Discussion of these works with others from diverse backgrounds and expertise has helped me to develop a more well-rounded understanding of molecular simulation. Not only has this community helped me develop a better understanding of simulation, but also helped me design tools from a simulator's perspective.

I am excited by the prospects of my work to persist and be used after I graduate, and I look forward to seeing future developments. Currently an undergraduate scientist in our lab is using `PlanckTon` to investigate the molecule Y6, a component in the current state-of-the-art OPV. The SMARTS definitions for all the atom types in Y6 are not yet present in the forcefields provided with `foyer`, so she is working to parameterize this compound through other methods and add these parameters to the custom GAFF forcefield shipped with `PlanckTon`. Although the workflow for adding Y6 is not as simple as other compounds, the modular framework and workflows are still in place to help her quickly spin up simulations once she has her parameters in place. As she has worked on this task, she has provided essential feedback on gaps in documentation. The work of Miller et al. [90] has shown the importance of polydisperse polymer lengths and their ability to form tie-chains for charge transport simulations. Another scientist in our lab has been developing tools for initializing a system of polymers with a distribution of lengths, and his work has included completely revamping `mBuild`'s polymer builder. And while these tools are still under development, they would be a great addition to `PlanckTon` [91]. Although `GRiTS` has some rudimentary fine-graining ability, further development of this method will be useful for its use in coarse-graining applications. The code and workflows that I leave behind have been designed for the next user, and I look forward to the science to which they can be applied. The order parameter workflow could easily be applied to new compounds and compound mixtures in order to find novel OPV compounds or determine the state at which self-assembly most robustly produces the morphologies best for charge transport. The most efficient way to sample this state space would be to start with

a small system size at a sparse selection of temperatures, solvent qualities, and densities, then increase the system size and state-space precision after high-order regions are identified. I also look forward to the conclusions of the reproducibility study. Already so many fixes and new features have been implemented as a result of the study's findings, and I hope that the MoSDeF community will continue to grow and strive for reproducibility between all the engines we support. These observations support our thesis that community-built open tools contribute to more efficient, correct scientific software development.

REFERENCES

[1] Monya Baker and Dan Penny. Is there a reproducibility crisis? *Nature*, 533 (7604):452–454, 2016. ISSN 14764687. doi: 10.1038/533452A.

[2] Carl W. Isaacson, Markus Kleber, and Jennifer A. Field. Quantitative Analysis of Fullerene Nanomaterials in Environmental Systems: A Critical Review. *Environ Sci Technol*, 43(17):6463–6474, 2009.

[3] The regents of the University of Michigan. HOOMD-Blue Tutorials, 2022.

[4] Justin Lemkul. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living Journal of Computational Molecular Science*, 1(1): 1–53, 2019. ISSN 25756524. doi: 10.33011/livecoms.1.1.5068. URL `https://www.livecomsjournal.org/article/5068-from-proteins-to-perturbed-hamiltonians-a-suite-of-tutorials-for-the-gromacs-2018-molecular-simulation-package-article-v1-0`.

[5] Justin A. Lemkul. GROMACS Tutorials, 2020. URL `http://www.mdtutorials.com/gmx/`.

[6] Computational Materials Science Lab of Boise State University. Notebook Tutorials, 2020. URL `https://github.com/cmelab/notebook{_}tutorials`.

[7] Greg Wilson, D a Aruliah, C Titus Brown, Neil P Chue Hong, Matt Davis, Richard T Guy, Steven H D Haddock, Kathryn D Huff, Ian M Mitchell, Mark D Plumbley, Ben Waugh, Ethan P White, Paul Wilson, and N P C Hong. Best Practices for Scientific Computing. *PLoS biology*, 340(May): e1001745, jan 2014. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001745. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3886731{&}tool=pmcentrez{&}rendertype=abstracthttp://arxiv.org/abs/1210.0530`.

[8] The Carpentries. Software Carpentry, 2022. URL `https://software-carpentry.org/`.

[9] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):1–4, 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003285.

[10] Matthew Gentzkow and Jesse M. Shapiro. Code and Data for the Social Sciences: A Practicioner's Guide. *University of Chicago Mimeo*, (January):1–44, 2014.

[11] Adam Rule, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih Cheng Huang, Rob Knight, Niema Moshiri, Mai H. Nguyen, Sara Brin Rosenthal, Fernando Pérez, and Peter W. Rose. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Computational Biology*, 15(7):1–8, 2019. ISSN 15537358. doi: 10.1371/journal.pcbi.1007007.

[12] Jürgen Cito and Harald C. Gall. Using docker containers to improve reproducibility in software engineering research. *Proceedings - International Conference on Software Engineering*, 1:906–907, 2016. ISSN 02705257. doi: 10.1145/2889160.2891057.

[13] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 129(12), 2008. ISSN 00219606. doi: 10.1063/1.2978177.

[14] Eric Jankowski, Neale Ellyson, Jenny W Fothergill, Michael M Henry, Mitchell H Leibowitz, Evan D Miller, Mone't Alberts, Samantha Chesser, Jaime D Guevara, Chris D. Jones, Mia Klopfenstein, Kendra K Noneman, Rachel Singleton, Ramon A Uriarte-Mendoza, Stephen Thomas, Carla E Estridge, and Matthew L Jones. Perspective on Coarse-Graining , Cognitive Load , and Materials Simulation. *Computational Materi-als Science*, 169(109129):109129, jan 2020. ISSN 09270256. doi: 10. 1016/j.commatsci.2019.109129. URL `https://linkinghub.elsevier.com/retrieve/pii/S0927025619304203`.

[15] Damien Irving. A minimum standard for publishing computational results in the weather and climate sciences. *Bulletin of the American Meteorological Society*, 97(7):1149–1158, 2016. ISSN 00030007. doi: 10.1175/BAMS-D-15-00010.1.

[16] John D. Blischak, Emily R. Davenport, and Greg Wilson. A Quick Introduction to Version Control with Git and GitHub. *PLoS Computational Biology*, 12 (1):1–18, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1004668.

[17] Inna Smirnova, Markus Reitzig, and Oliver Alexy. What makes the right OSS contributor tick? Treatments to motivate high-skilled developers. *Research Policy*, 51(1):104368, 2022. ISSN 00487333. doi: 10.1016/j.respol.2021.104368. URL https://doi.org/10.1016/j.respol.2021.104368.

[18] Dan Sholler, Igor Steinmacher, Denae Ford, Mara Averick, Mike Hoye, and Greg Wilson. Ten simple rules for helping newcomers become contributors to open projects. *PLoS Computational Biology*, 15(9):1–10, 2019. ISSN 15537358. doi: 10.1371/journal.pcbi.1007296. URL http://dx.doi.org/10.1371/journal.pcbi.1007296.

[19] Tsuyoshi Miyakawa. No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain*, 13(1):1–6, 2020. ISSN 17566606. doi: 10.1186/s13041-020-0552-2.

[20] Simon J.E. Taylor, Tillal Eldabi, Thomas Monks, Markus Rabe, and Adelinde M. Uhrmacher. Crisis, what crisis - Does reproducibility in modeling & simulation really matter? *Proceedings - Winter Simulation Conference*, 2018-Decem:749–762, 2019. ISSN 08917736. doi: 10.1109/WSC.2018.8632232.

[21] David L. Donoho, Arian Maleki, Morteza Shahram, Inam Ur Rahman, and Victoria Stodden. Reproducible research in computational harmonic analysis. *Computing in Science and Engineering*, 11(1):8–18, 2009. ISSN 15219615. doi: 10.1109/MCSE.2009.15.

[22] Matthew W. Thompson, Justin B. Gilmer, Ray A. Matsumoto, Co D. Quach, Parashara Shamaprasad, Alexander H. Yang, Christopher R. Iacovella, Clare

McCabe, and Peter T. Cummings. Towards Molecular Simulations that are Transparent, Reproducible, Usable By Others, and Extensible (TRUE). *arXiv*, pages 1–32, 2020. URL http://arxiv.org/abs/2003.02031.

[23] Ashley G. Smart. The war over supercooled water. *Physics Today*, aug 2018. ISSN 19450699. doi: 10.1063/PT.6.1.20180822a. URL https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full.

[24] Vanderbilt University. mBuild, 2020. URL https://github.com/mosdef-hub/mbuild.

[25] Christoph Klein, Matt Thompson, Andrew Z. Summers, Justin Gilmer, Co Quach, Alex Yang, Umesh Timalsina, Ray Matsumoto, Arjun Bansal, Andrew Ricardo Puente, Parashara Shamaprasad, Janos Sallai, Mike Henry, Jenny W. Fothergill, Peter Volgyesi, Ryan S. DeFever, Chris Iacovella, Ali Morshedifard, Bradley Dice, and Aryan Arora. mosdef-hub/foyer: Foyer 0.10.0 (0.10.0), 2021. URL https://github.com/mosdef-hub/foyer.

[26] The regents of the University of Michigan. Signac, 2020. URL https://github.com/glotzerlab/signac.

[27] Joshua A. Anderson, Bradley Dice, Jens Glaser, Tim Moore, Vyas Ramasubramani, Bryan VanSaders, Mike Henry, Jenny Fothergill, and Corwin Kerr. Fresnel, 2016. URL https://fresnel.readthedocs.io/en/stable/.

[28] The regents of the University of Michigan. HOOMD-blue, 2020. URL https: //github.com/glotzerlab/hoomd-blue/.

[29] Jenny W. Fothergill, Andres C. Hernandez, William B. Knowlton, Bernard Yurke, and Lan Li. Ab Initio Studies of Exciton Interactions of Cy5 Dyes. *The Journal of Physical Chemistry A*, 21, 2018. ISSN 1089-5639. doi: 10.1021/acs.jpca.8b05237.

[30] Shaun Millard, Jenny W. Fothergill, Zoe Anderson, Eric C. Brown, Matthew D. King, and Adam C. Colson. Supramolecular Interactions of Group VI Metal Carbonyl Complexes: The Facilitating Role of 1,3-Bis(p-isocyanophenyl)urea. *Inorganic Chemistry*, 58(12):8130–8139, 2019. ISSN 1520510X. doi: 10.1021/acs.inorgchem.9b00917.

[31] Jenny W. Fothergill, Adam C. Colson, and Matthew D. King. Cataloguing the Energetic Contributions to the Supramolecular Assembly of para-Substituted N, N'-Diphenylureas and Their Organometallic Derivatives in the Solid State: A Density Functional Theory Approach. *Crystal Growth and Design*, 21(1):563–571, 2021. ISSN 15287505. doi: 10.1021/acs.cgd.0c01386.

[32] Jenny W. Fothergill, Chris D. Jones, and Eric Jankowksi. GIXStapose, 2022. URL `https://github.com/cmelab/gixstapose`.

[33] Jenny Fothergill, Chris Jones, and Eric Jankowski. GIXStapose: An Interactive Morphology and Diffraction Viewer. In *Scipy 2020*, Austin, TX, 2020. URL `https://www.scipy2020.scipy.org/{_}files/ugd/2826fb{_}ca0896fe927d429eb36044795d9e18a0.pdf`.

[34] Michael R. Shirts, Christoph Klein, Jason M. Swails, Jian Yin, Michael K. Gilson, David L. Mobley, David A. Case, and Ellen D. Zhong. Lessons

learned from comparing molecular dynamics engines on the SAMPL5 dataset. *Journal of Computer-Aided Molecular Design*, 31(1):147–161, 2017. ISSN 15734951. doi: 10.1007/s10822-016-9977-1.

[35] Hannes H. Loeffler, Stefano Bosisio, Guilherme Duarte Ramos Matos, Donghyuk Suh, Benoit Roux, David L. Mobley, and Julien Michel. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *Journal of Chemical Theory and Computation*, 14(11):5567–5582, 2018. ISSN 15499626. doi: 10.1021/acs.jctc.8b00544.

[36] Michael Schappals, Andreas Mecklenfeld, Leif Kröger, Vitalie Botan, Andreas Köster, Simon Stephan, Edder J. García, Gabor Rutkai, Gabriele Raabe, Peter Klein, Kai Leonhard, Colin W. Glass, Johannes Lenhard, Jadran Vrabec, and Hans Hasse. Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom. *Journal of Chemical Theory and Computation*, 13(9):4270–4280, 2017. ISSN 15499626. doi: 10.1021/acs.jctc.7b00489.

[37] Jenny W. Fothergill. United-atom P3HT workspace (v0.0.1) [Data set]., 2022. URL `https://doi.org/10.5281/zenodo.5911940`.

[38] William Shockley and Hans J. Queisser. Detailed balance limit of efficiency of p-n junction solar cells. *Journal of Applied Physics*, 32(3):510–519, 1961. ISSN 00218979. doi: 10.1063/1.1736034.

[39] Ming Zhang, Lei Zhu, Guanqing Zhou, Tianyu Hao, Chaoqun Qiu, Zhe Zhao, Qin Hu, Bryon W. Larson, Haiming Zhu, Zaifei Ma, Zheng Tang, Wei

Feng, Yongming Zhang, Thomas P. Russell, and Feng Liu. Single-layered organic photovoltaics with double cascading charge transport pathways: 18% efficiencies. *Nature Communications*, 12(1):1–10, 2021. ISSN 20411723. doi: 10.1038/s41467-020-20580-8. URL http://dx.doi.org/10.1038/s41467-020-20580-8.

[40] Nationational Renewable Energy Laboratory. U.S. Department of Energy. Best Research-Cell Efficiency Chart, 2020.

[41] Evan Miller, Matthew Jones, Michael Henry, Paul Chery, Kyle Miller, and Eric Jankowski. Optimization and Validation of Efficient Models for Predicting Polythiophene Self-Assembly. *Polymers*, 10(12):1305, nov 2018. ISSN 2073-4360. doi: 10.3390/polym10121305. URL http://www.mdpi.com/2073-4360/10/12/1305.

[42] Evan D Miller, Michael M Henry, Matthew Lewis Jones, and Eric Jankowski. Planckton, mar 2019. URL https://doi.org/10.5281/zenodo.2616728.

[43] Joshua A. Anderson, Jens Glaser, and Sharon C. Glotzer. HOOMD-blue: A Python package for high-performance molecular dynamics and hard particle Monte Carlo simulations. *Computational Materials Science*, 173(October 2019): 109363, 2020. ISSN 09270256. doi: 10.1016/j.commatsci.2019.109363. URL https://doi.org/10.1016/j.commatsci.2019.109363.

[44] Ryan S. DeFever. Antefoyer, 2020. URL https://github.com/rsdefever/antefoyer.

[45] Eric Jankowski, Hilary S. Marsh, and Arthi Jayaraman. Computationally

linking molecular features of conjugated polymers and fullerene derivatives to bulk heterojunction morphology. *Macromolecules*, 46(14):5775–5785, jul 2013. ISSN 00249297. doi: 10.1021/ma400724e. URL http://pubs.acs.org/doi/abs/10.1021/ma400724e.

[46] Hilary S Marsh, Eric Jankowski, and Arthi Jayaraman. Controlling the Morphology of Model Conjugated Thiophene Oligomers through Alkyl Side Chain Length, Placement, and Interactions. *Macromolecules*, 47(8):2736–2747, apr 2014. ISSN 0024-9297. doi: 10.1021/ma5000267. URL http://pubs.acs.org/doi/abs/10.1021/ma5000267.

[47] Matthew Lewis Jones and Eric Jankowski. Computationally connecting organic photovoltaic performance to atomistic arrangements and bulk morphology. *Molecular Simulation*, 43(10-11):1–18, mar 2017. ISSN 10290435. doi: 10.1080/08927022.2017.1296958. URL http://dx.doi.org/10.1080/08927022.2017.1296958https://www.tandfonline.com/doi/full/10.1080/08927022.2017.1296958.

[48] Michael M. Henry, Matthew Lewis Jones, Stefan D. Oosterhout, Wade A. Braunecker, Travis W. Kemper, Ross E. Larsen, Nikos Kopidakis, Michael F. Toney, Dana C. Olson, and Eric Jankowski. Simplified Models for Accelerated Structural Prediction of Conjugated Semiconducting Polymers. *The Journal of Physical Chemistry C*, 121(47):26528–26538, nov 2017. ISSN 1932-7447. doi: 10.1021/acs.jpcc.7b09701. URL http://pubs.acs.org/doi/10.1021/acs.jpcc.7b09701http://pubs.acs.org/doi/abs/10.1021/acs.jpcc.7b09701.

[49] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101(5):4177–4189, 1994. ISSN 00219606. doi: 10.1063/1.467468.

[50] G. J. Martyna. Adiabatic path integral molecular dynamics methods. I. Theory. *Journal of Chemical Physics*, 104(5):2018–2027, 1996. ISSN 00219606. doi: 10.1063/1.470958.

[51] Gonzalo Nicolas-Barreales, Aaron Sujar, and Alberto Sanchez. A web-based tool for simulating molecular dynamics in cloud environments. *Electronics (Switzerland)*, 10(2):1–18, 2021. ISSN 20799292. doi: 10.3390/electronics10020185.

[52] Joshua A. Rackers, Zhi Wang, Chao Lu, Marie L. Laury, Louis Lagardère, Michael J. Schnieders, Jean Philip Piquemal, Pengyu Ren, and Jay W. Ponder. Tinker 8: Software Tools for Molecular Design. *Journal of Chemical Theory and Computation*, 14(10):5273–5289, 2018. ISSN 15499626. doi: 10.1021/acs.jctc.8b00529.

[53] Victor Rühle, Christoph Junghans, Alexander Lukyanov, Kurt Kremer, and Denis Andrienko. Versatile Object-Oriented Toolkit for Coarse-Graining Applications. *Journal of Chemical Theory and Computation*, 5(12):3211–3223, dec 2009. ISSN 1549-9618. doi: 10.1021/ct900369w. URL `http://pubs.acs.org/doi/abs/10.1021/ct900369w`.

[54] Evan D Miller, Michael M Henry, and Eric Jankowski. Diffractometer, aug 2018. URL `https://zenodo.org/record/1340716`.

[55] Sangwon Ko, Eric T Hoke, Laxman Pandey, Sanghyun Hong, Rajib Mondal, Chad Risko, Yuanping Yi, Rodrigo Noriega, Michael D McGehee, Jean-Luc Brédas, Alberto Salleo, and Zhenan Bao. Controlled conjugated backbone twisting for an increased open-circuit voltage while having a high short-circuit current in poly(hexylthiophene) derivatives. *Journal of the American Chemical Society*, 134(11):5222–5232, mar 2012. ISSN 1520-5126. doi: 10.1021/ja210954r. URL http://www.ncbi.nlm.nih.gov/pubmed/22385287.

[56] Duc T. Duong, Chenchen Wang, Erin Antono, Michael F. Toney, and Alberto Salleo. The chemical and structural origin of efficient p-type doping in P3HT. *Organic Electronics*, 14(5):1330–1336, 2013. ISSN 15661199. doi: 10.1016/j.orgel.2013.02.028. URL http://dx.doi.org/10.1016/j.orgel.2013.02.028.

[57] Andrea Rizzi, Travis Jensen, David R. Slochower, Matteo Aldeghi, Vytautas Gapsys, Dimitris Ntekoumes, Stefano Bosisio, Michail Papadourakis, Niel M. Henriksen, Bert L. de Groot, Zoe Cournia, Alex Dickson, Julien Michel, Michael K. Gilson, Michael R. Shirts, David L. Mobley, and John D. Chodera. *The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations*, volume 34. Springer International Publishing, 2020. ISBN 0123456789. doi: 10.1007/s10822-020-00290-5. URL https://doi.org/10.1007/s10822-020-00290-5.

[58] Jindal K. Shah, Eliseo Marin-Rimoldi, Ryan Gotchy Mullen, Brian P. Keene, Sandip Khan, Andrew S. Paluch, Neeraj Rai, Lucienne L. Romanielo, Thomas W Rosch, Brian Yoo, Edward J. Maginn, Brian Yoo Thomas W.

Rosch, and Edward J. Maginn. Cassandra: An Open Source Monte Carlo Package for Molecular Simulation. *Journal of Computational Chemistry*, 38(19):1727–1739, jul 2017. ISSN 01928651. doi: 10.1002/jcc.24807. URL http://dx.doi.org/10.1002/jcc.24807https://www.ncbi.nlm.nih.gov/pubmed/28436594http://doi.wiley.com/10.1002/jcc.24807.

[59] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilard Pall, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2015.06.001. URL https://www.sciencedirect.com/science/article/pii/S2352711015000059.

[60] Szilard Pall, Mark James Abraham, Carsten Kutzner, Berk Hess, and Erik Lindahl. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS BT - Solving Software Challenges for Exascale. pages 3–27, Cham, 2015. Springer International Publishing. ISBN 978-3-319-15976-8.

[61] Sander Pronk, Szilard Pall, Roland Schulz, Per Larsson, Pär Par Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, Erik Lindahl, Rossen Apostolov, Per Larsson, Peter M. Kasson, Roland Schulz, Sander Pronk, Erik Lindahl, Szilárd Páll, David van der Spoel, Michael R. Shirts, and Pär Par Bjelkmar. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013. ISSN 1367-

4803. doi: 10.1093/bioinformatics/btt055. URL https://doi.org/10.1093/bioinformatics/btt055.

[62] Erik Lindahl, Berk Hess, and David van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, 7(8):306–317, 2001. ISSN 0948-5023. doi: 10.1007/s008940100045. URL https://doi.org/10.1007/s008940100045.

[63] H J C Berendsen, D van der Spoel, and R van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1):43–56, 1995. ISSN 0010-4655. doi: https://doi.org/10.1016/0010-4655(95)00042-E. URL https://www.sciencedirect.com/science/article/pii/001046559500042E.

[64] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman J C Berendsen. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, 2005. doi: https://doi.org/10.1002/jcc.20291. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20291.

[65] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, mar 2008. ISSN 1549-9618. doi: 10.1021/ct700301q. URL https://doi.org/10.1021/ct700301q.

[66] Trung Dac Nguyen, Carolyn L Phillips, Joshua a. Anderson, and

Sharon C Glotzer. Rigid Body Constraints Realized in Massively-parallel Molecular Dynamics on Graphics Processing Units. *Computer Physics Communications*, 182(11):2307–2313, nov 2011. ISSN 00104655. doi: 10.1016/j.cpc.2011.06.005. URL http://www.sciencedirect.com/science/article/pii/S0010465511002153http://linkinghub.elsevier.com/retrieve/pii/S0010465511002153http://dx.doi.org/10.1016/j.cpc.2011.06.005.

[67] Jens Glaser, Xun Zha, Joshua A. Anderson, Sharon C. Glotzer, and Alex Travesset. Pressure in rigid body molecular dynamics. *Computational Materials Science*, 173(November 2019):109430, 2020. ISSN 09270256. doi: 10.1016/j.commatsci.2019.109430. URL https://doi.org/10.1016/j.commatsci.␣2019.109430.

[68] David N. Lebard, Benjamin G. Levine, Philipp Mertmann, Stephen A. Barr, Arben Jusufi, Samantha Sanders, Michael L. Klein, Athanas-sios Z. Panagiotopoulos, A Barr, Arben Jusufi, Samantha Sanders, and L Klein. Self-assembly of coarse-grained ionic surfactants accelerated by graphics processing units. *Soft Matter*, 8(8):2385–2397, 2012. ISSN 1744-683X. doi: 10.1039/C1SM06787G. URL http://xlink.rsc.org/?DOI=C1SM06787Ghttp://xlink.rsc.org/?DOI=c1sm06787ghttp://dx.doi.org/10.1039/C1SM06787G.

[69] A P Thompson, H M Aktulga, R Berger, D S Bolintineanu, W M Brown, P S Crozier, P J in 't Veld, A Kohlmeyer, S G Moore, T D Nguyen, R Shan, M J Stevens, J Tranchida, C Trott, and S J Plimpton. {LAMMPS} - a flexible

simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022. doi: 10.1016/j.cpc.2021.108171.

[70] Yangzesheng Sun, Robert F Dejaco, and J Ilja Siepmann. Deep neural network learning of complex binary sorption equilibria from molecular simulation data. *Chem. Sci.*, 10(16):4377–4388, 2019. ISSN 2041-6520. doi: 10.1039/C8SC05340E. URL http://dx.doi.org/10.1039/C8SC05340E.

[71] Tyler R Josephson, Ramanish Singh, Mona S Minkara, Evgenii O Fetisov, and J Ilja Siepmann. Partial molar properties from molecular simulation using multiple linear regression. *Molecular Physics*, 117(23-24):3589–3602, dec 2019. ISSN 0026-8976. doi: 10.1080/00268976.2019.1648898. URL https://doi.org/10.1080/00268976.2019.1648898.

[72] Marcus G. Martin and J. Ilja Siepmann. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *Journal of Physical Chemistry B*, 102(14):2569–2577, 1998. ISSN 15206106. doi: 10.1021/jp972543+.

[73] M. Yiannourakou, Ph Ungerer, V. Lachet, B. Rousseau, and J. M. Teuler. United atom forcefield for vapor-liquid equilibrium (VLE) properties of cyclic and polycyclic compounds from Monte Carlo simulations. *Fluid Phase Equilibria*, 481:28–43, 2019. ISSN 03783812. doi: 10.1016/j.fluid.2018.07.001. URL https://doi.org/10.1016/j.fluid.2018.07.001.

[74] H. J.C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987. ISSN 00223654. doi: 10.1021/j100308a038.

[75] William L. Jorgensen and Julian. Tirado-Rives. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, mar 1988. ISSN 0002-7863. doi: 10.1021/ja00214a001. URL http://pubs.acs.org/doi/abs/10.1021/ja00214a001.

[76] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993. ISSN 00219606. doi: 10.1063/1.464397.

[77] Carl S. Adorf, Paul M. Dodd, Vyas Ramasubramani, and Sharon C. Glotzer. Simple data and workflow management with the signac framework. *Computational Materials Science*, 146(C):220–229, 2018. ISSN 09270256. doi: 10.1016/j.commatsci.2018.01.035. URL https://doi.org/10.1016/j.commatsci.2018.01.035.

[78] Carl Simon Adorf, Vyas Ramasubramani, Bradley D Dice, Michael M Henry, Paul M Dodd, and Sharon C Glotzer. glotzerlab/signac, feb 2019. URL https://doi.org/10.5281/zenodo.2581327.

[79] Vyas Ramasubramani, Carl S Adorf, Paul M Dodd, Bradley D Dice, and Sharon C Glotzer. {s}ignac: {A} {P}ython framework for data and workflow management. In *Proceedings of the 17th Python in Science Conference*, pages 152–159, 2018. doi: 10.25080/Majora-4af1f417-016.

[80] Bradley D Dice, Brandon L Butler, Vyas Ramasubramani, Alyssa Travitz, Michael M Henry, Hardik Ojha, Kelly L Wang, Carl S Adorf, Eric Jankowski,

and Sharon C Glotzer. {s}ignac: {D}ata {M}anagement and {W}orkflows for {C}omputational {R}esearchers. In *Proceedings of the 20th Python in Science Conference*, pages 23–32, 2021. doi: 10.25080/majora-1b6fd038-003.

[81] John D. Chodera, William C. Swope, Jed W. Pitera, Chaok Seok, and Ken A. Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *Journal of Chemical Theory and Computation*, 3(1):26–41, 2007. ISSN 15499618. doi: 10.1021/ct0502864.

[82] John D. Chodera. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *Journal of Chemical Theory and Computation*, 12(4): 1799–1805, apr 2016. ISSN 1549-9618. doi: 10.1021/acs.jctc.5b00784. URL `https://pubs.acs.org/doi/10.1021/acs.jctc.5b00784`.

[83] Justin B. Gilmer, Ramanish Singh, Ryan W. Smith, Nicholas Craven, Jenny W. Fothergill, Co D. Quach, Brad Crawford, Eliseo Marin-Rimoldi, Timothy C. Moore, Zijie Wu, and Chris D. Jones. Reproducibility Study, 2022. URL `https://github.com/mosdef-hub/reproducibility{_}study`.

[84] José Mario Martínez and Leandro Martínez. Packing Optimization for Automated Generation of Complex System ' s Initial Configurations for Molecular Dynamics and Docking. *Journal of Computational Chemistry*, 24(7):819–825, 2003. doi: 10.1002/jcc.10216.

[85] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry*, 30(13):

2157–2164, oct 2009. ISSN 01928651. doi: 10.1002/jcc.21224. URL http://repositorio.uta.edu.ec/bitstream/123456789/27787/1/AL674.pdfhttp://doi.wiley.com/10.1002/jcc.21224.

[86] D Frenkel and B Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science. Elsevier Science, 2001. ISBN 9780080519982. URL https://books.google.com/books?id=5qTzldS9ROIC.

[87] H. Sun. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. *The Journal of Physical Chemistry B*, 102(38):7338–7364, 1998. ISSN 1520-6106. doi: 10.1021/jp980939v.

[88] Eric W. Lemmon, Ian H. Bell, Marcia L. Huber, and Mark O. McLinden. Thermophysical Properties of Fluid Systems. In P.J. Linstrom and W.G. Mallard, editors, *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology, Gaithersburg MD, 20899, 2021. doi: https://doi.org/10.18434/T4D303.

[89] Aidan P. Thompson, Steven J. Plimpton, and William Mattson. General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions. *Journal of Chemical Physics*, 131(15), 2009. ISSN 00219606. doi: 10.1063/1.3245303.

[90] Evan Miller, Matthew Jones, and Eric Jankowski. Tying Together Multiscale Calculations for Charge Transport in P3HT: Structural Descriptors, Morphology, and Tie-Chains. *Polymers*, 10(12):1358, dec 2018. ISSN 2073-4360. doi: 10.3390/polym10121358. URL http://www.mdpi.com/2073-4360/10/12/1358.

[91] Chris D. Jones, Jenny W. Fothergill, Rainier Barrett, and Eric Jankowski. PolyBinder, 2022. URL https://github.com/cmelab/polybinder.

[92] Scott E Page. *The Model Thinker: What You Need to Know to Make Data Work for You*. Hachette UK, 2018. ISBN 9780465094639.

[93] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604): 452–454, may 2016. ISSN 0028-0836. doi: 10.1038/533452a. URL http://www.nature.com/doifinder/10.1038/533452a.

[94] Megan Smith. Computer Science for All, 2016. URL https://www.whitehouse.gov/blog/2016/01/30/computer-science-all.

[95] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkens-Diehr. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering*, 16(5): 62–74, sep 2014. ISSN 1521-9615. doi: 10.1109/MCSE.2014.80. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6866038.

[96] Julie Wernert and Lorna Rivera. 2018 eXtreme Science and Engineering Discovery Environment (XSEDE) Annual User Satisfaction Survey Evaluation Report Prepared for : John Towns Prepared by Lizanne DeStefano Table of Contents. Technical Report September, 2018.

[97] Tracy K. Teal, Karen A. Cranston, Hilmar Lapp, Ethan P. White, Greg Wilson, Karthik Ram, and Aleksandra Pawlik. Data Carpentry: Workshops to

Increase Data Literacy for Researchers. *International Journal of Digital Curation*, 10:135–143, 2015. ISSN 1746-8256. doi: 10.2218/ijdc.v10i1.351. URL http://www.ijdc.net/index.php/ijdc/article/view/351.

[98] Aliaksei S Vasilevich, Aurélie Carlier, Jan de Boer, and Shantanu Singh. How Not To Drown in Data: A Guide for Biomaterial Engineers. *Trends in Biotechnology*, 35(8):743–755, aug 2017. ISSN 01677799. doi: 10.1016/j.tibtech.2017.05.007. URL http://dx.doi.org/10.1016/j.tibtech.2017.05.007https://linkinghub.elsevier.com/retrieve/pii/S0167779917301270.

[99] Efrem Braun, Justin Gilmer, Heather B. Mayes, David L. Mobley, Jacob I. Monroe, Samarjeet Prasad, and Daniel M. Zuckerman. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, 1(1), 2019. ISSN 25756524. doi: 10.33011/livecoms.1.1.5957. URL https://bit.ly/2HPkcmv.

[100] D. Frenkel. Simulations: The dark side. *The European Physical Journal Plus*, 128(1):10, jan 2013. ISSN 2190-5444. doi: 10.1140/epjp/i2013-13010-8. URL http://link.springer.com/10.1140/epjp/i2013-13010-8.

[101] John E Stone, David J Hardy, Ivan S Ufimtsev, and Klaus Schulten. GPU-accelerated molecular modeling coming of age. *Journal of Molecular Graphics and Modelling*, 29(2):116–125, sep 2010. ISSN 10933263. doi: 10.1016/j.jmgm.2010.06.010. URL http://dx.doi.org/10.1016/j.jmgm.2010.06.010https://linkinghub.elsevier.com/retrieve/pii/S1093326310000914.

[102] Joshua A. Anderson, Chris D. Lorenz, and A. Travesset. General Purpose Molecular Dynamics Simulations Fully Implemented on Graphics

Processing Units. *Journal of Computational Physics*, 227(10):5342–5359, may 2008. ISSN 00219991. doi: 10.1016/j.jcp.2008.01.047. URL http://www.sciencedirect.com/science/article/pii/S0021999108000818http://linkinghub.elsevier.com/retrieve/pii/S0021999108000818.

[103] Christian Robert Trott. *LammpsCuda-a new GPU accelerated Molecular Dynamics Simulations Package and its Application to Ion-Conducting Glasses*. PhD thesis, Technische Universitaet Ilmenau, 2011.

[104] Joshua A Anderson and Sharon C Glotzer. The development and expansion of HOOMD-blue through six years of GPU proliferation. *arXiv*, 1308.5587, aug 2013. URL http://arxiv.org/abs/1308.5587.

[105] Jens Glaser, Trung Dac Nguyen, Joshua A. Anderson, Pak Lui, Filippo Spiga, Jaime A. Millan, David C. Morse, and Sharon C. Glotzer. Strong Scaling of General-purpose Molecular Dynamics Simulations on GPUs. *Computer Physics Communications*, 192:97–107, dec 2014. ISSN 00104655. doi: 10.1016/j.cpc.2015.02.028. URL http://arxiv.org/abs/1412.3387http://dx.doi.org/10.1016/j.cpc.2015.02.028http://linkinghub.elsevier.com/retrieve/pii/S0010465515000867.

[106] Mo Zheng, Xiaoxia Li, and Li Guo. Algorithms of GPU-enabled reactive force field (ReaxFF) molecular dynamics. *Journal of molecular graphics & modelling*, 41C(2010):1–11, apr 2013. ISSN 1873-4243. doi: 10.1016/j.jmgm.2013.02.001. URL http://www.ncbi.nlm.nih.gov/pubmed/23454611.

[107] Greg Wilson. Software Carpentry: lessons learned. *F1000Research*, 3:62,

2014. ISSN 2046-1402. doi: 10.12688/f1000research.3-62.v1. URL http://f1000research.com/articles/3-62/v1.

[108] Susan A Ambrose, Michael W Bridges, Michele DiPietro, Marsha C Lovett, and Marie K Norman. *How Learning Works: Seven research-based principles for smart teaching*, volume 48. John Wiley & Sons, 2010. ISBN 9780470484104. doi: 10.1002/mop.21454. URL https://firstliteracy.org/wp-content/uploads/2015/07/How-Learning-Works.pdf.

[109] Daniel Kahneman. *Attention and effort*. Citeseer, 1973.

[110] Sapna Cheryan, Victoria C Plaut, Paul G Davies, and Claude M Steele. Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of personality and social psychology*, 97(6):1045, 2009.

[111] Jenessa R Shapiro and Amy M Williams. The Role of Stereotype Threats in Undermining Girls' and Women's Performance and Interest in STEM Fields. *Sex Roles*, 66(3-4):175–183, feb 2012. ISSN 0360-0025. doi: 10.1007/s11199-011-0051-0. URL http://link.springer.com/10.1007/s11199-011-0051-␣0.

[112] Jeffrey N. Schinske, Heather Perkins, Amanda Snyder, and Mary Wyer. Scientist spotlight homework assignments shift students' stereotypes of scientists and enhance science identity in a diverse introductory science class. *CBE Life Sciences Education*, 15(3):1–18, 2016. ISSN 19317913. doi: 10.1187/cbe.16-1-0002.

[113] Austin Cory Bart, Allie Sarver, Michael Friend, and Larry Cox II. Python-

Sneks: An Open-Source, Instructionally-Designed Introductory Curriculum with Action-Design Research. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education - SIGCSE '19*, pages 307–313, 2019. doi: 10.1145/3287324.3287428. URL http://dl.acm.org/citation.cfm?doid=3287324.3287428.

[114] Michelle Patrick Cook. Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education*, 90(6):1073–1091, nov 2006. ISSN 0036-8326. doi: 10.1002/sce.20164. URL http://onlinelibrary.wiley.com/doi/10.1002/sce.20164/abstracthttp://doi.wiley.com/10.1002/sce.20164.

[115] Alexandra Simperler and Greg Wilson. Software Carpentry get more done in less time. *Arxiv preprint*, jun 2015. URL http://arxiv.org/abs/1506.02575.

[116] Beth M Duckles. Value of Software Carpentry to Instructors Report. Technical Report January, Foundation, Software Carpentry Foundation, Data Carpentry, Portland, OR, 2016.

[117] Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. Integrating cognitive load theory and concepts of human-computer interaction. *Computers in Human Behavior*, 26(6):1278–1288, 2010. ISSN 07475632. doi: 10.1016/j.chb.2010.05.031. URL http://dx.doi.org/10.1016/j.chb.2010.05.031.

[118] Shawn A. Bohner and Sriram Mohan. Model-based engineering of software: Three productivity perspectives. *Proceedings - 33rd Annual IEEE Software En-*

gineering Workshop, SEW-33 2009, pages 35–44, 2010. doi: 10.1109/SEW.2009.19.

[119] Gergely Varga, Akos Ledeczi, Christopher R Iacovella, Peter T Cummings, Janos Sallai, Akos Ledeczi, Christopher R Iacovella, Clare McCabe, and Peter T Cummings. Enabling Cross-Domain Collaboration in Molecular Dynamics Workflows. *The Fourth International Conference on Advanced Collaborative Networks, Systems and Applications (COLLA 2014)*, pages 41–47, 2014.

[120] Daniel M. Oppenheimer. Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, 20(2):139–156, mar 2006. ISSN 0888-4080. doi: 10.1002/acp.1178. URL `http://doi.wiley.com/10.1002/acp.1178`.

[121] Jay W. Ponder and David A. Case. Protein Simulations. *Advances in Protein Chemistry*, 66:27–85, 2003. ISSN 00653233. doi: 10.1016/S0065-3233(03)66002-X. URL `http://www.sciencedirect.com/science/article/pii/S006532330366002X`.

[122] Murray S. Daw and M. I. Baskes. Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals. *Physical Review Letters*, 50(17): 1285–1288, 1983. ISSN 00319007. doi: 10.1103/PhysRevLett.50.1285.

[123] Norman L. Allinger, Kuohsiang Chen, and Jenn Huei Lii. An improved force field (MM4) for saturated hydrocarbons. *Journal of Computational Chemistry*, 17(5-6):642–668, 1996. ISSN 01928651. doi: 10.1002/(SICI)1096-987X(199604)17:5/6<642::AID-JCC6>3.0.CO;2-U.

[124] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general Amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004. ISSN 01928651. doi: 10.1002/jcc.20035.

[125] Samuel J. Keasler, Sophia M. Charan, Collin D. Wick, Ioannis G. Economou, and J. Ilja Siepmann. Transferable potentials for phase equilibria-united atom description of five- and six-membered cyclic alkanes and ethers. *Journal of Physical Chemistry B*, 116(36):11234–11246, 2012. ISSN 15205207. doi: 10.1021/jp302975c.

[126] Thomas A Halgren and Wolfgang Damm. Polarizable force fields. *Current Opinion in Structural Biology*, 11(2):236–242, apr 2001. ISSN 0959440X. doi: 10.1016/S0959-440X(00)00196-2. URL http://ac.els-cdn.com/S0959440X00001962/1-s2.0-S0959440X00001962-main.pdf?{_}tid=3e24fd1e-415c-11e7-bada-00000aacb35e{&}acdnat=1495725195{_}b9be23eab111032341bb59fde8955297.

[127] Eric von Hippel and Georg von Krogh. Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science. *Organization Science*, 14(2):209–223, apr 2003. ISSN 1047-7039. doi: 10.1287/orsc.14.2.209.14992. URL http://pubsonline.informs.org/doi/abs/10.1287/orsc.14.2.209.14992.

[128] Molly McLure Wasko and Samer Faraj. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly*, 29(1):35, 2005. ISSN 02767783. doi:

10.2307/25148667. URL http://www.jstor.org/stable/25148667https://www.jstor.org/stable/10.2307/25148667.

[129] Eric Jankowski. AIChE 2018 CoMSEF Hands-On Workshop Molecular Modeling Toolbox, 2018.

[130] Daniel S Katz, Lois Curfman McInnes, David E Bernholdt, Abigail Cabunoc Mayes, Neil P. Chue Hong, Jonah Duckles, Sandra Gesing, Michael A Heroux, Simon Hettrick, Rafael C. Jimenez, Marlon Pierce, Belinda Weaver, and Nancy Wilkins-Diehr. Community Organizations: Changing the Culture in Which Research Software Is Developed and Sustained. *Computing in Science & Engineering*, 21(2):8–24, mar 2019. ISSN 1521-9615. doi: 10.1109/MCSE.2018.2883051. URL https://ieeexplore.ieee.org/document/8565942/.

[131] Christoph Klein, Andrew Z. Summers, Matthew W. Thompson, Justin Gilmer, Clare McCabe, Peter T. Cummings, Janos Sallai, and Christopher R. Iacovella. Formalizing Atom-typing and the Dissemination of Force Fields with Foyer. *arXiv*, 167(May):1–38, 2018. ISSN 0927-0256. doi: 10.1016/j.commatsci.2019.05.026. URL http://arxiv.org/abs/1812.06779.

[132] Eric S. Harper, Matthew Spellings, Joshua Anderson, and Sharon C. Glotzer. harperic/freud: Zenodo DOI release, nov 2016. URL https://zenodo.org/record/166564.

[133] Christoph Klein, János Sallai, Trevor J. Jones, Christopher R. Iacovella, Clare McCabe, and Peter T. Cummings. A Hierarchical, Component Based Approach to Screening Properties of Soft Matter. In *Foundations of Molecular*

Modeling and Simulation, Molecular Modeling and Simulation: Applications and Perspectives, pages 79–92. 2016. doi: 10.1007/978-981-10-1128-3_5. URL http://link.springer.com/10.1007/978-981-10-1128-3{_}5.

[134] Richard Gowers, Max Linke, Jonathan Barnoud, Tyler Reddy, Manuel Melo, Sean Seyler, Jan Domański, David Dotson, Sébastien Buchoux, Ian Kenney, and Oliver Beckstein. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016. ISBN 1726-4642 (Electronic) 1726-4634 (Linking). doi: 10.25080/Majora-629e541a-00e. URL https://conference.scipy.org/proceedings/scipy2016/oliver{_}beckstein.html.

[135] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 109(8):1528–1532, 2015. ISSN 15420086. doi: 10.1016/j.bpj.2015.08.015.

[136] M. L. Jones. MorphCT - DOI: 10.5281/zenodo.1243843, 2018. URL https://doi.org/10.5281/zenodo.1243843.

[137] Alexander Stukowski. Visualization and analysis of atomistic simulation data with OVITO-the Open Visualization Tool. *Modelling and Simulation in Materials Science and Engineering*, 18(1), 2010. ISSN 09650393. doi: 10.1088/0965-0393/18/1/015012.

[138] Pascal T Merz and Michael R Shirts. Testing for physical validity in molecular simulations. *PLOS ONE*, 13(9):1–21, 2018. doi: 10.1371/journal.pone.0202764. URL https://doi.org/10.1371/journal.pone.0202764.

[139] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, feb 2014. ISSN 00104655. doi: 10.1016/j.cpc.2013.09.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0010465513003196.

[140] Michael T. Humbert, Yong Zhang, and Edward J. Maginn. PyLAT: Python LAMMPS Analysis Tools. *Journal of Chemical Information and Modeling*, page acs.jcim.9b00066, 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00066. URL http://pubs.acs.org/doi/10.1021/acs.jcim.9b00066.

[141] M. L. Jones. Rhaco - DOI: 10.5281/zenodo.1308187, 2018. URL https://doi.org/10.5281/zenodo.1308187.

[142] Hythem Sidky, Yamil J. Colón, Julian Helfferich, Benjamin J. Sikora, Cody Bezik, Weiwei Chu, Federico Giberti, Ashley Z. Guo, Xikai Jiang, Joshua Lequieu, Jiyuan Li, Joshua Moller, Michael J. Quevillon, Mohammad Rahimi, Hadi Ramezani-Dakhel, Vikramjit S. Rathee, Daniel R. Reid, Emre Sevgen, Vikram Thapar, Michael A. Webb, Jonathan K. Whitmer, and Juan J. De Pablo. SSAGES: Software Suite for Advanced General Ensemble Simulations. *Journal of Chemical Physics*, 148(4), 2018. ISSN 00219606. doi: 10.1063/1.5008853.

[143] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, feb 1996. ISSN 02637855. doi: 10.1016/0263-7855(96)00018-5. URL http://www.sciencedirect.com/science/article/pii/0263785596000185http://linkinghub.elsevier.com/retrieve/pii/0263785596000185.

[144] Mary Jo Foley. GitHub to give users of its free plan access to unlimited private repositories, 2019. URL https://zd.net/2VaIrhY.

[145] V Käfer, D Graziotin, I Bogicevic, S Wagner, and J Ramadani. Poster: Communication in Open-Source Projects–End of the E-mail Era? In *2018 IEEE/ACM 40th International Conference on Software Engineering: Companion (ICSE-Companion)*, pages 242–243, may 2018.

[146] Project Jupyter, Matthias Bussonnier, Jessica Forde, Jeremy Freeman, Brian Granger, Tim Head, Chris Holdgraf, Kyle Kelley, Gladys Nalvarte, Andrew Osheroff, M Pacer, Yuvi Panda, Fernando Perez, Benjamin Ragan-Kelley, and Carol Willing. {B}inder 2.0 - {R}eproducible, interactive, sharable environments for science at scale. In Fatih Akici, David Lippa, Dillon Niederhut, and M Pacer, editors, *{P}roceedings of the 17th {P}ython in {S}cience {C}onference*, pages 113–120, 2018. doi: 10.25080/Majora-4af1f417-011.

[147] Andrew Z Summers, Chrisiacovella, and Justin Gilmer. mosdef-hub/mosdef_tutorials: MoSDeF Tutorials 0.0.0, mar 2019. URL https://doi.org/10.5281/zenodo.2617384.

[148] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):1–20, 2017. doi:

10.1371/journal.pone.0177459. URL https://doi.org/10.1371/journal.pone.0177459.

[149] Dirk Merkel. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.*, 2014(239), mar 2014. ISSN 1075-3583. URL http://dl.acm.org/citation.cfm?id=2600239.2600241.

[150] Tom Crick, Benjamin A Hall, and Samin Ishtiaq. Reproducibility in Research: Systems, Infrastructure, Culture. *arXiv*, pages 1–12, mar 2015. URL http: //arxiv.org/abs/1503.02388.

[151] Daniel S. Katz, Kyle E. Niemeyer, Sandra Gesing, Lorraine Hwang, Wolfgang Bangerth, Simon Hettrick, Ray Idaszak, Jean Salac, Neil Chue Hong, Santiago Núñez Corrales, Alice Allen, R. Stuart Geiger, Jonah Miller, Emily Chen, Anshu Dubey, and Patricia Lago. Report on the Fourth Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE4). *arXiv*, 1(1):0–3, may 2017. doi: 10.5334/jors.184. URL http://arxiv.org/abs/1705.02607http://dx.doi.org/10.5334/jors.184.

[152] D. Richter, A. Baumgärtner, K. Binder, B. Ewen, and J. B. Hayter. Dynamics of Collective Fluctuations and Brownian Motion in Polymer Melts. *Physical Review Letters*, 47(2):109–113, jul 1981. ISSN 0031-9007. doi: 10.1103/PhysRevLett.47.109. URL https://link.aps.org/doi/10.1103/PhysRevLett.47.109.

[153] Kurt Kremer, Gary S. Grest, Kurt Kremer, Institutfor Festkorperforschung, Forschungszentrum Julich, West Germanjf, Institutfor Physik, Universitat

Mainz, D Mainz, West Germany, and Gary S. Grest. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *The Journal of Chemical Physics*, 92(8):5057–5086, 1990. ISSN 00219606. doi: 10.1063/1.458541.

[154] Thomas E. Gartner and Arthi Jayaraman. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules*, 52:755–786, 2019. ISSN 15205835. doi: 10.1021/acs.macromol.8b01836.

[155] Gregory Voth, editor. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. CRC Press, sep 2008. ISBN 978-1-4200-5955-7. doi: 10.1201/9781420059564. URL `https://www.taylorfrancis.com/books/9781420059564`.

[156] Cecilia Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Current Opinion in Structural Biology*, 18(1):10–15, 2008. ISSN 0959440X. doi: 10.1016/j.sbi.2007.10.005.

[157] Ron Elber. Long-timescale simulation methods. *Current Opinion in Structural Biology*, 15(2):151–156, 2005. ISSN 0959440X. doi: 10.1016/j.sbi.2005.02.004.

[158] John L. Klepeis, Kresten Lindorff-Larsen, Ron O. Dror, and David E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, 19(2):120–127, 2009. ISSN 0959440X. doi: 10.1016/j.sbi.2009.03.004.

[159] Shina C.L. Kamerlin, Spyridon Vicatos, Anatoly Dryga, and Arieh Warshel. Coarse-Grained (Multiscale) Simulations in Studies of Biophysical and

Chemical Systems. *Annual Review of Physical Chemistry*, 62(1):41–64, 2011. ISSN 0066-426X. doi: 10.1146/annurev-physchem-032210-103335.

[160] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, 116(14):7898–7936, 2016. ISSN 15206890. doi: 10.1021/acs.chemrev.6b00163.

[161] Siewert J Marrink and D Peter Tieleman. Perspective on the Martini model. *Chemical Society reviews*, 42(16):6801–6822, aug 2013. ISSN 1460-4744. doi: 10.1039/c3cs60093a. URL http://www.ncbi.nlm.nih.gov/pubmed/23708257.

[162] Jason D Perlmutter and Michael F Hagan. Mechanisms of virus assembly. *Annual Review of Physical Chemistry*, 66:217–239, 2014. ISSN 0066-426X. doi: 10.1146/annurev-physchem-040214-121637. URL http://arxiv.org/abs/1407.3856{%}0Ahttp://dx.doi.org/10.1146/annurev-physchem-040214-121637.

[163] Timothy C. Moore, Christopher R. Iacovella, and Clare McCabe. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *The Journal of Chemical Physics*, 140(22):224104, jun 2014. ISSN 0021-9606. doi: 10.1063/1.4880555. URL http://aip.scitation.org/doi/10.1063/1.4880555http://www.ncbi.nlm.nih.gov/pubmed/24929371.

[164] Lanyuan Lu, James F. Dama, and Gregory a. Voth. Fitting coarse-grained distribution functions through an iterative force-matching method. *Journal of Chemical Physics*, 139, 2013. ISSN 00219606. doi: 10.1063/1.4811667.

[165] Lee-ping Wang, Todd J Martinez, and Vijay S Pande. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The Journal of Physical Chemistry Letters*, 5(11):1885–1891, jun 2014. ISSN 1948-7185. doi: 10.1021/jz500737m. URL http://pubs.acs.org/doi/10.1021/jz500737m.

[166] M. Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *Journal of Chemical Physics*, 129(14), 2008. ISSN 00219606. doi: 10.1063/1.2992060.

[167] Aviel Chaimovich and M. Scott Shell. Relative entropy as a universal metric for multiscale errors. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 81(6):1–4, 2010. ISSN 15502376. doi: 10.1103/PhysRevE.81.060104.

[168] Brooke E Husic and Vijay S Pande. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, 140(7):2386–2396, feb 2018. ISSN 0002-7863. doi: 10.1021/jacs.7b12191. URL http://pubs.acs.org/doi/10.1021/jacs.7b12191.

[169] Andrew L Ferguson. Machine learning and data science in soft materials engineering. *Journal of Physics: Condensed Matter*, 30(4):043002, jan 2018. ISSN 0953-8984. doi: 10.1088/1361-648X/aa98bd. URL http://stacks.iop.org/0964-1726/26/i=10/a=105037?key=crossref.7455f3bb928620d7b120b4c9d6c542f4http://stacks.iop.org/0953-8984/30/i=4/a=043002?key=crossref.8eb5b295cc38868eda810456a0c743d4.

[170] Nieves Espinosa, Markus Hösel, Dechan Angmo, and Frederik C. Krebs. Solar Cells with One-Day Energy Payback for the Factories of the Future.

Energy & Environmental Science, 5(1):5117, 2012. ISSN 1754-5692. doi: 10.1039/c1ee02728j. URL http://xlink.rsc.org/?DOI=C1EE02728J.

[171] Katherine a. Mazzio and Christine K Luscombe. The future of organic photovoltaics. *Chemical Society Reviews*, 44(1):78–90, sep 2015. ISSN 0306-0012. doi: 10.1039/C4CS00227J. URL http://dx.doi.org/10.1039/C4CS00227Jhttp://xlink.rsc.org/?DOI=C4CS00227J.

[172] Koen Vandewal, Scott Himmelberger, and Alberto Salleo. Structural factors that affect the performance of organic bulk heterojunction solar cells. *Macromolecules*, 46:6379–6387, 2013. ISSN 00249297. doi: 10.1021/ma400924b.

[173] Tracey M. Clarke and James R. Durrant. Charge Photogeneration in Organic Solar Cells. *Chemical Reviews*, 110(11):6736–6767, nov 2010. ISSN 00092665. doi: 10.1021/cr900271s. URL http://pubs.acs.org/doi/abs/10.1021/cr900271s.

[174] Thomas F. Harrelson, Adam J. Moulé, and Roland Faller. Modeling organic electronic materials: bridging length and time scales. *Molecular Simulation*, 43 (10-11):730–742, 2017. ISSN 10290435. doi: 10.1080/08927022.2016.1273526. URL http://dx.doi.org/10.1080/08927022.2016.1273526.

[175] Chris Groves. Developing understanding of organic photovoltaic devices: Kinetic Monte Carlo models of geminate and non-geminate recombination, charge transport and charge extraction. *Energy and Environmental Science*, 6 (11):3202–3217, 2013. ISSN 17545692. doi: 10.1039/c3ee41621f.

[176] C. Groves. Simulating charge transport in organic semiconduc-

tors and devices: A review. *Reports on Progress in Physics*, 80(2): 026502, feb 2017. ISSN 00344885. doi: 10.1088/1361-6633/80/2/026502. URL http://stacks.iop.org/0034-4885/80/i=2/a=026502?key=crossref.44d1e9a057340b9d03e5e225c0a84eab.

[177] Jan-Michael Y. Carrillo, Rajeev Kumar, Monojoy Goswami, Bobby G. Sumpter, and W. Michael Brown. New insights into the dynamics and morphology of P3HT:PCBM active layers in bulk heterojunctions. *Physical Chemistry Chemical Physics*, 15(41):17873, nov 2013. ISSN 1463-9076. doi: 10.1039/c3cp53271b. URL http://xlink.rsc.org/?DOI=c3cp53271bhttp://pubs.rsc.org/en/Content/ArticleLanding/2013/CP/c3cp53271bhttp://www.ncbi.nlm.nih.gov/pubmed/24056906.

[178] Margherita Moreno, Mosè Casalegno, Guido Raos, Stefano V Meille, and Riccardo Po. Molecular modeling of crystalline alkylthiophene oligomers and polymers. *The journal of physical chemistry. B*, 114(4):1591–1602, feb 2010. ISSN 1520-5207. doi: 10.1021/jp9106124. URL http://www.ncbi.nlm.nih.gov/pubmed/20058895.

[179] David M. Huang, Roland Faller, Khanh Do, and Adam J. Moule. Coarse-grained computer simulations of polymer/fullerene bulk heterojunctions for organic photovoltaic applications. *Journal of Chemical Theory and Computation*, 6(2):526–537, feb 2010. ISSN 15499618. doi: 10.1021/ct900496t. URL http://pubs.acs.org/doi/abs/10.1021/ct900496thttps://pubs.acs.org/doi/10.1021/ct900496t.

[180] Cheng Kuang Lee, Chun Wei Pao, and Chih Wei Chu. Multiscale molecular

simulations of the nanoscale morphologies of P3HT:PCBM blends for bulk heterojunction organic photovoltaic cells. *Energy and Environmental Science*, 4(10):4124–4132, 2011. ISSN 17545692. doi: 10.1039/c1ee01508g. URL http://xlink.rsc.org/?DOI=c1ee01508g.

[181] Ram S Bhatta, Yeneneh Y Yimer, David S Perry, and Mesfin Tsige. Improved Force Field for Molecular Modeling of Poly(3-hexylthiophene). *The Journal of Physical Chemistry B*, 117(34):10035–10045, aug 2013. ISSN 1520-6106. doi: 10.1021/jp404629a. URL http://www.ncbi.nlm.nih.gov/pubmed/23899343http://pubs.acs.org/doi/abs/10.1021/jp404629a.

[182] Gabriele D'Avino, Sébastien Mothy, Luca Muccioli, Claudio Zannoni, Linjun Wang, Jérôme Jéroîme Cornil, David Beljonne, Frédéric Castet, Gabriele D Avino, Sébastien Mothy, Luca Muccioli, Claudio Zannoni, Linjun Wang, Jérôme Jéroîme Cornil, David Beljonne, and Frédéric Castet. Energetics of electron-hole separation at P3HT/PCBM heterojunctions. *Journal of Physical Chemistry C*, 117(25):12981–12990, 2013. ISSN 19327447. doi: 10.1021/jp402957g.

[183] Orestis Alexiadis and Vlasis G Mavrantzas. All-Atom Molecular Dynamics Simulation of Temperature Effects on the Structural, Thermodynamic, and Packing Properties of the Pure Amorphous and Pure Crystalline Phases of Regioregular P3HT. *Macromolecules*, 46(6):2450–2467, mar 2013. ISSN 0024-9297. doi: 10.1021/ma302211g. URL http://pubs.acs.org/doi/abs/10.1021/ma302211g.

[184] M. L. Jones, D. M. Huang, B. Chakrabarti, and Chris Groves. Relating Molec-

ular Morphology to Charge Mobility in Semicrystalline Conjugated Polymers. *Journal of Physical Chemistry C*, 120(8):4240–4250, 2016. ISSN 19327455. doi: 10.1021/acs.jpcc.5b11511.

[185] C. Scherer and D. Andrienko. Comparison of systematic coarse-graining strategies for soluble conjugated polymers. *European Physical Journal: Special Topics*, 225(8-9):1441–1461, 2016. ISSN 19516401. doi: 10.1140/epjst/e2016-60154-9.

[186] Cristina Greco, Anton Melnyk, Kurt Kremer, Denis Andrienko, and Kostas Ch Daoulas. Generic Model for Lamellar Self-Assembly in Conjugated Polymers: Linking Mesoscopic Morphology and Charge Transport in P3HT. *Macromolecules*, 52:968–981, 2019. ISSN 15205835. doi: 10.1021/acs.macromol.8b01863.

[187] Evan D Miller, Matthew Lewis Jones, and Eric Jankowski. Enhanced Computational Sampling of Perylene and Perylothiophene Packing with Rigid-Body Models. *ACS Omega*, 2(1):353–362, jan 2017. ISSN 2470-1343. doi: 10.1021/acsomega.6b00371. URL `http://pubs.acs.org/doi/abs/10.1021/acsomega.6b00371http://pubs.acs.org/doi/10.1021/acsomega.6b00371`.

[188] Jyh Luen Chen and Feng Chih Chang. Phase Separation Process in Poly(-caprolactone)-Epoxy Blends. *Macromolecules*, 32(2):5348–5356, 1999. ISSN 0024-9297. doi: 10.1021/ma981819o. URL `http://pubs.acs.org/doi/pdf/10.1021/ma981819o`.

[189] Chunyu Li and Alejandro Strachan. Molecular scale simulations on thermoset polymers: A review. *Journal of Polymer Science Part B: Polymer Physics*,

53(2):103–122, jan 2015. ISSN 08876266. doi: 10.1002/polb.23489. URL http://doi.wiley.com/10.1002/polb.23489.

[190] Adri C T van Duin, Siddharth Dasgupta, Francois Lorant, and William A. Goddard. ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A*, 105(41):9396–9409, oct 2001. ISSN 1089-5639. doi: 10.1021/jp004368u. URL http://pubs.acs.org/doi/abs/10.1021/jp004368u.

[191] Stephen Thomas, Monet Alberts, Michael M Henry, Carla E Estridge, and Eric Jankowski. Routine million-particle simulations of epoxy curing with dissipative particle dynamics. *Journal of Theoretical and Computational Chemistry*, 17(03):1840005, may 2018. ISSN 0219-6336. doi: 10.1142/S0219633618400059. URL https://www.worldscientific.com/doi/abs/10.1142/S0219633618400059.

[192] Pavel V Komarov, Chiu Yu-tsung, Chen Shih-ming, Pavel G Khalatur, and Peter Reineker. Highly Cross-Linked Epoxy Resins: An Atomistic Molecular Dynamics Simulation Combined with a Mapping/Reverse Mapping Procedure. *Macromolecules*, 40(22):8104–8113, 2007. doi: 10.1021/ma070702+. URL http://dx.doi.org/10.1021/ma070702+.

[193] Jacob R Gissinger, Benjamin D Jensen, and Kristopher E Wise. Modeling chemical reactions in classical molecular dynamics simulations. *Polymer*, 128:211–217, 2017. ISSN 0032-3861. doi: 10.1016/j.polymer.2017.09.038. URL https://doi.org/10.1016/j.polymer.2017.09.038.

[194] Michael Langeloth, Taisuke Sugii, Michael C. Böhm, and Florian Müller-Plathe. The glass transition in cured epoxy thermosets: A comparative

molecular dynamics study in coarse-grained and atomistic resolution. *The Journal of Chemical Physics*, 143(24):243158, 2015. ISSN 0021-9606. doi: 10.1063/1.4937627. URL http://scitation.aip.org/content/aip/journal/jcp/143/24/10.1063/1.4937627.

[195] Chunyu Li, Grigori A. Medvedev, Eun Woong Lee, Jaewoo Kim, James M. Caruthers, and Alejandro Strachan. Molecular dynamics simulations and experimental studies of the thermomechanical response of an epoxy thermoset polymer. *Polymer (United Kingdom)*, 53(19):4222–4230, 2012. ISSN 00323861. doi: 10.1016/j.polymer.2012.07.026. URL http://dx.doi.org/10.1016/j.polymer.2012.07.026.

[196] Ketan S. Khare and Frederick R. Phelan. Quantitative Comparison of Atomistic Simulations with Experiment for a Cross-Linked Epoxy: A Specific Volume-Cooling Rate Analysis. *Macromolecules*, 51(2):564–575, 2018. ISSN 15205835. doi: 10.1021/acs.macromol.7b01303.

[197] Gokhan Kacar, Elias A J F Peters, and Gijsbertus De With. Multi-scale simulations for predicting material properties of a cross-linked polymer. *Computational Materials Science*, 102:68–77, 2015. ISSN 09270256. doi: 10.1016/j.commatsci.2015.02.021. URL http://dx.doi.org/10.1016/j.commatsci.␣2015.02.021.

[198] Hong Liu, Min Li, Zhong-Yuan Lu, Zuo-Guang Zhang, Chia-Chung Sun, and Tian Cui. Multiscale Simulation Study on the Curing Reaction and the Network Structure in a Typical Epoxy System. *Macromolecules*, 44(21):8650–

8660, nov 2011. ISSN 0024-9297. doi: 10.1021/ma201390k. URL http://pubs.acs.org/doi/abs/10.1021/ma201390k.

[199] Stephen Thomas. *New Methods for Understanding and Controlling the Self-Assembly of Reacting Systems Using Coarse-Grained Molecular Dynamics*. PhD thesis, Boise State University, 2018.

[200] Thomas A. Knotts, Nitin Rathore, David C. Schwartz, and Juan J. de Pablo. A coarse grain model for DNA. *The Journal of Chemical Physics*, 126(8): 084901, feb 2007. ISSN 00219606. doi: 10.1063/1.2431804. URL http: //aip.scitation.org/doi/10.1063/1.2431804.

[201] Herminia Ibarra. Provisional Selves: Experimenting with Image and Identity in Professional Adaptation. *Administrative Science Quarterly*, 44(4):764, dec 1999. ISSN 00018392. doi: 10.2307/2667055. URL https://www.jstor.org/stable/2667055?origin=crossref.

[202] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State by Fast Computing Machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953. doi: 10.1063/1.1699114. URL http://link.aip.org/link/?JCP/21/1087/1.

[203] B.J. Alder and T.E. Wainbright. Phase Transition for a Hard Sphere System. *Journal of Chemical Physics*, 27(1957):1208, 1957. ISSN 00219606. doi: 10.1063/1.1743957.

[204] Brittany L. Cannon, Donald L. Kellis, Lance K. Patten, Paul H. Davis, Jeunghoon Lee, Elton Graugnard, Bernard Yurke, and William B. Knowl-

ton. Coherent Exciton Delocalization in a Two-State DNA-Templated Dye Aggregate System. *The Journal of Physical Chemistry A*, 121(37):6905–6916, sep 2017. ISSN 1089-5639. doi: 10.1021/acs.jpca.7b04344. URL `http://pubs.acs.org/doi/abs/10.1021/acs.jpca.7b04344`.

[205] Gregory S. Engel, Tessa R. Calhoun, Elizabeth L. Read, Tae Kyu Ahn, Tomáš Mančal, Yuan Chung Cheng, Robert E. Blankenship, and Graham R. Fleming. Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature*, 446(7137):782–786, 2007. ISSN 14764687. doi: 10.1038/nature05678.

[206] Elisabetta Collini, Cathy Y. Wong, Krystyna E. Wilk, Paul M.G. Curmi, Paul Brumer, and Gregory D. Scholes. Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature. *Nature*, 463(7281):644–647, 2010. ISSN 00280836. doi: 10.1038/nature08811.

[207] Francesca Fassioli, Rayomond Dinshaw, Paul C. Arpin, and Gregory D. Scholes. Photosynthetic light harvesting: Excitons and coherence. *Journal of the Royal Society Interface*, 11(92), 2014. ISSN 17425662. doi: 10.1098/rsif.2013.0901.

[208] J. Frenkel. On the transformation of light into heat in solids. II. *Physical Review*, 37(10):1276–1294, 1931. ISSN 0031899X. doi: 10.1103/PhysRev.37.17.

[209] A.S. Davydov. The theory of molecular excitons. *Uspekhi Fizicheskih Nauk*, 82 (3):393–448, 1964. ISSN 0042-1294. doi: 10.3367/ufnr.0082.196403a.0393.

[210] A. Eisfeld and J. S. Briggs. The shape of the J-band of pseudoisocyanine.

Chemical Physics Letters, 446(4-6):354–358, 2007. ISSN 00092614. doi: 10.1016/j.cplett.2007.07.110.

[211] A. Eisfeld and J. S. Briggs. The J- and H-bands of organic dye aggregates. *Chemical Physics*, 324(2-3):376–384, 2006. ISSN 03010104. doi: 10.1016/j.chemphys.2005.11.015. URL https://ac.els-cdn.com/S030101040500563X/1-s2.0-S030101040500563X-main.pdf?{_}tid=2d1ca7fa-d6fb-4a50-80ea-de78fdfe0437{&}acdnat=1524852266{_}077988b158caec34ea697d7480e3134f.

[212] Kenneth D. Philipson, Vicki L. Sato, and Kenneth Sauer. Exciton Interaction in the Photosystem I Reaction Center from Spinach Chloroplasts. Absorption and Circular Dichroism Difference Spectra. *Biochemistry*, 11(24):4591–4595, 1972. ISSN 15204995. doi: 10.1021/bi00774a027.

[213] Rudolph J. Marcus and Gilbert R. Haugen. Resonance fluorescense in Chlorophyll a solutions. *Photochemistry and Photobiology*, 4(2):183–192, 1965. ISSN 17511097. doi: 10.1111/j.1751-1097.1965.tb05735.x.

[214] Seth Lloyd. Quantum coherence in biological systems. *Journal of Physics: Conference Series*, 302(1), 2011. ISSN 17426596. doi: 10.1088/1742-6596/302/1/012037.

[215] Cathal Smyth, Francesca Fassioli, and Gregory D. Scholes. Measures and implications of electronic coherence in photosynthetic light-harvesting. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1972):3728–3749, 2012. ISSN 1364503X. doi: 10.1098/rsta.2011.0420.

[216] Fausto Rossi. The Excitonic Quantum Computer. *IEEE Transactions on Nanotechnology*, 3(1):165 – 172, 2004. doi: 10.1109/TNANO.2004.824018. URL http://ieeexplore.ieee.org/ielx5/7729/28578/01278286.pdf?tp={&}arnumber=1278286{&}isnumber=28578.

[217] Andrew M. Childs, David Gosset, and Zak Webb. Universal computation by multiparticle quantum walk. *Science*, 339(6121):791–794, 2013. ISSN 10959203. doi: 10.1126/science.1229957.

[218] Frank Würthner, Theo E. Kaiser, and Chantu R. Saha-Möller. J-Aggregates: From Serendipitous Discovery to Supramolecular Engineering of Functional Dye Materials. *Angewandte Chemie International Edition*, 50(15):3376–3410, apr 2011. doi: 10.1002/anie.201002307. URL http://doi.wiley.com/10.1002/anie.201002307.

[219] Michael Kasha. Energy Transfer Mechanisms and the Molecular Exciton Model for Molecular Aggregates. *Radiat. Res.*, 20(1):55–70, 1963.

[220] Gregory D. Scholes, Tihana Mirkovic, Daniel B. Turner, Francesca Fassioli, and Andreas Buchleitner. Solar light harvesting by energy transfer: From ecology to coherence. *Energy and Environmental Science*, 5(11):9374–9393, 2012. ISSN 17545692. doi: 10.1039/c2ee23013e.

[221] Gregory D. Scholes, Graham R. Fleming, Alexandra Olaya-Castro, and Rienk Van Grondelle. Lessons from nature about solar light harvesting. *Nature Chemistry*, 3(10):763–774, 2011. ISSN 17554330. doi: 10.1038/nchem.1145. URL http://dx.doi.org/10.1038/nchem.1145.

[222] Gregory D. Scholes. Coherence in photosynthesis. *Nature Physics*, 7(6):448–449, 2011. ISSN 17452481. doi: 10.1038/nphys2013. URL http://dx.doi.org/10.1038/nphys2013.

[223] Elisabet Romero, Vladimir I. Novoderezhkin, and Rienk Van Grondelle. Quantum design of photosynthesis for bio-inspired solar-energy conversion. *Nature*, 543(7645):355–365, 2017. ISSN 14764687. doi: 10.1038/nature22012.

[224] Paul W K Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, mar 2006. ISSN 1476-4687. doi: 10.1038/nature04586. URL http://dx.doi.org/10.1038/nature04586http://www.nature.com/nature/journal/v440/n7082/abs/nature04586.htmlhttp://www.ncbi.nlm.nih.gov/pubmed/16541064.

[225] Yonggang Ke, Luvena L. Ong, William M. Shih, and Peng Yin. Three-dimensional structures self-assembled from DNA bricks. *Science*, 338(6111):1177–1183, 2012. ISSN 10959203. doi: 10.1126/science.1227268.

[226] Larysa I. Markova, Vladimir L. Malinovskii, Leonid D. Patsenker, and Robert Häner. J- vs. H-type assembly: Pentamethine cyanine (Cy5) as a near-IR chiroptical reporter. *Chemical Communications*, 49(46):5298–5300, 2013. ISSN 1364548X. doi: 10.1039/c3cc42103a.

[227] Francesca Nicoli, Matthias K. Roos, Elisa A. Hemmig, Marco Di Antonio, Regina De Vivie-Riedle, and Tim Liedl. Proximity-Induced h-Aggregation of cyanine dyes on DNA-Duplexes. *Journal of Physical Chemistry A*, 120(50):9941–9947, 2016. ISSN 15205215. doi: 10.1021/acs.jpca.6b10939.

[228] Loni Kringle, Nicolas P.D. Sawaya, Julia Widom, Carson Adams, Michael G. Raymer, Alán Aspuru-Guzik, and Andrew H. Marcus. Temperature-dependent conformations of exciton-coupled Cy3 dimers in double-stranded DNA. *Journal of Chemical Physics*, 148(8), 2018. ISSN 00219606. doi: 10.1063/1.5020084. URL http://dx.doi.org/10.1063/1.5020084.

[229] Brittany L. Cannon, Lance K. Patten, Donald L. Kellis, Paul H. Davis, Jeunghoon Lee, Elton Graugnard, Bernard Yurke, and William B. Knowlton. Large Davydov Splitting and Strong Fluorescence Suppression: An Investigation of Exciton Delocalization in DNA-Templated Holliday Junction Dye Aggregates. *The Journal of Physical Chemistry A*, page acs.jpca.7b12668, 2018. ISSN 1089-5639. doi: 10.1021/acs.jpca.7b12668. URL http://pubs.acs.org/doi/10.1021/acs.jpca.7b12668.

[230] Oliver Kühn, Thomas Renger, and Volkhard May. Theory of exciton-vibrational dynamics in molecular dimers. *Chemical Physics*, 204:99–114, 1996. URL https://ac.els-cdn.com/0301010495004483/1-s2.0-0301010495004483-main.pdf?{_}tid=81b712fc-4ccc-4893-9d14-62bf9ca395d9{&}acdnat=1526925670{_}5d93c8765391da6a30c8bb81db8ef85b.

[231] V. Czikklely, H. D. Forsterling, and H. Kuhn. Extended dipole model for aggregates of dye molecules. *Chemical Physics Letters*, 6(3):207–210, 1970.

[232] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, and G. A. Petersson. Gaussian 09, 2009.

[233] James J.P. Stewart. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling*, 13(12):1173–1213, 2007. ISSN 16102940. doi: 10.1007/s00894-007-0233-4.

[234] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. ISSN 00219606. doi: 10.1063/1.464913.

[235] Takeshi Yanai, David P. Tew, and Nicholas C. Handy. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chemical Physics Letters*, 393(1-3):51–57, 2004. ISSN 00092614. doi: 10.1016/j.cplett.2004.06.011.

[236] Jeng Da Chai and Martin Head-Gordon. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Physical Chemistry Chemical Physics*, 10(44):6615–6620, 2008. ISSN 14639076. doi: 10.1039/b810189b.

[237] G. A. Petersson, Andrew Bennett, Thomas G. Tensfeldt, Mohammad A. Al-Laham, William A. Shirley, and John Mantzaris. A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *The Journal of Chemical Physics*, 89(4):2193–2218, 1988. ISSN 00219606. doi: 10.1063/1.455064.

[238] G. A. Petersson and Al-Laham Mohammad A. A Complete Basis Set Model Chemistry. II. The Total Energies of open-Shell Atoms and Hydrides of the First-Row Atoms. *J. Chem. Phys.*, 9(September 1990):6081–6090, 1991.

[239] Michał Maj, Jonggu Jeon, Robert W. Góra, and Minhaeng Cho. Induced optical activity of DNA-templated cyanine dye aggregates: Exciton coupling theory and TD-DFT studies. *Journal of Physical Chemistry A*, 117(29):5909–5918, 2013. ISSN 10895639. doi: 10.1021/jp309807y.

[240] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum mechanical continuum solvation models. *Chemical Reviews*, 105(8):2999–3093, 2005. ISSN 00092665. doi: 10.1021/cr9904009.

[241] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society*, 114(25): 10024–10035, 1992. ISSN 15205126. doi: 10.1021/ja00051a040.

[242] Marcus D. Hanwell, Donald E. Curtis, David C. Lonie, Tim Vandermeerschd, Eva Zurek, and Geoffrey R. Hutchison. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(8):1–17, 2012. ISSN 17582946. doi: 10.1186/1758-2946-4-17.

[243] Francisco José Avila Ferrer and Fabrizio Santoro. Comparison of vertical and adiabatic harmonic approaches for the calculation of the vibrational structure of electronic spectra. *Physical Chemistry Chemical Physics*, 14(39):13549–13563, 2012. ISSN 14639076. doi: 10.1039/c2cp41169e.

[244] Carlo Adamo and Denis Jacquemin. The calculations of excited-state properties with time-dependent density functional theory. *Chemical Society Reviews*, 42(3):845–856, 2013. ISSN 14604744. doi: 10.1039/c2cs35394f.

[245] Alberto Baiardi, Julien Bloino, and Vincenzo Barone. General time dependent approach to vibronic spectroscopy including franck-condon, herzberg-teller, and duschinsky effects. *Journal of Chemical Theory and Computation*, 9 (9):4097–4115, 2013. ISSN 15499618. doi: 10.1021/ct400450k.

[246] T. E. Sharp and H. M. Rosenstock. Franck-condon factors for polyatomic molecules. *The Journal of Chemical Physics*, 41(11):3453–3463, 1964. ISSN 00219606. doi: 10.1063/1.1725748.

[247] J. Franck. Elementary Processes of Photochemical Reactions. *Transactions of the Faraday Society*, 21:536–542, 1926. doi: 10.1201/ebk1420076370-c5.

[248] E. Condon. A theory of intensity distribution in band systems. *Physical Review*, 28:1182 – 1201, 1926. doi: doi:10.1103/PhysRev.28.1182.

[249] G. Herzberg and E. Teller. Fluctuation structure of electron transfer in multiatomic molecules. *Z. Phys. Chemie*, 21:410 – 446, 1933.

[250] A. H. Herz. Dye-Dye Interaction of Cyanines in Solution and at AgBr Surfaces. *Photographic Science and Engineering*, 18(3):323 – 335, 1974.

[251] H. Von Berlepsch, C. Böttcher, and L. Dähne. Structure of J-aggregates of pseudoisocyanine dye in aqueous solution. *Journal of Physical Chemistry B*, 104(37):8792–8799, 2000. ISSN 15206106. doi: 10.1021/jp000085q.

[252] Christian Friedl, Thomas Renger, Hans V. Berlepsch, Kai Ludwig, Marcel Schmidt am Busch, and Jörg Megow. Structure Prediction of Self-Assembled Dye Aggregates from Cryogenic Transmission Electron Microscopy, Molec-

ular Mechanics, and Theory of Optical Spectra. *Journal of Physical Chemistry C*, 120(34):19416–19433, 2016. ISSN 19327455. doi: 10.1021/acs.jpcc.6b05856.

[253] Stefan Grimme. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *Journal of computational chemistry*, 27(15):1787–1799, 2006. ISSN 1096-987X. doi: 10.1002/jcc.

[254] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *Journal of computational chemistry*, 32(7):1457–1465, 2011. ISSN 1096-987X. doi: 10.1002/jcc.

[255] M. Born and R. Oppenheimer. On the quantum theory of molecules. In *World Scientific Series in 20th Century Chemistry: Quantum Chemistry*, volume 8, pages 1–24. World Scientific Publishing Co. Pte. Ltd., 2000. doi: 10.1063/1.4755287.

[256] Fabrizio Santoro, Roberto Improta, Alessandro Lami, Julien Bloino, and Vincenzo Barone. Effective method to compute Franck-Condon integrals for optical spectra of large molecules in solution. *Journal of Chemical Physics*, 126 (8), 2007. ISSN 00219606. doi: 10.1063/1.2437197.

[257] Fabrizio Santoro, Alessandro Lami, Roberto Improta, Julien Bloino, and Vincenzo Barone. Effective method for the computation of optical spectra of large molecules at finite temperature including the Duschinsky and Herzberg-Teller effect: The Qx band of porphyrin as a case study. *Journal of Chemical Physics*, 128(22), 2008. ISSN 00219606. doi: 10.1063/1.2929846.

[258] Adèle D. Laurent, Carlo Adamo, and Denis Jacquemin. Dye chemistry

with time-dependent density functional theory. *Physical Chemistry Chemical Physics*, 16(28):14334–14356, 2014. ISSN 14639076. doi: 10.1039/c3cp55336a.

[259] Benoît Champagne, Maxime Guillaume, and Freddy Zutterman. TDDFT investigation of the optical properties of cyanine dyes. *Chemical Physics Letters*, 425:105 – 109, 2006. doi: 10.1016/j.cplett.2006.05.009. URL http://ac.els-cdn.com/S0009261406006555/1-s2.0-S0009261406006555-main.pdf?{_}tid=346444c8-9a4f-11e7-ae0f-00000aab0f26{&}acdnat=1505505243{_}31e161606f613f8f8d2240eef8bc61a3.

[260] Boris Le Guennic and Denis Jacquemin. Taking Up the Cyanine Challenge with Quantum Tools. *Accounts of Chemical Research*, 48(3):530–537, 2015. ISSN 15204898. doi: 10.1021/ar500447q.

[261] Heinz Mustroph and Andrew Towns. Fine Structure in Electronic Spectra of Cyanine Dyes: Are Sub-Bands Largely Determined by a Dominant Vibration or a Collection of Singly Excited Vibrations? *ChemPhysChem*, 19(9):1016–1023, 2018. ISSN 14394235. doi: 10.1002/cphc.201701300. URL http://doi.wiley.com/10.1002/cphc.201701300.

[262] ThermoFisher Scientific. Fluorescence SpectraViewer, 2018.

[263] Dushyant B. Varshey, John R. G. Sander, Tomislav Friščić, and Leonard R. MacGillivray. Supramolecular Interactions. *Supramolecular Chemistry*, 2012. doi: 10.1002/9780470661345.smc003.

[264] Hans-Jörg Schneider. Binding Mechanisms in Supramolecular Complexes. *Angewandte Chemie International Edition*, 48(22):3924–3977, may 2009.

doi: 10.1002/anie.200802947. URL http://doi.wiley.com/10.1002/anie.200802947.

[265] P. A. Gale and J. W. Steed. *Supramolecular chemistry: from molecules to nanomaterials*. John Wiley & Sons, Ltd, Oxford, 2012.

[266] Gautam R. Desiraju. Crystal engineering: A holistic view. *Angewandte Chemie - International Edition*, 46(44):8342–8356, 2007. ISSN 14337851. doi: 10.1002/anie.200700534.

[267] Gautam R. Desiraju. Crystal engineering: From molecule to crystal. *Journal of the American Chemical Society*, 135(27):9952–9967, 2013. ISSN 00027863. doi: 10.1021/ja403264c.

[268] Arijit Mukherjee. Building upon Supramolecular Synthons: Some Aspects of Crystal Engineering. *Crystal Growth & Design*, 15(6):3076–3085, jun 2015. ISSN 1528-7483. doi: 10.1021/acs.cgd.5b00242. URL http://pubs.acs.org/doi/10.1021/acs.cgd.5b00242.

[269] Mérina K. Corpinot and Dejan Krešimir Bučar. A Practical Guide to the Design of Molecular Crystals. *Crystal Growth and Design*, 19(2):1426–1453, 2019. ISSN 15287505. doi: 10.1021/acs.cgd.8b00972.

[270] Rajesh Chakrabarty, Partha Sarathi Mukherjee, and Peter J. Stang. Supramolecular coordination: Self-assembly of finite two- and three-dimensional ensembles. *Chemical Reviews*, 111(11):6810–6918, 2011. ISSN 00092665. doi: 10.1021/cr200077m.

[271] Timothy R. Cook, Yao-Rong Zheng, and Peter J. Stang. Metal-Organic Frameworks and Self-Assembled Supramolecular Coordination Complexes: Comparing and Contrasting the Design, Synthesis and Functionality of Metal-Organic Materials. *Chem Rev*, 113(1):734–777, 2013. ISSN 15378276. doi: 10.1021/cr3002824.Metal-Organic. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

[272] Sougata Datta, Manik Lal Saha, and Peter J. Stang. Hierarchical Assemblies of Supramolecular Coordination Complexes. *Accounts of Chemical Research*, 51(9):2047–2063, 2018. ISSN 15204898. doi: 10.1021/acs.accounts.8b00233.

[273] Shinji Tanaka, Hayato Tsurugi, and Kazushi Mashima. Supramolecular assemblies of multi-nuclear transition metal complexes: Synthesis and redox properties. *Coordination Chemistry Reviews*, 265(1):38–51, 2014. ISSN 00108545. doi: 10.1016/j.ccr.2014.01.012. URL http://dx.doi.org/10.1016/j.ccr.2014.01.012.

[274] Dario Braga, Fabrizia Grepioni, Lucia Maini, and Simone D'Agostino. Making crystals with a purpose; A journey in crystal engineering at the University of Bologna. *IUCrJ*, 4:369–379, 2017. ISSN 20522525. doi: 10.1107/S2052252517005917.

[275] Julio Zukerman-Schpector, Ionel Haiduc, and Edward R.T. Tiekink. The metal-carbonyl...pi(aryl) interaction as a supramolecular synthon for the stabilisation of transition metal carbonyl crystal structures. *Chemical Communications*, 47(47):12682–12684, 2011. ISSN 1364548X. doi: 10.1039/c1cc15579b.

[276] Kai Chi Chang, Shih Sheng Sun, and Alistair J. Lees. Anion sensing by rhenium(I) carbonyls with polarized N-H recognition motifs. *Inorganica Chimica Acta*, 389:16–28, 2012. ISSN 00201693. doi: 10.1016/j.ica.2012.02.001. URL http://dx.doi.org/10.1016/j.ica.2012.02.001.

[277] Arumugam Ramdass, Veerasamy Sathish, Murugesan Velayudham, Pounraj Thanasekaran, Kuang Lieh Lu, and Seenivasan Rajagopal. Monometallic rhenium(I) complexes as sensor for anions. *Inorganic Chemistry Communications*, 35(2013):186–191, 2013. ISSN 13877003. doi: 10.1016/j.inoche.2013.06.24. URL http://dx.doi.org/10.1016/j.inoche.2013.06.024.

[278] Laura Ion, Dolores Morales, Sonia Nieto, Julio Pérez, Lucía Riera, Víctor Riera, Daniel Miguel, Richard A. Kowenicki, and Mary McPartlin. Biimidazole and bis(amide)bipyridine molybdenum carbonyl complexes as anions receptors. *Inorganic Chemistry*, 46(7):2846–2853, 2007. ISSN 00201669. doi: 10.1021/ic0621650.

[279] Silverio Coco, Carlos Cordovilla, Cristina Domínguez, Bertrand Donnio, Pablo Espinet, and Daniel Guillon. Columnar mesophases in Hybrid organic-inorganic supramolecular aggregates: Liquid crystals of Fe, Cr, Mo, and W at room temperature, built from triazines and metalloacid complexes. *Chemistry of Materials*, 21(14):3282–3289, 2009. ISSN 08974756. doi: 10.1021/cm900811j.

[280] S. Karthikeyan, R. Nagarajaprakash, Garisekurthi Satheesh, Chowan Ashok Kumar, and Bala Manimaran. Self-assembly of a fac-Mn(CO)3-core contain-

ing dinuclear metallacycles using flexible ditopic linkers. *Dalton Transactions*, 44(39):17389–17398, 2015. ISSN 14779234. doi: 10.1039/c5dt01866h.

[281] B. H.S. Thimmappa. Low valent metal clusters - an overview. *Coordination Chemistry Reviews*, 143(C):1–34, 1995. ISSN 00108545. doi: 10.1016/0010-8545(94)07002-2.

[282] G. Longoni, C. Femoni, M. C. Iapalucci, and P. Zanello. Electron-Sink Features of Homoleptic Transition-Metal Carbonyl Clusters. In P. Braunstein, L. A. Oro, and P. R. Raithby, editors, *Metal Clusters in Chemistry*, pages 1137–1158. Wiley-VCH Verlag GmbH, 1999.

[283] Allen C. Sarapu and Richard F Fenske. Bonding Properties of the Methyl Isocyanide Ligand. A Single-Crystal X-Ray of Bromotricarbonylbis(methyl isocyanide) manganese, Mn(C0)3(CNCH3)2Br. *Inorganic Chemistry*, 11(12): 3021–3025, 1972.

[284] R. B. King and Mohan S. Saran. Isocyanide-metal complexes. II. Carbonyl and cyanide stretching modes in tertbutyl isocyanide derivatives of the octahedral metal carbonyls. *Inorganic Chemistry*, 13(1):74–78, 1974.

[285] Albert Cotton and Franco Zingales. The Donor-Acceptor Properties of Isonitriles. *Journal of the American Chemical Society*, 83:351–355, 1960. URL `https://pubs.acs.org/doi/pdf/10.1021/ja01463a022`.

[286] Allen C. Sarapu and Richard F. Fenske. The Transition Metal-Isocyanide Bond. An Approximate Molecular Orbital Study. *Inorganic Chemistry*, 14(2): 247–253, 1975. ISSN 1520510X. doi: 10.1021/ic50144a006.

[287] Yasuhiro Yamamoto. ZEROVALENT TRANSITION Metal complexes of organic isocyanides. *Coordination Chemistry Reviews*, 32:193–233, 1980.

[288] Lothar Weber. Homoleptic Isocyanide Metalates. *Angewandte Chemie International Edition*, 37(11):1515–1517, 1998. ISSN 14337851. doi: 10.1002/(sici)1521-3773(19980619)37:11<1515::aid-anie1515>3.3.co;2-0.

[289] Patricia A. Leach, Steven J. Geib, Joseph A. Corella, Garry F. Warnock, and N. John Cooper. Synthesis and Structural Characterization of [Co{CN(2,6-C6H3Me2)}4]-, the First Transition Metal Isonitrilate. *Journal of the American Chemical Society*, 116(19):8566–8574, 1994. ISSN 15205126. doi: 10.1021/ja00098a017.

[290] Garry F. Warnock and N. John Cooper. The first transition-metal isonitrilate: synthesis and characterization of cobalt complex K[Co(2,6-Me2C6H3NC)4]. *Organometallics*, 8(7):1826–1827, 1989.

[291] Alex E. Carpenter, Arnold L. Rheingold, and Joshua S. Figueroa. A Well-Defined Isocyano Analogue of HCo(CO)4. 1: Synthesis, Decomposition, and Catalytic 1,1-Hydrogenation of Isocyanides. *Organometallics*, 35(14):2309–2318, 2016. ISSN 15206041. doi: 10.1021/acs.organomet.6b00297.

[292] F. Ekkehardt Hahn and Matthias Tamm. Chelate Complexes with Aliphatic Tripodal Triisocyanide Ligands: Synthesis and Crystal Structures of fac-{[HC(CH2CH2CH2NC)3]W(CO)3}, fac-{[N(CH2CH2CH2NC)3]W(CO)3}, and the Dimer fac,fac-{[N(CH2CH2CH2NC)3]W(CO)3}2. *Organometallics*, 13(8):3002–3008, 1994. ISSN 15206041. doi: 10.1021/om00020a016.

[293] F. Ekkehardt Hahn and Matthias Tamm. Chelate Complexes of Tripodal, Aliphatic Triisocyanide Ligands. *Angewandte Chemie International Edition in English*, 31(9):1212–1214, 1992. ISSN 15213773. doi: 10.1002/anie.199212121.

[294] E. Ekkehardt Hahn and Mutthias Tamm. Chelate Complexes with Triisocyanide Ligands: synthesis and crystal structures of fac-{[N(CH2CH2CH2O-3-CNC6H4)3]W(CO)3}.cntdot.0.5CH2Cl2 and fac-{[N(CH2CH2CH2O-2-CNC6H4)3]W(CO)3.cntdot.HCl}.cntdot.2.5CH2Cl2. *Angewandte Chemie International Edition in English*, 30(2):203–205, 1992. ISSN 15213773. doi: 10.1002/anie.199102031.

[295] N. L. Wagner, F. E. Laib, and D. W. Bennett. Conformational isomerism in (p-RC6H4NC)2W(dppe)2: Substantial structural changes resulting from subtle differences in the π-acidity of p-RC6H4NC. *Journal of the American Chemical Society*, 122(44):10856–10867, 2000. ISSN 00027863. doi: 10.1021/ja0018466.

[296] Cantwell G. Carson, Rosario A. Gerhardt, and Rina Tannenbaum. Chemical stability and characterization of rhodium-diisocyanide coordination polymers. *Journal of Physical Chemistry B*, 111(51):14114–14120, 2007. ISSN 15206106. doi: 10.1021/jp0765852.

[297] Rina Tannenbaum. Three-Dimensional Coordination Polymers of Their Catalytic Activityt. *Chem. Mater.*, (6):550–555, 1994.

[298] Alex E. Carpenter, Charles C. Mokhtarzadeh, Donald S. Ripatti, Irena Havrylyuk, Ryo Kamezawa, Curtis E. Moore, Arnold L. Rheingold, and Joshua S. Figueroa. Comparative measure of the electronic influence of highly sub-

stituted aryl isocyanides. *Inorganic Chemistry*, 54(6):2936–2944, 2015. ISSN 1520510X. doi: 10.1021/ic5030845.

[299] Robert J. Angelici and Mihaela Lazar. Isocyanide ligands adsorbed on metal surfaces: Applications in catalysis, nanochemistry, and molecular electronics. *Inorganic Chemistry*, 47(20):9155–9165, 2008. ISSN 00201669. doi: 10.1021/ic800513t.

[300] Sarah L. Horswell, Christopher J. Kiely, Ian A. O'Neil, and David J. Schiffrin. Alkyl Isocyanide-Derivatized Platinum Nanoparticles. *J. Am. Chem. Socm*, 121(23):5573–5574, 1999. doi: https://doi.org/10.1021/ja984284n.

[301] Radu Custelcean. Crystal engineering with urea and thiourea hydrogen-bonding groups. *Chemical Communications*, (3):295–307, 2008. ISSN 13597345. doi: 10.1039/b708921j.

[302] Nicole Volz and Jonathan Clayden. The urea renaissance. *Angewandte Chemie - International Edition*, 50(51):12148–12155, 2011. ISSN 14337851. doi: 10.1002/anie.201104037.

[303] Lucile Fischer and Gilles Guichard. Folding and self-assembly of aromatic and aliphatic urea oligomers: Towards connecting structure and function. *Organic and Biomolecular Chemistry*, 8(14):3102–3117, 2010. ISSN 14770520. doi: 10.1039/c001090a.

[304] K. Pulka-Ziach and S. Sek. α-Helicomimetic foldamers as electron transfer mediators. *Nanoscale*, 9(39):14913–14920, 2017. ISSN 20403372. doi: 10.1039/c7nr05209j.

[305] Vesna Blažek Bregović, Nikola Basarić, and Kata Mlinarić-Majerski. Anion binding with urea and thiourea derivatives. *Coordination Chemistry Reviews*, 295:80–124, 2015. ISSN 00108545. doi: 10.1016/j.ccr.2015.03.011. URL http://dx.doi.org/10.1016/j.ccr.2015.03.011.

[306] Ai-Fang Li, Jin-He Wang, Fang Wang, and Yun-Bao Jiang. Anion complexation and sensing using modified urea and thiourea-based receptors. *Chem. Soc. Rev*, 39(39):3581–4008, 2010. ISSN 0306-0012.

[307] Lukas Pfeifer, Keary M. Engle, George W. Pidgeon, Hazel A. Sparkes, Amber L. Thompson, John M. Brown, and Véronique Gouverneur. Hydrogen-Bonded Homoleptic Fluoride-Diarylurea Complexes: Structure, Reactivity, and Coordinating Power. *Journal of the American Chemical Society*, 138(40): 13314–13325, 2016. ISSN 15205126. doi: 10.1021/jacs.6b07501.

[308] Massimo Boiocchi, Laura Del Boca, David Esteban-Gómez, Luigi Fabbrizzi, Maurizio Licchelli, and Enrico Monzani. Nature of Urea-Fluoride Interaction: Incipient and definitive proton transfer. *J. Am. Chem. Soc.*, 126(50): 16507–16514, 2004. ISSN 0002-7863. URL http://www.ncbi.nlm.nih.gov/pubmed/15600354.

[309] Valeria Amendola, Greta Bergamaschi, Massimo Boiocchi, Luigi Fabbrizzi, and Lorenzo Mosca. The interaction of fluoride with fluorogenic ureas: An ON^1-OFF-ON^2 response. *J. Am. Chem. Soc*, 135(16):6345–6355, 2013. ISSN 00027863. doi: 10.1021/ja4019786.

[310] David Esteban-Gómez, Luigi Fabbrizzi, and Maurizio Licchelli. Why, on interaction of urea-based receptors with fluoride, beautiful colors develop.

Journal of Organic Chemistry, 70(14):5717–5720, 2005. ISSN 00223263. doi: 10.1021/jo050528s.

[311] Valeria Amendola, Luigi Fabbrizzi, and Lorenzo Mosca. Anion recognition by hydrogen bonding: Urea-based receptors. *Chemical Society Reviews*, 39 (10):3889–3915, 2010. ISSN 14604744. doi: 10.1039/b822552b.

[312] Radu Custelcean, Bruce A. Moyer, Vyacheslav S. Bryantsev, and Benjamin P. Hay. Anion coordination in metal-organic frameworks functionalized with urea hydrogen-bonding groups. *Crystal Growth and Design*, 6(2):555–563, 2006. ISSN 15287483. doi: 10.1021/cg0505057.

[313] Nicholas H. Evans and Paul D. Beer. Advances in anion supramolecular chemistry: From recognition to chemical applications. *Angewandte Chemie - International Edition*, 53(44):11716–11754, 2014. ISSN 15213773. doi: 10.1002/anie.201309937.

[314] Philip A. Gale. Anion receptor chemistry. *Chemical Communications*, 47(1): 82–86, 2011. ISSN 1364548X. doi: 10.1039/c0cc00656d.

[315] Nathalie Busschaert, Claudia Caltagirone, Wim Van Rossom, and Philip A. Gale. Applications of Supramolecular Anion Recognition. *Chemical Reviews*, 115(15):8038–8155, 2015. ISSN 15206890. doi: 10.1021/acs.chemrev.5b00099.

[316] Radu Custelcean. Anions in crystal engineering. *Chemical Society Reviews*, 39 (10):3675–3685, 2010. ISSN 14604744. doi: 10.1039/b926221k.

[317] L. S. Reddy, S. K. Chandran, S. George, N. J. Babu, and A. Nangia. Crystal Structures of N-Aryl-N'-4-Nitrophenyl Ureas: Molecular Conformation and

Weak Interactions Direct the Strong Hydrogen Bond Synthon. *Crystal Growth and Design*, 7(12):2675–2690, 2007. doi: 10.1021/cg070155j.

[318] Margaret C. Etter, Mohammad Zia-Ebrahimi, Zofia Urbañczyk-Lipkowska, and Thomas W. Panunto. Hydrogen Bond Directed Cocrystallization and Molecular Recognition Properties of Diarylureas. *Journal of the American Chemical Society*, 112(23):8415–8426, 1990. ISSN 15205126. doi: 10.1021/ja00179a028.

[319] Margaret C. Etter. Encoding and Decoding Hydrogen-Bond Patterns of Organic Compounds. *Accounts of Chemical Research*, 23(4):120–126, 1990. ISSN 15204898. doi: 10.1021/ar00172a005.

[320] Margaret C. Etter and Thomas W. Panunto. 1,3-Bis(m-nitrophenyl)urea: An exceptionally good complexing agent for proton acceptors. *J. Am. Chem. Soc*, 110:5896–5897, 1988.

[321] Pavel Majer and Ramnarayan S. Randad. A safe and efficient method for preparation of N,N'-Usymmetrically disubstituted ureas utilizing triphosgene. *J. Org. Chem.*, 59(13):1937–1938, 1994.

[322] Neil J Coville. A Synthetic Route to M(CO)6-n(RNC)n (M = Cr, MO, W; n = 1-6) from M(CO)6 and Isonitriles. *Inorganica Chimica Acta*, 65(1982):7–8, 2015.

[323] Michel O. Albers, Neil J. Coville, and Eric Singleton. Catalysed and non-catalysed reaction between [Fe(CO)5] and isonitriles. *Journal of the Chemical Society, Dalton Transactions*, (6):1069–1079, 1982. ISSN 14727773. doi: 10.1039/DT9820001069.

[324] Wesley Sattler and Gerard Parkin. Synthesis of transition metal isocyanide compounds from carbonyl complexes via reaction with Li[Me3SiNR]. *Chemical Communications*, (48):7566–7568, 2009. ISSN 13597345. doi: 10.1039/b917156h.

[325] Christopher C. Page, Christopher C. Moser, Xiaoxi Chen, and P. Leslie Dutton. Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature*, 402(6757):47–52, 1999. ISSN 00280836. doi: 10.1038/46972.

[326] Christopher C. Moser, Christopher C. Page, and P. Leslie Dutton. Darwin at the molecular scale: Selection and variance in electron tunnelling proteins including cytochrome c oxidase. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1472):1295–1305, 2006. ISSN 09628436. doi: 10.1098/rstb.2006.1868.

[327] Marina A. Solomos, Taylor A. Watts, and Jennifer A. Swift. Predicting Cocrystallization Based on Heterodimer Energies: Part II. *Crystal Growth and Design*, 17(10):5073–5079, 2017. doi: 10.1021/acs.cgd.7b00922.

[328] Christoph Janiak. A critical account on n-n stacking in metal complexes with aromatic nitrogen-containing ligands. *Journal of the Chemical Society, Dalton Transactions*, (21):3885–3896, 2000. ISSN 1470479X. doi: 10.1039/b003010o.

[329] Y. N. Imai, Y. Inoue, I. Nakanishi, and K. Kitaura. Amide-π interactions between formamide and benzene. *Journal of Computational Chemistry*, 30(14):2267–2276, 2009. doi: doi:10.1002/jcc.21212.

[330] Koushik Kasavajhala, Swetha Bikkina, Indrajit Patil, Alexander D. Jr. MacKerell, and U. Deva Priyakumar. Dispersion Interactions between Urea and Nucleobases Contribute to the Destabilization of RNA by Urea in Aqueous Solution Koushik. *J Phys Chem B*, 119(9):3755–3761, 2015. doi: 10.1021/jp512414f.Dispersion.

[331] Siddharth Goyal, Aditya Chattopadhyay, Koushik Kasavajhala, and U. Deva Priyakumar. Role of Urea-Aromatic Stacking Interactions in Stabilizing the Aromatic Residues of the Protein in Urea-Induced Denatured State. *Journal of the American Chemical Society*, 139(42):14931–14946, 2017. ISSN 15205126. doi: 10.1021/jacs.7b05463.

[332] Tadashi Ema, Takayuki Yamasaki, Sagiri Watanabe, Mahoko Hiyoshi, and Kazuto Takaishi. Cross-Coupling Approach to an Array of Macrocyclic Receptors Functioning as Chiral Solvating Agents. *Journal of Organic Chemistry*, 83(18):10762–10769, 2018. ISSN 15206904. doi: 10.1021/acs.joc.8b01327.

[333] Pall Thordarson. Determining association constants from titration experiments in supramolecular chemistry. *Chemical Society Reviews*, 40(3):1305–1323, 2011. ISSN 14604744. doi: 10.1039/c0cs00062k.

[334] D. Brynn Hibbert and Pall Thordarson. The death of the Job plot, transparency, open science and online tools, uncertainty estimation methods and other developments in supramolecular chemistry data analysis. *Chemical Communications*, 52(87):12792–12805, 2016. ISSN 1364548X. doi: 10.1039/c6cc03888c.

[335] Paul J Smith, Mark V Reddington, and Craig S Wilcox. Ion pair binding by a urea in chloroform solution. *Tetrahedron Letters*, 33(41):6085–6088, 1992.

[336] Martín Regueiro-Figueroa, Kristina Djanashvili, David Esteban-Gómez, Andrés De Blas, Carlos Platas-Iglesias, and Teresa Rodríguez-Blas. Towards selective recognition of sialic acid through simultaneous binding to its cis-diol and carboxylate functions. *European Journal of Organic Chemistry*, (17): 3237–3248, 2010. ISSN 1434193X. doi: 10.1002/ejoc.201000186.

[337] John P. Clare, Anna Statnikov, Vincent Lynch, Andrew L. Sargent, and John W. Sibert. "Wurster-type" ureas as redox-active receptors for anions. *Journal of Organic Chemistry*, 74(17):6637–6646, 2009. ISSN 00223263. doi: 10.1021/jo9011392.

[338] Israel Carreira-Barral, Marta Mato-Iglesias, Andrés De Blas, Carlos Platas-Iglesias, Peter A. Tasker, and David Esteban-Gómez. Ditopic receptors containing urea groups for solvent extraction of Cu(ii) salts. *Dalton Transactions*, 46(10):3192–3206, 2017. ISSN 14779234. doi: 10.1039/c7dt00093f.

[339] David Esteban Gómez, Luigi Fabbrizzi, Maurizio Licchelli, and Enrico Monzani. Urea vs. thiourea in anion recognition. *Organic and Biomolecular Chemistry*, 3(8):1495–1500, 2005. ISSN 14770520. doi: 10.1039/b500123d.

[340] Katja Heinze and Volker Jacob. Stepwise Assembly of Mixed Metal Dinuclear Carbonyl Complexes. *European Journal of Inorganic Chemistry*, (21): 3918–3923, 2003. ISSN 14341948. doi: 10.1002/ejic.200300272.

[341] Michael J. Hynes. EQNMR: A computer program for the calculation of sta-

bility constants from nuclear magnetic resonance chemical shift data. *Journal of the Chemical Society, Dalton Transactions*, (2):311–312, 1993. ISSN 14727773. doi: 10.1039/DT9930000311.

[342] George M. Sheldrick. SADABS: Area Detector Absorption Correction, 1996.

[343] Oleg V. Dolomanov, Luc J. Bourhis, Richard J. Gildea, Judith A.K. Howard, and Horst Puschmann. OLEX2: A complete structure solution, refinement and analysis program. *Journal of Applied Crystallography*, 42(2):339–341, 2009. ISSN 00218898. doi: 10.1107/S0021889808042726.

[344] George M. Sheldrick. Crystal structure refinement with SHELXL. *Acta Crystallographica Section C: Structural Chemistry*, 71(Md):3–8, 2015. ISSN 20532296. doi: 10.1107/S2053229614024218.

[345] George M. Sheldrick. SHELXT - Integrated space-group and crystal-structure determination. *Acta Crystallographica Section A: Foundations of Crystallography*, 71(1):3–8, 2015. ISSN 16005724. doi: 10.1107/S2053273314026370.

[346] George M. Sheldrick. A short history of SHELX. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):112–122, 2008. ISSN 16005724. doi: 10.1107/S0108767307043930.

[347] Bruker. APEX3, 2016.

[348] Clare F. Macrae, Ian J. Bruno, James A. Chisholm, Paul R. Edgington, Patrick McCabe, Elna Pidcock, Lucia Rodriguez-Monge, Robin Taylor, Jacco Van De Streek, and Peter A. Wood. Mercury CSD 2.0 - New features for the visualiza-

tion and investigation of crystal structures. *Journal of Applied Crystallography*, 41(2):466–470, 2008. ISSN 00218898. doi: 10.1107/S0021889807067908.

[349] Roberto Dovesi, Roberto Orlando, Alessandro Erba, Claudio M. Zicovich-Wilson, Bartolomeo Civalleri, Silvia Casassa, Lorenzo Maschio, Matteo Ferrabone, Marco De La Pierre, Philippe D'Arco, Yves Noël, Mauro Causà, Michel Rérat, and Bernard Kirtman. CRYSTAL14: A program for the ab initio investigation of crystalline solids. *International Journal of Quantum Chemistry*, 114(19):1287–1317, 2014. ISSN 1097461X. doi: 10.1002/qua.24658.

[350] R Dovesi, V R Saunders, C Roetti, R Orlando, Claudio M. Zicovich-Wilson, F Pascale, B Civalleri, K Doll, N M Harrison, I J Bush, Philippe D'Arco, M Llunell, Mauro Causà, and Yves Noël. CRYSTAL 14: User's Manual. *University of Torino, Torino*, pages 1–382, 2016.

[351] Stefan Grimme. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 25(12):1463–1473, 2004. ISSN 01928651. doi: 10.1002/jcc.20078.

[352] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, jan 1988. ISSN 0163-1829. doi: 10.1103/PhysRevB.37.785. URL `https://link.aps.org/doi/10.1103/PhysRevB.37.785`.

[353] Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics*, 90(2):1007–1023, 1989. ISSN 00219606. doi: 10.1063/1.456153.

[354] T. H. Dunning and P. J. Hay. Gaussian Basis Sets for Molecular Calculations. In H. F.. Schaefer III, editor, *Modern Theoretical Chemistry, Vol. 3: Methods of Electronic Structure Theory*, pages 1–27. Plenum Press, 1977.

[355] P. Jeffrey Hay and Willard R. Wadt. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *The Journal of Chemical Physics*, 82(1):270–283, 1985. ISSN 00219606. doi: 10.1063/1.448799.

[356] G Gilat. Analysis of methods for calculating spectral properties in solids. *Journal of Computational Physics*, 10(3):432–465, dec 1972. ISSN 0021-9991. doi: 10.1016/0021-9991(72)90046-0. URL https://www.sciencedirect.com/science/article/pii/0021999172900460.

[357] Hendrik J. Monkhorst and James D. Pack. Special points for Brillouin-zone integrations. *Physical Review B*, 13(12):5188–5192, 1976. ISSN 01631829. doi: 10.1103/PhysRevB.16.1748.

[358] S.F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics*, 19:553–566, 1970. doi: 10.1080/00268977000101561. URL https://www.tandfonline.com/doi/abs/10.1080/00268977000101561.

[359] Markus Albrecht. Supramolecular chemistry - General principles and selected examples from anion recognition and metallosupramolecular chemistry. *Naturwissenschaften*, 94(12):951–966, 2007. ISSN 00281042. doi: 10.1007/s00114-007-0282-7.

[360] Ze Fan Yao, Jie Yu Wang, and Jian Pei. Control of π-π Stacking via Crystal Engineering in Organic Conjugated Small Molecule Crystals. *Crystal Growth and Design*, 18(1):7–15, 2018. ISSN 15287505. doi: 10.1021/acs.cgd.7b01385.

[361] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016. ISSN 20525206. doi: 10.1107/S2052520616003954.

[362] M. F. Peintinger, D. V. Oliveira, and T. Bredow. Consistent Gaussian basis sets of triple-zeta valence with polarization quality for solid-state calculations. *J. Comput. Chem.*, 34:451–459, 2013. doi: 10.1002/jcc.23153.

[363] Ansgar Schäfer, Christian Huber, and Reinhart Ahlrichs. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *The Journal of Chemical Physics*, 100(8):5829–5835, 1994. ISSN 00219606. doi: 10.1063/1.467146.

[364] K.L. Schuchardt, B.T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, and T.L. Windus. Basis Set Exchange: A Community Database for Computational Sciences. *J. Chem. Inf. Model.*, 47(3):1045–1052, 2007. doi: 10.1021/ci600510j.

APPENDIX A:

NOTEBOOK EXAMPLE FOR P3HT VALIDATION

A.1 P3HT Validation

[1]:
```
import os


from cmeutils.structure import order_parameter
from fresnel import camera, pathtrace, light
from gixstapose.diffractometer import Diffractometer, get_angle
from gixstapose.draw_scene import get_scene
import gsd.hoomd
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image
from planckton.utils.units import string_to_quantity,␣
 ↪kelvin_from_reduced
from scipy.interpolate import RectBivariateSpline
import signac
import unyt as u
```

In the following notebook, we will demonstrate the recreation of part a of the following figure from Miller 2018. This figure shows the order parameter of a OPLS-UA P3HT system at different temperatures and solvent parameters (ϵ_s or e_factor).

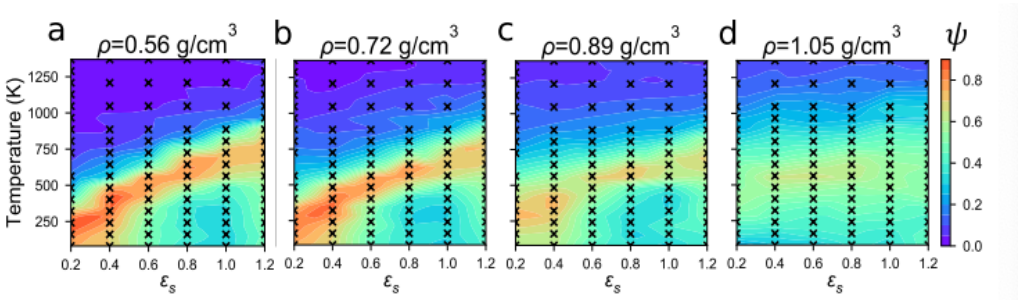
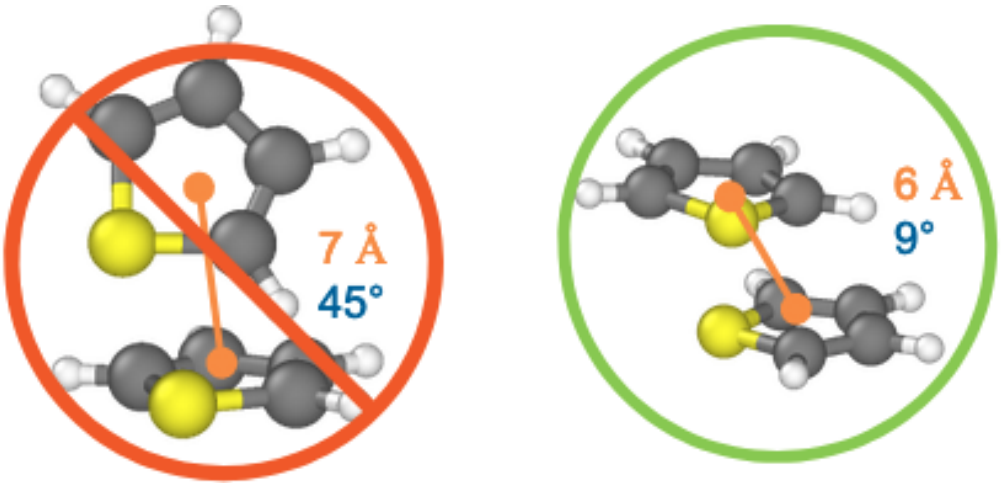


Figure 3. The degree of ordering, ψ, for Protocol (1) shows the most robust assembly occurs at lower densities, with more temperature-solvent combinations resulting in high ψ. (**a**) $\rho = 0.56$ g/cm^3; (**b**) $\rho = 0.72$ g/cm^3; (**c**) $\rho = 0.89$ g/cm^3; (**d**) $\rho = 1.05$ g/cm^3. Red regions denote order, whereas blue denotes disorder. Each black "x" indicates a measurement from an MD trajectory, and ψ values between measurements are linearly interpolated.

The order parameter is used to quantify the degree of ordering in the system by grouping structures into clusters, and then calculating the ratio of the structures in a "large" clusters over the total number of structures in the system. In the case of a P3HT polymer, the structure we use for the order parameter analysis is the thiophene moeity. The clustering criteria depends on the angle between the planes of the thiophenes and the distance between the thiophene centers. A thiophene pair must meet both clustering criteria to be considered cluster neighbors. For example, if the angle cutoff is 10° and the distance cutoff is 6Å, the image below shows a thiophene pair which does not meet the criteria (encircled and crossed in red) and another which does (encircled in green).



For a cluster to be considered "large", it must contain at least 6 thiophenes.

All of this analysis, from the simulations to the functions used to run the analysis itself has been redone using new tools. The goal is to show that not only can we recreate our previous work, but also the tools we've developed make this analysis more transparent, reproducible, usable, and extensible.

The dataset used in this notebook was generated using PlanckTon, which provides a way to easily and reproducibly interface with the HOOMD-blue molecular dynamics engine, and PlanckTon-flow, which uses Signac to initialize and submit simulations across the desired parameter space. The workspace with all data used in this notebook can be found here.

First, assuming the tarball from the above link is unpacked in this directory, we can use signac to explore this workspace:

[2]:
```
p = signac.get_project("p3ht-ua")
```

```
p.detect_schema()
```

[2]:
```
ProjectSchema(<len=15>)
{
 'density': 'str([0.56_g-cm**3], 1)',
 'dt': 'float([0.001], 1)',
 'e_factor': 'float([0.2, 0.4, 0.6, 0.8, 1.0], 5)',
 'forcefield': 'str([gaff-custom], 1)',
 'input': 'tuple([('P3HT-16-gaff',)], 1)',
 'kT': 'tuple([(1,), (1.5,), (2,), ..., (4,), (5,)], 8)',
 'mode': 'str([gpu], 1)',
 'n_compounds': 'tuple([(100,)], 1)',
 'n_steps': 'tuple([(100000000.0,)], 1)',
 'r_cut': 'float([2.5], 1)',
 'remove_hydrogens': 'bool([True], 1)',
 'shrink_kT': 'int([5], 1)',
 'shrink_steps': 'float([100000.0], 1)',
 'shrink_tau': 'int([1], 1)',
 'tau': 'tuple([(0.3,)], 1)',
}
```

The above schema shows that we are looking at a dataspace of P3HT 16-mers, at a density of 0.56 g/cm^3 at 8 different temperatures and 5 different e_factors. In order to run the order parameter analysis, there is a little bookkeeping to be done with units.

PlanckTon runs its simulations in reduced units to reduce floating point error. This means all lengths are scaled such that the largest sigma value in the non-bonded potential is equal to 1, and so on with the largest epsilon value in the non-bonded potential, and the largest particle mass.

This information is stored in the job document and can be converted to a unyt quantity using the `string_to_quantity` function.

In the cell below we set our angle and distance cutoffs for the order parameter analysis to 10° and 6Å scaled by the unit length of our simulation.

[3]:
```
for job in p:
    if job.doc.get("done"):
        break

ref_energy = string_to_quantity(job.doc["ref_energy"])
ref_mass = string_to_quantity(job.doc["ref_mass"])
ref_distance = string_to_quantity(job.doc["ref_distance"])
print(f"{ref_distance:.3f} {ref_mass:.2e} {ref_energy:.3f}")

r_max = float(6 * u.Angstrom / ref_distance)
a_max = 10
print(f"r_max: {r_max:.2f}, a_max: {a_max}")
```

```
3.564 Å~ 5.32e-26 kg 0.250 kcal/mol
r_max: 1.68, a_max: 10
```

In the following cell, the order parameter analysis is performed over the entire

workspace. Previously, this was done using manual indexing and was hardcoded for our P3HT input files. But this workflow has been updated to use GRiTS which uses SMARTS chemical grammar to detect chemical patterns and map the particles in that pattern to a bead. The `order_parameter` function in cmeutils then takes this mapping and the bead positions from GRiTS and calculates the order parameter of the last 10 frames of the trajectory.

[4]:
```
%%time


o_dict = {}
orders = []
job_strs = []
ts = []
efs = []
keys = []
for (ef, kt), jobs in p.groupby(["e_factor", "kT"]):
    for job in jobs:
        if job.doc.get("done"):
            t_si = int(kelvin_from_reduced(kt[0],ref_energy))
            key = (ef, t_si)
            if key not in keys:
                gsdfile = job.fn("trajectory.gsd")
                cg_gsdfile = job.fn("cg-trajectory.gsd")
                mapfile = job.fn("mapping.json")
```

```
            if job.doc.get("order") is None:
                if not (exists(mapfile) and exists(cg_gsdfile)):
                    system = CG_System(
                        gsdfile,
                        beads={"_B" : "c1cscc1"},
                        conversion_dict=amber_dict,
                        add_hydrogens=True,
                    )
                    mapping = system.mapping["_B...c1cscc1"]
                    system.save_mapping(mapfile)
                    system.save(cg_gsdfile)
                    print("\tCG_System created")
                else:
                    with open(mapfile) as f:
                        d = json.load(f)
                    mapping = np.stack(d["_B...c1cscc1"])
                    print("\tUsing mapping")
                order, _ = order_parameter(gsdfile, cg_gsdfile,␣
↪mapping, r_max, a_max)
                order = np.mean(order)
                job.doc["order"] = order
            else:
                order = job.doc.order
```

```
                o_dict[key] = order
                orders.append(order)
                keys.append(key)
                ts.append(t_si)
                efs.append(ef)
                job_strs.append(str(job))
```

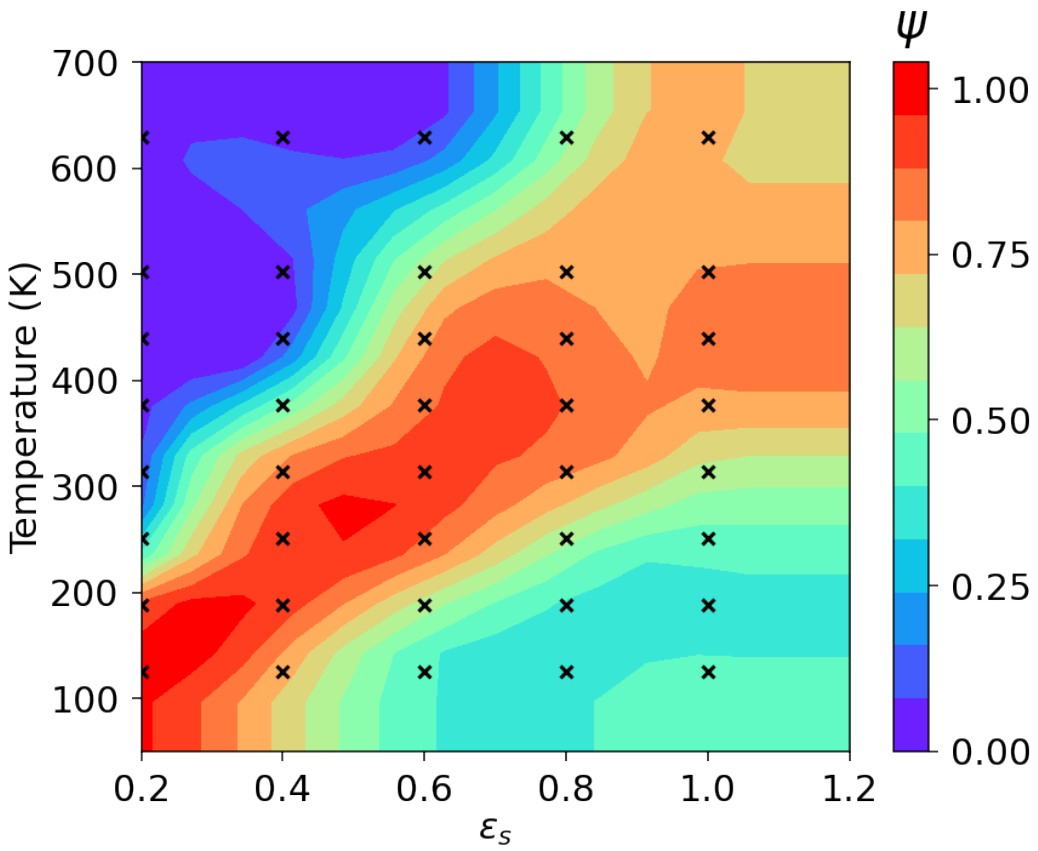
```
CPU times: user 29.2 ms, sys: 2.94 ms, total: 32.2 ms
Wall time: 30 ms
```

Next we will plot the order parameter vs T and e_factor using an interpolating function from scipy.

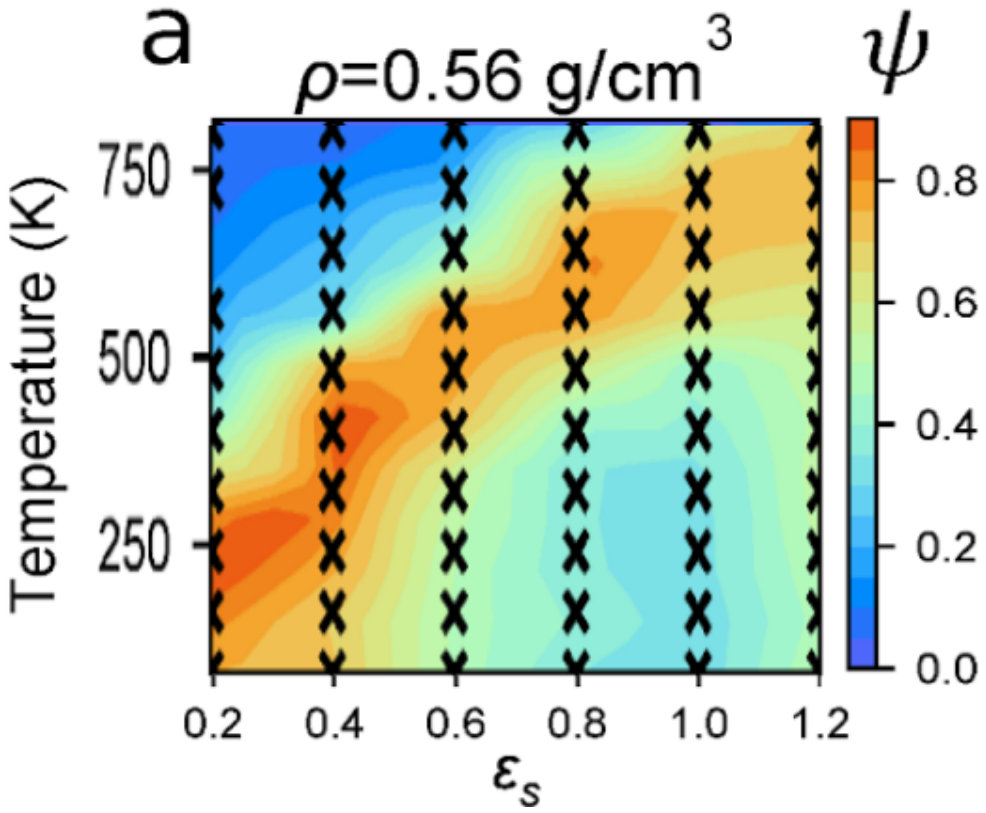
[5]:
```
x = np.array(efs)
y = np.array(ts)
z = np.array(orders)


ux = np.unique(x)
uy = np.unique(y)
uz = np.zeros((len(uy),len(ux)))


# key = (ef, t_si)
for i, xi in enumerate(ux):
    for j, yj in enumerate(uy):
        uz[j,i] = o_dict[(xi, yj)]
```

```
f = RectBivariateSpline(uy,ux,uz)


xs = np.linspace(0.2, 1.2, 15)
ys = np.linspace(50, 700, 15)
zs = f(ys, xs)


zs[np.where(zs > 1)] = 1
zs[np.where(zs < 0)] = 0


plt.contourf(xs, ys, zs, 15, cmap="rainbow", vmax=1, vmin=0)


cbar = plt.colorbar(ticks=np.linspace(0, 1, 5, endpoint=True))
cbar.ax.set_title(r"$\psi$", fontsize=40, pad=20)


# Show the positions of the sample points, just to have some␣
 ↪reference
plt.scatter(x, y, c="k", marker="x", s=100)
plt.xlabel(r"$\epsilon_{s}$")
plt.ylabel("Temperature (K)")
plt.xticks(np.linspace(0.2, 1.2, 6))
fig = plt.gcf()
fig.set_size_inches(12, 10)
```

```
plt.savefig("order_parameter.pdf")
```



Even with the changes to our analysis method (using GRiTS instead of manual indexing), the forcefield (using a GAFF-UA model with flexible thiophenes instead of an OPLS-UA model with rigid thiophenes), and the clustering criteria (10° instead of 20°), the trend in the order parameter across temperature and e_factor space appears pretty robust! I have included only part a from Miller 2018 below with the plot skewed so that the bounds are the same.



We can show that our order parameter metric is robust, but how does it relate to any physically measurable analysis? To address this question, let's look at some diffraction patterns for the highest and lowest order jobs to see if our order parameter metric is correlated with differences in the periodic structure. We'll be using GIXStapose to visualize the real space structure and its simulated diffraction pattern.

[6]:
```
highest_order_jobs = [job_strs[orders.index(i)] for i in␣
 ↪sorted(orders)[-3:]]
lowest_order_jobs = [job_strs[orders.index(i)] for i in␣
 ↪sorted(orders)[:4]]
```

```
print("Higest order")
for job_id in highest_order_jobs:
    job = p.open_job(id=job_id)
    T_SI = int(kelvin_from_reduced(job.sp.kT[0],ref_energy))
    print(f"e_factor: {job.sp.e_factor}, T: {T_SI}K")
    print(f"\t{job}")

print("\n\nLowest order")
for job_id in lowest_order_jobs:
    job = p.open_job(id=job_id)
    T_SI = int(kelvin_from_reduced(job.sp.kT[0],ref_energy))
    print(f"e_factor: {job.sp.e_factor}, T: {T_SI}K")
    print(f"\t{job}")
```

```
Higest order
e_factor: 0.4, T: 188K
        6289cb58ab91a927f702b7cf238820aa
e_factor: 0.4, T: 251K
        16db9573e2e2c6a027abad40cf87733e
e_factor: 0.2, T: 125K
        c1e543f38d9f49677b3e3649018152b8
```

```
Lowest order
e_factor: 0.2, T: 629K
        c0fd30e6247f95e9c26984ec5f24d99e
e_factor: 0.4, T: 629K
        e8e42f76f02e02e686332ce8127e1b24
e_factor: 0.2, T: 503K
        349a22231886a7a46aeb7f20c95204e0
e_factor: 0.6, T: 629K
        d2da0d6355f3d11fd0132bcf8433df7e
```

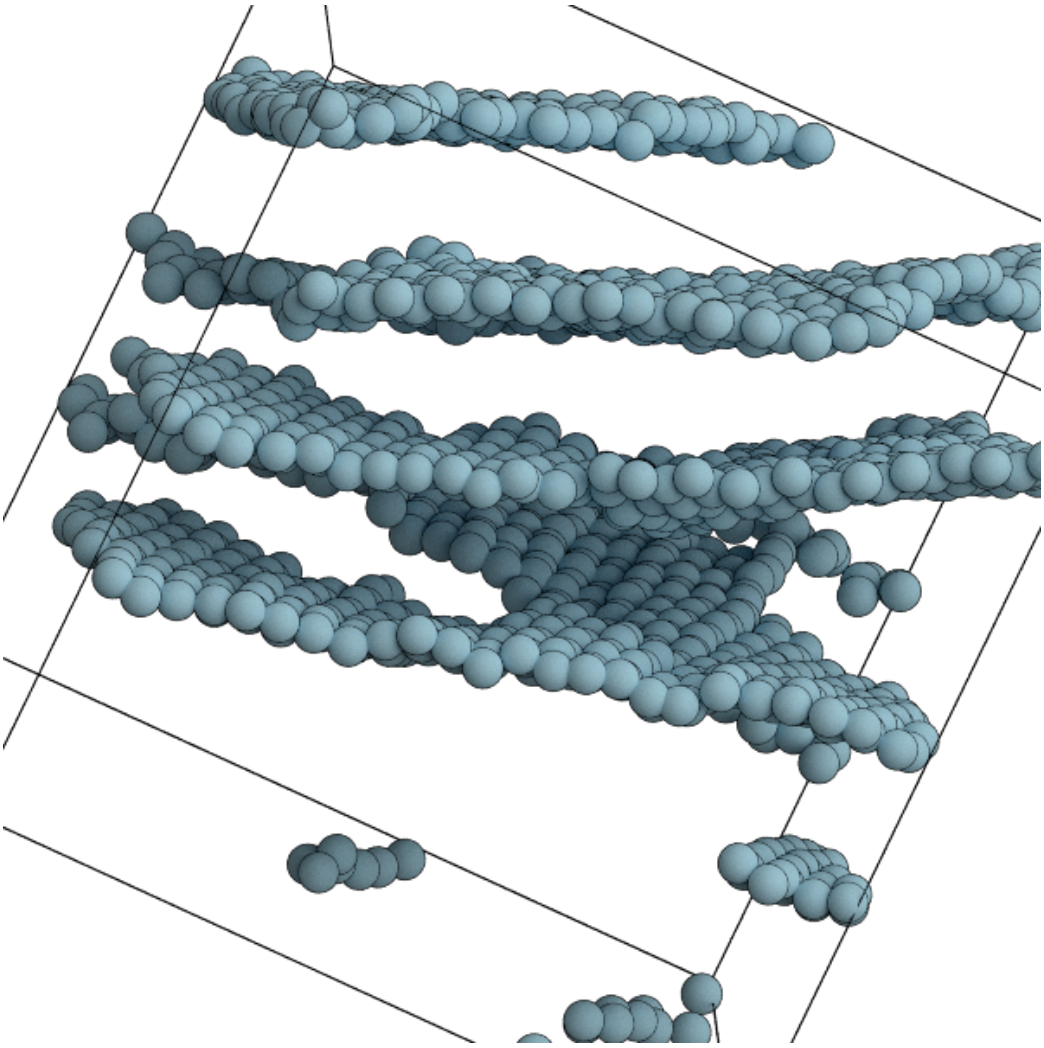
First let's look at one of the high order jobs (e_factor: 0.4, T: 251K). We'll first visualize the centers of the thiophenes. (It is a little less busy than the atomistic sturcture and allows us to more easily see the lamellar spacing.)

The camera position was chosen using the GIXStapose GUI to rotate the structure as to best see the lamellae.

[7]:
```
job = p.open_job(id='16db9573e2e2c6a027abad40cf87733e')
cg_gsd = job.fn("cg-trajectory.gsd")
scene, info = get_scene(cg_gsd, color={"_B": "lightblue"},␣
 ↪scale=ref_distance)
cam = camera.Orthographic(
    position = [8.133, 5.203, 43.865],
    look_at =  [0.000, 0.000, 0.000],
    up =       [0.900, 0.424, -0.101],
    height =   31.896
```

```
)
scene.camera = cam


scene.geometry[0].radius[:] *= 3


output = pathtrace(scene, light_samples=40, w=600, h=600)


image = Image.fromarray(output[:], mode="RGBA")
image.save("cg-trajectory_scene.png")


output
```

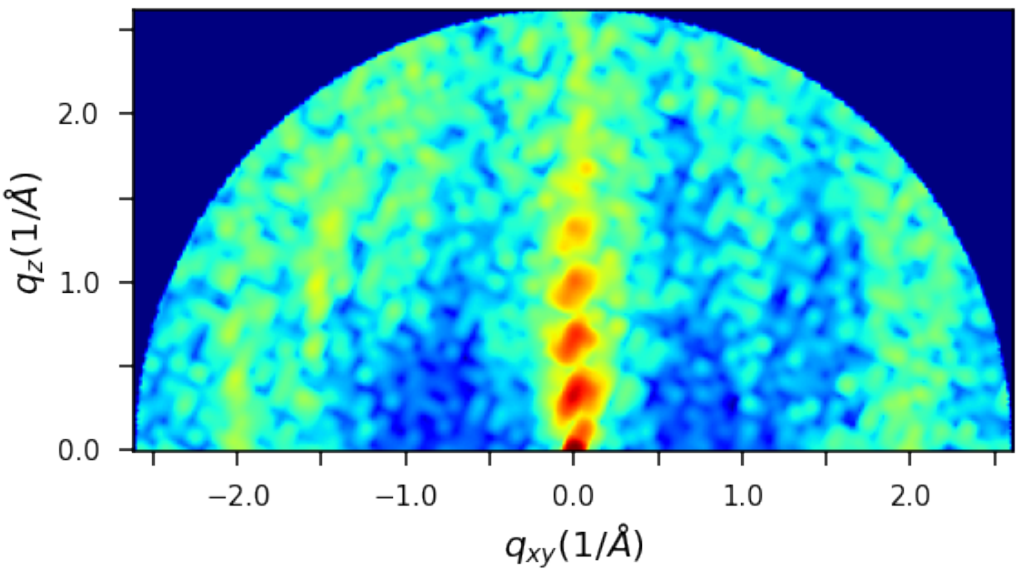
[7]:



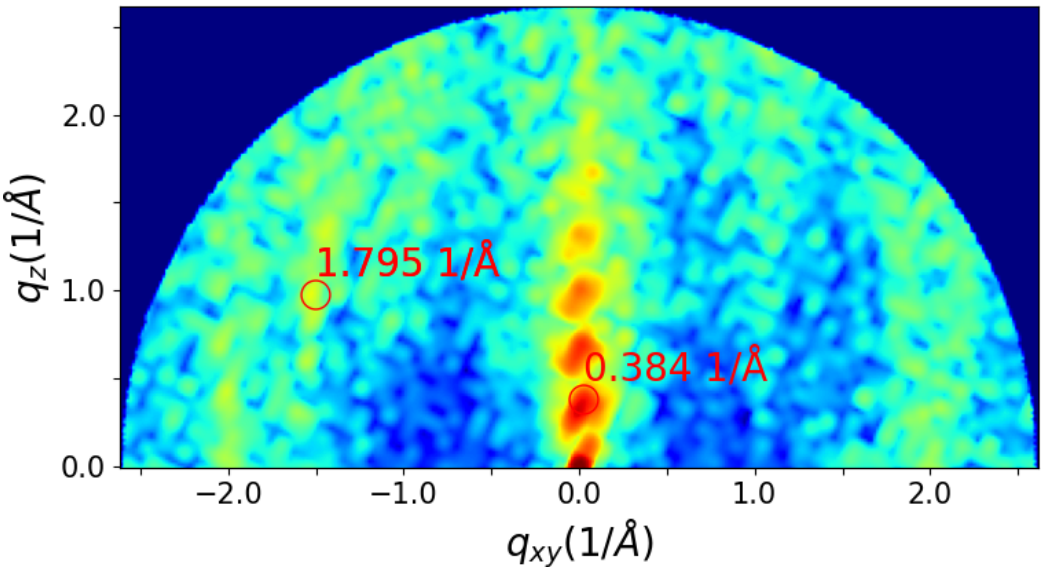
We can see that this high order structure has clear repeated planes of thiophenes, even within these planes there appears to be some ordering most likely due to pi orbital overlap in the thiophenes. Let's look at the diffraction pattern of this view:

[8]:
```
d = Diffractometer(length_scale=float(ref_distance))
d.load(info["positions"], info["box"][:3])
d.diffract_from_camera(cam)
```

```
fig, ax = d.plot(cmap="jet", crop=2.6)
y_min, y_max = ax.get_ylim()
ax.set_ylim((-0.01, y_max))
[l.set_visible(False) for l in ax.xaxis.get_ticklabels()[::2]]
[l.set_visible(False) for l in ax.yaxis.get_ticklabels()[::2]]

plt.show()
```



In this view, we can see bright peaks in the q_z-direction corresponding to the lamellar spacing. The pi-stacking can also faintly be seen in the q_{xy}-direction. We can label these peaks with their distance from the origin using the interactive PeakLabeller class provided in GIXStapose. (To see a demo of how to use this class check out the GIXStapose example.)



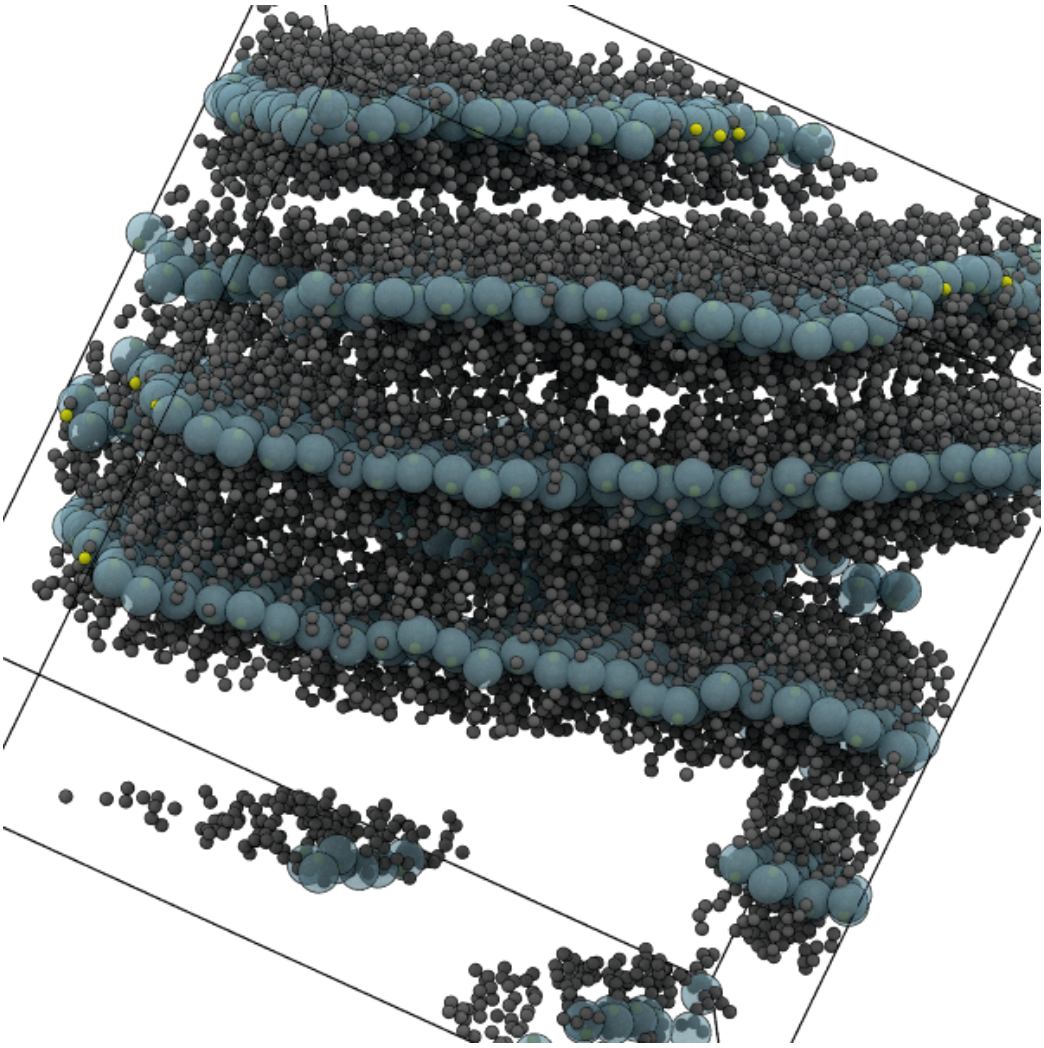
These peaks correspond to periodic distances of 17.24 Å and 3.61 Å which are in the right ballpark for P3HT: Duong 2013 reports a lamellar spacing of 16.5 Å and a pi-stacking spacing of 3.83 Å for neat P3HT.

In the next cell, we'll visualize the overlay of the coarse grain structure with the atomistic one:

[9]:
```
ua_gsd = job.fn("trajectory.gsd")
ua_scene, _ = get_scene(
    ua_gsd,
    color={'c3': "grey", 'cc': "grey", 'cd': "grey", 'ss':␣
 ↪"yellow"},
    scale=ref_distance,
    scene=scene
)
```

```
ua_scene.camera = cam


ua_scene.geometry[0].material.spec_trans = 0.5


ua_scene.lights = light.cloudy()


output = pathtrace(ua_scene, light_samples=40, w=600, h=600)


image = Image.fromarray(output[:], mode="RGBA")
image.save("cg-overlay_scene.png")


output
```

[9]:



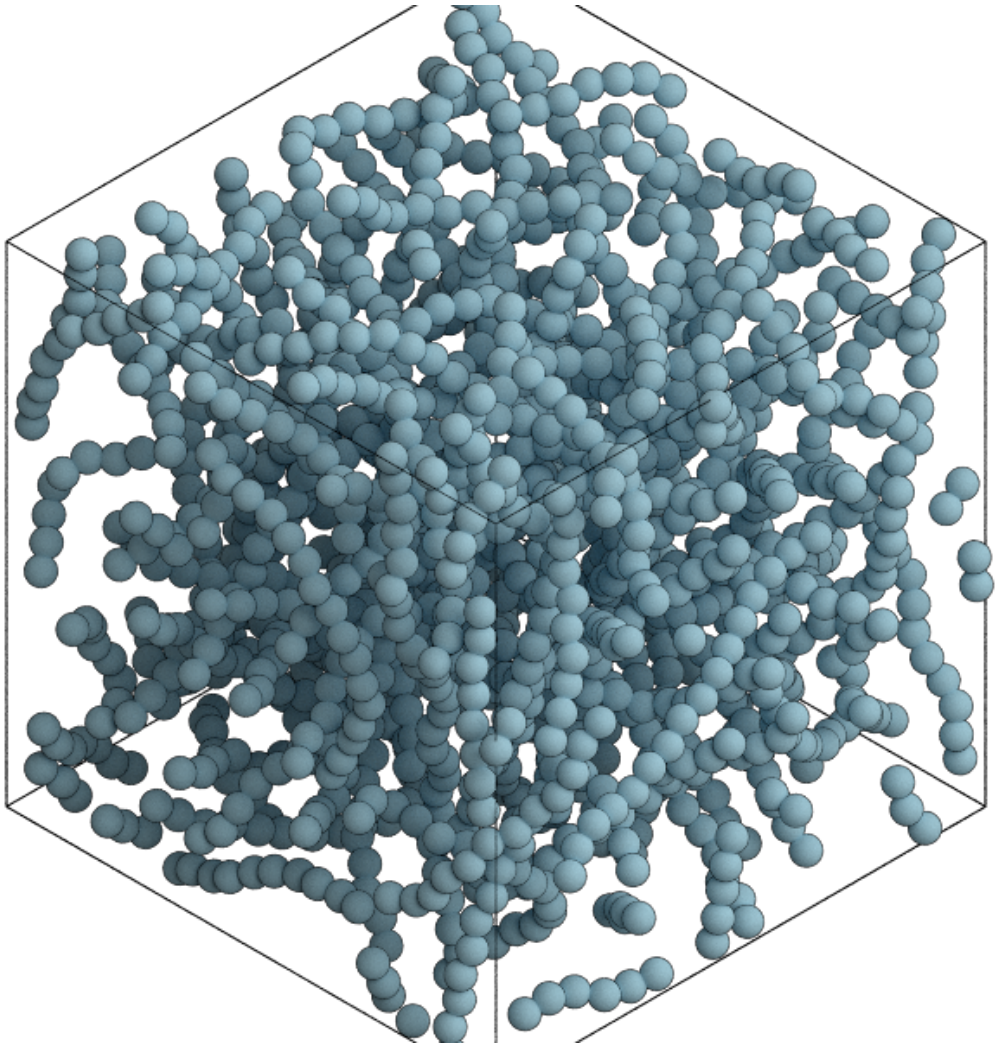
When we view this overlay (thiophene center beads show in translucent blue, atomistic carbons in grey and sulfur in yellow), it becomes more clear that this lamellar spacing is due to the regions with the alkyl tails interacting.

Finally let's look at one of the low order jobs (e_factor: 0.2, T: 629K):

[10]:
```
job = p.open_job(id='c0fd30e6247f95e9c26984ec5f24d99e')
cg_gsd = job.fn("cg-trajectory.gsd")
```

[10]:

```
scene, info = get_scene(cg_gsd, color={"_B": "lightblue"},␣
 ↪scale=ref_distance)


scene.geometry[0].radius[:] *= 3


output = pathtrace(scene, light_samples=40, w=600, h=600)


image = Image.fromarray(output[:], mode="RGBA")
image.save("cg-trajectory-amorphous_scene.png")


output
```

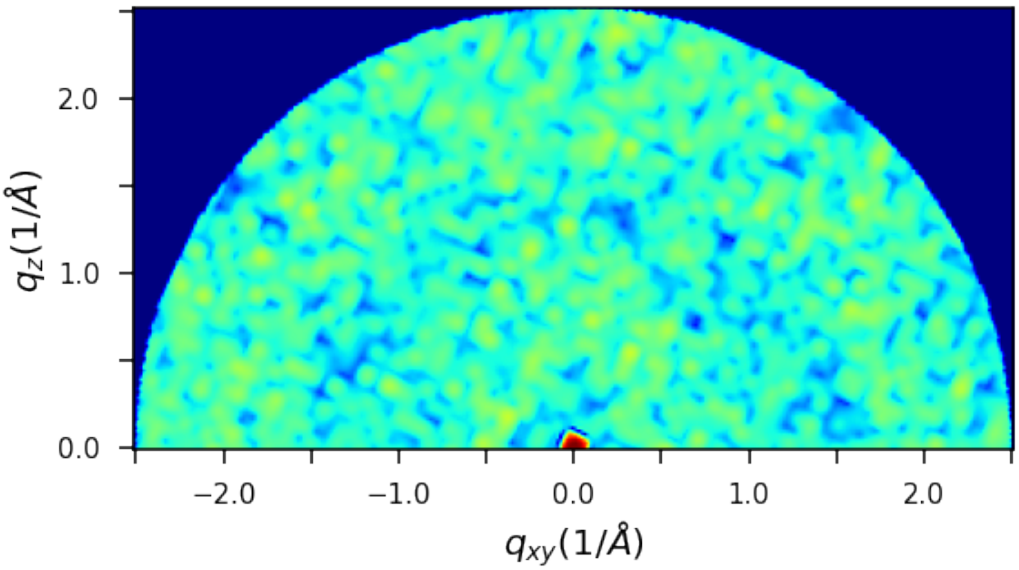


This view shows that the packing looks basically random.

And we can confirm this by examining the diffraction pattern:

[11]:
```
d = Diffractometer(length_scale=float(ref_distance))
d.load(info["positions"], info["box"][:3])
d.diffract_from_camera(cam)
fig, ax = d.plot(cmap="jet", crop=2.6)
```

```
y_min, y_max = ax.get_ylim()
ax.set_ylim((-0.01, y_max))
[l.set_visible(False) for l in ax.xaxis.get_ticklabels()[::2]]
[l.set_visible(False) for l in ax.yaxis.get_ticklabels()[::2]]


plt.show()
```



The general lack of peaks suggests that there are few periodic features to be found.

In conclusion, this analysis has shown that we can reproduce our prior work with updated tools, the order parameter metric is robust across different forcefields and relates to the prominence of peaks in the diffraction pattern.

[ ]:

APPENDIX B:

SUPPORTING INFORMATION FOR REPRODUCIBILITY STUDY

B.1 Forcefield parameters

Table B.1: Non-bonded parameters for SPC/E water

	epsilon (kJ/mol)	sigma (nm)	charge (e)	mass (amu)
OW	0.650194	0.316557	-0.8476	15.99940
HW	0.0	0.0	0.4238	1.00800

Table B.2: Non-bonded parameters for TraPPE-UA benzene

	epsilon (kJ/mol)	sigma (nm)	mass (amu)
CHE	0.419880362216737	0.3695	13.01900

B.2 Potential energy plots

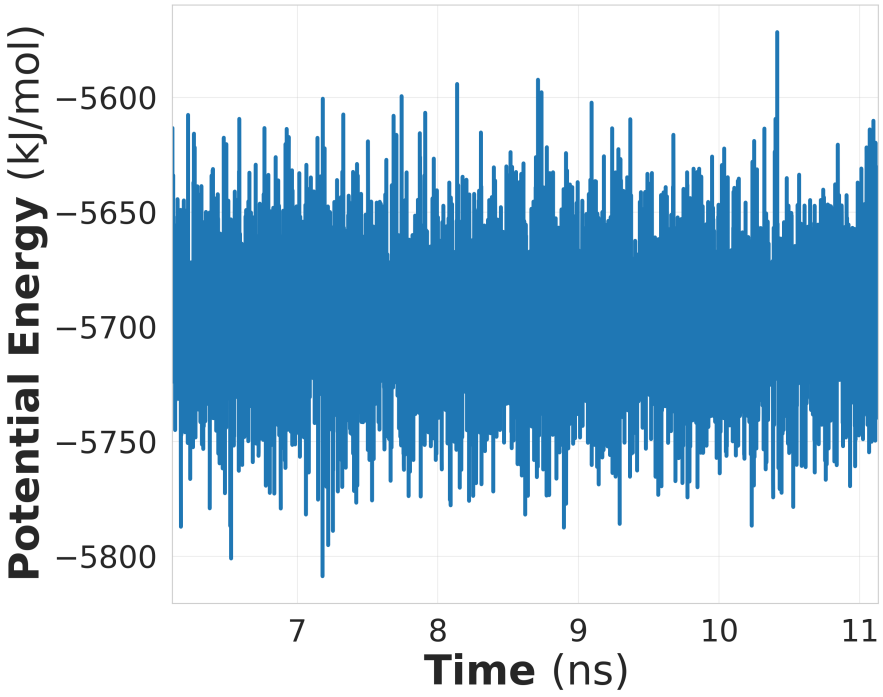


Figure B.1: Potential energy over time for methaneUA NVT ensemble.

B.3 Single-point Energies

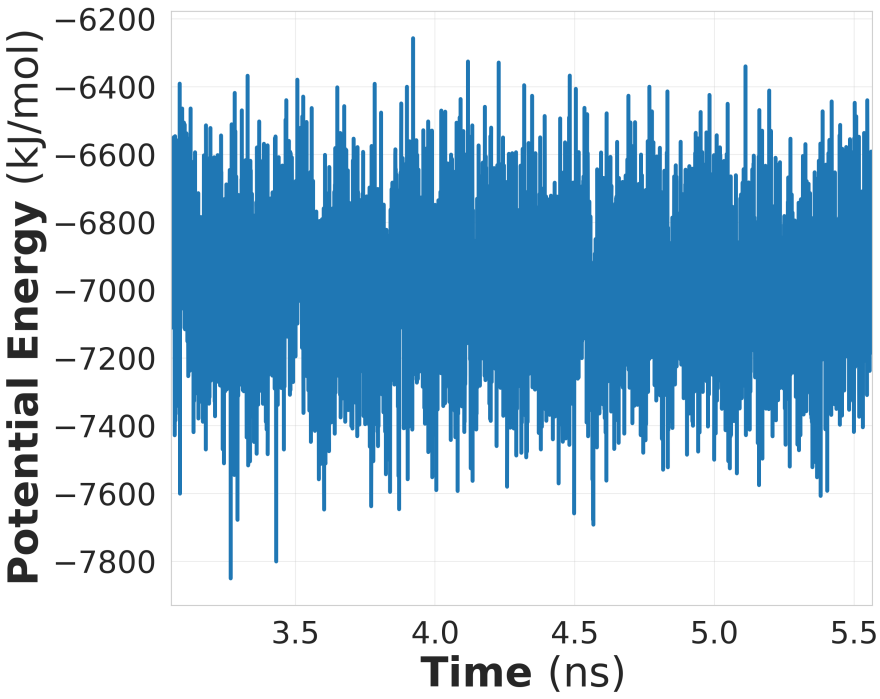


Figure B.2: Potential energy over time for ethanolAA NVT ensemble.

Table B.3: Single-point energy breakdown for methaneUA

Engine	Potential	VDW	Tail Correction
LAMMPS-VU	5.367E+05	5.37E+05	-1.28E+02
LAMMPS-UD	5.369E+05	5.37E+05	-1.19E+02
MCCCS	5.367E+05	5.37E+05	-1.28E+02
HOOMD	5.368E+05	5.37E+05	-1.19E+02
GROMACS	5.368E+05	5.37E+05	-1.19E+02
GOMC	5.367E+05	-	-1.28E+02
Cassandra	5.367E+05	5.37E+05	-1.28E+02

Table B.4: Single-point energy breakdown for pentaneUA

Engine	Potential	VDW	Tail Correction	Bond	Angle	Dihedral
LAMMPS-VU	5.377E+05	5.38E+05	-1.81E+02	0	0	0
LAMMPS-UD	5.366E+05	5.37E+05	-1.68E+02	5.50E-07	6.30E-08	2.90E-09
MCCCS	5.377E+05	5.38E+05	-1.81E+02	0	0	0
HOOMD	5.378E+05	5.38E+05	-1.68E+02	0	0	0
GROMACS	5.378E+05	5.38E+05	-1.68E+02	0	0	1.00E-02
GOMC	5.377E+05	-	-1.81E+02	-	-	-
Cassandra	5.377E+05	5.38E+05	-1.81E+02	0	6.37E-08	2.94E-09

Table B.5: Single-point energy breakdown for benzeneUA

Engine	Potential	VDW	Tail Correction
LAMMPS-VU	3.889E+05	3.89E+05	-2.51E+02
LAMMPS-UD	3.890E+05	3.89E+05	-2.34E+02
MCCCS	3.889E+05	3.89E+05	-2.51E+02
HOOMD	3.891E+05	3.89E+05	-2.34E+02
GROMACS	3.891E+05	3.89E+05	-2.34E+02
Cassandra	3.889E+05	3.89E+05	-2.51E+02
GOMC	3.889E+05	-	-2.51E+02

Table B.6: Single-point energy breakdown for SPC/E water

Engine	Potential	VDW	Tail Correction	Total Electrostatic	Short Range Electrostatic	Long Range Electrostatic	Bond	Angle
LAMMPS-VU	5.718E+04	7.67E+04	-2.76E+02	-1.95E+04	2.93E+05	-3.12E+05	0	0
LAMMPS-UD*	5.977E+04	7.68E+04	-2.49E+02	-1.70E+04	2.90E+05	-3.07E+05	0	0
MCCCS	5.717E+04	7.67E+04	-2.76E+02	-1.95E+04	-	-	0	0
HOOMD	5.718E+04	7.67E+04	-2.76E+02	-1.43E+04	-5.18E+03	5.72E+04	0	0
GROMACS*	5.977E+04	7.68E+04	-2.50E+02	-1.70E+04	-2.37E+04	6.64E+03	0	0
GOMC	5.718E+04	-	-2.76E+02	-	-	-	-	-
Cassandra	5.732E+04	7.68E+04	-2.76E+02	-1.95E+04	3.25E+05	-3.44E+05	0	7.27E-09

*The discrepancies seen in LAMMPS-UD and GROMACS are most likely due to a different size simulation box.

Table B.7: Single-point energy breakdown for OPLSAA ethanol

Engine	Potential	VDW	Tail Correction	Total Electrostatic	Short Range Electrostatic	Long Range Electrostatic	Bond	Angle	Dihedral
LAMMPS-VU	3.141E+04	1.94E+04	-4.18E+02	3.58E+03	8.13E+04	-7.77E+04	0	7.23E+03	1.19E+03
LAMMPS-UD*	3.153E+04	1.96E+04	-3.83E+02	3.55E+03	8.06E+04	-7.71E+04	5.96E-05	7.23E+03	1.19E+03
MCCCS	3.141E+04	1.94E+04	-4.18E+02	3.58E+03	-	-	0	7.23E+03	1.19E+03
HOOMD	3.141E+04	1.94E+04	-4.18E+02	2.31E+03	1.27E+03	2.30E+04	7.23E+03	1.19E+03	8.43E+03
GROMACS*	3.153E+04	1.96E+04	-3.83E+02	3.55E+03	1.69E+04	1.39E+03	0	7.23E+03	1.19E+03
GOMC	3.141E+04		-4.18E+02	-	-	-	-	-	-
Cassandra	3.141E+04	1.94E+04	-4.18E+02	3.58E+03	9.56E+04	-9.21E+04	-	7.23E+03	1.19E+03

*The discrepancies seen in LAMMPS-UD and GROMACS are most likely due to a different size simulation box.

APPENDIX C:

PERSPECTIVE ON COARSE-GRAINING, COGNITIVE LOAD, AND MATERIALS SIMULATION

The following chapter was published in Computational Materials Science under authors Eric Jankowski, Neale Ellyson, Jenny W. Fothergill, Michael M. Henry, Mitchell H. Leibowitz, Evan D. Miller, Mone't Alberts, Samantha Chesser, Jaime D. Guevara, Chris D. Jones, Mia Klopfenstein, Kendra K. Noneman, Rachel Singleton, Ramon A. Uriarte-Mendoza, Stephen Thomas, Carla E. Estridge, and Matthew L. Jones. My contributions to this paper were writing about my experience as a new student, editing, and figure creation.

C.1 Abstract

The predictive capabilities of computational materials science today derive from overlapping advances in simulation tools, modeling techniques, and best practices. We outline this ecosystem of molecular simulations by explaining how important contributions in each of these areas have fed into each other. The combined output of these tools, techniques, and practices is the ability for researchers to advance understanding by efficiently combining simple models with powerful software. As specific examples, we show how the prediction of organic photovoltaic morphologies have improved by orders of magnitude over the last decade, and how the processing of reacting epoxy thermosets can now be investigated

with million-particle models. We discuss these two materials systems and the training of materials simulators through the lens of cognitive load theory.

For students, the broad view of ecosystem components should facilitate understanding how the key parts relate to each other first, followed by targeted exploration. In this way, the paper is organized in loose analogy to a coarse-grained model: The main components provide basic framing and accelerated sampling from which deeper research is better contextualized. For mentors, this paper is organized to provide a snapshot in time of the current simulation ecosystem and an on-ramp for simulation experts into the literature on pedagogical practice.

C.2 A vibrant ecosystem

This perspective describes four issues in computational materials, the vibrant ecosystem in which they are being solved (Figure C.1), and a review of recent advances and best practices for studying materials self-assembly. A central theme of this work is the use of simplified models[92] to provide accessible on-ramps for deeper investigation. The four issues are as follows:

1. Understanding materials behavior through computer simulation
2. Reproducibility of research
3. Accessibility of materials simulation tools
4. Demand for computationally literate researchers

These issues overlap: Reproducible results better advance understanding of materials. Accessible tools facilitate reproducibility. Students with molecular simulation expertise have transferable, in-demand skills. By discussing these issues in the context of the molecular simulation ecosystem, we show how components of the ecosystem are related and are advancing materials research.

The problems of research reproducibility and demand for computationally lit-

erate researchers are broad, encompassing more than the molecular simulation community. In 2016, 52% of researchers agreed there is a "crisis" of reproducibility [93] and more than 600,000 high-paying tech jobs went unfilled in the US [94]. Who will fill these jobs and who will ensure research is reproducible? One candidate population is the pool of XSEDE [95] supercomputer users. These researchers (2,186 undergraduate and 8,409 graduate students in 2017) use nationally-available high performance computing (HPC) facilities to perform scientific research [96] and develop expertise with automating repeatable tasks, managing software stacks, using parallel hardware, and writing software to extract understanding from data. Such computational researchers have the opportunity to demonstrate leadership with reproducibility because the entire research apparatus of one user, including the hardware, software, and pseudorandom number generator seeds used to perform a computation can be replicated exactly by another user—luxuries that are generally not available to non-computational research. However, the fact that only 0.011% of the 19.8 million US undergraduates in 2017 were XSEDE users gives a sense for how rare such leaders might be and the gaps that exist in training computationally literate scientists. Researchers themselves are aware of the gaps: 60% of those surveyed in 2015 reported computational training as their greatest need [97]. In part, this is due to increased data ubiquity and the associated data science and HPC skills needed to manage it [98].

Because computational materials researchers develop XSEDE-user skills, understanding the computational materials ecosystem of tools, techniques, and practices can inform modern workforce training more broadly. We aim for materials simulations that are *transferable*, *reproducible*, *usable*, and *extensible* (*TRUE*). In this work we describe best practices and computational tools that enable TRUE simulations. These practices and tools help researchers waste less time, enhance

research reproducibility, and prepare them for in-demand technical roles.

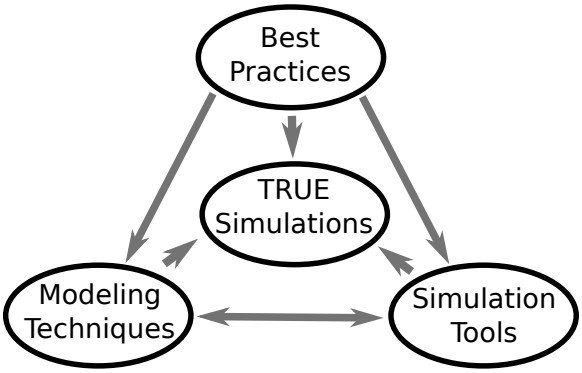


Figure C.1: Molecular simulations are becoming more informative and reproducible due to overlapping advances in modeling techniques and simulation tools through best practices in teaching, sharing, and software development.

"Best practices" refers to the use of open software, software engineering practices, and pedagogy for teaching computing generally and molecular simulations specifically. These practices have both a) enabled the creation of open source tools used broadly by the molecular simulation community, and b) been used within the molecular simulation community to advance simulation tools and modeling techniques (Figure C.1). The "simulation tools" discussed are used primarily to perform molecular dynamics simulations using pairwise potentials to model the interactions between simulation elements. Each of the main simulation engines is a significant feat of software engineering towards meeting their users' demands of application-specificity and performance. "Modeling techniques" refers to the algorithms used within simulation engines, interaction potentials (force fields), statistical sampling techniques, and theory. Explaining how practices, techniques, and engines are connected to each other is important because each area in isolation has near-infinite depth that can hinder accessibility to new researchers.

C.3 Timescale problems

Molecular simulations predict the structure and properties of materials using computer implementations of physics-based descriptions of matter. The recent re-

view article by Braun *et al.* provides a comprehensive overview of the components and considerations for molecular dynamics (MD) simulations [99], which we also focus on here. MD simulations suffer from two *scaling* timescale problems: The more atoms needed to represent a system, (1) the more calculation time is required to generate the next configuration, and (2) the more configurations need to be sampled before equilibrium is achieved. In other words, it takes a lot more time to simulate larger systems. These timescale problems derive from algorithmic scaling of calculating interactions between N simulation elements (often atoms) and because larger systems have more configurations (microstates) [100, 41]. Graphics processing units (GPUs) represent a major advance in computing hardware for ameliorating these scaling problems, and we recommend the 2010 review by Stone *et al.* as a starting point [101]. Because of the performance benefits of GPUs, all open-source MD packages now offer GPU support [102, 103, 104, 105, 106].

We introduce here the *training* timescale problem that MD and MC simulation techniques also suffer from: Researchers spend more time making and fixing modeling errors as the number of software dependencies and scientific topics needed for the model increases, especially if any of them are new to the researcher. The importance of the training timescale problem explains the growing efforts around training computational researchers [107]. Because scaling and training problems are obstacles to performing TRUE simulations, it is important for researchers to be mindful of tradeoffs between them when making modeling choices.

C.4 Best practices and cognitive load

Evidence-based instructional practices are being applied within communities of scientific software developers to create tools and training materials that feed back into these communities. Ambrose *et al.* provides a comprehensive review of the science of teaching, and is an accessible introduction to research around *cognitive*

load that we focus on here [108]. The basic idea of cognitive load is that the mental faculties of learners are finite, and their performance on a task (e.g., testing a new MD package) is hindered when they are asked to do more than one thing at a time [109]. The lens of cognitive load provides an accessible introduction to the research around stereotype threat and inclusivity, major barriers to participation of historically underrepresented groups in science, technology, engineering, and mathematics [110, 111, 112]. Reduction of cognitive load is a principle of course design [113, 114, 115, 97, 7, 116], human computer interactions [117], model-based computing [118, 119], and efforts to make academic writing more accessible [120]. In particular, Software Carpentry, Data Carpentry, and Library Carpentry (The Carpentries) are community-driven projects that apply the science of teaching (especially cognitive load reduction) to empower individuals to use computing in support of their professions [7, 107, 115, 97, 116]. We focus on cognitive load because of its centrality to tool accessibility and inclusive research communities.

For a sense of the ubiquity of cognitive *over*load in materials simulation, consider a novice simulator investigating how metal nanoparticles sinter on a surface during additive manufacturing in an atmosphere with alkanes. They begin with an xml file and discover they need to use a command-prompt to get it "in" to their lab's simulation engine. They review the literature to find dozens of seemingly appropriate forcefields with different parameterizations and functional forms [121]. After selecting the embedded atom model [122] to represent the metal atoms, they find difficulty choosing a forcefield for the atmosphere from MM4 [123], OPLS-AA [75], GAFF [124], COMPASS [87], and TraPPE [125], all with different models for the same compounds—how can this be? They consider the importance of charges, leading to Ewald summation [68] and polarizable force fields [126]. They begin to despair and wonder if compiling a density functional theory package will be faster. It isn't. Now with six unresolved lines of questioning and a command-

prompt that beeps at the letter 'f', the simulator feels like they're moving backward.

Beeping prompts and sad students are finding help in the modern pedagogy summarized above, facilitated by the development of open-source tools. It came as a surprise to many that large, decentralized software projects could be successful despite lack of private return [127]. However, people enjoy helping each other online, both for the enjoyment of sharing their experiences and building professional reputation, and researchers further derive utility from software that helps with research [128]. This is apparent in the development logs of Carpentries lessons: As one example, there are 1,942 commits from about 230 individuals since 2013, all attempting to make a better lesson for teaching the basics of version control (Software Carpentry commit log). Open-source, GPU-accelerated MD engines have experienced growth in community development over the same time frame (Figure C.2).

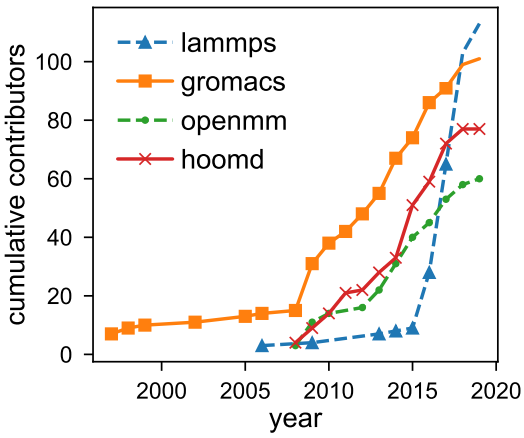


Figure C.2: Number of unique authors of four popular MD simulation engines over the last two decades. The increased growth around 2010 coincides with maturation of GPU technologies for MD and growth in Software Carpentry efforts. Numbers are approximate, as a few authors in each community may be double-counted if they commit with multiple pseudonyms.

This is not to say any one of the Carpentries, GPUs, or GitHub explains the recent

growth in open source science software, but instead emphasizes that coincident contributions to pedagogical practices, hardware advances, and online development communities are important in understanding this ecosystem.

Simulators are exchanging information beyond individual packages, now sharing teaching material under version control (David Kofke's molecular simulation course), pre-packaged virtual machines for workshops [129], and new journals for living documents of best practices, tutorials, and perpetual reviews. A number of organizations have grown around the support of sustainable software development for science and the work of Katz *et al.* provides a broad overview [130]. A list of tools for molecular simulation is included in Table C.1, and The OpenScience Project catalogs hundreds of open projects across disciplines.

Beyond tools designed for molecular simulation, there are important categories of tools for lowering the cognitive load of software development and fostering collaboration. GitHub, Bitbucket, and Gitlab are the three largest platforms [144] for collaborating on code repositories, and offer extra useful features at no cost for academic use. Messaging products Slack and Gitter are now popular for their integration with hosted repositories [145] and lower the barrier to entry for discussing issues and getting help. To lower the cognitive load of getting someone else's code to run, myBinder [146] enables users to launch Jupyter notebooks supporting multiple languages with pre-made enviroments for the code in question. As examples, the MoSDeF tutorials[147] use myBinder to spin up a Jupyter notebooks enabling users to begin tutorials without touching the software environment on their own computer. For solving the software stack problem on HPC clusters, singularity [148] enables users to deploy portable "containers" from open-source Dockerfiles[149] across multiple clusters, works with NVIDIA GPUS, and on many XSEDE resources.

Recurring themes in recommended readings [150, 7, 151] around best practices

Table C.1: Open source software helpful for materials simulations.

Package Name	Description
Diffractometer[47, 54]	Python code for generating scattering patterns from MD snapshots
foyer[131]	Python package for atom-typing
freud[132]	Python exposure of C++ analysis: RDF, order parameters, correlation functions
mBuild[133]	Python package for system initialization with reusable, hierarchical components enabling complex initialization from simple building blocks
MDAnalysis[134]	Python package for MD trajectory analysis supporting many file formats
MDTraj[135]	Python package for analyzing and extracting information from MD trajectories
MorphCT[47, 136]	Python package for obtaining and aggregating charge transport properties from MD snapshots
OVITO[137]	Python explosure of C++ analysis: Visualization, structure determination
packmol[85]	Library for initializing configurations of simulation elements
physical_validation[138]	Python package for performing thermodynamic consistency checks
Planckton[42]	Python package for initializing and executing HOOMD-Blue (`hoomd`) simulations
PLUMED[139]	Software for advanced sampling, using collective variables
pyLAT[140]	Python package used to manage LAMMPS output
Rhaco[141]	Python package for initializing and simulating molecules and atoms at surfaces
SSAGES[142]	Use collective variables and advanced sampling methods with amny engines
signac-flow[77, 78]	Python package for automating workflows including HPC schedulers and job submission
signac[77, 78]	Python package used to manage multi-dimensional data spaces and general workflows at scale
VMD[143]	Interactive and scriptable visualize and analysis of simulations
VOTCA[53]	Package for automating coarse-graining and charge transport calculations

and considerations for community-driven scientific software development include:

1. take into account cognitive load
2. use version control
3. automate repetitive tasks
4. collaborate on and share open code
5. write code in the highest-level language possible
6. software development is a fundamental literacy for engineers and researchers.

We also recommend Ref. [134] as an example of a clearly described scientific software package (MDAnalysis) with relevance to materials simulation.

C.5 Modeling Techniques

We now return to the original problem of advancing understanding of materials from molecular simulations and describe techniques for extracting more information from each step of an MD trajectory. Both coarse-grained models and advanced sampling help with scaling timescale problems by focusing on the key features of the phenomenon of interest, spending less effort on irrelevant details. In practice, implementing these techniques can lead to increased cognitive load if we are unaware of available infrastructure, so we organize key sources for learning more.

Briefly, by representing a collection of atoms with "coarse" simulation elements, significantly longer timescales are accessible because a) less computation is needed to compute the next configuration, and b) dynamics are accelerated because the underlying energy landscape is smoothed. Simplified models of polymers are among the first systems studied with molecular simulations [152, 153]

and the literature around coarse-graining is now extensive. For a polymer focus, see the recent perspective by Gartner and Jayaraman [154]. For biomolecules and protein folding there are many good sources including the reviews of Voth [155], Clementi [156], Elber [157], Klepeis [158], Kamerlin [159], and Kmiecik [160]. The MARTINI model stands out as a broadly successful coarse biomolecular model [161]. The specific problem of virus capsid self-assembly is reviewed in Perlmutter *et al.* [162]. For multiscale modeling, coarse-grained potentials can be derived by matching structure [163] or forces [164, 165], and the relative entropy framework of Shell and Chaimovich [166, 167] provides a measure of information loss through coarse-graining.

The calculation of free energy differences, rare events, and alternative approaches to sampling long dynamics can be accomplished by applying statistical mechanics to simulated trajectories. SSAGES [142] provides a comprehensive overview of advanced sampling techniques as well as open software for deploying them. Markov state models (MSM) [168] are statistical tools for describing the coarse dynamics of MD trajectories and provide a way of aggregating information from multiple short runs. MSMs themselves are a coarse-graining approach that has benefited from and contributed to the molecular simulation ecosystem. Machine learning approaches provide opportunities for extracting collective variables, trends, and patterns from materials simulations, and the review by Ferguson [169] provides a current, comprehensive view.

C.6 Organic photovoltaic structure and performance

In this section we review key topics in simulations of organic photovoltaics (OPVs) and describe our recent work in this context. OPVs convert photons into electrical current and engineering their structure to improve performance is an active area of research. For a more detailed picture of why OPVs are a promising technology for sustainable energy generation, start with [170, 171]. OPV

performance is strongly dependent on the morphology, and Refs. [172, 173] provide overviews of the key factors governing charge generation, separation, and transport. A review summarizing computational OPV morphology prediction at different length-scales is presented in Ref. [174]. Computational predictions of OPV morphologies are used as inputs into charge transport simulations that link OPV structure to metrics determining their efficiency. Refs. [173] and [175] explain charge generation and transport, while [176] explains how these properties can be simulated with kinetic Monte Carlo algorithms. We summarize recent morphology and charge transport predictions of the benchmark OPV material poly(3-hexylthiophene) (P3HT) in Table C.2. Combining hardware, software, and coarse-graining advances, routine simulations of P3HT have improved by roughly four orders of magnitude over the last decade (0.6 monomer-μs in 2010 vs. 6900 monomer-μs in 2018).

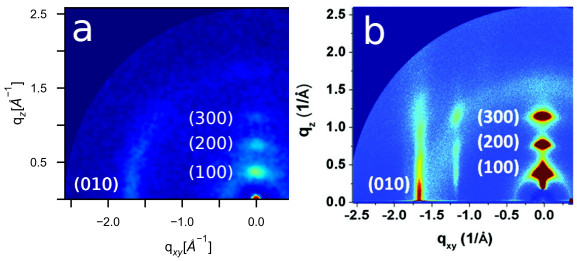
Table C.2: Overview of recent computational studies of P3HT, including method (MD or MC), resolution (AA - all-atom, UA - united-atom, 3CG - coarse-grained with 3 simulation elements per repeat unit, 1CG - coarse-grained with one element per repeat unit), approximate number of repeat units simulated, simulation time, computational effort (estimated largest product of Repeats Units $\times$ Time), and if the structures were used for charge-transport calculations. The work of Carrillo et al.[177] is a notable outlier, having successfully combined coarse models of millions of repeat units with development access to the then-most-powerful supercomputer on the planet. *Explicit numbers were not provided in the report, but are estimated here.

Year	Study	Method	Model	Repeat Units	Simulation Time	Effort (μs)	CT
2010	Moreno[178]	MD	AA	300	2 ns	6.0×10^{-1}	No
2010	Huang[179]	MD	AA	720	5 - 35 ns	2.5×10^{1}	No
2010	Huang[179]	MD	3CG	36864	10 ns	3.7×10^{2}	No
2011	Lee[180]	MD	1CG	X	10 ns	X	No
2013	Bhatta[181]	MD	AA	320-1280	5 ns	6.4×10^{0}	No
2013	D'Avino[182]	MD	UA	1600	60 ns	9.6×10^{0}	Yes
2013	Alexiadis[183]	MD	AA	2700*	20-45 ns	1.2×10^{2}	No
2013	Jankowski[45]	MD	3CG	2250-3750	1.7 μs	6.4×10^{3}	No
2013	Carrillo[177]	MD	1CG	3.2×10^{6}	400 ns	1.3×10^{6}	No
2016	Jones[184]	MD	3CG	4600-17000*	8 ns	1.4×10^{2}	Yes
2016	Scherer[185]	MD	3CG	8000	80 ns	6.4×10^{2}	No
2017	Jones[47]	MD	3CG	2250-3750	1.7 μs	6.4×10^{3}	Yes
2018	Miller[41, 90]	MD	UA	1500-15000	0.3-3 μs	6.9×10^{3}	Yes
2019	Greco[186]	MC	1CG	8000-16000	N/A	N/A	Yes

A challenge to making efficient OPVs is determining which combinations of photoactive compounds and thermodynamic conditions (temperature, pressure, concentrations) will result in a favorable morphology—a task well-suited to MD simulation. Probing this vast data space requires organizing ensembles of simulations, distributing them on high-performance computing clusters, retrieving the data, and then distilling the data into understandable chunks.

In our recent work [187, 48, 41, 90], we use HOOMD-Blue (`hoomd`) to predict morphologies of perylene, perylothiophene, P3HT, and poly(benzodithiophene-thienopyrrolo-dione) (BDT-TPD) oligomers using simplified (united atom) models. Neglecting partial charges and treating conjugated systems as rigid are two assumptions that lower cognitive load associated with force fields, avoiding the first-principles calculation of unknown charge, dihedral, and angle parameterizations missing from OPLS-UA or GAFF. These simplifications helped with both training and scaling timescale problems, resulting in morphology predictions in agreement with experiments [187, 48, 41] (see Figure C.3 and Figure C.4).

Figure C.3: a) Simulated and b) experimental grazing incident X-Ray scattering of P3HT show near identical features and wavenumbers along the (010) and (100) planes. The agreement indicates the same structures are being probed in both cases. Figure adapted with permission from Ref. [41].



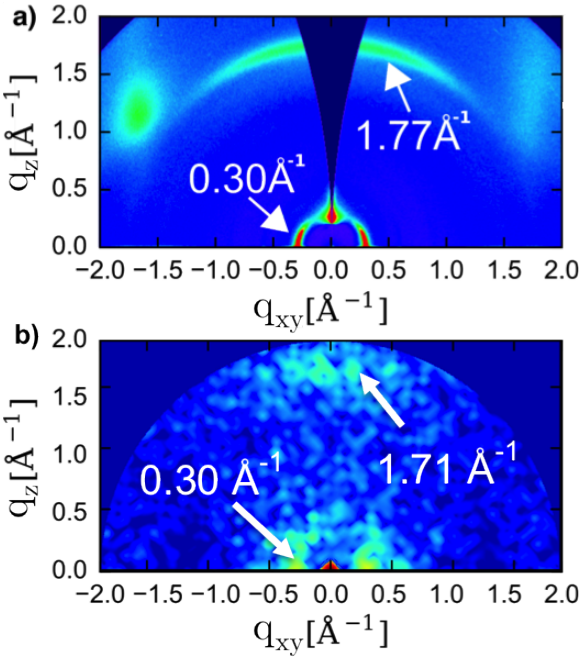


Figure C.4: a) Experimental and b) simulated grazing incident X-Ray scattering of BDT-TPD showing agreement. The agreement validates our simplified model. Reprinted with permission from[48]. Copyright 2017 American Chemical Society.

An example of developing transferable skills to deal with combinatorial explosion occured in this P3HT work: Only 14 temperatures, 6 solvent strengths, and 5 densities equates to 420 unique simulations. In each of these 420 cases, we aim to understand how the proximity and orientation of thiophene rings correlates with charge transport. A single structural descriptor is applied to each case, and a "phase-diagrams" is constructed for each density, providing a handful of figures summarizing large data spaces (Figure C.5).

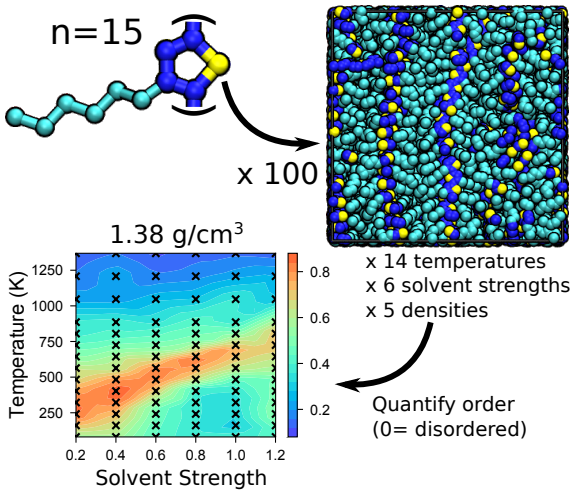


Figure C.5: 100 P3HT chains of 15 repeat units are represented as three simulation species: Yellow (sulfur), blue (aromatic carbon), and cyan (aliphatic carbon). Temperature, solvent strength (specified as a scaling of the Lennard-Jones well depth) and density, determine thermodynamic self-assembly of 420 unique structures. The self-assembled structures are quantified for ordering on the interval $[0, 1]$ - where 0 is completely disordered and 1 is completely ordered, according to the clustering of neighboring repeat units based on relative distances and orientations. Ordering is then summarized into "phase-diagrams" depicting the order as a function of these three variables; demonstrating how hundreds of simulations can be distilled to a few, quickly interpretable figures.

Testing transferability, reproducibility, usability, and extendibility of OPV modeling techniques is an exciting area of future work. For P3HT in particular, the number of models, simulation engines, and sampling schemes used makes it a good candidate for evaluating TRUE-ness. With a myriad of open approaches for coarse-graining, there is no fundamental reason why multiscale efforts validated by different groups could not be a test-bed for testing reproducibility these scales. Similarly, the availability of various charge-transport calculation approaches provides for opportunities to reproduce predictions of how charge transport depends on morphology and chemistry. Such cross-validation teams would help accelerate improvements around charge transfer calculations themselves, where opportunities exist to improve understanding of this broadly applicable phenomenon.

C.7 Predicting crosslinking dynamics

In this section we review computational approaches to predicting the crosslinked networks of toughened thermosets and discuss our recent work in this context. Thermosets are strong, low-density materials formed by the covalent bonding of liquid precursors into a 3D network that can be made less brittle through the introduction of a thermoplastic "toughener". In the fabrication of composite materials made from carbon fibers impregnated with toughened thermosets the network is "cured" through the heating and cooling of a part over time. The temperature history experienced by the part during curing influences the rates of diffusion and reaction, and therefore its resulting nanostructure and residual stresses. As the thermoset precursors crosslink, there is an entropic driving force for the phase-separation of the thermoplastic [188], which complicates nanostructure evolution. For a review of the key concepts in modeling thermosets (cure fraction, gelation, and glass transition temperature) see Li and Strachan [189]. The challenge focused on here is using molecular simulations to predict how thermoset formulation, toughener chemistry, and temperature history determine the cured nanostructure.

The central problems are those of scaling and training timescales, plus the fact that reacting systems are not in equilibrium. On the sampling side, the slow dynamics of gelling, glassy thermosets make relaxation intractably long even for small systems. Further, validating simulations against experimental systems with 1nm-100nm phase-separated length scales demands large simulated volumes. On the training side the main tensions are between faithful representation of reaction kinetics, coarse models that enable access to long timescales, and implementing these simultaneously. The knowledge that the equilibrium integration schemes available in `hoomd` and `lammps` are in conflict with the exothermic formation of bonds, and that using ReaxFF [190] won't permit sufficient volumes to be accessed

is liberating: It allows the question to be reframed as "How predictive of nanostructure can a simplified model of crosslinking thermosets be?"

To advance towards the goal of large, fast, predictive thermoset simulations we develop `epoxpy` as detailed in Ref. [191]. This was the first project in which we employed continuous integration into model development. Sanity checks built around initialization of the tougheners and the generation of trajectories with and without the reaction algorithms implemented allowed the submission of large ensembles of jobs to multiple clusters with confidence. This further allowed the research team to quickly progress through a series of models in support of the science question:

- DPD models are good for simplicity and performance, but not for representing entangled glasses
- LJ potentials and bond constraints can be parameterized to model entangled glasses
- Angle constraints are needed here for T_g measurements to fit the DiBenedetto expression
- In some cases, million-particle systems are needed to capture the microphase separated morphologies (Figure C.6)
- Bond-forming models can be made with `hoomd` plugins and calibrated against reaction kinetic models with and without heats of reaction

The results of our approach are summarized with recent simulations in Table C.3. The two distinguishing features of our recent work ([191]) are (1) the ability to investigate structural evolution while the model epoxies cure and (2) ability to do so for million-particle volumes in a few days or weeks on a single GPU.

Table C.3: Overview of recent reacting epoxy models, sorted by system size. Most efforts do not capture dynamic bonding during MD integration nor validate glass transition (T_g) as a function of cure fraction α (DiBenedetto expression).

Model (Chemistry)	Bond Probability	System Size	$T_g(\alpha)$ Validation
CGMD[192] EP/CA	Arbitary	5.0×10^3	No
AAMD[193] (DGEBA/DETA)	Arbitary	5.9×10^3	No
CGMD/AAMD[194] (DGEBA/DETA)	1	2.2×10^4	No
AAMD[195] (DGEBA/33DDS)	1	6.9×10^4	No
AAMD[196] (DGEBA/44DDS)	1	9.8×10^4	No
DPD[197] (DGEBA/DETA)	1	1.1×10^5	No
DPD[198] (DDS/RA/SA)	0.001	2.5×10^5	No
CGMD/DPD[199] (DGEBA/44DDS/PES)	$\sim \exp(\frac{E_a}{k_B T})$	4.0×10^6	Yes

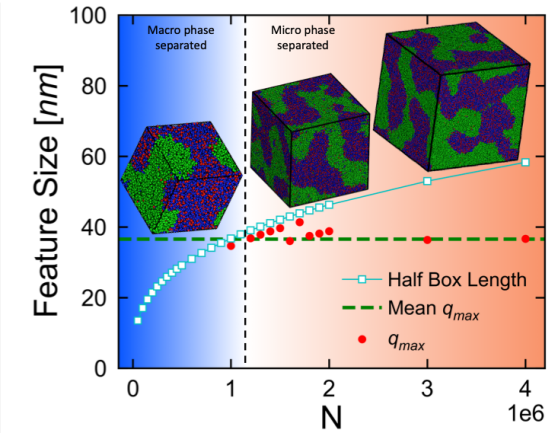


Figure C.6: The ability to perform curing simulations of million-particle toughened thermoset models enables the identification of sufficient box sizes. Here, divergence of the low-wavenumber structure factor is used to identify macrophase separation, and for small volumes (blue, left) of this CG model the morphologies appear macrophase separated. Large volumes (orange, right) of the same model show a local maximum in the structure factor (q_{max}) indicating microphase separation is observable when the length-scales of separation are smaller than half the simulation box length. Here, 1.2e6 particles (35nm boxes) are needed, reinforcing the importance of fast, large simulations for studying toughened epoxy thermosets.

Despite the simplifications, our coarse simulations match experimental reaction dynamics and glass transition temperatures [199]. Further, because we can vary temperature over the course of these reacting systems we can for the first time use MD to investigate how nanostructure depends upon temperature history during curing. Here we present new results (Figure C.7 and Figure C.8) summarizing the evolution of structure in two types of curing simulations where the primary activation energy is 2.1 dimensionless energy units, the secondary $E_A = 2.52$, $N = 400000$, $L = 73.7$nm and $dt = 0.01$, using the Lennard-Jones parameters from Table 5.1 and the fiducial simulation parameters from table 5.2 of [199]. Specifically, the ratio of coarse amine, epoxy, and toughener (A, B, and C) considered here is 1:2:2, with $\varepsilon_{AA} = 0.9216, \varepsilon_{BB} = 1.0, \varepsilon_{CC} = 0.8840$, $\varepsilon_{AB,AC,BC}$ are obtained using the Lorentz-Berthelot mixing rule (e.g. $\varepsilon_{AB} = \sqrt{\varepsilon_{AA}\varepsilon_{BB}}$), harmonic bond $r_0 = 1.0$ and $k = 100\frac{\varepsilon}{\sigma^2}$. The Langevin thermostat with drag parameter $\gamma = 4.5$ is used to ad-

vance simulation trajectories. Simulations were performed on NVIDIA K40 cards using hoomd 2.2.1 (commit hash f664aebdf55e44f10cdd6d5edc3a090f1bca713b),

In both figures, the wavenumber associated with microphase separation is plotted vs. time, with the simulation temperature (red) overlaid, along with the system's glass transition temperature (purple) which is a function of the degree-of-cure. In Figure C.7, the sample being cured at 0.8 kT (above T_g) is quenched to 0.5 kT (below T_g) before gelation, and in Figure C.8 the quench occurs after the onset of gelation. In both cases, the morphologies achieve the same degree of cure ($\alpha = 0.78$), and the standard error of five independent simulations are plotted with the grey error bars. We observe that curing post-gelation narrows the variance in cured structure, and these results demonstrate the importance of temperature history on cured morphology.

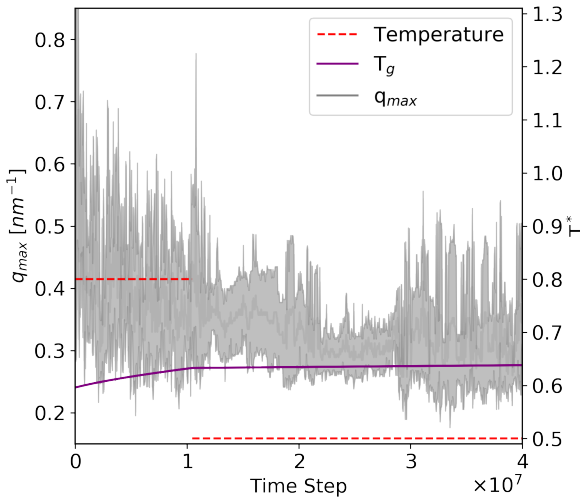


Figure C.7: Time evolution of the dominant length scale measured by the toughener-toughener structure factor for toughened reacting epoxy thermosets quenched below T_g (solid line) before gelation at time step 10,480,000. Curing temperature is shown by the dotted line. Error bars represent standard error from five independent simulations

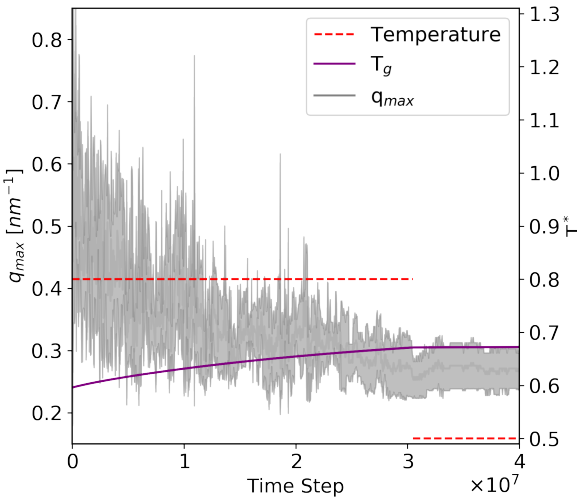


Figure C.8: Time evolution of the dominant length scale measured by the toughener-toughener structure factor for toughened reacting epoxy thermosets quenched below T_g (solid line) after gelation at time step 30,480,000. Curing temperature is shown by the dotted line. Error bars represent standard error from five independent simulations

The present example combining simplified models with continuous integration—two of the best practices from section C.4—demonstrates improved understanding of materials behavior. Because the code is open, the data is available, and because experimentation in this area is active, we identify epoxy thermosets as an area where we expect development around TRUE simulations to accelerate. Community validation of morphology predictions from simulations with varied temperature histories offers opportunity to increase the industrial impact of molecular simulations.

C.8 Training new simulators

In this section we describe several examples of on-boarding students to new projects wherein open tools (namely `hoomd`, mBuild, foyer, signac) and Software Carpentry pedagogy are used to reduce cognitive load and aid reproducibility. In addition to using the aforementioned tools directly in python scripts, we de-

velop two packages, Rhaco [141] and Planckton [42], that combine these tools to accomplish common tasks for specific systems. Rhaco facilitates the initialization and simulation of matter near surfaces and Planckton provides infrastructure for coupling MD simulations of OPVs to charge transport simulations (Table C.1). Examples of the diverse surface systems investigated with Rhaco are summarized in Figure C.9 and have enabled broad testing of forcefield compatibilities and qualitative investigation of surface phenomena while teaching students from multiple disciplines. In Figure C.10 we summarize DNA, fullerene, OPV material, asphaltene, and patchy particle models developed by new students leveraging mBuild and `hoomd`.

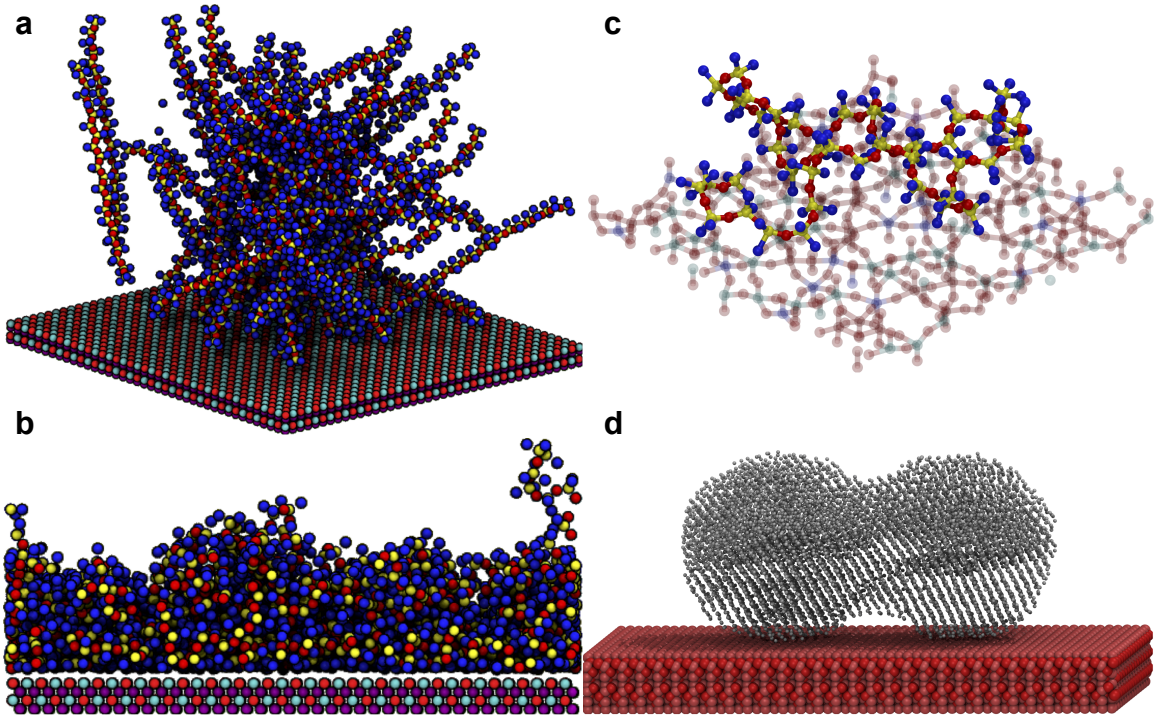


Figure C.9: a) PDMS chains initialized over NiMnGa, b) configuration of PDMS on NiMnGa by combining UFF and an OPLS-UA-derived potential, c) PDMS initialized on an M1 surface, d) Sintering silver nanoparticles on a corundrum surface combining EAM and UFF.

These examples are representative of students being able to initialize and debug models in weeks. In the case of Figure C.10a, quickly moving past the initialization step of building coarse-grained DNA identified issues with our imple-

mentation of the Knotts model. In Figure C.10b, the minimal physics for polar fullerene oxides solubilizing $C_{60}$was identified by trying seven coarse models. In Figure C.10c, initializing the complex truxene (a new candidate molecule for OPVs) with foyer immediately identified needed dihedral constraints. The asphaltenes components in Figure C.10d can be tuned for molecular weights and number of chains and was accomplished by an REU team. The patchy trapezoids in Figure C.10e permit quick testing of how patch size and shape influences assembly propensity.

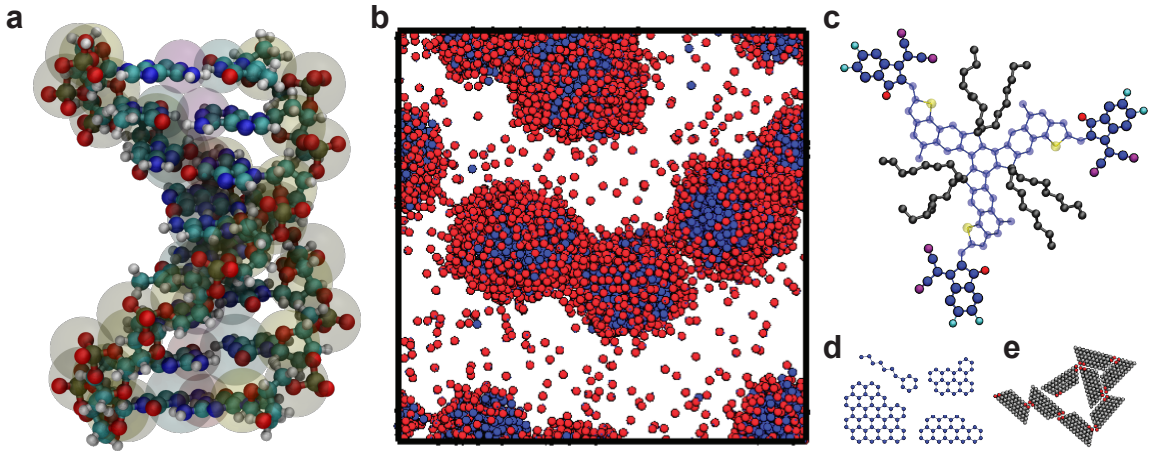


Figure C.10: a) Coarse-grained DNA initialized directly from sequence of nucleobases using the model from [200], b) Micelle self-assembly from a coarse-grained model of fullerene and their oxides, c) truxene molecule with electroactive components core and three functional groups, d) examples of programmatically-generated asphaltene components, e) 2D patchy particles designed to self-assemble terminal structures.

By combining tools for particular applications, we create "higher-level" languages for describing the system initializations in these examples. This enables the concepts of the models to be probed faster, lowering the load associated with initialization and parameterization, and the management of conversion between units and dimensionless quantities used in `hoomd`. Further, the code development benefits are bidirectional: By engaging with developer communities, students broaden their network of support and provide feedback that informs tool development.

All of the repositories mentioned here (`hoomd`, mBuild, foyer, signac, and Software Carpentry's instructor training) have now merged student pull requests originating from missing features or bugs encountered en route to the above work. These examples dispel the notion that aiming for simplicity necessarily creates "black-boxes" that limit student understanding of details, or the notion that higher-level languages necessarily create dependencies on established techniques. Rather, aiming for low cognitive load can focus development of new techniques and features with the best payoff for the researcher. Students anecdotally report increased interest and confidence, and this area represents a potential opportunity for future studies of professional identity, retention, and long-term career outcomes [201].

Related to the professional identities of materials simulators is the idea that certain scientist roles might enhance TRUE simulations. For example, should training prioritize the development of tool developers and tool users separately, or jack-of-all-trades scientists that can do everything? We find guidance to answering this question from our experiences with distributed code development: It is impossible for any individual to master everything, so leveraging the knowledge embedded in diverse development communities should be prioritized. Prioritizing diverse development communities highlights the need for the individuals in these communities to communicate and collaborate. The existence of and adherence to community codes of conduct (see The Carpentries code of conduct, for example) helps to ensure communities practice inclusivity and fosters collaboration. Within such communities we do not know if the prioritization of particular roles is yet warranted to enhance the realization of TRUE simulations, but we hold the optimistic opinion that communities offer the opportunity for the ensemble of individual interests and strengths to overlap in a way that makes any individual's shortcomings irrelevant.

C.9 Outlook

This instant in time represents a period of major change in the materials simulation ecosystem. Distributed coding communities have appeared, grown (Figure C.2), and disrupted the development practices and availability of simulation tools, continuing progress towards TRUE simulations since the field's origins [202, 203]. In this perspective we make the case for thoughtful training to take a central role in enhancing research reproducibility while simultaneously training researchers for broadly-needed technical roles. We also make the case that *en route* to TRUE simulations, these simulations should begin as less "true": Lowering cognitive load by sacrificing completeness now is made up for by increased efficiency and correctness later. However, TRUE simulations are not yet the norm. Materials simulators are in a position of opportunity and responsibility: We can demonstrate how reproducible science can be performed through increased engagement with our peers in the coming years. If the current momentum around communication and collaboration wanes, however; if the community becomes less inclusive rather than more, we may expect the amount of beeping prompts to increase. We take a more optimistic view: There has never been a better time to be a molecular simulator because of how active the community is with helping train its members to do transferable, reproducible, usable, extensible science.

C.10 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. 1835593, 1653954, 1658076, and 1229709. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575 [95]. This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. The authors gratefully acknowledge

sponsorship through the Laboratory Directed Research and Development (LDRD) program of the Idaho National Laboratory. This material is based upon work supported by The Boeing Company under contract BRT-LO118-0015. We acknowledge high-performance computing support of the R2 compute cluster (DOI: 10.18122/B2S41H) provided by Boise State University's Research Computing Department. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. KN, BA, MA, JG and EJ thank the Shodor Education Foundation and the Blue Waters Student Internship Program for support of this work.

The raw and processed data required to reproduce these findings are available to download from https://doi.org/10.5281/zenodo.2619585.

APPENDIX D:

AB-INITIO STUDIES OF EXCITON INTERACTIONS OF CY5 DYES

The following chapter was published in Journal of Physical Chemistry A in 2018 under authors Jenny W. Fothergill, Andres Correa Hernandez, William B. Knowlton, Bernard Yurke, and Lan Li. My contributions to this paper under the guidance from Drs Li, Knowlton, and Yurke were all DFT, python, and java calculations, data curation and analysis, all figure creation, and all draft writing.

This was my first published paper and I have learned a lot since its publishing, so I'd like to give a little further context with guidance for how it could be improved. At this time I wrote some python code which used singular value decomposition to fit the mostly-planar dye molecules to a vector and a plane. I showed a graphic for how this was done in the SI, but did not include any code or any details of what software versions I used. The method is not theoretically complex and if I recall correctly I used a function from the `Scipy` library, but this work would have been more reproducible had the exact code and its dependencies been provided in the supporting information (SI). At the time, however I was not aware of version control using git or GitHub and instead used Dropbox's history. I did my best to report all details for the computations I performed although they were done in a proprietary package (Gaussian09), so the source code is not available. Along with the SI, PDB files for all DFT optimized structures were included, but

I did not provide any of the raw input or output files used by Gaussian, nor did I report how I converted these structures to PDB. Another bespoke code used in the publication, referred to as the KRM code, was developed in Java by Dr Yurke and its theory is given in Cannon et al. [204, Supporting Information], but to my knowledge the source code is not publicly available or under version control, so I could not report its version or give instructions for how this work could be reproduced. This critique of my past work is by no means a condemnation of it; my effort and the guidance I received from my collaborators were an invaluable learning experience. Learning to use git and GitHub is an initial barrier which adds to cognitive load—a novice computational scientist needs to learn the theory and background of their method, how to navigate the shell, how to use a new software, and perhaps a scripting language, so although using version control is a vital part of writing reproducible scientific code, initially learning how to use it may seem like an unwanted additional hurdle. I was also somewhat ignorant of what details I should report—at the time I don't think I knew how to determine what version of python or java or which additional libraries I was using. Additionally, DFT was used to get the optimized orientations of the oblique dimer in DNA and these orientations were used as input to the KRM code in order to get an absorbance spectrum to compare with experiment. DFT does not take into account temperature, and due to computational limitations the DNA cannot be included in the model, so perhaps a better method for obtaining the average orientations of these dye molecules would've been to use molecular dynamics, which could more easily model this larger system.

D.1 Abstract

The excited state properties of cyanine dyes and the orientations of their aggregates were studied using density functional theory (DFT). The effects of exchange-correlation functional and solvent model on the absorption spectrum of Cy5 was

investigated. Using the 6-31+G(d,p) basis set and B3LYP exchange-correlation functional with IEF-PCM (water) solvent, the predicted spectrum achieved a maximum absorbance within 0.007 eV of experiment. An in-house program based on the theoretical model of Kühn, Renger, and May (KRM), which predicts the orientation of dyes within an aggregate from its absorbance and circular dichroism (CD) spectra or vice versa, was used to investigate the orientation of an experimentally observed dimer. The absorbance spectrum predicted using the KRM model of the dimer structure optimized with the 6-31+G(d,p) basis set, ωB97XD exchange-correlation functional, and IEF-PCM (water) solvent agrees with experimental data.

D.2 Introduction

Chromophore aggregates in the light-harvesting complexes of photosynthetic organisms have been shown to exhibit exciton delocalization, in which an electron-hole pair is delocalized over spatially separated chromophores [205, 206, 207]. Exciton delocalization plays a role in the energy transfer to the reaction center in photosynthesis and was first observed in nonbiological molecular crystals [208, 209, 210, 211]. Early steady-state absorption measurements suggested exciton delocalization was present in chlorophyll from spinach photosystem I [212]. Later, femtosecond two-dimensional spectroscopy was used to observe the indicative quantum "beating" between chromophores in bacteriochlorophyll at 77K, a result that was later confirmed at room temperature [213, 214]. These excitonic phenomena are of considerable interest due to their potential for applications in the realms of solar energy harvesting and quantum optics [215, 216, 217]. A signature of exciton delocalization in dye aggregates is the shifted absorption maxima (relative to the monomer) due to molecular transition dipole interactions. When two chromophores come close enough that their transition dipoles interact, the excited state energy is split. Two commonly formed types of aggregate

are known as J- and H- aggregates [218]. J-aggregates, named after the chemist E. E. Jelley, display a narrow, intense, bathochromic absorbance (i.e., red shift) with nearly resonant fluorescence, while H-aggregates display a hypsochromic shift (i.e., blue shift) in absorbance maxima and quenched fluorescence as explained by the molecular exciton theory of Kasha [219]. Figure D.1 illustrates the energy split-

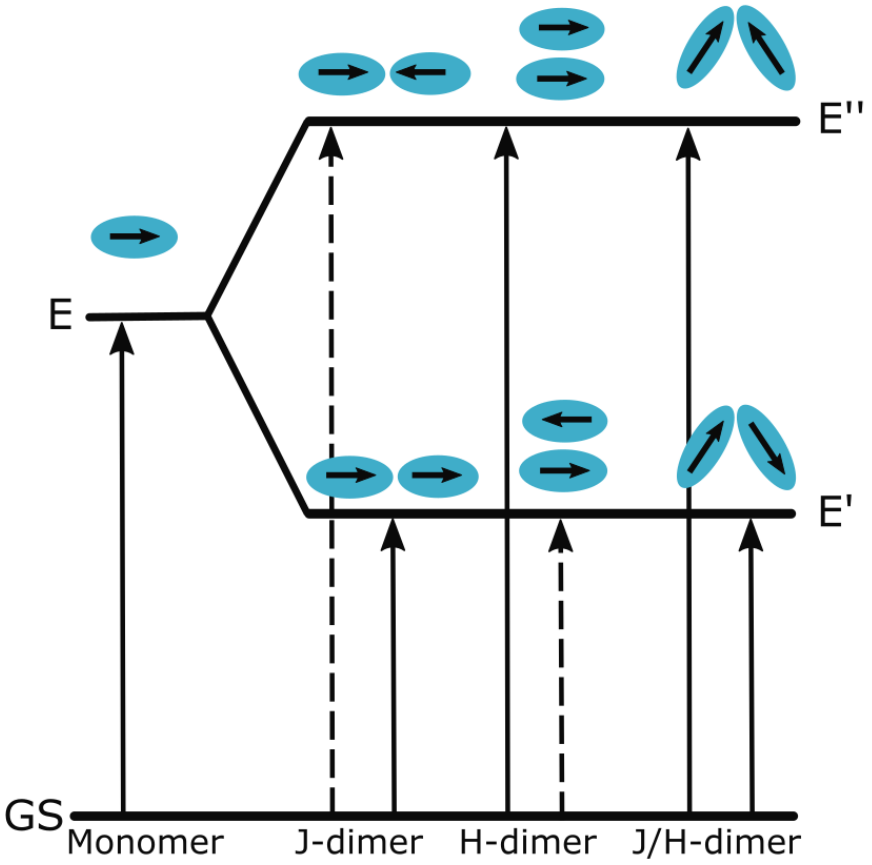


Figure D.1: Energy diagram based on molecular exciton theory of Kasha showing the excitation pathways for the J-dimer, H-dimer, and oblique dimer relative to the monomer [218]. The allowed (solid) and forbidden (dashed) transitions result from the orientations of the molecular transition dipole moments.

ting and allowed states of different dye dimers. The transition dipole moment is assumed to be parallel to the long axis of the dye molecule. The selection rules for light absorption involve taking the vector sum of the transition dipoles, so only transitions with net non-zero vectors are allowed. Oblique aggregates, in which

the alignment of the transition dipole moments of the dyes is at some angle between the H- or J- configurations, have allowed transitions to both the higher and lower energy states, so the absorbance spectra show Davydov splitting of their absorbance maxima [209]. Due to their excitonic properties, dye aggregates have been proposed to be used for light harvesting and excitonic quantum computing applications [220, 221, 222, 223]. However, the chromophores must be precisely spaced in order to achieve predictable exciton delocalization behavior as small changes in the orientation of the chromophores will cause large shifts in the absorption spectra.

Recently, cyanine dyes covalently incorporated into duplex DNA have been studied for their exciton delocalization properties that include J- and H-aggregate behavior and Davydov splitting that manifest as red shifts, blue shifts, and simultaneous red and blue shifts in absorbance, respectively, as compared to the dye monomer. The self-assembly properties of DNA can bring dyes within separation distances that induce shifts in the absorption maxima. Using DNA as a scaffold allows for manipulation of the orientation of cyanine dye molecules. DNA nanotechnology enables the precise construction of nanodevices due to the self-assembly of nucleic acids determined by Watson-Crick base pairing. The use of DNA as a building material has demonstrated precise control of 2D and 3D nanoscale shapes [224, 225]. It has been shown that Cy5 dimers covalently bound into the sugar-phosphate backbone of duplex DNA adopt a J-dimer configuration in 100 mM sodium chloride [226]. Another study found that Cy3 dimers covalently attached to DNA showed H-dimer absorbance in which the intensity varied with base-pair separation distance and the rigidity of the DNA scaffold [227]. The exciton-coupling strength of Cy3 dimers in double-stranded DNA has been found to decrease with an increase in temperature [228]. It has also been found that varying the salinity of a solution containing Cy5 in DNA with magnesium chloride

(from 0 mM to 100 mM) as well as the DNA concentration resulted in a conformational change from a J- dimer to H- tetramer configuration [204]. The relative orientation of the dyes in the DNA scaffold has been investigated by fitting the absorption and circular dichroism (CD) spectra using an in-house program based on the theoretical model of Kühn, Renger, and May (KRM) [204, 229, 230]. In this model, the dominant vibrational mode of the electronic ground state and the electronic excited state of each molecule in the aggregate is treated non-perturbatively by including its Hamiltonian in the system Hamiltonian that is diagonalized on a truncated Hilbert space in order to obtain the system energy eigenvalues. To go beyond the approximation in which the exchange interaction arises from point dipoles, the program, from here on referred to as the KRM code, treats the interaction using the extended dipole model [231]. In this manner, the program is able to more accurately model aggregates of rod shaped molecules, like Cy5, for which nearest neighbor distances can be shorter than the length of the molecules. Cannon et al. used the KRM code to find the orientation of the chromophores by varying the input orientations and fitting parameters until a good fit was obtained between the theoretical and experimental spectra [204, 229]. Having determined the values of the fitting parameters for Cy5, the KRM code can be used to predict the absorbance and CD spectrum of a given Cy5 aggregate structure. Further detail about the methods implemented in the code can be found in Cannon et al. [204].

This paper investigates the excited state and intermolecular Cy5 (see Figure D.2 for structure) dye interactions using ab-initio density functional theory (DFT)-based approach. Determining the position and orientation of these molecules is a vital first step to study their excitonic behavior. This work is in support of a previously published article in which the orientation of dyes covalently bound to duplex DNA was investigated by analysis of the absorption and circular dichro-

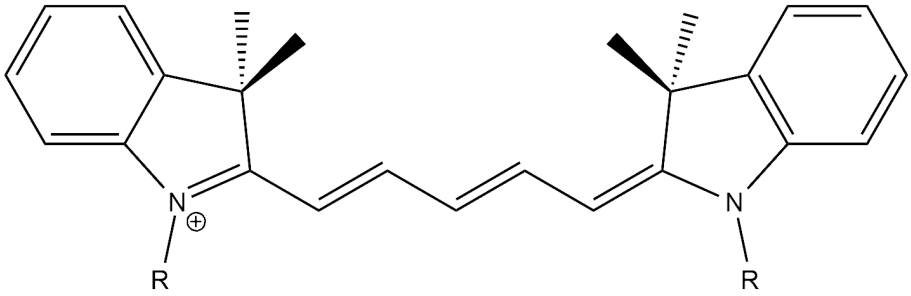


Figure D.2: The molecular structure of Cy5 (or 1-1′-dimethyl-3,3,3′,3′-tetramethylindocarbo-cyanine). Linkers are attached at R groups: for H-dimers and monomer, R = methyl, for oblique dimer, R = propyl chain.

ism spectra using the KRM model [204]. The dimer orientation predicted by Cannon et al. for a Cy5 dimer in duplex DNA in 0 mM MgCl2 from analysis with the KRM code of the absorbance and circular dichroism data (this dimer structure is from here on referred to as the oblique dimer) was also optimized using different exchange-correlation functionals [204]. The predicted spectra of the optimized oblique dimer structures using the KRM code were compared to the experimental spectrum obtained by Cannon et al. [204]. Vibrationally resolved absorbance spectra for the Cy5 monomer were generated using each exchange-correlation functional and compared to the experimental spectra. The computational results advanced our understanding of exchange-correlation functional effect on the structural stability and excitonic phenomenon of the Cy5 materials.

For the Cy5-DNA materials system to be considered as a viable candidate for excitonic applications, control of position of Cy5 dyes in different DNA assemblies must be demonstrated. In our work, DFT-based electronic structure calculations helped determine the orientation of Cy5 dyes in DNA at the ground state. TD-DFT modelled a system in the excited state, revealing the effect of exchange-correlation functional and solvent on the absorbance spectra.

D.3 Methods

All ab-initio calculations were performed via the Gaussian09 quantum chemistry package [232]. Molecules were first optimized using the semi-empirical PM6 method [233]. These structures were then optimized to a residual force of 4.50x10-4 Hartree/Bohr (2.31x10-2 eV/Å) using the Kohn-Sham formulation of DFT with the B3LYP, CAM-B3LYP, or ωB97-XD hybrid exchange functionals and 6-31+G(d,p) basis set [234, 235, 236, 237, 238]. B3LYP is a three-parameter hybrid, combining Hartree-Fock, GGA, and LDA with no long-range or dispersion correction [234]. CAM-B3LYP also considers long-range correction by the Coulomb attenuating method [235]. ωB97-XD has both long-range and damped empirical dispersion corrections [236]. The effect of basis set extension was also investigated using the 6-311++G(2df,2pd) basis set. These basis sets and exchange-correlation functionals were chosen based on a study which found them to work well for the Cy5-DNA system [239]. Because the research focus is on the excited state energies, the diffuse functions (represented by +) and the polarization functions (represented by the p, d, or f orbital designation) in the basis set were used to better approximate the higher energy molecular orbitals. All solvated structures were optimized using the polarizable continuum model of water solvent using the integral equation formalism (IEF-PCM) [240]. Structures were confirmed to be minima on the potential energy surface by the lack of negative vibrational frequencies.

The terms flipped and stacked denote the position of the tertiary amine groups in the dimers: stacked refers to the amine groups of each dye molecule being on the same side while flipped refers to the amine groups being on opposite sides (see Figure D.3). The initial positions of the flipped and stacked H-dimer were chosen by placing the centers of the molecules within 1 nm of each other but no closer than 5 Åbecause any closer is not physically likely to occur. The atoms in

the H-dimer were fully geometrically relaxed.

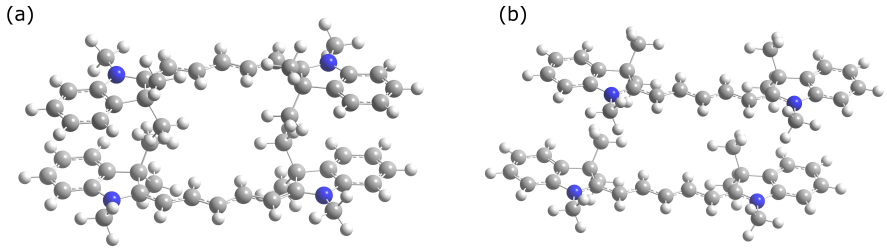


Figure D.3: The initial structures for the flipped (a) and stacked (b) Cy5 H-dimers. The blue atom represents the nitrogen in the amine group while carbon is in grey and hydrogen is in white.

The initial position of the oblique dimer was obtained by matching the dimer position with the orientation vectors obtained by Cannon et al. using the KRM code [204]. In the experiments of Cannon et al., the oblique dimer was covalently bound into DNA, so to model this situation, the initial orientation of the oblique dimer from the KRM code was covalently bound into the DNA backbone with propyl linkers and optimized using the universal force field (UFF) in the Avogadro molecule editor [241, 242]. Then, the position of the terminal carbon atom on the linker group was frozen to simulate binding to DNA, and the DNA was removed to reduce computational time before optimization using PM6 followed by DFT optimization with a hybrid functional.

To investigate the effect of solvent on the excited state energies, a linear response formalism which adds the solvent terms to the excited state equations was employed, and the geometry of the lowest excited state was optimized with IEF-PCM water solvent. To account for the Duschinsky effect (i.e., the change in vibrational modes upon electronic transition), the Adiabatic Hessian approach was used to expand the excited state potential energy surface around the equilibrium excited state geometry [243, 244]. To obtain a vibrationally resolved absorption spectra the magnitude of the transition dipole moment between vibrational lev-

els in the ground and excited states was approximated using the Franck-Condon (FC) with and without the Herzberg-Teller (HT) approximation for comparison [245, 246, 247, 248, 249]. The resulting stick spectra were broadened using Gaussian functions with a half-width at half-maximum (HWHM) of 300 cm-1.

D.4 Results and Discussion

Table D.1 and Table D.2 show the solvation energy for the Cy5 monomer and H-dimer and the dye-dye interaction energy of the H-dimers in consideration of various basis set and exchange-correlation functional combinations. The solvation energy is calculated as follows:

$$E_{sol} = E_T - E_v \tag{D.1}$$

where E_T is the energy of the solvated molecule and E_v is the energy of the relaxed molecule structure in vacuum. The interaction energy of the dimer is calculated as follows:

$$E_{int} = E_{dimer} - 2 \times E_{monomer} \tag{D.2}$$

where E_{dimer} is the energy of the dimer and $E_{monomer}$ is the energy of the monomer.

Table D.1: Solvation energy of relaxed Cy5 monomer for given basis sets and exchange-correlation (xc) functionals

Basis Set	xc-functional	Solvation Energy (eV)
6-31+G(d,p)	B3LYP	-1.488
6-31+G(d,p)	CAM-B3LYP	-1.501
6-31+G(d,p)	ωB97-XD	-1.506
6-311++G(2df,2pd)	B3LYP	-1.477
6-311++G(2df,2pd)	CAM-B3LYP	-1.492
6-311++G(2df,2pd)	ωB97-XD	-1.493

The solvation energy is used as a qualitative measure to determine which dye orientation will be more stable as the experiment we aim to support takes place in aqueous solution [204]. The more negative the solvation energy, the more energet-

Table D.2: Solvation energy and interaction energy of relaxed Cy5 H-dimers (flipped and stacked) for given basis sets and exchange-correlation (xc) functionals

Structure	Basis Set	xc-functional	Solvation Energy (eV)	Interaction Energy (eV)
H-dimer (flipped)	6-31+G (d,p)	B3LYP	-4.234	0.002
	6-31+G (d,p)	CAM-B3LYP	-4.343	-0.039
	6-31+G (d,p)	ωB97-XD	-4.460	-0.636
	6-311++G (2df,2pd)	B3LYP	-4.104	0.008
	6-311++G (2df,2pd)	CAM-B3LYP	-4.334	-0.025
	6-311++G (2df,2pd)	ωB97-XD	-4.427	-0.572
H-dimer (stacked)	6-31+G(d,p)	B3LYP	-4.413	-0.003
	6-31+G(d,p)	CAM-B3LYP	-4.638	-0.069
	6-31+G(d,p)	ωB97-XD	-4.690	-1.054
	6-311++G(2df,2pd)	B3LYP	-4.239	0.009
	6-311++G(2df,2pd)	CAM-B3LYP	-4.624	-0.044
	6-311++G(2df,2pd)	ωB97-XD	-4.663	-1.033

ically favorable for the constituent to exist in the solvent. Table D.1 and Table D.2 show that the solvation energy of the dimer is more negative per chromophore than that of the monomer, suggesting that the dimer structure would be more energetically favorable in a polar solvent such as water. We suggest the reason for this is the reduction of hydrophobic interactions with the solvent. The monomer calculations show no significant difference in either structure or solvation energy as exchange-correlation functional varies. For the dimer structures, however, the way in which the functional models consider long-range and dispersion interaction differs and thus impacts the results. Compared to the flipped structure, the stacked H-dimer structure overall has more negative solvation energy because in the optimized geometry this dimer is more compact and thus required a smaller cavity in the solvent.

The interaction energy provides insight into the strength of the intermolecular interaction of the two dyes. Table D.2 shows that the interaction energy between the H-dimers using B3LYP is very small or even positive suggesting that

this functional underestimates the intermolecular dye interactions because similar cationic cyanogen dyes are known to aggregate in aqueous solution [250, 251, 252]. Considering the long-range correction (CAM-B3LYP) in the exchange-correlation functional lowers the energy, but adding dispersion corrections (ωB97-XD) results in the lowest energy (i.e., the most favorable interaction) for both the flipped and stacked H-dimer. The interaction and solvation energies indicate that the stacked H-dimer is more energetically favorable, i.e., more stable, than the flipped structure, even though stacking the methyl groups could result in steric hindrance; potentially, favorable pi-stacking of the aromatic rings contributes to the lower interaction energy.

As Cy5 is a cationic dye, a preliminary investigation into the effect of the presence of an explicit (chlorine) counter ion was done (see Supporting Information Table E.1). However, analysis with the KRM code suggests that the oblique dimer structure which was relaxed with counter ions is not a better fit to experiment (see Figure E.1).

Comparison of the solvation energy or the interaction energy of the H-dimer structures optimized using the same exchange-correlation function and either the small (6-31+G(d,p)) or large (6-311++G(2df,2pd)) basis sets shows that these energies do not differ between basis sets. A good agreement between the relaxed dimer structures within the same exchange-correlation functional and between basis sets further confirms that the basis set superposition error (BSSE) is negligible in the small basis set. (For a detailed orientation of each dye within the dimer structures, see Table E.2 and Table E.2.) Finally, this similarity of the optimized structures between basis sets can be visualized using the absorbance spectra predicted by the KRM code, where small changes in dye orientation will induce large shifts in the absorbance (see Figure E.2).

To compare the spectra, and thus the difference between the dimer structures, the

root mean square error (RMSE) is provided as a measure of the difference between two absorbance data as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{1i} - y_{2i})^2}{n}} \tag{D.3}$$

where y_1 and y_2 are the y-axis (absorbance) values corresponding to the first and second data sets and n is the number of points. The lower the RMSE, the more similar the spectra, and if the two data sets are equal, the RMSE will be zero.

Including the DNA in the DFT calculation makes the system unmanageably large, so investigation of the effect of the DNA scaffold on the Cy5 dimer structure was done in stages. First, the initial structure of the oblique dimer, which was experimentally observed in duplex DNA, was designed to fit the orientation vectors found by Cannon et al. using the KRM code [204]. As confirmation, following the UFF relaxation of the dye dimer in DNA structure, the vector fit of the relaxed dimer structure was re-entered into the KRM code to ensure that the predicted spectrum continued to match the experimental spectrum. For information on the vector fitting employed, see Figure E .3. Figure D.4 confirms that the generated and experimental spectra are in good agreement. (RMSE = 0.0543)

The next stage in the investigation of the effect of DNA scaffolding is based on the simplifying assumption that the main contribution of the DNA scaffold is the position restraint imposed by the alkyl linker chains. This assumption is based on necessity (the system is too large for DFT) and the observation the absorbance of the monomer does not change appreciably when bound to DNA (see Figure E.4). After relaxation in DNA with UFF, the oblique dimer with fixed linker chains was optimized using PM6 and then DFT using the 6-31+G(d,p) basis set (chosen based on the results of the H-dimer calculations which showed that the BSSE using this basis set for the Cy5 dimers was not significant) and various hybrid functionals

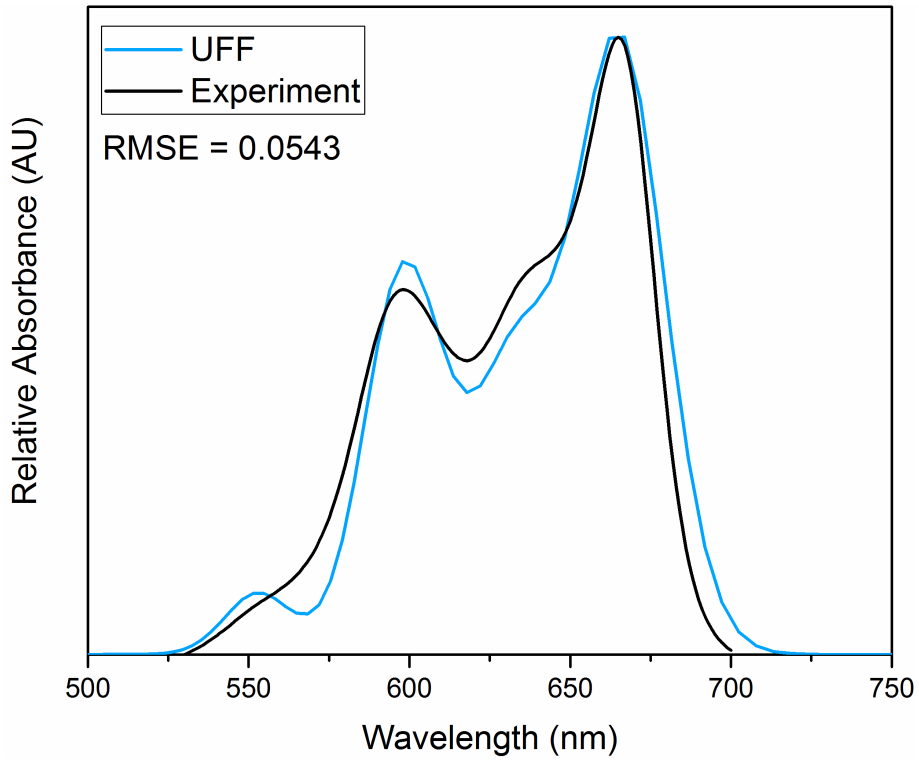


Figure D.4: Theoretical absorbance spectrum generated using KRM code for the oblique dimer structure relaxed using UFF compared to experimental absorbance of oblique dimer obtained by Cannon et. al. Initial structure for oblique dimer was designed based on vectors determined by Cannon et. al., so as expected the spectra show good agreement between the theoretical and experimental [204]. RMSE value provided for quantification of difference.

(B3LYP, CAM-B3LYP, and ωB97XD). Figure D.5 shows the predicted spectra of the relaxed structures using the KRM code. For the corresponding orientation of the dyes after relaxation, see Table E.4.

The small peak at 550 nm in all predicted spectra corresponds to a vibronic transition which is not observed in experiment most likely due to temperature or solvent induced peak broadening. Comparison of the predicted spectra in Figure D.5 with the experimental spectra suggest that the dimer drifts away from the orientation found by experiment after (a) the PM6 optimization (the RMSE value increases from 0.0543 to 0.1195) and drifts even further upon optimization using (b) B3LYP (the RMSE value increases from 0.1195 to 0.1336); however, after optimization with the long range corrected (c) CAM-B3LYP and dispersion corrected (d)

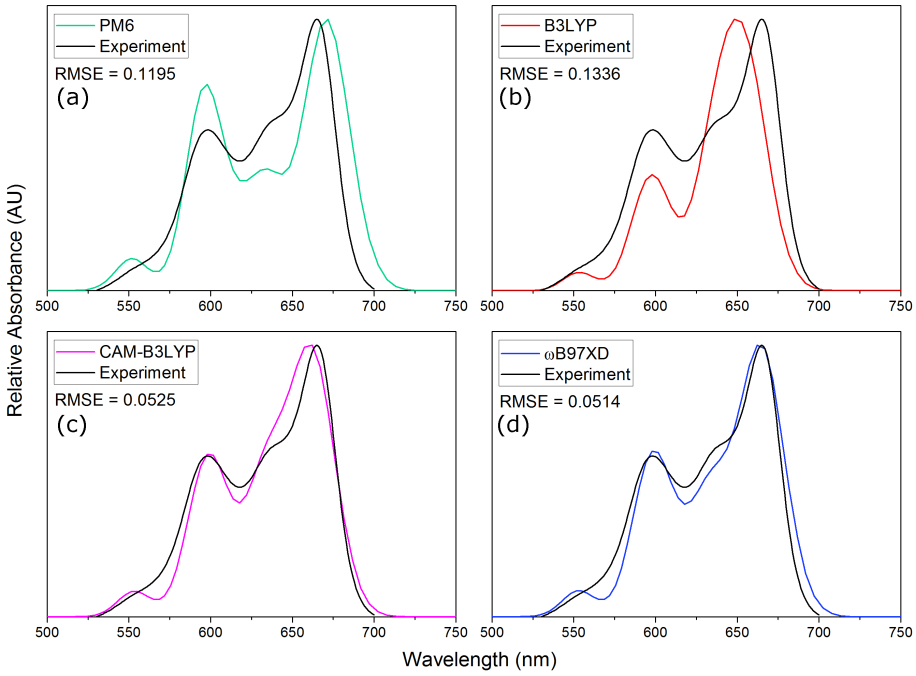


Figure D.5: Theoretical absorbance spectra for the relaxed oblique dimer generated using KRM code compared to experimental absorbance of oblique dimer obtained by [204]. All structures were relaxed with the (a) PM6 semi-empirical method before further relaxation with a hybrid functional: (b) B3LYP, (c) CAM-B3LYP, or (d) ωB97XD. RMSE value provided for quantification of difference.

ωB97XD functionals, the predicted spectra become a better fit to the experimental spectrum (the RMSE decreases from 0.1195 to 0.0525 and 0.0514 for CAM-B3LYP and ωB97XD, respectively), and even show a better agreement with the experimental spectrum than the initial position which was designed to be a good fit (Figure D.4, RMSE=0.0543). Although the KRM method requires reducing the molecules to vectors and does not involve any ab-initio energy calculations from atomic positions, this result suggests that long-range and dispersion correction are vital to accurately modelling the intermolecular interactions of the dye aggregates. Additionally, this system was also optimized using dispersion corrected B3LYP, but analysis of the resulting structures with the KRM code suggests they are not as good a fit to experiment [253, 254] (see Figure E.5). The structure of the oblique dimer optimized using ωB97XD, which provides the best fit of the predicted spectrum using the KRM code to the experimental spectrum, is shown in

Figure D.6.

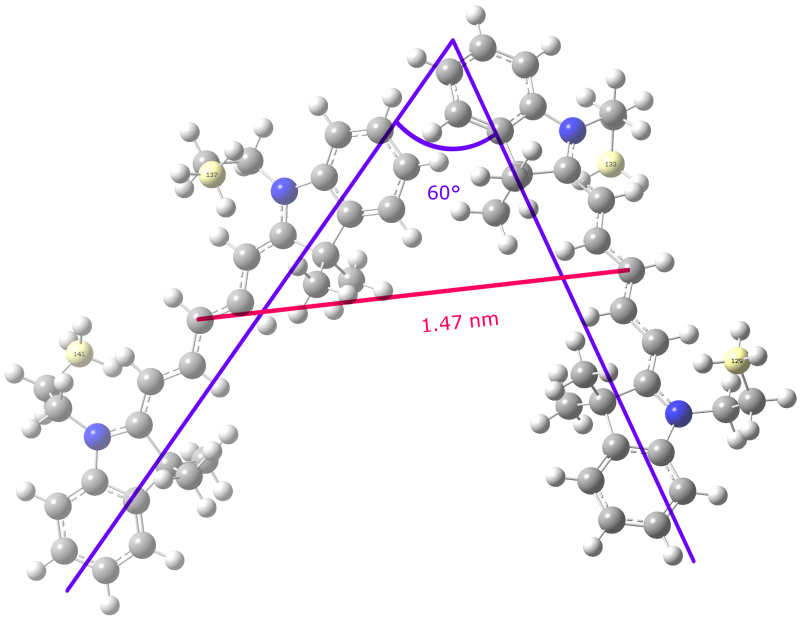


Figure D.6: Structure of the Cy5 oblique dimer with propyl linkers optimized using ωB97XD functional. The terminal carbon atoms of the linker chain (highlighted in yellow) were "frozen" during relaxation. See Figure D.5(d) for the predicted spectrum from this structure using the KRM code.

In addition to the spectra generated using the KRM code, an ab-initio calculation of the absorbance spectrum of the monomer was performed using the Franck-Condon approximation with and without the Herzberg-Teller approximation on vibrational modes determined using DFT and TD-DFT (see Figure D.7 for molecular structure). The intensity of absorption depends on the square of the electronic transition dipole moment and the radiation frequency, and it is also assumed that the Born-Oppenheimer approximation, where nuclear motions are much slower than electronic transitions, holds true, so during an electronic transition the nuclei can be considered static [255]. The Franck-Condon (FC) approximation further assumes that the electronic transition dipole also remains static, while the Herzberg-

Teller (HT) approximation allows for linear variation of the dipole moment with respect to the nuclear coordinates during the transition [247, 248, 256, 257]. The FC approximation can predict fully dipole-allowed transitions while the HT approximation can better predict weakly allowed or dipole forbidden transitions. Table D.3 shows the shift in absorption maxima (relative to experiment) for spectra generated using the FC and FC with HT approximations.

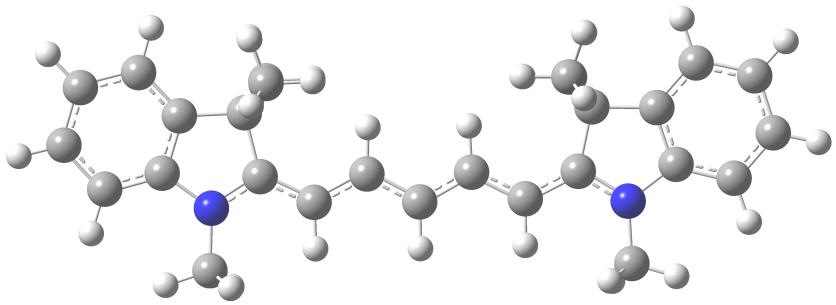


Figure D.7: Molecular structure of Cy5 monomer optimized using 6-31+G(d,p) B3LYP in IEF-PCM water solvent. The average C=C bond length in the methine chain is 1.4 Å.

Although the optimized geometries of the Cy5 monomers do not vary appreciably between exchange-correlation functionals (see Table E.5), the transition energies from the ground to excited state are found to depend strongly on the optimization conditions. Table D.3 and Figure D.8 show that the conditions which bring the wavelength of maximum absorbance, λ_{max}, of the predicted spectrum closest to that observed in experiment are obtained using the Franck-Condon approximation on structures optimized with the B3LYP functional in IEF-PCM water solvent ($\Delta\lambda_{max}$x 0.007 eV or 2.2 nm). The consideration of long-range (CAM-B3LYP) and

Table D.3: Comparison of the difference in maximum absorbance ($\Delta\lambda_{max}$) of experimental Cy5 monomer spectrum [204] to absorption spectra generated using the Franck-Condon (FC) approximation with or without contribution from the Herzberg-Teller (HT) approximation. Approximation schemes use TD-DFT and DFT results relaxed using the 6-31+G(d,p) basis set and three different xc-functionals in IEF-PCM (water) solvent.

xc-functional	Approximation	$\Delta\lambda_{max}$ (eV)
B3LYP	FC	0.007
B3LYP	FCHT	0.080
CAM-B3LYP	FC	0.188
CAM-B3LYP	FCHT	0.207
ωB97XD	FC	0.215
ωB97XD	FCHT	0.222

dispersion correction (ωB97XD) overestimate the ground to excited state energy transition of the monomer.

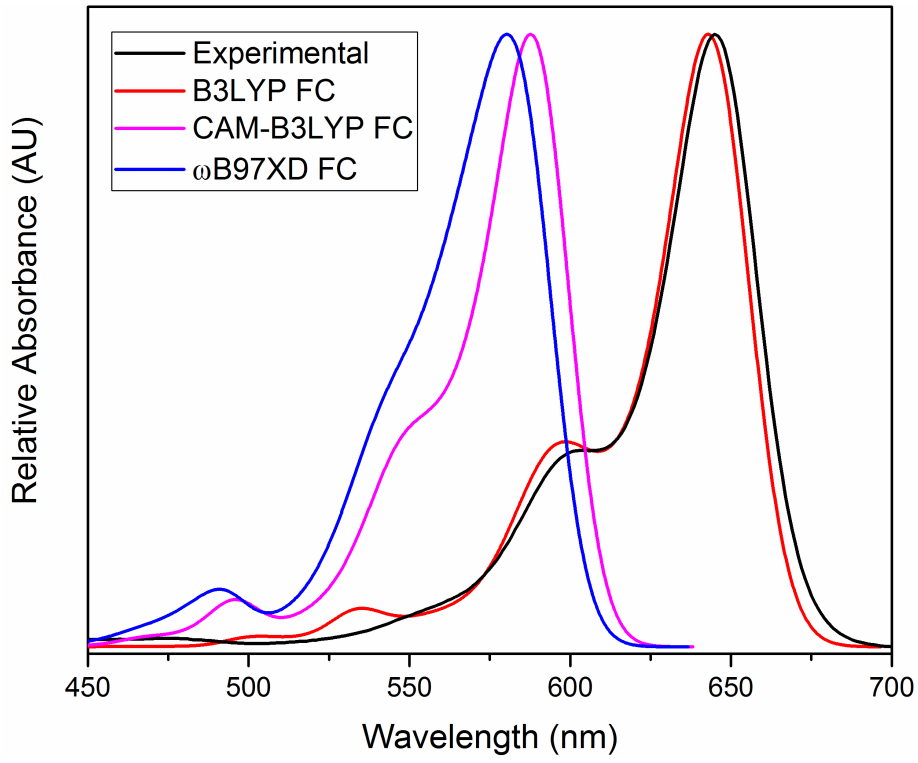


Figure D.8: Comparison of spectra generated using the FC approximation to experimental spectrum obtained by [204]. The almost complete overlap of the predicted absorbance spectrum generated using B3LYP and the FC approximation (red line) with the experimental spectrum (black line) suggest that the conditions not only accurately predict λ_{max} but also the vibronic structure. All structures were optimized using 6-31+G(d,p) basis set in IEF-PCM (water) solvent.

For all spectra the predicted energies of the transitions are higher than observed in experiment. This is expected; TD-DFT using hybrid functionals has been shown to overestimate the vertical absorbance transition energy of cyanine dyes [258, 259, 260]. The addition of solvent is known to improve the prediction of the transition energy [259] and appears to red-shift the energies and improve the calculation of the most intense transition, λ_{max}, for all xc-functionals regardless of whether only the FC or also the HT approximations are considered. (See Figure E.6 for spectra from vacuum calculations.) Figure D.8 provides a comparison of the Cy5 monomer spectra generated with various exchange-correlation functionals in IEF-PCM water solvent using the FC approximation. The good agreement between the absorbance spectrum generated using the B3LYP exchange-correlation functional with the FC approximation and the experimental spectrum shows that not only does this method accurately predict λ_{max}, but the relative strength of the vibronic peaks also seem to be in good agreement.

The use of the HT approximation does not improve the calculation of the most intense transition which suggests that the contribution of weakly-allowed or dipole-forbidden transitions to the absorption spectrum of a Cy5 molecule is negligible. Figure D.9 compares the spectra for the same structures generated using the FC approximation with the HT approximation and shows that this method is not as useful for predicting the absorption spectrum for this system as the FC approximation alone. As opposed to the HT principle, using the FC principle the symmetric ground state vibration can only couple with symmetric vibrations; the intensity distribution of the band shapes will be dominated by one vibrational mode [261]. The good agreement of the absorbance spectrum generated using the FC principle with experiment suggests that primarily one vibrational mode contributes to the vibrational profile of Cy5.

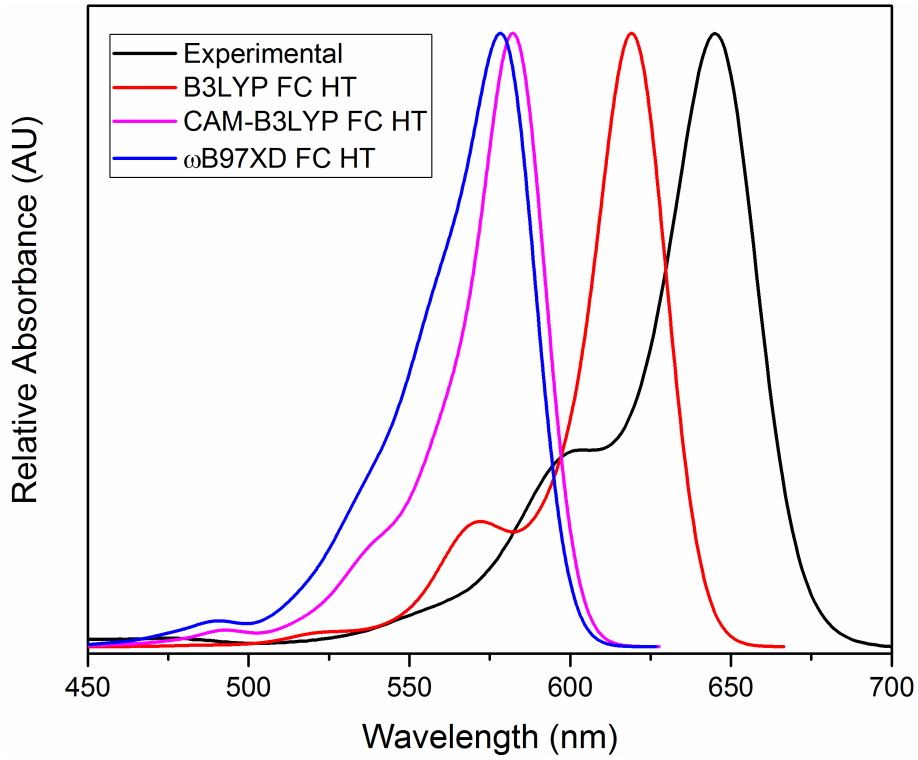


Figure D.9: Comparison of spectra generated using the FC approximation with the HT approximation to experimental spectrum obtained by [204]. Comparison of these spectra with those in Figure D.8 show that adding the HT approximation does not improve accuracy. All structures were optimized using 6-31+G(d,p) basis set in IEF-PCM (water) solvent.

D.5 Conclusions

Comparison of the H-dimer structures within different exchange-correlation functionals shows that the structures optimized using the smaller 6-31+G(d,p) basis set are not significantly impacted by BSSE. By comparing the spectra calculated using the in-house KRM code with the experimental dimer spectrum, the oblique dimer structure optimized using the ωB97XD functional in IEF-PCM water solvent provides the best agreement with the experiment. It suggests that the long-range and dispersion corrections imposed by the exchange-correlation functional are needed to accurately estimate the dye-dye interactions. Comparison of the vibrationally resolved electronic absorption spectra of the monomer produced using the FC and HT approximations shows that the spectrum obtained using the structures

optimized with the B3LYP functional in IEF-PCM water solvent and the FC approximation agrees well with the experimental monomer spectrum. The B3LYP functional works well for a single molecule while the long-range and dispersion correction could over-estimate the transition energy for a single molecule. For future work, we will combine quantum mechanical and molecular mechanical (QM/MM) calculations to incorporate DNA into the chromophore system. We will also calculate vibrationally resolved absorption spectra of the oblique dimer structures to continue revealing the effect of exchange-correlation functional. Our work aims to investigate the effects that various simulation conditions have on the unobservable atomic structures and, in turn, the effects that the atomic structure has on the observable spectra. By understanding what factors are most important when simulating the system, we hope to contribute our understanding to the knowledge of how to best optimize this system in experiment.

D.6 Associated Content

D.6.1 Supporting information

The following files are available free of charge: Molecular structure files for oblique dimers relaxed with UFF, PM6, B3LYP, CAM-B3LYP, and ωB97XD. (obdimer-UFF.pdb, obdimer-PM6.pdb, obdimer-B3LYP.pdb, obdimer-CAM-B3LYP.pdb, obdimer-wB97XD.pdb) Information about calculations including counter ions and additional dispersion corrected xc-functionals; tabulated orientation vectors and centers of mass for the H- and oblique dimers, and spectra predicted using the KRM code for the H-dimers; information about vector fitting; comparison of the absorbance spectrum of free and bound Cy5; tabulated structural information for Cy5 monomer; comparison of Cy5 monomer spectra obtained without solvent model to experimental spectrum obtained by [204] (Appendix E)

D.7 Author Information

D.7.1 Corresponding Author

*Boise State University C/O Lan Li, 1910 University Drive, Boise, ID 83725-2090

lanli@boisestate.edu

D.7.2 Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

D.8 Acknowledgements

The research was supported by National Science Foundation INSPIRE Grant No. 1648655. We would like to acknowledge the efforts of Dr. Brittany L. Cannon, Lance Patten, Allison Christy, and Donald Kellis, who provided their absorbance data and a basis for study of this system. We would like to thank our research group members, Matt Lawson, Ember Sikorski, and Thiago da Silva, for their support and helpful discussions of DFT theory. We would also like to thank Dr. Matt King and Sean Ruettgers for helpful discussion of Gaussian software and DFT theory. This research made use of the resources of the High Performance Computing Center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. Department of Energy and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517. We would like to acknowledge high-performance computing support of the R2 computer cluster (DOI: 10.18122/B2S41H) provided by Boise State University's Research Computing Department.

D.9 Abbreviations

DFT density functional theory; TD-DFT time-dependent density functional theory; KRM Kühn, Renger, and May; BSSE basis set superposition error; IEF-

PCM integral equation formalism polarizable continuum model; UFF universal force field; FC Franck-Condon; HT Herzberg-Teller; HWHM half-width at half-maximum; RMSE root mean square error; QM/MM quantum mechanical and molecular mechanical; CD circular dichroism

APPENDIX E:

SUPPORTING INFORMATION FOR AB-INITIO STUDIES OF EXCITON INTERACTIONS OF CY5 DYES

As Cy5 is a cationic dye, the presence of an explicit (chlorine) counter ion will likely affect the solvation and interaction energy. It was found that inclusion of the counter ion(s) resulted in a less negative solvation energy and a more negative interaction energy for the dimer structures and a more negative solvation energy for the Cy5 monomer (see Table E.1). This does not necessarily suggest that the structure without the counter ion is more soluble, but instead because the vacuum structure is more stable with the counter ion, the solvent does not need to do as much to stabilize the charge.

Table E.1: Solvation energy and interaction energy of relaxed Cy5 H-dimers (flipped and stacked) and monomer for given basis sets and exchange-correlation (xc) functionals with explicit counter ion(s). The energies of the same structure optimized at the same level of theory without the counter ion(s) are also included for comparison.

Structure			Solvation Energy (eV)	Interaction Energy (eV)
Cy5 monomer	6-31+G(d,p)	B3LYP	-1.488	–
Cy5 monomer + $1Cl^-$	6-31+G(d,p)	B3LYP	-1.947	–
H-dimer (flipped)	6-31+G(d,p)	ωB97-XD	–	-0.636
H-dimer (flipped) + $2Cl^-$	6-31+G(d,p)	ωB97-XD	–	-0.665
H-dimer (stacked)	6-31+G(d,p)	CAM-B3LYP	-4.638	-0.069
	6-31+G(d,p)	ωB97-XD	-4.690	-1.055
H-dimer (stacked) + $2Cl^-$	6-31+G(d,p)	CAM-B3LYP	-3.322	-0.088
	6-31+G(d,p)	ωB97-XD	-2.980	-1.170

The relaxed structure for oblique dimer with chlorine counter ions was also calculated using the 6-31+G(d,p) basis and ωB97XD functional. Analysis of this structure with the KRM code (see Figure E.1) shows that it is not as good of fit as the oblique dimer without the chlorine counter ions. (The RMSE between the predicted spectrum of the oblique dimer without chlorine and experiment with this functional is 0.0514.)

As our method used a finite basis set, the overlap of basis functions can cause an increase in the effective basis set which in turn result in a lower energy solution to

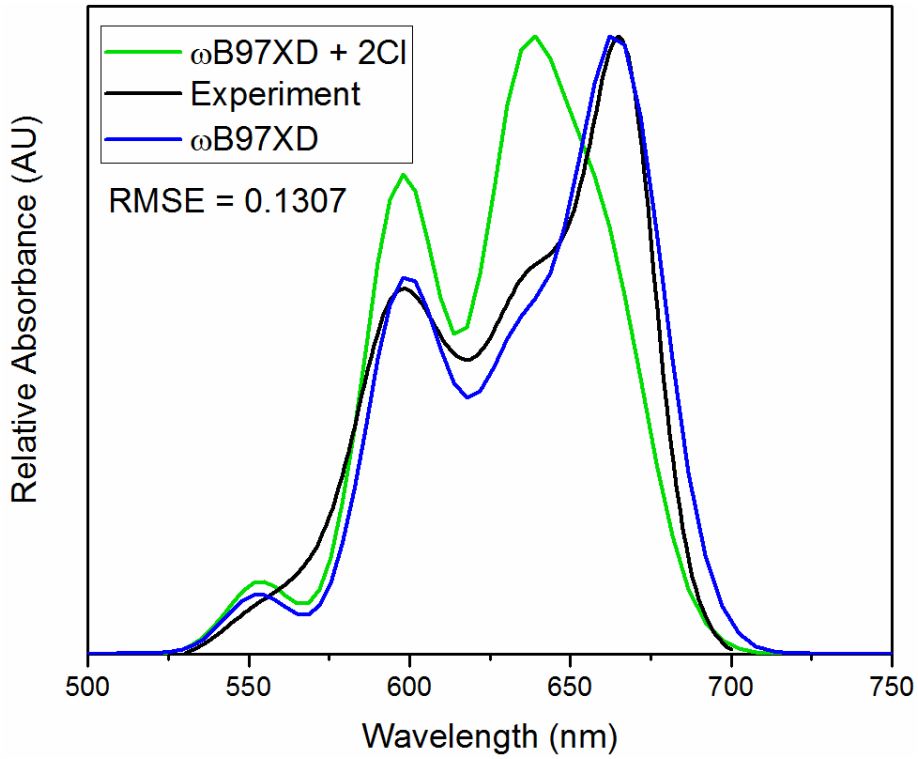


Figure E.1: Theoretical absorbance spectrum generated using KRM code for the oblique dimer with chlorine counter ions relaxed using ωB97XD xc-functional compared to experimental absorbance of oblique dimer obtained by Cannon et al. [204] and oblique dimer without chlorine ions. RMSE value provided for quantification of difference.

the Schrödinger equation. The difference in energy between the finite overlapping basis set and the theoretical infinite basis set is called basis set superposition error (BSSE). By comparing the energies and geometries of the systems optimized with a small basis set to those optimized using a large basis set, the amount of BSSE present can be qualitatively described. Using a series of basis sets and exchange-correlation functionals, the optimized geometries of the Cy5 monomer and H-dimer were compared to determine the extent of BSSE.

Except for the case of the B3LYP exchange-correlation functional (Figure E.2 a and b), there is good agreement of the spectra generated using the H-dimer structures relaxed using the 6-311++G(2df,2pd) and 6-31+G(d,p) basis sets, suggesting the degree of BSSE present in the 6-31+G(d,p) basis set does not impact the energy or

Table E.2: Orientation of the relaxed Cy5 stacked H-dimer structures for given basis sets and exchange-correlation (xc) functionals including the zenith (θ) and azimuth (ϕ) angles for the vectors which lie along the long axis of the chromophores (arbitrarily labelled dye 1 and 2), the zenith (θ_p) and azimuth (ϕ_p) angles for the perpendicular vector which points in the direction of the methyl groups connected to the tertiary amine (see Figure E.3 for a graphical representation of these vectors), and the centers of mass of the chromophores.

Basis Set	xc-functional	dye	θ (°)	ϕ (°)	θ_p (°)	ϕ_p (°)	center of mass		
							x (nm)	y (nm)	z (nm)
6-31+G (d,p)	B3LYP	1	71.00	173.64	93.93	-97.64	-0.033	-0.006	0.411
		2	72.48	48.20	84.11	139.97	0.033	0.006	-0.411
6-31+G (d,p)	CAM-B3LYP	1	72.77	-179.77	95.74	-91.47	-0.013	0.018	0.347
		2	71.25	41.80	83.60	133.86	0.013	-0.018	-0.347
6-31+G (d,p)	ωB97-XD	1	81.42	-173.25	97.43	-84.36	0.008	0.013	0.273
		2	78.15	35.08	81.03	126.92	-0.008	-0.013	-0.273
6-311++G (2df,2pd)	B3LYP	1	69.90	179.28	92.62	-91.62	-0.053	-0.023	0.488
		2	75.83	43.05	92.71	132.39	0.053	0.023	-0.488
6-311++G (2df,2pd)	CAM-B3LYP	1	72.75	-179.44	95.47	-91.06	-0.012	0.016	0.348
		2	71.23	41.41	83.44	133.52	0.012	-0.016	-0.348
6-311++G (2df,2pd)	ωB97-XD	1	80.80	-173.96	97.28	-85.13	0.008	0.004	0.267
		2	78.14	35.71	83.95	126.95	-0.008	-0.004	-0.267

structure enough to justify using the 6-311++G(2df,2pd) basis set. The disagreement occurs in the B3LYP case due to the absence of long range and/or dispersion correction in the exchange-correlation functional, showing failure to accurately estimate the inter-dye interactions. These results suggest that the magnitude of basis set BSSE in the small basis set does not significantly impact the resulting energies enough to justify using the large basis set, so the small basis set was used for further calculations to reduce computational time.

In order to succinctly represent the relative orientation of the dyes in a dimer structure, the dye molecules are simplified to vectors. To calculate these vectors, the atom positions in dye molecules are fit to a plane, using the method of least squares, and a vector, using a singular value decomposition (Figure E.3). As the molecules are longer than their width, only the vector which lies along the long axis of the molecule is used to represent the molecule in the KRM code. A vec-

Table E.3: Orientation of the relaxed Cy5 flipped H-dimer structures for given basis sets and exchange-correlation (xc) functionals including the zenith (θ) and azimuth (ϕ) angles for the vectors which lie along the long axis of the chromophores (arbitrarily labelled dye 1 and 2), the zenith (θ_p) and azimuth (ϕ_p) angles for the perpendicular vector which points in the direction of the methyl groups connected to the tertiary amine (see Figure E.3 for a graphical representation of these vectors), and the centers of mass of the chromophores.

Basis Set	xc-functional	dye	θ (°)	ϕ (°)	θ_p (°)	ϕ_p (°)	center of mass		
							x (nm)	y (nm)	z (nm)
6-31+G (d,p)	B3LYP	1	89.72	11.53	79.78	101.58	-0.176	-0.027	0.315
		2	94.92	29.68	106.35	121.06	0.176	0.027	-0.315
6-31+G (d,p)	CAM-B3LYP	1	94.30	20.99	85.06	110.62	-0.155	-0.033	0.246
		2	93.96	19.61	104.34	110.59	0.155	0.033	-0.246
6-31+G (d,p)	ωB97-XD	1	95.08	20.31	85.37	109.90	-0.139	-0.037	0.208
		2	94.77	20.05	101.54	111.01	0.139	0.037	-0.208
6-311++G (2df,2pd)	B3LYP	1	89.84	9.56	79.50	99.59	-0.228	-0.040	0.383
		2	97.12	31.23	106.69	123.27	0.228	0.040	-0.383
6-311++G (2df,2pd)	CAM-B3LYP	1	94.25	20.87	84.98	110.49	-0.155	-0.033	0.245
		2	93.93	19.79	104.16	110.75	0.155	0.033	-0.245
6-311++G (2df,2pd)	ωB97-XD	1	95.13	20.22	85.47	109.81	-0.138	-0.036	0.208
		2	94.68	20.17	101.86	111.13	0.138	0.036	-0.208

tor which points in the direction of the methyl groups connected to the tertiary amine, lies in the plane of the molecule, and is perpendicular to the long axis of the molecule was also defined to provide the absolute orientation of the chromophores. The zenith and azimuth angles (θ and ϕ, respectively) for these vectors for the H- and oblique dimers in spherical coordinates are given in Table E.2, Table E.2, and Table E.4.

Assuming that the DNA scaffold does not have a large impact on the electrostatic environment of the Cy5 dyes may be a gross oversimplification; however, this assumption is supported by the observation that the absorbance spectrum of Cy5 does not exhibit a shift in absorbance maxima or noticeable change in vibronic structure when bound to DNA (see Figure E.4).

The structure of the oblique dimer was also relaxed using the long range dispersion correction implemented by Grimme [253]. Two- (D2) and three-body (D3)

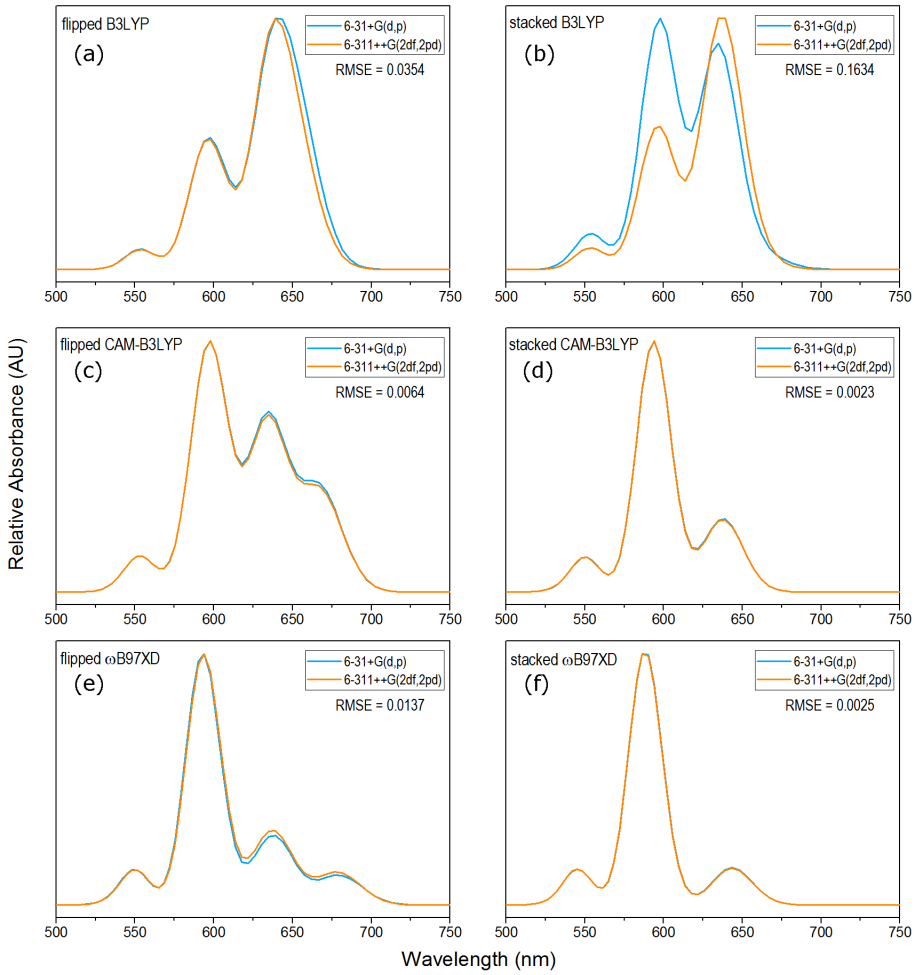


Figure E.2: Comparison of basis set effect on theoretical absorbance spectra of (a)(c)(e) the flipped and (b)(d)(f) stacked H-dimer structures predicted using KRM code. With the exception of the stacked structure relaxed using (b) B3LYP, the spectra show significant overlap for the 6-311++G(2df,2pd) and 6-31+G(d,p) basis sets suggesting the smaller basis set is adequate. RMSE value provided for quantification of difference.

dispersion effects have been considered, the latter also includes rational damping according to the formula of Becke and Johnson (BJ-damping) [254]. It was found that the absorbance spectra predicted using the KRM model are not more similar to experiment than ωB97XD (see Figure E.5). Interestingly, however, the oblique dimer structures optimized using dispersion-corrected B3LYP (D2 and D3-BJ) are more plane-parallel than those optimized using any other method, which may mean that this method models pi-stacking, which should require parallel stacking of the conjugated bonds, better than other dispersion-corrected methods.

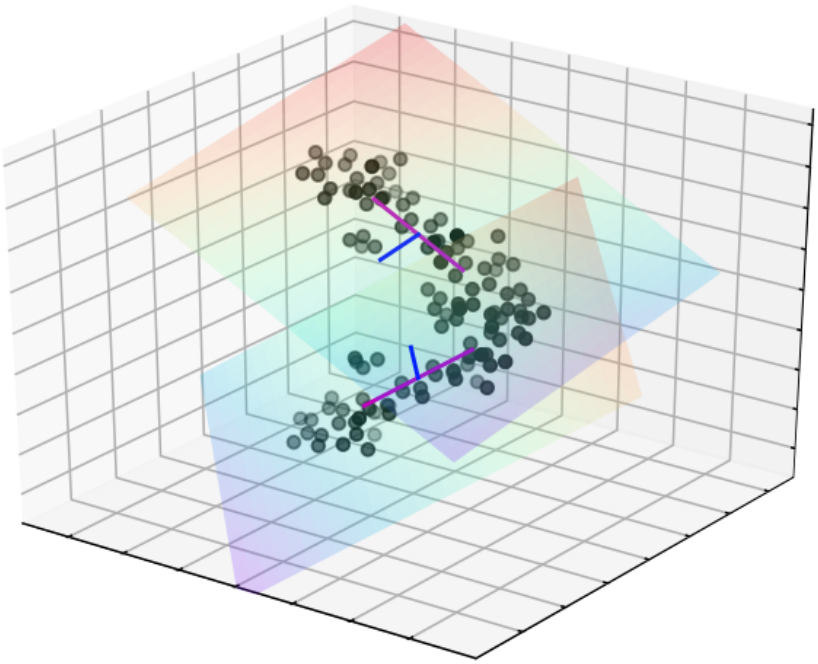


Figure E.3: Graphic showing the vector fit to long axes of the molecules (magenta), the plane of the molecule (rainbow), and the perpendicular vector pointing in the direction of the tertiary amine (blue).

Table E.4: Orientation of the Cy5 oblique dimer optimized using the 6-31+G(d,p) and specified exchange-correlation (xc) functionals including the zenith (θ) and azimuth (ϕ) angles for the vectors which lie along the long axis of the chromophores (arbitrarily labelled dye 1 and 2), the zenith (θ_p) and azimuth (ϕ_p) angles for the perpendicular vector which points in the direction of the methyl groups connected to the tertiary amine (see Figure E.3 for a graphical representation of these vectors), and the centers of mass of the chromophores.

xc-functional	dye	θ (°)	ϕ (°)	θ_p (°)	ϕ_p (°)	center of mass		
						x (nm)	y (nm)	z (nm)
B3LYP	1	105.67	-21.32	118.72	45.21	0.507	-0.343	0.308
	2	29.20	33.96	147.32	-137.60	-0.507	0.343	-0.308
CAM-B3LYP	1	103.82	-18.46	111.98	30.32	0.495	-0.282	0.313
	2	21.97	32.08	159.01	-148.70	-0.495	0.282	-0.313
ωB97-XD	1	100.96	-13.05	11.72	5.39	0.506	-0.235	0.325
	2	19.55	33.20	109.59	-151.56	-0.506	0.235	-0.325

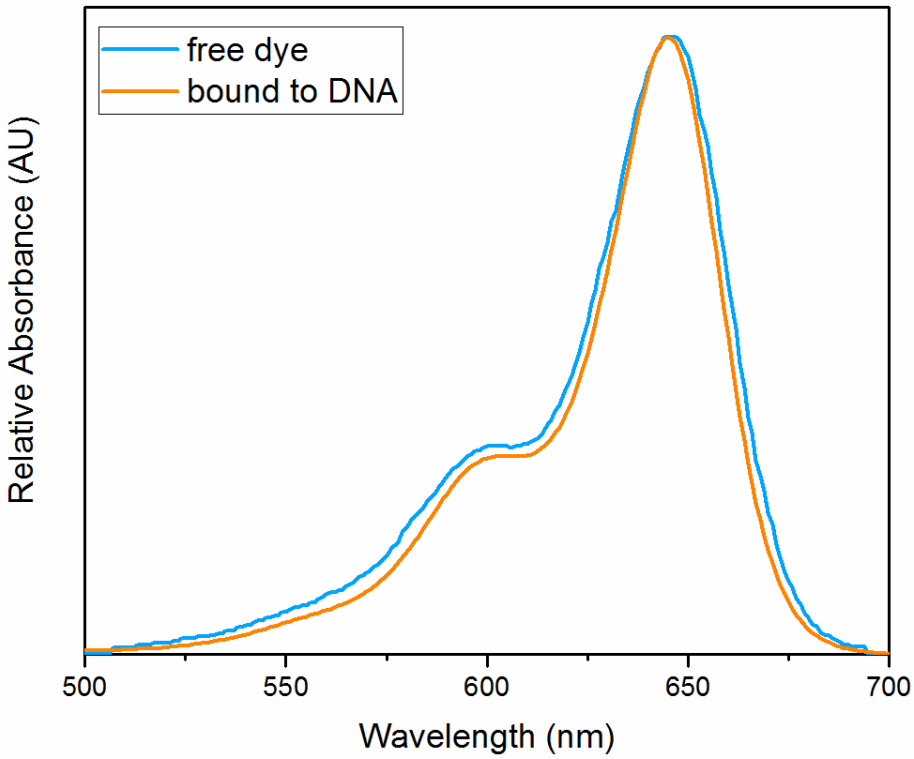


Figure E.4: Comparison of absorbance spectra of Cy5 dye when free in solution and when bound to duplex DNA [204, 262].

Table E.5: Structural information for Cy5 monomers optimized with given basis set and exchange correlation (xc) functionals, showing little significant difference between structures. The molecule length is measured as the greatest distance between any two atoms in the molecule and the methine chain is the greatest distance between carbon atoms in the methine chain.

Basis Set	xc-functional	Molecule length (Å)	Methine chain length (Å)	Average C=C bond length (Å)
6-31+G(d,p)	B3LYP	18.698	7.436	1.398
6-31+G(d,p)	CAM-B3LYP	18.572	7.386	1.392
6-31+G(d,p)	ωB97-XD	18.495	7.365	1.393
6-311++G (2df,2pd)	B3LYP	18.616	7.403	1.391
6-311++G (2df,2pd)	CAM-B3LYP	18.480	7.349	1.386
6-311++G (2df,2pd)	ωB97-XD	18.396	7.328	1.387

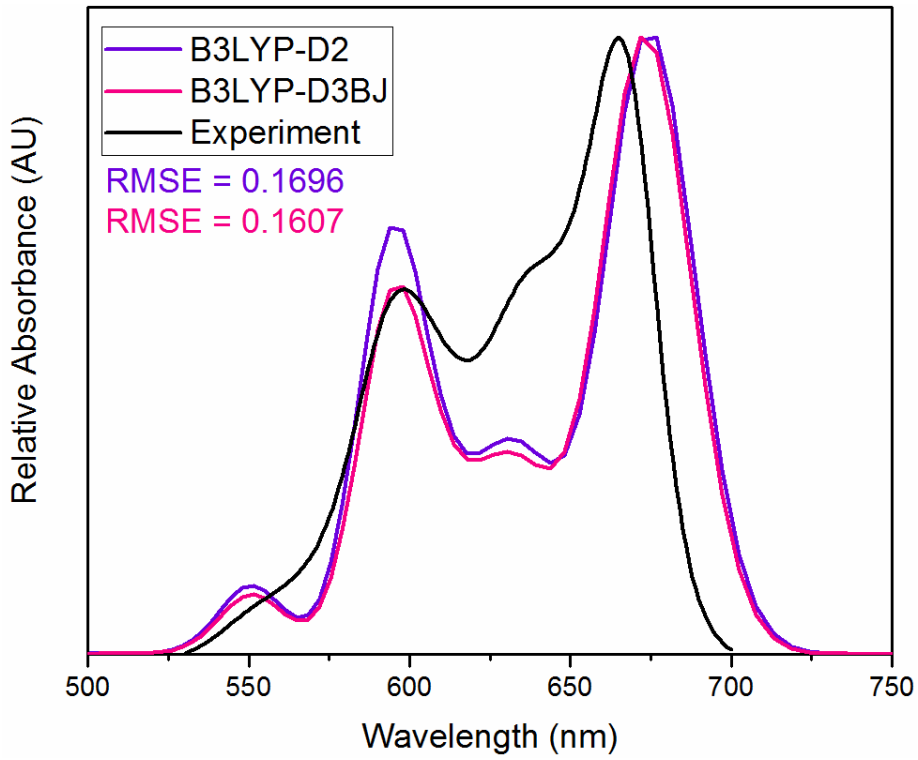


Figure E.5: Comparison of theoretical absorbance spectrum generated using KRM code for the oblique dimer structures relaxed using B3LYP-D2 and B3LYP-D3BJ xc-functionals compared to experimental absorbance of oblique dimer obtained by Cannon et al. [204]. RMSE values provided for quantification of difference.

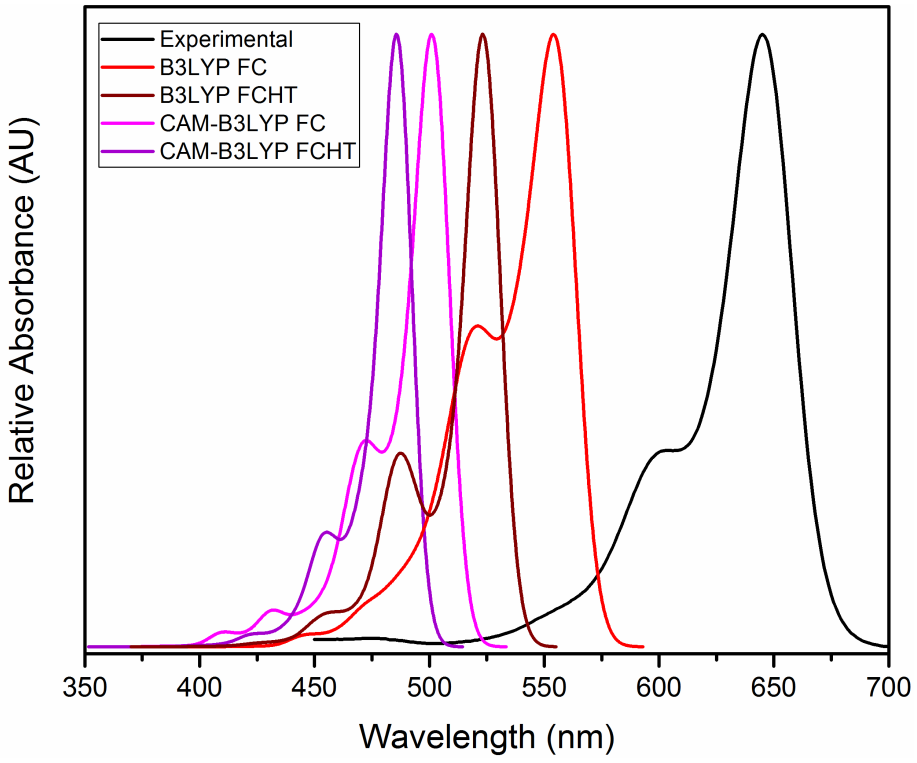


Figure E.6: Comparison of Cy5 monomer spectra generated using the FC approximation with and without the HT approximation to experimental spectrum obtained by Cannon et al. [204]. Comparison of the vacuum spectra to those obtained using a solvent model (Figure D.8 and Figure D.9) show that the bathochromic shift induced by the solvent provides a more accurate prediction of the maximum absorbance

APPENDIX F:

SUPRAMOLECULAR INTERACTIONS OF GROUP VI METAL CARBONYL COMPLEXES: THE FACILITATING ROLE OF 1,3-BIS(*P*-ISOCYANOPHENYL)UREA

The following chapter was published in Inorganic Chemistry in 2019 under authors Shaun Millard, Jenny W. Fothergill, Zoe Anderson, Dr. Eric C. Brown, Dr. Matthew D. King, and Dr. Adam C. Colson. My contributions to this paper were DFT calculations under the guidance of Dr. King.

F.1 Abstract

An investigation of supramolecular phenomena involving zero valent transition metal complexes was facilitated by the production of the ditopic isocyanide ligand 1,3-bis(*p*-isocyanophenyl)urea, which was synthesized *via* sub-stoichiometric phosgenation of 4-isocyanophenylamine and used to coordinate Group VI metal carbonyl fragments. The resulting binuclear organometallic complexes were observed to pack into ladder-like anisotropic arrays in the solid state. Crystallographic and computational evidence suggests that this packing motif can be attributed to a combination of intermolecular π-π and urea-π interactions. Similar to other *N,N'*-diarylureas bearing electron withdrawing groups, 1,3-bis(*p*-isocyanophenyl)urea and the organometallic complexes prepared therefrom also

exhibit an affinity towards anion binding in non-aqueous solution. Equilibrium constants (K) for the formation of host-guest complexes between the organometallic derivatives of 1,3-bis(p-isocyanophenyl)urea and chloride, nitrate, and acetate anions exceed 10^3, 10^4, and 10^5 M^{-1}, respectively.

F.2 Introduction

Non-covalent interactions are foundational to the field of supramolecular chemistry and have enabled emerging applications in chemical sensing, molecular recognition, and the self assembly of hierarchal materials [263, 264, 265]. Such interactions are also exploited in the field of crystal engineering to direct the organization of molecular solids in which constituent molecules are arrayed according to desired structural motifs [266, 267, 268, 269]. Although research efforts in these fields often emphasize the assembly of purely organic molecules, the integration of organometallic and coordination complexes into supramolecular structures and engineered molecular solids is also well-documented [270, 271, 272, 273, 274]. Less common, however, are accounts of supramolecular phenomena or crystal engineering involving low- or zero-valent transition metal carbonyl complexes, despite the fact that such complexes have long attracted interest due to their unique structural and electronic properties as well as the convenient spectroscopic monitoring afforded by the metal-bound carbon monoxide ligands [275, 276, 277, 278, 279, 280, 281, 282].

One challenge associated with the study of supramolecular behavior in low-valent transition metal complexes is the incorporation of organic ligands capable of binding to the electron-rich metal centers while simultaneously participating in non-covalent intermolecular interactions. To address this challenge, we have designed a ditopic organic ligand (**1**) bearing isocyanide functional groups appended to the well known *N,N'*-diarylurea moiety (Figure F.1). Organic isocyanides are isolobal analogues of carbon monoxide and their ability to stabilize

neutral and anionic transition metal complexes has been broadly demonstrated [283, 284, 285, 286, 287, 288, 289, 290]. Unlike carbon monoxide, additional functionality can be conferred upon isocyanides by modification of the appended organic component, a useful feature that has been exploited in the preparation of chelating ligands, coordination polymers, and functional components in advanced materials [291, 292, 293, 294, 295, 296, 297, 298, 299, 300].

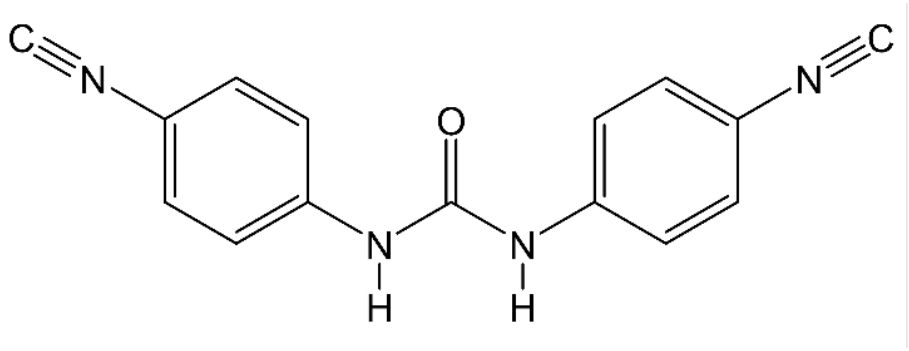


Figure F.1: 1,3-bis(*p*-isocyanophenyl)urea (1)

Our selection of the *N,N'*-diarylurea moiety as a facilitating agent for supramolecular chemistry and crystal engineering was motivated by the versatility of the urea group in promoting multiple intermolecular interactions [264, 301, 302, 303, 304]. Examples of three prominent supramolecular motifs associated with the *N,N'*-diarylurea moiety are depicted in Figure F.2. The most intuitive non-covalent interactions are those between the N-H hydrogen bond donors and C=O acceptors of the urea group. In the absence of competing donor or acceptor species, these interactions can drive the formation of ribbon- or tape-like supramolecular structures (Figure F.2a)[301]. In the presence of alternative hydrogen bond acceptors such as halides or oxoanions, strong interactions are often observed between the urea host and the anionic guest (Figure F.2b), presenting opportunities for applications in anion detection or anion-templated supramolecular assembly [301, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316]. Perhaps less intuitively, the urea group is also known to facilitate π-π interactions between neigh-

boring *N,N′*-diarylurea subunits (Figure F.2c). The relative co-planarity of the aryl groups in *N,N′*-diarylurea compounds is largely determined by the extent of intramolecular hydrogen bonding between the oxygen atom of the urea moiety and the *ortho* proton located on the adjacent *N*-aryl ring; experimental and computational evidence indicates that this planarity is most pronounced when the aryl rings bear electron withdrawing functional groups [312, 317, 318, 319, 320].

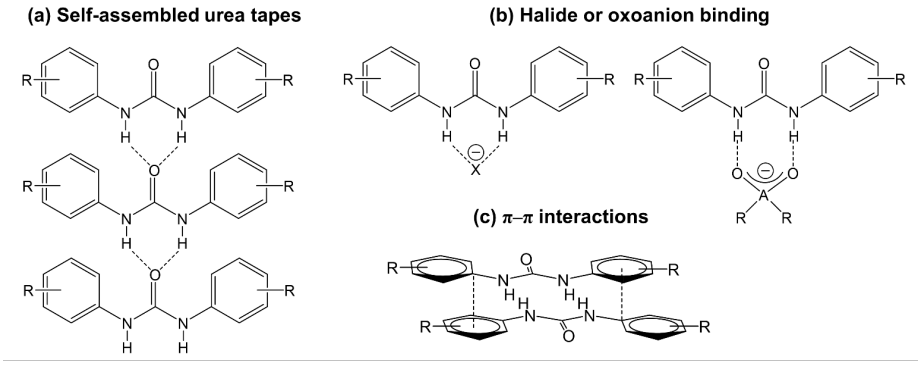


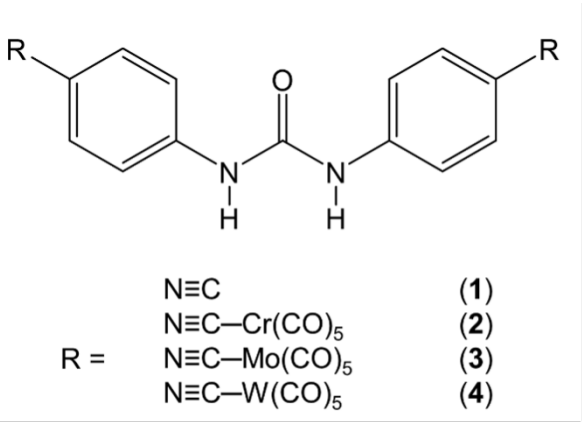
Figure F.2: Major supramolecular motifs associated with *N,N′*-diarylurea compounds

The objective of the present study was to identify which—if any—of the supramolecular motifs depicted in Figure F.2 might be observed in *N,N′*-diarylureas appended with zero-valent metal carbonyl subunits. Accordingly, we describe the synthesis of 1,3-bis(*p*-isocyanophenyl)urea and its subsequent coordination to $M(CO)_5$ (M = Cr, Mo, and W) metal carbonyl fragments. A notable structural feature of these complexes is the nearly planar conformation of the *N,N′*-diarylurea moiety, which facilitates molecular packing into ordered stacks through a combination of π-π and urea-π interactions. Like their purely organic counterparts, the metal bound derivatives of 1,3-bis(*p*-isocyanophenyl)urea also exhibit an affinity towards anion binding in acetonitrile solution.

F.3 Results and Discussion

The ditopic 1,3-bis(*p*-isocyanophenyl)urea ligand (**1**) was prepared by treating 4-isocyanophenylamine with a sub-stoichiometric quantity of triphosgene. The

Table F.1: Diagnostic infrared C≡N and C≡O stretching frequencies for compounds 1-4.



	ν C≡N (cm^{-1})	ν C≡O (cm^{-1})
1	2127	N/A
2	2143	2058, 1955
3	2143	2063, 1956
4	2144	2059, 1950

*Recorded in CH_2Cl_2

formation of **1** presumably proceeds through the intermediacy of an isocyanate species, although no attempts were made to isolate or characterize the latter [321]. The PdO-catalyzed ligand substitution method described by Coville [322] was used to append $Cr(CO)_5$ and $W(CO)_5$ fragments to the isocyanide groups of **1**, producing homobimetallic complexes **2** and **4**, respectively. The Mo-containing variant (**3**) was prepared *via* chemical decarbonylation of $Mo(CO)_6$ using trimethylamine *N*-oxide in the presence of **1**. Complexes **2**, **3**, and **4** are readily identifiable by their diagnostic infrared spectra, as summarized in Table F.1. It is worth noting that the C≡N stretching frequency of **1** increases upon metal coordination, suggesting that the isocyanide acts primarily as a σ-donor with very little π-accepting character. This observation is consistent with previous reports of aromatic isocyanide coordination in heteroleptic metal carbonyls in which pronounced shifting of the C≡N stretch to lower frequencies was not observed until two or more isocyanide ligands were attached to a metal core [322, 323, 324].

Single crystals of complexes **2**, **3**, and **4** suitable for X-ray diffraction studies were grown from acetonitrile solutions at -20 °C. As a representative example of the series, the molecular structure of **3** is shown in Figure F.3. The most prominent structural feature is the nearly planar configuration of the *N,N′*-diarylurea moiety. The C_{ortho}-C_{ipso}-$N_{urea-}C_{urea}$ torsion angles of 0.46° and 7.5° are similar to those reported by Reddy et al. [317] for other *N,N′*-diarylureas bearing para-substituted electron withdrawing groups. This planarity may be attributed to intermolecular C-H· · ·O hydrogen bonding between ortho aryl protons and the urea oxygen atom, as proposed by Etter et al. [318]; indeed, the C_{ortho} · · O_{urea} distances and C_{ortho}-H· · · O_{urea} angles in complexes **2**, **3**, and **4** range from 2.84 to 2.87 Å and 121 to 122°, respectively, and are consistent with the analogous distances (2.83 to 2.94 Å) and angles (119 to 122°) observed in other *N,N′*-diarylureas of comparable planarity [317, 318]. Salient structural features for complexes **2** and **4** are presented in Table F.2 and Figure G.1 and Figure G.2 in the Supporting Information.

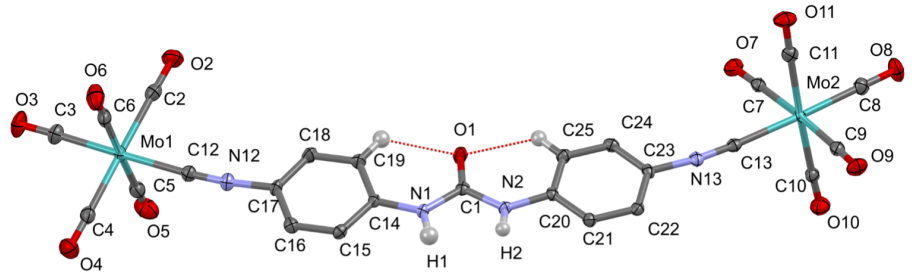


Figure F.3: Molecular structure of complex 3. Thermal ellipsoids are rendered at the 50% probability level. Only urea N-H atoms and aromatic hydrogen atoms participating in hydrogen bonding are shown.

The packing of **3** within the molecular solid is depicted in Figure F.4a. Despite the relative steric bulk of the $Mo(CO)_5$ subunits, the planarity of the *N,N′*-diarylurea moiety facilitates packing into ordered stacks, resulting in staggered, ladder-like arrays of metal atoms (Figure F.4b). It is tempting to view the orientation and proximity of these metal sites and speculate on the potential development of charge transfer materials, especially considering that adjacent metal atoms are

Table F.2: Selected bond distances, close contacts, and torsion angles for complexes 2, 3, and 4

	2 (M = Cr)	**3** (M = Mo)	**4** (M = W)
		Distances (Å)	
M - $C_{isocyanide}$	1.981(2), 1.979(2)	2.127(2), 2.130(2)	2.119(3), 2.113(3)
C_{ortho} - H· · ·O_{urea}	2.24, 2.22	2.25, 2.23	2.25, 2.23
C_{ortho} · · ·O_{urea}	2.86, 2.84	2.86, 2.84	2.87, 2.84
π-π contacts ‡	3.31	3.33	3.32
urea-π contacts ‡	3.29	3.32	3.31
		Angles (°)	
C_{ortho}-H· · ·O_{urea}	121.8, 121.3	121.8. 121.0	121.0, 122.0
C_{ortho}-C_{ipso}-N_{urea}-C_{urea}	1.1, 6.1	0.46, 7.5	0.45, 7.8

‡distance between adjacent mean *N,N'*-diarylurea molecular planes

separated by only 6.0 to 7.6 Å. This separation is well within the 4 to 14 Å range over which the majority of electron transfer processes occur in benchmark biological systems [325, 326]. In their present form, however, complexes **2**, **3**, and **4** undergo electrochemically irreversible oxidation and are therefore unlikely candidates for such applications (Figure G.3).

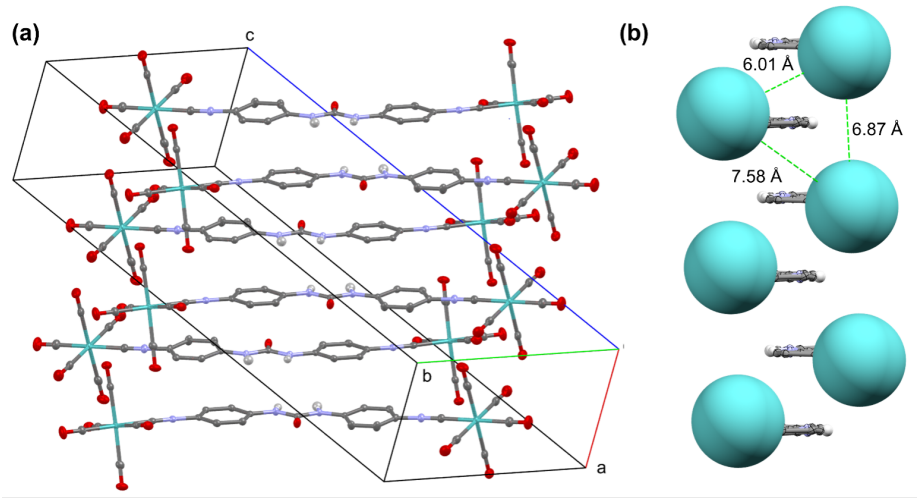


Figure F.4: (a) Molecular packing of 3 showing the orientation of the molecules relative to the unit cell axes. (b) Alternative representation of the molecular packing of 3 viewed through the *N,N'*-diarylurea planes with the CO ligands omitted and Mo atoms depicted at full Van der Waals radius.

The molecular packing of **3** is reminiscent of the π-stacking motif introduced in

Figure F.2c in which planar *N,N′*-diarylurea moieties are arranged parallel to one another in an offset or "slipped" orientation; however, close inspection of the interactions between molecules of **3** reveals several subtle deviations from the archetypal packing motif. As shown in Figure F.5a, the separation between the adjacent *N,N′*-diarylurea planes of complex **3** is ∼3.32 Å, while the distance between the centroids of interacting aromatic rings alternates between 3.76 and 4.79 Å. By comparison, previous structural studies of *N,N′*-diarylureas bearing elec-tron withdrawing functional groups have reported interplanar distances of 3.4 to 3.5 Å, with the aromatic centroids separated by no more than 4.2 Å [317, 327] Al-though the interplanar distances associated with complex **3** are consistent with other π-π interactions reported in the literature, the significant separation be-tween alternating pairs of ring centroids suggest that π-π interactions are not the only non-covalent interactions contributing to the close contact between *N,N′*-diarylurea planes [328]. Additional insight into the nature of a secondary interac-tion can be gained by viewing adjacent pairs of *N,N′*-diarylurea moieties from a direction normal to the molecular planes (Figure F.5b). From this perspective, it becomes apparent that the interactions between an arbitrary *N,N′*-diarylurea moi-ety of complex **3** and the molecules positioned directly above and below are not equivalent—one face of the aromatic ring experiences substantial overlap with the *aromatic ring* of a neighboring molecule, while the opposite face interacts with the *nitrogen atom* of a different neighbor. In the latter case, the distance between the aromatic ring centroid and the urea nitrogen atom is only 3.34 Å, suggesting the existence of a non-covalent urea-π interaction. Similar close contacts are also observed in the Cr- and W-containing derivatives (**2** and **4**), as tabulated in Table F.2.

Interactions between aromatic π-systems and urea or amide moieties have been predicted computationally and observed experimentally, although reports of such interactions outside of a biological context are scarce [329, 330, 331, 332]. Because

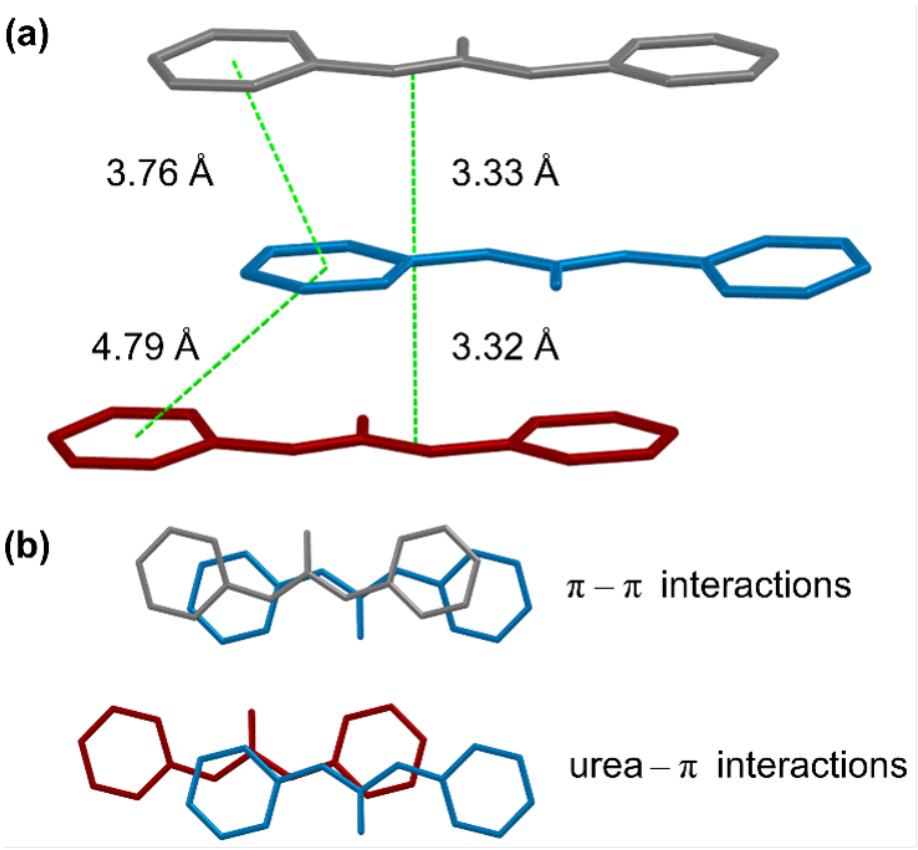


Figure F.5: (a) Distances between adjacent *N,N′*-diarylurea planes and the aromatic ring centroids of complex 3. (b) Alternating stacking motifs viewed normal to the *N,N′*-diarylurea planes. Hydrogen atoms and all ring substituents are omitted for clarity.

urea-π interactions are not commonly observed in synthetic systems, computational methods were employed to determine whether urea-π interactions truly contribute to the molecular packing observed for complexes **2**, **3**, and **4**, or if the packing motif is simply an artifact of the steric encumbrance imposed by the $Mo(CO)_5$ groups.

Initial solid-state DFT calculations were performed on the crystal structure of complex **3** to assess lattice energy in terms of cohesive and conformational strain energies associated with crystal packing. The basis set superposition error-corrected lattice energy was determined to be -123.20 kcal mol^{-1}, which accounts for the balance of -172.86 kcal mol^{-1} of cohesive energy offset with 49.65 kcal

Table F.3: Calculated energies per molecule (kcal mol^{-1}) of complex 3 in the crystal and dimer conformations.

Complex 3	
Total Lattice Energy	-123.20
Cohesive Energy	-172.86
Conformational Energy	49.65

Complex 3 Dimers		
	π - π	urea - π
Total Binding Energy	-118.43	-111.64
Relative Binding Energy	0.000	6.797

***N,N'*-diarylurea Dimers**		
	π - π	urea - π
Total Binding Energy	-57.92	-55.62
Relative Binding Energy	0.000	2.295

mol^{-1} of adverse molecular strain energy due to deformations emerging from crystal packing forces (Table F.3). This significant contribution of molecular strain to the overall lattice energy, primarily in the *N,N'*-diarylurea moieties, likely results from steric effects of geometric packing of the bulky $Mo(CO)_5$ end groups.

To allocate the relative contributions of the substituent π-π and urea-π stacking motifs described in Figure F.5 to the complex **3** lattice energy, appropriate dimer configurations were isolated from the DFT-optimized solid-state structure. Single-point energy calculations were performed to obtain dimer and single-molecule energies for the models representative of their strained solid-state molecular conformations. It was found that the two motifs are comparable in interaction energy, with total binding energies of -118.43 and -111.64 kcal mol^{-1} for the π-π and urea-π stacking motifs, respectively, corresponding to a ΔE_{bind} of only 6.80 kcal mol^{-1} favoring the π-π intermolecular interactions. These stacked dimer models, however, still contained interacting $Mo(CO)_5$ end groups, thereby not adequately representing the true nature or relative consequence of the two participating *N,N'*-diarylurea stacking motifs.

Subsequent calculations isolated the interaction energy contributions to the *N,N'*-diarylurea π-π and urea-π stacking motifs, for which the $Mo(CO)_5$ groups were removed from the molecular models. The *p*-substituent isocyanides groups were terminated with H atoms to provide a better representation of the bonding configuration of the *N,N'*-diarylurea moieties as they exist in the complex **3** crystal structure. The resulting calculations produced absolute *N,N'*-diarylurea dimer interaction energies of -57.92 and -55.62 kcal mol^{-1}, respectively. Remarkably, this amounts to an approximate difference of only 4% (*i.e.* 2.30 kcal mol^{-1}) between the π-π and urea-π stacking interactions of the *N,N'*-diarylurea groups, despite the reduction of surface contact area from 94.8 to 76.4 Å[266, 267, 268, 269] for the urea-π stacked dimer (based on MSMS surface area calculated using 1.5-Å probe sphere). These results demonstrate that the urea-π stacking is indeed a robust intermolecular interaction and plays an important role in molecular ordering during crystal growth, and that the association is not an ancillary product of crystal packing.

The underlying nature of the prominent urea-π stacking interactions observed in the synthetic complexes can be explained through observation of electrostatic potentials and molecular orbital interactions. An electrostatic potential map of the urea-π stacked *N,N'*-diarylurea dimer shows that the positive electrostatic potential arising from the urea nitrogen atom overlaps with the negative potential central to an aromatic ring of the adjacent molecule (Figure F.6). The proximity of the molecules in the urea-π dimer pair, separated by 3.32 Å, provides a favorable distance for strong electrostatic interaction between these substituent groups. In addition, inspection of molecular orbitals reveals that the *N,N'*-diarylurea HOMO and LUMO share significant constructive spatial overlap in their relative dimer positions (Figure F.7). The individual molecule orbital energies were calculated at -6.344 and -1.601 eV for the HOMO and LUMO, respectively. The resulting

HOMO of the dimer complex, representative of the single-molecule orbital combination formed in the crystalline orbital, was calculated at -6.439 eV. The combination of advantageous electrostatic interactions and molecular orbital interactions provides the physical description for the strong interaction evident in the urea-π stacking of the *N,N'*-diarylurea dimer in the crystal structure of complex **3**.

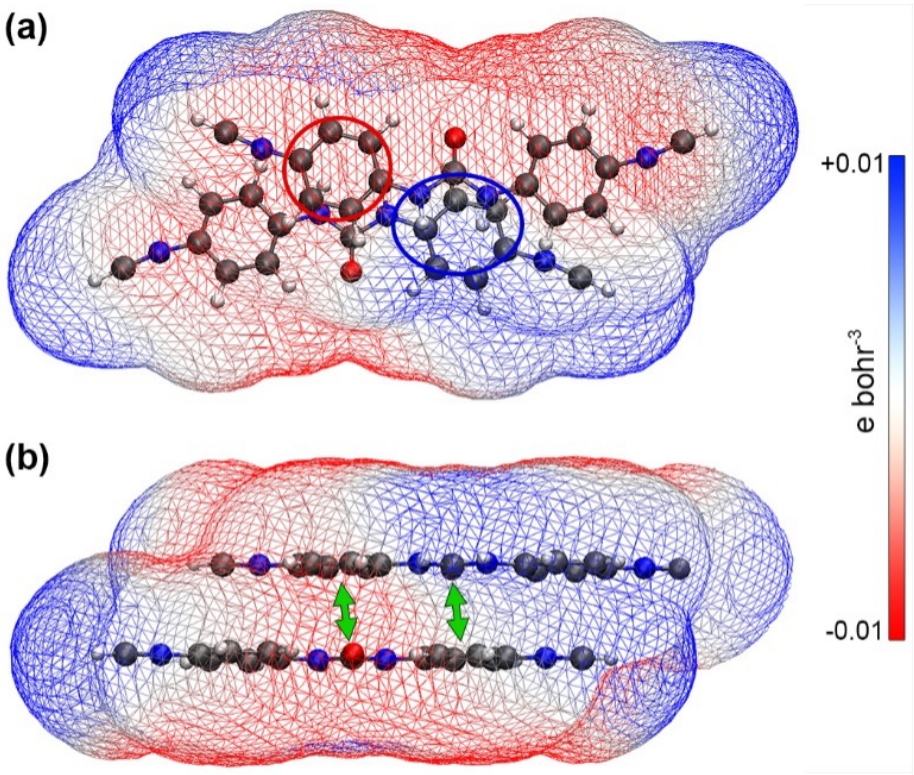


Figure F.6: Electrostatic potentials mapped onto the electron density isosurface (isovalue: 1.0 × 10^{-6} *e* bohr^{-3}) of the *N,N'*-diarylurea urea-π stacked dimer. (a) Top view of urea-π dimer indicating urea nitrogen atoms (blue circle) and the aromatic ring (red circle) of the upper molecule involved in electrostatic binding interactions. (b) Locations of the two nitrogen-ring interactions (green arrows) in the *N,N'*-diarylurea urea-π dimer.

Having investigated the intermolecular interactions between complexes **2**-**4** in the solid state, a short study of the anion-binding behavior of **1**-**4** in non-aqueous solution was initiated. Urea containing compounds are known to act as effective receptors or hosts for anionic guests [301, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316]. The host-guest interaction of interest can be represented by the following equilibrium relationship: $host + guest^{-} \rightleftharpoons [host \cdot guest]^{-}$. The corre-

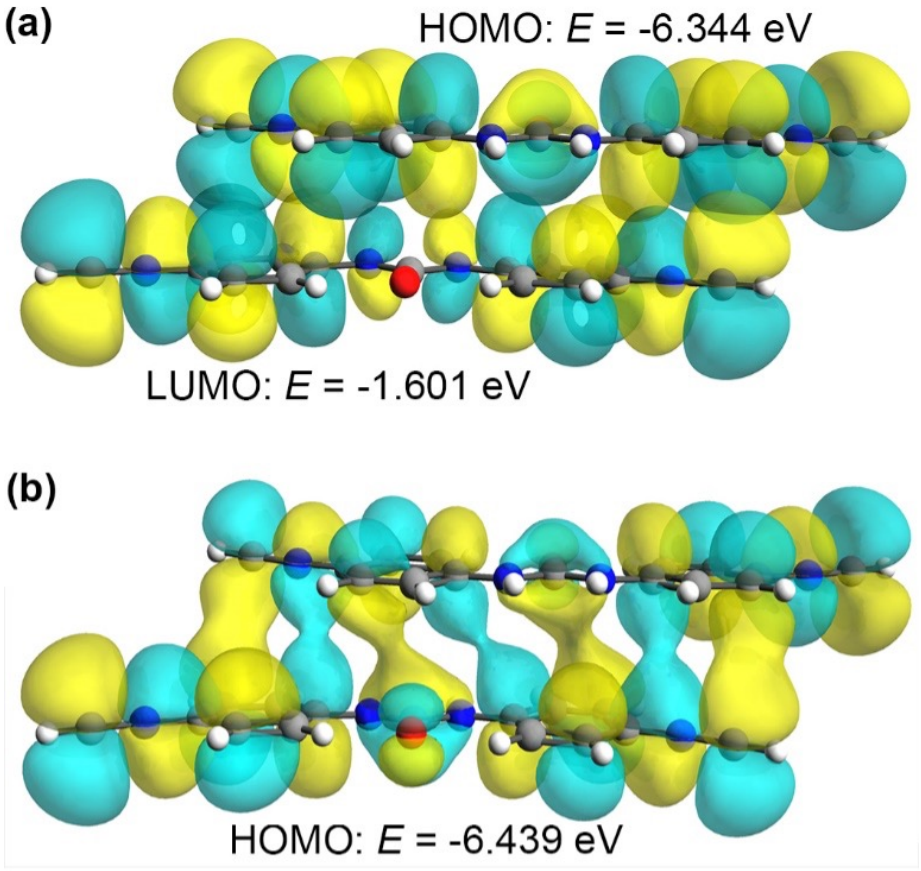


Figure F.7: Molecular orbital surface representations of *N,N′*-diarylurea (a) HOMO and LUMO orbitals for non-interacting molecules in the urea-π stacked dimer orientation, and (b) the resulting HOMO of the *N,N′*-diarylurea urea-π dimer.

sponding equilibrium constant (K) may be estimated by performing spectroscopic titration experiments followed by fitting of the resultant data to established binding models [333, 334]. In the present study, solutions of **1-4** were titrated with the tetrabutylammonium salts of chloride, nitrate, and acetate as representative examples of the halides, inorganic oxoanions, and organic oxoanions, respectively.

^{1}H NMR spectroscopy was selected as the primary method for estimating equilibrium binding constants, as incremental titration of urea hosts 1-4 with anionic guests consistently produced a detectable response in the spectral features. Under the dilute conditions of the titration experiments ($\approx$ 0.1 mM), broadening of the urea N-H proton NMR signals prevented unambiguous chemical shift assign-

ment. Consequently, the ^{1}H NMR signals corresponding to the aromatic ring protons were used to probe anion binding behavior. Two unique aromatic proton resonances were observed for compounds 1-4, a downfield signal corresponding to the pair of protons nearest the urea moiety (H_α) and an upfield signal representing the pair of protons nearest the isocyanide functional group (H_β). The latter signal shifts upfield upon titration and was used to determine K for all host-guest complexes, as the former signal (H_α) was prone to considerable broadening, especially during titrations with acetate. As a representative example, Figure F.8 a illustrates the effects of chloride titration on the aromatic ^{1}H NMR signals of **1**, while Figure F.8 b compares the magnitude of the ^{1}H NMR chemical shifts of H_β observed upon titration of **1** with nitrate, chloride, and acetate anions.

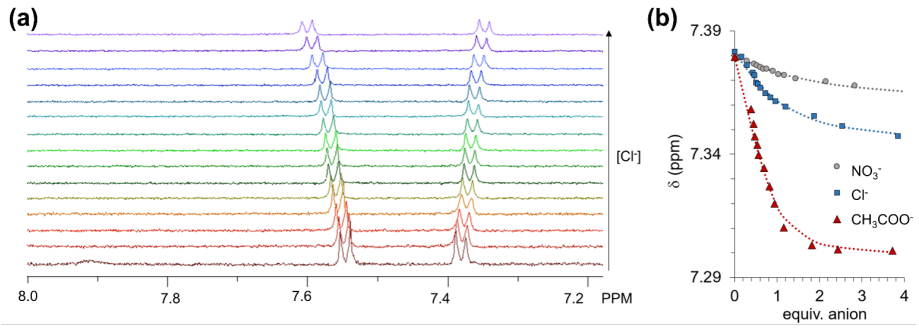


Figure F.8: (a) Overlay of ^{1}H NMR spectra obtained during the titration of 1 (0.90 x 10^{-4} M in CD_3CN) with $[Bu_4N]Cl$. (b) Comparison of ^{1}H NMR chemical shifting observed during titration of 1 with nitrate, chloride, and acetate anions. The dotted lines represent the results of non-linear fitting to a 1:1 host-guest binding model.

The respective upfield and downfield shifts observed for the H_β and H_α ^{1}H NMR signals are consistent with previous observations of signal shifting during titration of *N,N'*-diarylureas with anions [308, 335, 336, 337]. It should be noted that deprotonation of *N,N'*-diarylureas has been reported during titrations with basic anions such as fluoride or acetate, especially when the urea (or thiourea) hosts are particularly acidic [311, 338]. As a means of confirming that the observed ^{1}H NMR chemical shifts may be attributed to host-guest complex formation rather

than deprotonation of the *N,N'*-diarylurea host, UV-Vis spectroscopy was used to probe the magnitude of the bathochromic spectral shift accompanying addition of excess acetate anion. Formation of hydrogen bonding host-guest complexes is typically characterized by a modest spectral shift (< 50 nm), while formal deprotonation produces an absorption band that is generally red-shifted by more than 100 nm [308, 310, 339]. Addition of excess acetate to **1** produced a bathochromic spectral shift of only 15 nm with no additional features appearing at higher wavelengths, indicating that the urea moiety remains protonated under the titration conditions (Figure G.5).

Equilibrium constants for the formation of host-guest complexes of **1**-**4** with chloride, nitrate, and acetate were derived from ^{1}H NMR titration profiles and are reported in Table F.4. All values of K assume a 1:1 binding ratio between the urea host and anion guest, an assumption substantiated by direct observation of the 1:1 host-guest complexes using electrospray ionization mass spectrometry as well as the consistency of the 1:1 binding model in accounting for the experimental data (Figure G.6-Figure G.10)[333]. Compounds **1**-**4** exhibit similar affinities for each of the anions studied, the values of K becoming progressively larger with increasing anion basicity. The equilibrium constants for the formation of host-guest complexes of **1**-**4** with chloride and nitrate anions are comparable to those reported by Boiocchi et al. [308] for the 1,3-bis(*p*-nitrophenyl)urea receptor, although a higher affinity towards acetate was measured for the latter ($\log K = 6.61$). The results presented in Table F.4 suggest that the isocyanide functionalized *N,N'*-diarylurea (**1**) is capable of strong anion-binding and that attachment of organometallic subunits does not substantially impact the ability of the urea moiety to act as an anion receptor.

Table F.4: Equilibrium constants (log K) for formation of host-guest complexes of 1-4 with selected anions

	log K^a		
Urea Host	NO_3^-	Cl^-	CH_3COO^-
1	3.62(5)	4.42(3)	5.30(8)
2	3.52(3)	4.35(3)	5.41(7)
3	3.60(3)	4.35(8)	5.50(3)
4	3.70(3)	4.58(3)	5.66(4)

[a] In CD_3CN solution at 25 °C. Values in parentheses indicate uncertainty in the last figure

F.4 Conclusions

The synthesis of 1,3-bis(*p*-isocyanophenyl)urea, its coordination to Group VI metal carbonyl fragments, and the structural characterization of the binuclear organometallic products have been reported. The nearly planar configuration of the *N,N'*-diarylurea moiety enables the packing of the organometallic complexes into ladder like anisotropic arrays in which the zero valent metal atoms are separated by 6.0 to 7.6 Å. Crystallographic and computational evidence suggests that that the formation of these arrays can be attributed to a combination of intermolecular π-π and urea-π interactions. Although the metal carbonyl fragments employed in this exploratory study undergo irreversible electrochemical oxidation, the methods and observations reported herein might be extended to produce similar molecular solids containing more electrochemically-robust organometallic fragments; specimens of the latter may in turn find application as charge transfer materials. Similar to other *N,N'*-diarylureas bearing electron withdrawing groups, 1,3-bis(*p*-isocyanophenyl)urea and its organometallic derivatives were also found to behave as anion receptors in non-aqueous solvent. Equilibrium constants (K) for the formation of host-guest complexes of 1,3-bis(*p*-isocyanophenyl)urea with chloride, nitrate, and acetate exceed 10^3, 10^4, and 10^5 M^{-1}, respectively. Complexes of 1,3-bis(*p*-isocyanophenyl)urea with Group VI metal carbonyl complexes

exhibit similar anion binding behaviors, presenting opportunities for additional research into anion detection or anion-templated supramolecular assembly using low-valent organometallic species.

F.5 Experimental

F.5.1 General Considerations

All synthetic operations were carried out under a nitrogen atmosphere using standard Schlenk techniques to exclude moisture and oxygen. Nitrogen was prepurified by passage through columns of activated copper catalyst (BASF PuriStar R3-11G) and molecular sieves (RCI-DRI 13X). Glassware was dried in an oven at 130 °C, assembled while hot, and allowed to cool under reduced pressure. All solvents were dried according to published procedures and degassed with nitrogen prior to use.21 $Cr(CO)_6$ (Beantown Chemical, 99%), $Mo(CO)_6$ (Acros, 98%), $W(CO)_6$ (Beantown Chemical, 97%), PdO (Acros), trimethylamine *N*-oxide dihydrate (Beantown Chemical, 98%), and triphosgene (Chem Impex, 99%) were used as received without further purification. 4-Isocyanophenylamine was prepared according to literature procedures and sublimed prior to use [340].

Infrared spectra were obtained using a Thermo Scientific Nicolet iS5 FTIR spectrometer equipped with a 0.2 mm BaF_2 liquid cell. ^1H and ^{13}C NMR data were recorded on a 600 MHz Bruker AVANCE III spectrometer. Electrospray ionization mass spectrometry (ESI-MS) was carried out using a Bruker HCTultra CTD II spectrometer in negative ion mode. Samples of **1-4** were dissolved in CH_3CN and treated with the tetrabutylammonium salts of chloride, nitrate, and acetate prior to injection into the mass spectrometer. Elemental analyses were performed by Atlantic Microlab, Inc in Norcross, GA, USA.

F.5.2 Synthesis of 1,3-bis(*p*-isocyanophenyl)urea (1)

4-isocyanophenylamine (3.00 g, 25.4 mmol) was dissolved in 80 mL of anhydrous dichloromethane, followed by addition of 7.8 mL of triethylamine. The solution was cooled to 0 °C and triphosgene (1.20 g, 4.04 mmol) was slowly introduced into the reaction vessel. (*CAUTION: Triphosgene is toxic and its reaction with 4-isocyanophenylamine generates considerable heat and an abundance of hydrogen chloride. Triphosgene should be added very slowly and in several portions to allow for sufficient heat exchange with the cooling media.*) The light yellow reaction mixture was magnetically stirred for 3 hours at 0 °C, then stirred for an additional 45 hours at 25 °C. Methanol (10 mL) was added to the reaction mixture and stirring was continued for an additional hour. Organic solvents were removed under reduced pressure, and the residues were dissolved in 60 mL of dimethyl formamide (DMF). Deionized water (60 mL) was slowly added, and the reaction vessel was gently heated to ensure that the solution remained clear. After addition of deionized water, the solution was allowed to cool slowly to room temperature, whereupon an off-white precipitate formed. The precipitate was filtered and washed with three 20 mL portions of water, followed by 20 mL of diethyl ether and 20 mL of hexanes, respectively. After drying under reduced pressure for one day, **1** was obtained with sufficient purity for further experimentation. Yield: 2.82 g (84.6%). IR (CH_2Cl_2, cm^{-1}): ν_{CN} 2027 (vs). ^{1}H NMR (CD_3CN, 20 °C): δ 7.38 (d, 4H, Ar-*H*, *J* = 8.86), 7.53 (d, 4H, Ar-*H*, *J* = 8.86), 7.60 (br s, 2H, N-*H*). ^{13}C NMR (CD_3CN, 20 °C): δ 119.3, 127.1, 140.3, 151.9, 163.5, *ipso C* not observed. MS(ESI): *m/z* 297 $[M + Cl^-]^-$, 324 $[M + NO3^-]^-$, 321 $[M + CH_3COO^-]^-$. Anal. Calcd for $C_{15}H_{10}N_4O$: C, 68.69; H, 3.84; N, 21.36. Found: C, 68.47; H, 4.05; N, 21.17.

F.5.3 Synthesis of Complex 2

$Cr(CO)_6$ (317 mg, 1.44 mmol) and **1** (182 mg, 0.694 mmol) were combined with 25 mL of DMF and heated to 90 °C, whereupon PdO (14 mg, 0.12 mmol) was added

to the reaction vessel. The reaction mixture was magnetically stirred at 90 °C for 15 minutes, then allowed to cool to room temperature. DMF was removed by vacuum distillation, leaving behind an oily residue. The oily residue was extracted with dichloromethane and filtered to remove insoluble impurities. Hexanes were added slowly to the dichloromethane filtrate until the solution became cloudy, then the solution was centrifuged at 5000 rpm for 2 minutes, after which the clear supernatant was decanted and dried under reduced pressure. The residual pale yellow solid was dissolved in warm acetonitrile, then cooled slowly to -20 °C to yield crystals of **2** · CH_3CN. Yield: 242 mg (50.7%). IR (CH_2Cl_2, cm^{-1}): ν_{CN} 2143 (m), ν_{CO} 2058 (s), 1955 (vs). ^{1}H NMR (CD_3CN, 20 °C): δ 7.44 (d, 4H, Ar-*H*, J = 8.75), 7.58 (d, 4H, Ar-*H*, J = 8.89), 7.71 (br s, 2H, N-*H*). ^{13}C NMR (CD_3CN, 20 °C): δ 120.2, 127.9, 141.1, 153.0, isocyanide and adjacent *ipso* carbons not observed. MS(ESI): *m/z* 681 $[M + Cl^-]^-$, 708 $[M + NO_3^-]^-$, 705 $[M + CH_3COO^-]^-$. Anal. Calcd for $C_{27}H_{13}Cr_2N_5O_{11}$: C, 47.18; H, 1.91; N, 10.19. Found: C, 47.23; H, 1.83; N, 10.16.

F.5.4 Synthesis of Complex 3

$Mo(CO)_6$ (811 mg, 3.07 mmol) and **1** (403 mg, 1.54 mmol) were dissolved in 25 mL of tetrahydrofuran (THF). A dropping funnel charged with trimethylamine N oxide dihydrate (342 mg, 3.07 mmol), THF (10 mL), and methanol (10 mL) was attached to the reaction flask, the contents of which were added dropwise to the reaction mixture over the course of 1 hour. The reaction mixture was magnetically stirred for 6 hours at room temperature, after which the solvents were removed under reduced pressure. The residues were extracted with dichloromethane and filtered to remove insoluble impurities. Hexanes were added slowly to the dichloromethane filtrate until the solution became cloudy, then the solution was centrifuged at 5000 rpm for 2 minutes, after which the clear supernatant was decanted and dried under reduced pressure. The residual off-white solid was dissolved in warm acetonitrile, then cooled slowly to 20 °C to yield crystals of **3**.

Yield: 871 mg (77.0%). IR (CH_2Cl_2, cm^{-1}): ν_{CN} 2143 (m), ν_{CO} 2063 (s), 1956 (vs). ^1H NMR (CD_3CN, 20 °C): δ 7.43 (d, 4H, Ar-*H*, *J* = 8.86), 7.57 (d, 4H, Ar-*H*, *J* = 8.82), 7.69 (br s, 2H, N-*H*). ^{13}C NMR (CD_3CN, 20 °C): δ 119.3, 127.2, 140.4, 151.8, isocyanide and adjacent *ipso* carbons not observed. MS(ESI): *m/z* 769 $[M + Cl^-]^-$, 796 $[M + NO_3^-]^-$, 793 $[M + CH_3COO^-]^-$. Anal. Calcd for $C_{25}H_{10}Mo_2N_4O_{11}$: C, 40.89; H, 1.37; N, 7.63. Found: C, 41.03; H, 1.38; N, 8.15.

F.5.5 Synthesis of Complex 4

$W(CO)_6$ (369 mg, 1.05 mmol) and **1** (132 mg, 0.503 mmol) were combined with 25 mL of DMF and heated to 90 °C, whereupon PdO (10 mg, 0.082 mmol) was added to the reaction vessel. The reaction mixture was magnetically stirred at 90 °C for 5 minutes, then cooled to room temperature. DMF was removed by vacuum distillation, leaving behind an oily residue. The oily residue was extracted with dichloromethane and filtered to remove insoluble impurities. Hexanes were added slowly to the dichloromethane filtrate until the solution became cloudy, then the solution was centrifuged at 5000 rpm for 2 minutes, after which the clear supernatant was decanted and dried under reduced pressure. The residual yellow solid was dissolved in warm acetonitrile, then cooled slowly to -20 °C to yield crystals of **4** · CH_3CN. Yield: 211 mg (44.1%). IR (CH_2Cl_2, cm^{-1}): ν_{CN} 2144 (m), ν_{CO} 2059 (s), 1950 (vs). ^1H NMR (CD_3CN, 20 °C): δ 7.43 (d, 4H, Ar-*H*, *J* = 8.93), 7.57 (d, 4H, Ar-*H*, *J* = 9.00), 7.69 (br s, 2H, N-*H*). ^{13}C NMR (CD_3CN, 20 °C): δ 119.3, 127.3, 140.4, 151.8, isocyanide and adjacent *ipso* carbons not observed. MS(ESI): *m/z* 945 $[M + Cl^-]^-$, 972 $[M + NO_3^-]^-$, 969 $[M + CH_3COO^-]^-$. Anal. Calcd for $C_{27}H_{13}N_5O_{11}W_2$: C, 34.10; H, 1.38; N, 7.36. Found: C, 34.29; H, 1.31; N, 7.41.

F.5.6 Determination of Equilibrium Formation Constants (*K*) by ^1H NMR

In typical titration experiments, CD_3CN solutions of urea hosts **1**-**4** (0.75 mL, 0.10 mM) were loaded into standard NMR tubes and initial ^1H NMR spectra were col-

lected. Aliquots of an anion-containing solution were then delivered to the NMR tubes using a microsyringe, the mass of each aliquot being recorded on a microbalance. The first ten aliquots of titrant were taken from a stock solution of the anion guest (2.0 mM) prepared by dissolving a known quantity of the appropriate tetrabutylammonium salt in a 0.10 mM solution of the urea host, thereby minimizing host dilution effects. Subsequent aliquots of titrant were taken from a stock solution of anion guest (4.0 mM) prepared in the same manner. Sufficient anion was delivered during each titration step to enable collection of ^{1}H NMR spectra at the following approximate [anion]/[urea] ratios: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 2.0, 3.0, and 4.0. The upfield shifting of the aromatic proton signals centered between 7.38-7.44 ppm was recorded and values of K were calculated by non-linear fitting to a 1:1 binding model using the WinEQNMR2 software package [341].

F.5.7 Crystal Structure Determination of Complexes 2-4

X-ray diffraction data were collected at 100 K on a Bruker D8 Venture using MoKα -radiation (λ = 0.71073 Å). Data were corrected for absorption effects using the SADABS area detector absorption correction program [342]. The structures were solved by direct methods using Olex2 with the SHELXT structure solution program and refined with the SHELXL refinement package using least squares minimization [343, 344, 345]. All non-hydrogen atoms were refined with anisotropic thermal parameters. Hydrogen atoms attached to heteroatoms were identified from the residual density maps and refined with isotropic thermal parameters. All other hydrogen atoms in the investigated structure were located from difference Fourier maps, but their positions were ultimately placed in geometrically calculated positions and refined using a riding model. Additional calculations and refinement of structures were carried out using APEX3 and SHELXTL software [346, 347]. Graphical representations of crystallographic data were generated us-

ing the Mercury software package [348]. X-ray data collection and refinement parameters are tabulated in Table G.1.

F.5.8 Computational Methods

All DFT calculations were performed using the Crystal14 software package [349, 350]. Calculations utilized a pairwise dispersion-corrected B3LYP-D2 den-sity functional with atom-centered Dunning cc-pVDZ basis sets on all non-metal atoms[234, 351, 253, 352, 353]. Mo atoms were treated with LANL2DZ effective core potentials with standard Dunning D95V valence orbitals [354, 355]. Atomic coordinates from the X-ray structural determination of complex **3** were taken as the initial geometry for the crystal structural optimization. The energy minimization allowed full relaxation of atom positions and lattice parameters within the constraints of space group symmetry. A shrinking factor of 4 was used in defining the k point sampling of the density matrix and commensurate grid in reciprocal space to achieve precision on energy convergence of 10^{-6} hartree for all geometry optimizations and single-point energy calculations [356, 357]. A pruned (75,974) integration grid was used to define the radial and angular grid-point distribution. Corrections for basis set superposition error were performed using a counterpoise method [358]. Interaction energies for stacked dimer structures were calculated using dimers extracted from the expanded DFT optimized complex **3** crystal structure. Single-point energies were used to determine interaction energies from $E_{int}^{AB} = E(AB) - [E(A) + E(B)]$, where single-molecule energies for A and B were calculated in the full AB basis.

Calculated and experimental molecular bond lengths, bond angles, dihedral angles, and RMSDs for complex **3** are provided in Table G.2, Table G.3, and Table G.4.

F.6 Supporting Information

Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.inorgchem.9b00917.

Molecular structures of complexes **2** and **4**, electrochemical methods and data, ESI-MS data, tables of X-ray data collection and refinement parameters, and calculated and experimental molecular bond lengths, bond angles, dihedral angles, and RMSDs for complex **3** (Appendix G)

F.7 Acsession Codes

CCDC 1905535, 1905536, and 1905537 contain the supplementary crystallographic data for this paper. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre at www.ccdc.cam.ac.uk/data_request/cif or by emailing data_request@ccdc.cam.ac.uk.

F.8 Author Information

* Email: adamcolson@boisestate.edu

F.9 Notes

The authors declare no competing financial interest.

F.10 Acknowledgements

The authors acknowledge Boise State University for financial support and Prof. Orion Berryman and the Center for Biomolecular Structure and Dynamics CoBRE (National Institutes of Health, CoBRE NIGMS P20GM103546) at the University of Montana for facilitating X-ray data collection. Single crystal X-ray diffraction data were collected using instrumentation supported by NSF MRI CHE-1337908.

APPENDIX G:

SUPPORTING INFORMATION FOR SUPRAMOLECULAR INTERACTIONS OF GROUP VI METAL CARBONYL COMPLEXES: THE FACILITATING ROLE OF 1,3-BIS(*P*-ISOCYANOPHENYL)UREA

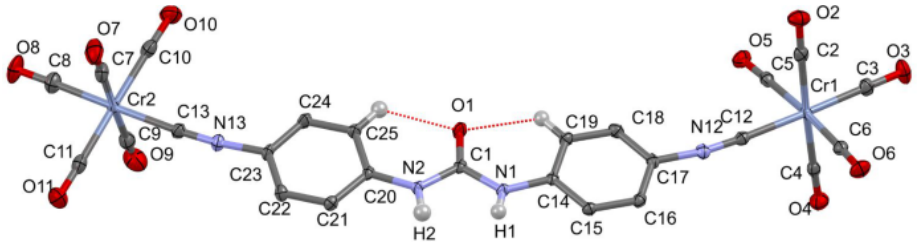


Figure G.1: Molecular structure of complex 2. Thermal ellipsoids are rendered at the 50% probability level. Only urea N–H atoms and aromatic hydrogen atoms participating in hydrogen bonding are shown.

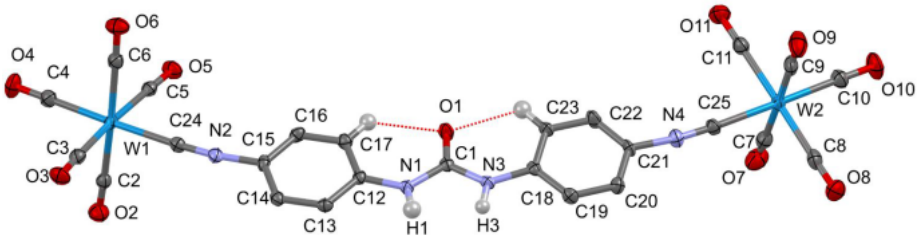


Figure G.2: Molecular structure of complex 4. Thermal ellipsoids are rendered at the 50% probability level. Only urea N–H atoms and aromatic hydrogen atoms participating in hydrogen bonding are shown.

G.1 Electrochemical Measurements

Cyclic voltammograms were recorded in 0.1 M $[Bu_4N][PF_6]$ DMF solutions at ν = 100 mV/sec with a Princeton Applied Research VersaSTAT 3 potentiostat. All experiments were performed using a standard three-electrode configuration under an atmosphere of pure nitrogen. Glassy carbon working electrodes (3 mm, CH Instruments) were used for all measurements and were polished with aqueous slurries of 0.3 μm and 0.05 μm alumina powder, sequentially. After polishing, the electrodes were rinsed with Milli-Q water, methanol, and dichloromethane and dried in a stream of air. Working electrodes were preconditioned by performing three cyclical scans from 2.0 to -2.5 V at 250 mV/sec in a DMF solution of $[Bu_4N][PF_6]$ (0.1 M). A graphite rod served as the counter electrode and a silver wire immersed in a 0.1 M DMF solution of $[Bu_4N][PF_6]$ and separated from the cell compartment by a porous glass frit (CoralPor 1000) was employed as a

Ag^+/Ag pseudoreference electrode. Measured potentials are reported relative to the ferrocenium(1+)/ferrocene(0) redox couple, which was achieved by addition of ferrocene at the end of each set of scans.

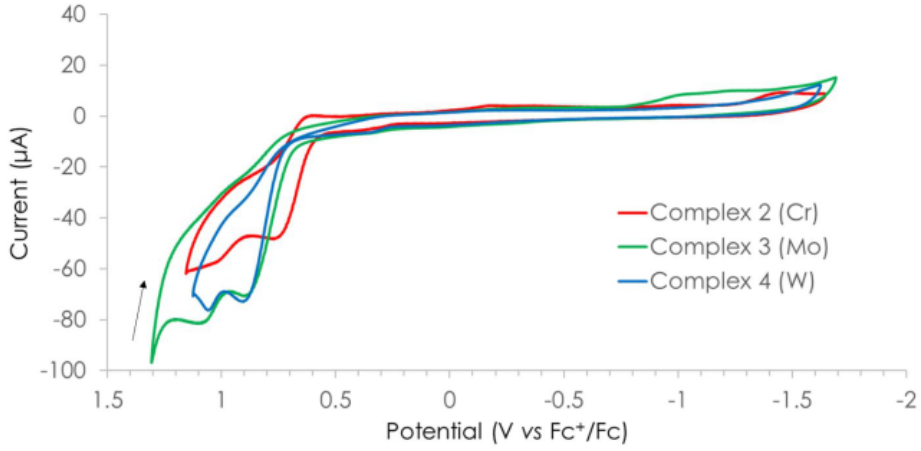


Figure G.3: Cyclic voltammograms of complexes 2-4 (≈ 1 mM) recorded in 0.1 M $[Bu_4N][PF_6]$ DMF solution at ν = 100 mV/sec with a glassy carbon working electrode, graphite rod counter electrode, and a silver wire pseudoreference electrode.

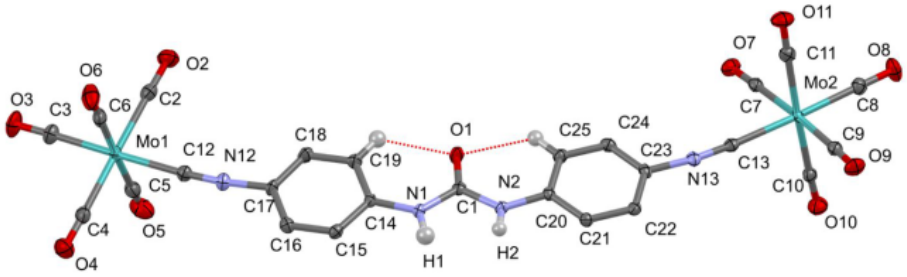


Figure G.4: Atom labels used in structure analysis of complex 3 provided in Table G.2, Table G.3, and Table G.4.

Table G.2: Heavy atom bond lengths (Å) and RMSD for experimental X-ray and calculated crystal structures of complex 3.

Bond	exp.	DFT
C12-Mo1	2.127	2.1778
C14-C15	1.395	1.4104
C14-C19	1.398	1.4094

C15-C16	1.381	1.3892
C16-C17	1.386	1.4013
C17-C18	1.385	1.4021
C18-C19	1.384	1.3907
C20-C21	1.401	1.4084
C20-C25	1.394	1.4085
C21-C22	1.378	1.3885
C22-C23	1.383	1.4003
C23-C24	1.390	1.4038
C24-C25	1.378	1.3904
C26-C27	1.453	1.4564
Mo1-C2	2.052	2.1024
Mo1-C3	2.020	2.0675
Mo1-C4	2.056	2.1078
Mo1-C5	2.063	2.1201
Mo1-C6	2.046	2.0903
Mo2-C10	2.053	2.0989
Mo2-C11	2.052	2.0963
Mo2-C13	2.130	2.1826
Mo2-C7	2.044	2.0925
Mo2-C8	2.022	2.0745
Mo2-C9	2.068	2.1158
N1-C1	1.381	1.3827
N1-C14	1.397	1.3956
N12-C12	1.155	1.1747
N12-C17	1.400	1.3832
N13-C13	1.161	1.1749

N13-C23	1.396	1.3825
N2-C1	1.373	1.3840
N2-C20	1.397	1.3958
N3-C26	1.136	1.1619
O1-C1	1.216	1.2293
O10-C10	1.135	1.1509
O11-C11	1.137	1.1506
O2-C2	1.137	1.1495
O3-C3	1.143	1.1560
O4-C4	1.138	1.1488
O5-C5	1.138	1.1472
O6-C6	1.138	1.1524
O7-C7	1.137	1.1496
O8-C8	1.147	1.1550
O9-C9	1.134	1.1463
RMSD		**0.0279**

Table G.3: Heavy atom bond angles (°) and RMSD for experimental X-ray and calculated crystal structures of complex 3.

Bond Angle	**exp.**	**DFT**
C1-N1-C14	127.10	127.47
C1-N2-C20	128.00	128.30
C10-Mo2-C11	176.44	178.47
C10-Mo2-C13	92.66	91.79
C11-Mo2-C13	90.38	89.43
C12-N12-C17	172.00	170.06
C13-N13-C23	171.90	171.30

C14-C15-C16	120.90	120.95
C14-C19-C18	119.70	119.87
C15-C14-C19	119.30	119.19
C15-C16-C17	119.10	119.29
C16-C17-C18	120.80	120.38
C17-C18-C19	120.10	120.30
C2-Mo1-C12	86.43	88.10
C2-Mo1-C3	92.27	91.92
C2-Mo1-C4	178.49	177.19
C2-Mo1-C5	89.50	89.71
C2-Mo1-C6	89.54	87.64
C20-C21-C22	120.90	121.04
C20-C25-C24	119.50	119.55
C21-C20-C25	119.50	119.37
C21-C22-C23	119.00	119.24
C22-C23-C24	120.80	120.23
C23-C24-C25	120.40	120.56
C3-Mo1-C12	178.49	177.46
C3-Mo1-C4	88.90	90.00
C3-Mo1-C5	89.89	89.28
C3-Mo1-C6	90.82	91.79
C4-Mo1-C12	92.41	90.08
C4-Mo1-C5	91.45	92.36
C4-Mo1-C6	89.49	90.26
C5-Mo1-C12	89.33	88.18
C5-Mo1-C6	178.83	177.17
C6-Mo1-C12	89.94	90.74

C7-Mo2-C10	89.96	90.02
C7-Mo2-C11	88.42	89.14
C7-Mo2-C13	85.38	85.77
C7-Mo2-C8	92.48	91.77
C7-Mo2-C9	175.60	175.80
C8-Mo2-C10	87.78	88.11
C8-Mo2-C11	89.12	90.64
C8-Mo2-C13	177.82	177.54
C8-Mo2-C9	91.83	92.40
C9-Mo2-C10	91.10	90.63
C9-Mo2-C11	90.76	90.29
C9-Mo2-C13	90.30	90.06
Mo1-C12-N12	175.60	176.06
Mo1-C2-O2	178.40	177.36
Mo1-C3-O3	178.50	177.88
Mo1-C4-O4	178.60	178.49
Mo1-C5-O5	178.70	177.70
Mo1-C6-O6	179.20	177.40
Mo2-C10-O10	176.70	177.09
Mo2-C11-O1	178.00	179.22
Mo2-C13-N13	175.50	174.57
Mo2-C7-O7	176.60	176.05
Mo2-C8-O8	179.20	178.48
Mo2-C9-O9	178.70	177.70
N1-C1-N2	111.50	111.70
N1-C14-C15	117.00	116.85
N1-C14-C19	123.70	123.96

N12-C17-C16	120.30	120.85
N12-C17-C18	118.90	118.76
N13-C23-C22	120.50	120.81
N13-C23-C24	118.70	118.96
N2-C20-C21	116.70	116.54
N2-C20-C25	123.80	124.09
N3-C26-C27	178.80	179.85
O1-C1-N1	124.10	124.39
O1-C1-N2	124.40	123.91
RMSD		**0.86**

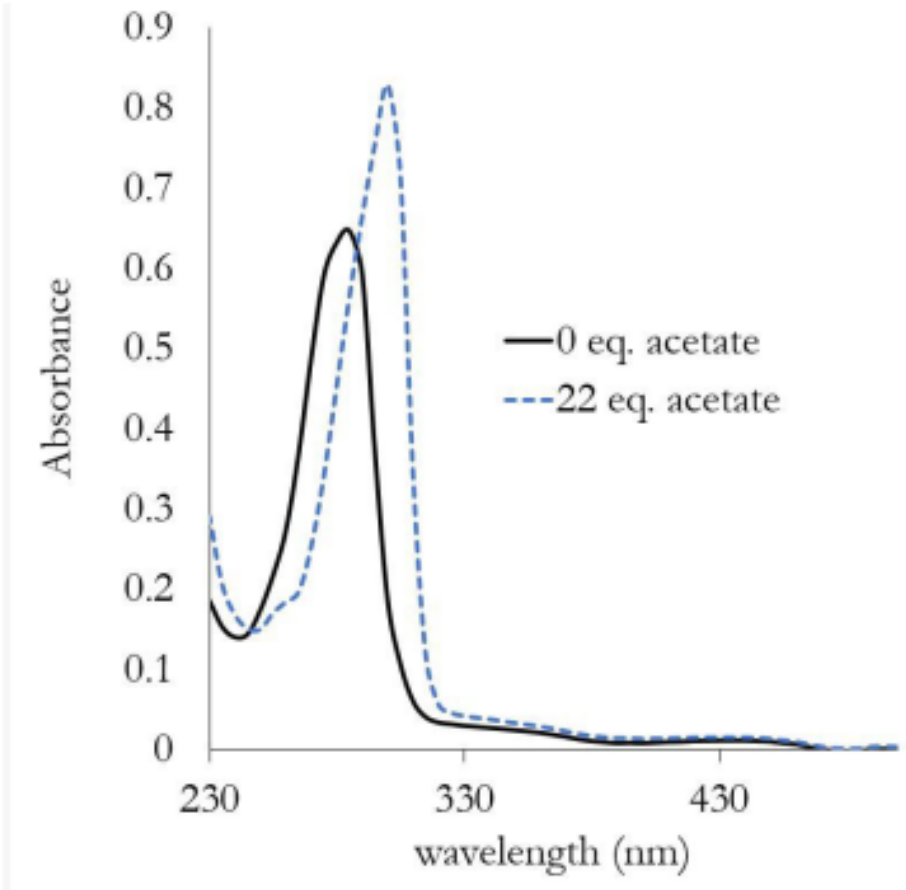


Figure G.5: UV-Vis spectrum of 1 (10.8 μM in CD_3CN) in the absence and presence of excess acetate anion.

Table G.1: X-ray data collection and refinement parameters for complexes 2-4.

Compound	**2**	**3**	**4**
Formula	$C_{25}H_{10}N_4O_{11}Cr_2$ ·CH_3CN	$C_{25}H_{10}N_4O_{11}Mo_2$ ·CH_3CN	$C_{25}H_{10}N_4O_{11}W_2$ ·CH_3CN
Formula weight	687.42	775.30	951.12
Temperature (K)	100	100	100
Crystal system	monoclinic	monoclinic	monoclinic
Space group	$P2_1/c$	$P2_1/c$	$P2_1/c$
a (Å)	6.8126(2)	6.8694(5)	6.8575(4)
b (Å)	13.8536(5)	14.0047(10)	13.9691(9)
c (Å)	32.1439(11)	32.536(2)	32.525(2)
α (deg)	90	90	90
β (deg)	93.8630(10)	93.115(2)	93.145(2)
γ (deg)	90	90	90
Volume ($Å^3$)	3026.82(17)	3125.5(4)	3110.9(3)
Z	4	4	4
density (g/cm^3)	1.509	1.648	2.031
abs coeff (mm^{-1})	0.784	0.867	7.454
F(000)	1384	1528	1784
Crystal size (mm)	0.42 × 0.18 × 0.12	0.4 × 0.05 × 0.05	0.17 × 0.14 × 0.05
λ (MoKα) (Å)	0.71073	0.71073	0.71073
2θ range (deg)	5.87 to 55.068	5.818 to 61.12	5.804 to 54.968
reflns (coll)	40623	120953	50308
reflns (unique)	6937	9565	7120
Data/restraints/ parameters	6937/0/415	9565/0/415	7120/0/415
GOF (on F^2)	1.131	1.072	1.206
Final R indexes [$I \geq 2\sigma$ (I)]	R_1 = 0.0356, wR_2 = 0.0848	R_1 = 0.0327, wR_2 = 0.0578	R_1 = 0.0242, wR_2 = 0.0463
Final R indexes [all data]	R_1 = 0.0433, wR_2 = 0.0877	R_1 = 0.0555, wR_2 = 0.0623	R_1 = 0.0312, wR_2 = 0.0477
Largest diff. peak/hole (e $Å^{-3}$)	0.45/-0.23	0.58/-0.47	0.88/-0.43

Table G.4: Heavy atom dihedral angles (°) and RMSD for experimental X-ray and calculated crystal structures of complex 3.

Dihedral Angle	exp.	DFT
C1-N1-C14-C15	173.1	-179.16
C1-N1-C14-C19	-7.5	1.15
C1-N2-C20-C21	179.3	177.82
C1-N2-C20-C25	-0.5	-2.82
C14-C15-C16-C17	0.1	0.19
C14-N1-C1-N2	-179.2	177.72
C14-N1-C1-O1	1.0	-2.02
C15-C14-C19-C18	-3.2	-1.12
C15-C16-C17-C18	-2.8	-0.83
C15-C16-C17-N12	176.2	178.44
C16-C17-C18-C19	2.5	0.50
C17-C18-C19-C14	0.5	0.49
C19-C14-C15-C16	2.9	0.79
C20-C21-C22-C23	-0.2	-0.06
C20-N2-C1-N1	-178.0	179.74
C20-N2-C1-O1	1.9	-0.52
C21-C20-C25-C24	0.5	1.20
C21-C22-C23-C24	0.0	0.55
C21-C22-C23-N13	-179.1	-178.62
C22-C23-C24-C25	0.4	-0.16
C23-C24-C25-C20	-0.6	-0.73
C25-C20-C21-C22	-0.1	-0.82
N1-C14-C15-C16	-177.6	-178.92
N1-C14-C19-C18	177.4	178.57
N12-C17-C18-C19	-176.5	-178.79
N13-C23-C24-C25	179.5	179.03
N2-C20-C21-C22	-179.9	178.57
N2-C20-C25-C24	-179.7	-178.14
RMSD		**2.46**

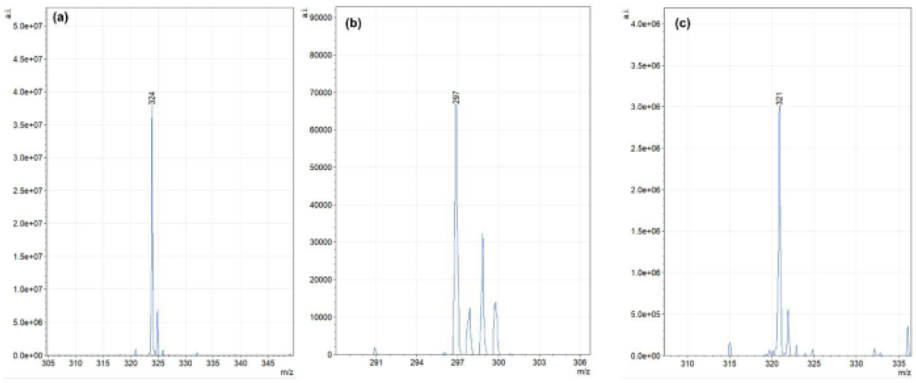


Figure G.6: ESI-MS data for 1:1 host–guest complexes of 1 with (a) NO_3^-, (b) Cl^-, and (c) CH_3COO^-.

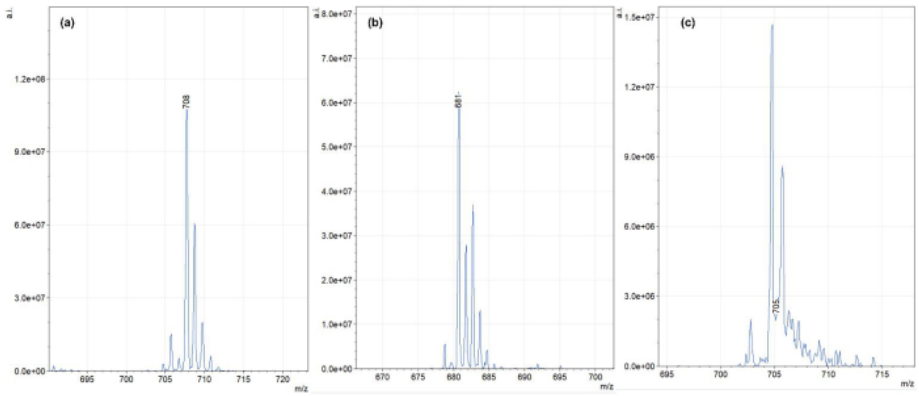


Figure G.7: ESI-MS data for 1:1 host–guest complexes of 2 with (a) NO_3^-, (b) Cl^-, and (c) CH_3COO^-.

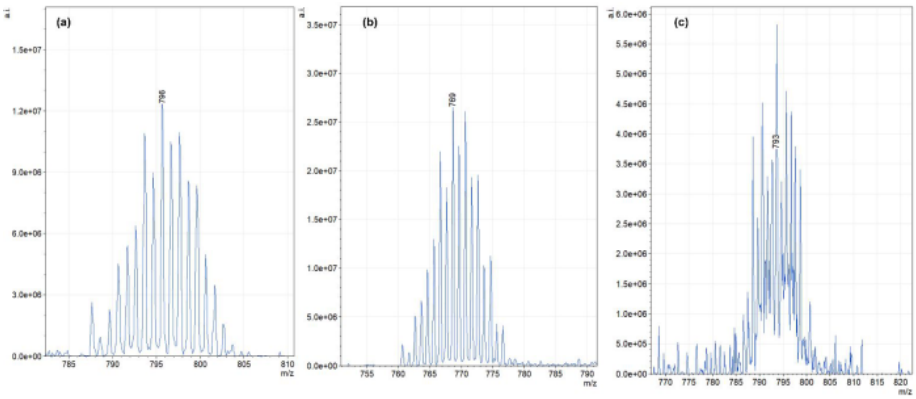


Figure G.8: ESI-MS data for 1:1 host–guest complexes of 3 with (a) NO_3^-, (b) Cl^-, and (c) CH_3COO^-.

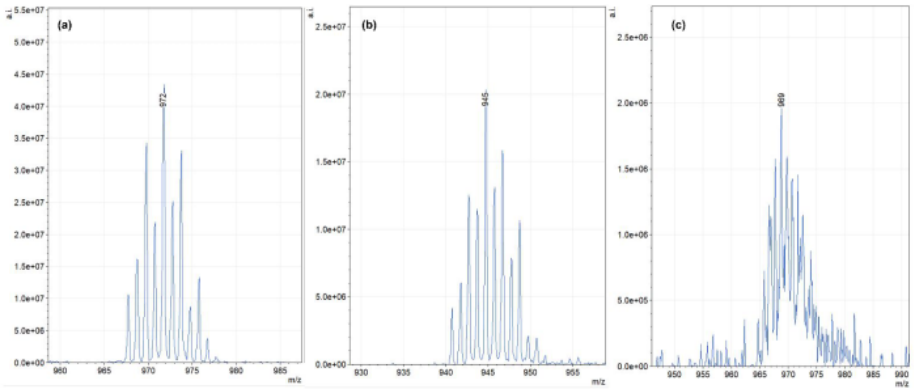


Figure G.9: ESI-MS data for 1:1 host–guest complexes of 4 with (a) NO_3^-, (b) Cl^-, and (c) CH_3COO^-.

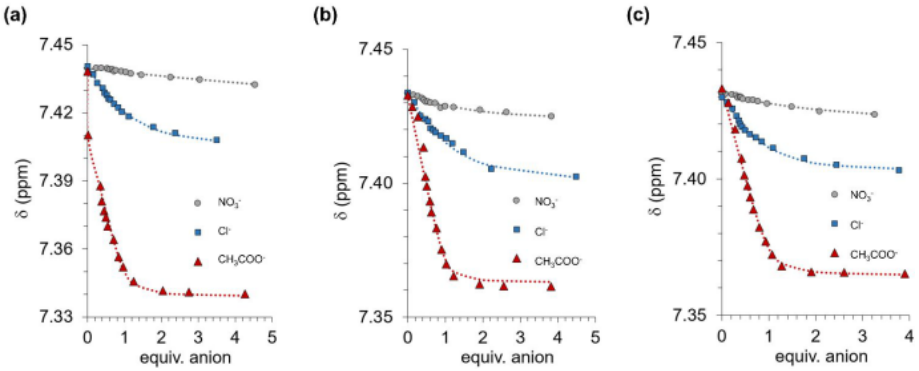


Figure G.10: ^{1}H NMR chemical shifting observed during titration of (a) 2, (b) 3, and (c) 4 ($\sim$0.1 mM in CD_3CN) with nitrate, chloride, and acetate anions. The dotted lines represent the results of non-linear fitting to a 1:1 host–guest binding model.

APPENDIX H:

CATALOGUING THE ENERGETIC CONTRIBUTIONS TO THE SUPRAMOLECULAR ASSEMBLY OF *P*-SUBSTITUTED *N,N'*-DIPHENYLUREAS AND THEIR ORGANOMETALLIC DERIVATIVES IN THE SOLID STATE: A DENSITY FUNCTIONAL THEORY APPROACH

The following chapter was published in Crystal Growth and Design under authors Jenny W. Fothergill, Dr. Adam C. Colson, and Dr. Matthew D. King. My contributions to this paper were DFT calculations under the guidance of Dr. King, writing of the initial draft, and figure creation.

H.1 Abstract

Crystal engineering relies on the predictability of the elaborate interplay of cohesive and conformational energies driven by both intra- and intermolecular interactions of the constituent molecules. In an effort to better understand these influences on crystal packing of *p*-substituted *N,N'*-diphenylureas (*p*DPUs) and organometallic derivatives, we present a detailed computational investigation of *p*DPU species utilizing solid-state density functional theory (DFT), and demon-

strate the applicability of predictive supramolecular synthons applied towards growth of related organometallic complexes. Dominant noncovalent interactions of *p*DPUs can be tuned by altering the electron-withdrawing character of the *p*-substituents. The strength of this electron-withdrawing nature governs the inclination of the molecules to form either dominant electrostatic or π-stacking intermolecular interactions in the crystal structure due to potential molecular conformational stabilization through intramolecular C-H$\cdots$O electrostatic interactions between *ortho* phenyl hydrogens and the urea oxygen atoms. This propensity is also influenced by the symmetry of *p*-substitutions in mono- and di-substituted DPUs. The results of the holistic DFT investigation show a relationship between gas-phase and solid-state conformations, and also presents evidence of mechanisms leading to deviations in predicted crystallization behaviors based on the balance of intra- and inter-molecular interactions. The foundational computational study was expanded to build on previous experimental and theoretical work involving zerovalent transition metal complexes in which *p*-isocyanophenyl DPUs were appended with group IV metal carbonyl fragments. In this study, we synthesized an asymmetric analogue of the latter in which *N*-(*p*-isocyanophenyl)-*N*′-phenylurea (*p*CNHDPU) was appended to a $Mo(CO)_5$ metal carbonyl fragment, allowing us to associate the crystallization behaviors and interactions of organometallic DPU derivatives with those of simpler *p*DPUs. It was observed that the supramolecular assembly of the organometallic complexes display similar predictive patterns, as well as additional complexities to molecular packing arising of the bulkier metal carbonyl substituents. An inclusive computational categorization of DPU-based systems in complement to experimental data will aid in the advancement of design rules for patterned crystal growth of DPU and related systems for the development of innovative materials having unique solid-state properties.

H.2 Introduction

The properties of supramolecular aggregates and molecular crystals may differ considerably from those of their isolated molecular constituents, and developing a deeper understanding of the noncovalent interactions that guide molecular assembly could contribute to advances in various applications, including the self-assembly of light-emitting diodes (LEDs), photovoltaic arrays, and field-effect transistors (FETs) [266, 269, 359, 268, 264, 360]. The *N,N′*-diphenylurea (DPU) moiety is a particularly versatile synthon for supramolecular assembly because it can participate in multiple types of noncovalent interactions. *N,N′* diphenylureas bearing electron-donating functional groups can act as both hydrogen bond donors and acceptors, resulting in the formation of supramolecular "ribbons" or "tapes" (Figure H.1a) [301, 320]. Alternatively, aromatic rings bearing electron-withdrawing functional groups can participate in π-π interactions, resulting in the slip-stack packing of relatively planar *N,N′* diphenylurea units, as depicted in Figure H.1b [360].

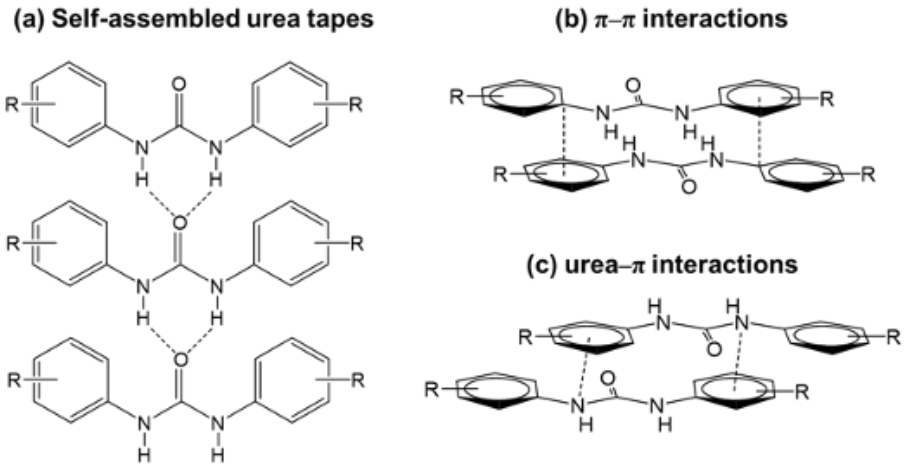


Figure H.1: Supramolecular motifs associated with *N,N′*-diphenylurea compounds.

In addition to π-π interactions, we have recently reported that π-urea interactions can facilitate the assembly of anisotropic molecular arrays in the solid

state [30]. Organometallic DPUs prepared by tethering $Mo(CO)_5$ fragments to 1,3-bis(*p*-isocyanophenyl)urea were observed to exhibit an unusual ladder-like packing motif consisting of alternating π-π and urea-π interactions. Intriguingly, solid state DFT calculations revealed that the cohesive contributions from the π-π and urea-π interactions to the total lattice energy were nearly equal in magnitude. This holistic computational treatment of lattice energy provided important insights into the subtleties of the intermolecular interactions governing molecular assembly in the solid state. The present work describes a more expansive computational effort to survey the lattice energies of DPU systems bearing electron-withdrawing functional groups and to identify the cohesive and conformational strain energy contributions to their overall lattice energies. Additionally, a mono-substituted $Mo(CO)_5$ derivative of the previously reported metal-bound 1,3-bis(*p*-isocyanophenyl)urea was synthesized and crystallized to relate observed supramolecular behavior of these organometallic complexes with the predictability discerned from the explored *p*DPUs. Ultimately, we anticipate that the findings presented in this work will inform the development of design rules for the patterned crystal growth of DPU-based systems.

H.3 Materials and Methods

H.3.1 Theoretical

The DPU molecules selected for this study include 1,3-bis(*p*-cyanophenyl)urea (*p*CyDPU), 1-(*p*-cyanophenyl)-3-(*p*-nitrophenyl)urea (*p*CyNDPU), 1,3-bis(*p*-trifluoromethylphenyl)urea (*p*CF_3DPU), 1-(*p*-cyanophenyl)-3-phenylurea (*p*CyHDPU), 1-(*p*-nitrophenyl)-3-phenylurea (*p*NHDPU), and 1-(*p*-chlorophenyl)-3-phenylurea (*p*ClHDPU), the crystal structures of which have previously been archived in the Cambridge Crystallographic Data Centre (*p*CyDPU, 1554302; *p*CyNDPU, 676835; *p*CF_3DPU, 1554299; *p*CyHDPU, 1554301; *p*NHDPU, 676839; *p*ClHDPU, 1554300) [361, 327]. All DFT calculations were performed using

the Crystal14 software package [349, 350]. Calculations utilized a pairwise dispersion-corrected B3LYP-D2 density functional with the atom-centered triple-ζ basis with polarization functions, pob-TZVP, which is an Aldrich's triple-ζ basis set modified for solid-state calculations [234, 351, 352, 362]. Dimer calculations used Ahlrich's valence triple-ζ basis [363, 364]. Atomic coordinates determined using single crystal X-ray diffraction were used as the initial geometry for the crystal structure optimization. The energy minimization allowed full relaxation of the atom positions within the constraints of the space group symmetry to a residual force of 10^{-6} hartree. Shrinking factors of 3 (*p*CF_3DPU), 4 (*p*CyDPU, *p*CyNDPU, *p*NHDPU, *p*ClHDPU), and 6 (*p*CyHDPU) were used in defining the **k** point sampling of the density matrix in reciprocal space [356, 357]. Corrections for basis set superposition error (BSSE) were performed using the counterpoise method [358]. The lattice energy was calculated using the energy of the optimized crystal structure (E_{cell}) and the BSSE-corrected energy of a single molecule ($E_{moleculeBSSE}$) as follows:

$$E_{Lattice} = \frac{E_{cell} - Z \times E_{moleculeBSSE}}{Z} \tag{H.1}$$

where Z is the number of molecules in the unit cell. The strain energy of the crystal packing was calculated using the single point energies of a molecule extracted from the crystal structure ($E_{moleculeSP}$) and the energy of the molecule after geometry optimization ($E_{moleculeopt}$) as follows:

$$E_{strain} = E_{moleculeSP} - E_{moleculeopt} \tag{H.2}$$

The cohesive energy was calculated as follows:

$$E_{cohesive} = E_{lattice} - E_{strain} \tag{H.3}$$

Interaction energies for the dimer structures were calculated using single-point energies of dimers extracted from the optimized crystal structure (E_{dimer}), along with BSSE-corrected molecular single-point energies as follows:

$$E_{interaction} = E_{dimer} - 2 \times E_{moleculeBSSE} \tag{H.4}$$

H.3.2 Experimental

General Considerations

All synthetic operations were carried out under a nitrogen atmosphere using standard Schlenk techniques to exclude moisture and oxygen. Nitrogen was prepurified by passage through columns of activated copper catalyst (BASF PuriStar R3-11G) and molecular sieves (RCI-DRI 13X). Glassware was dried in an oven at 130 °C, assembled while hot, and allowed to cool under reduced pressure. All solvents were dried according to published procedures and degassed with nitrogen prior to use. $Mo(CO)_6$ (Acros, 98%), trimethylamine *N*-oxide dihydrate (Beantown Chemical, 98%), and phenyl isocyanate (Acros, 99%) were used as received without further purification. 4-Isocyanophenylamine was prepared according to literature procedures and sublimed prior to use [340]. Infrared spectra were obtained using a Thermo Scientific Nicolet iS5 FTIR spectrometer equipped with a 0.2 mm BaF_2 liquid cell. ^1H and ^{13}C NMR data were recorded on a 600 MHz Bruker AVANCE III spectrometer. Electrospray ionization mass spectrometry (ESI-MS) was carried out using a Bruker HCTultra CTD II spectrometer in negative ion mode. CH_3CN solutions of analytes were treated with the tetrabutylammonium salts of chloride and nitrate prior to injection into the mass spectrometer. Elemental analyses were performed by Atlantic Microlab, Inc in Norcross, GA, USA.

Synthesis of *N*-(*p*-isocyanophenyl)-*N'*-phenylurea (*p*CNHDPU)

p-Isocyanophenylamine (318 mg, 2.69 mmol) and phenyl isocyanate (300 mg, 2.52 mmol) were dissolved in 20 mL of CH_3CN. The reaction mixture was heated at reflux for three hours, after which the solvent was removed under reduced pressure. The solid residues were washed with toluene followed by hexanes and dried under reduced pressure. The product was isolated as a pale yellow solid (433 mg, 72% yield). IR (CH_2Cl_2, cm^{-1}): ν_{CN} 2064 (vs). ^1H NMR (CD_3CN, 20 °C): δ 7.05 (t, J = 7.41 Hz, 1H, Ar H), 7.31 (t, J = 8.15 Hz, 2H, Ar H), 7.37 (m, 3H), 7.44 (d, J = 7.61 Hz, 2H, Ar H), 7.53 (d, J = 8.59 Hz, 2H Ar H), 7.55 (br s, 1H, NH). ^{13}C NMR (CD_3CN, 20 °C): δ 117.3, 119.0, 122.9, 127.1, 128.8, 139.1, 140.7, 152.3, 163.4, 206.5. MS(ESI): m/z 272 $[M + Cl^-]^-$, 299 $[M + NO3^-]^-$. Anal. Calcd for $C_{14}H_{11}N_3O$: C, 70.87; H, 4.67; N, 17.71. Found: C, 70.58; H, 4.63; N, 17.91.

Synthesis of $Mo(CO)_5$(*p*CNHDPU)

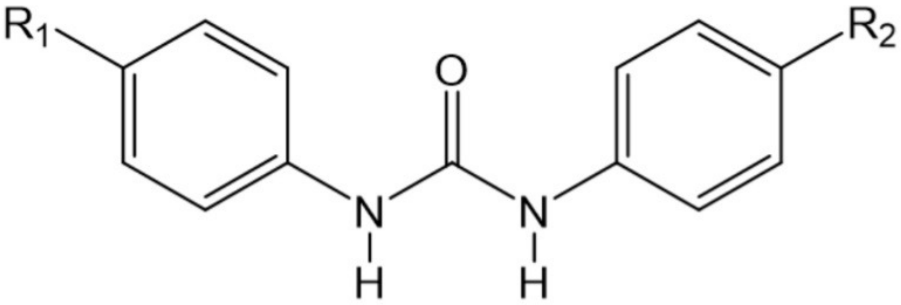
$Mo(CO)_6$ (556 mg, 2.11 mmol) and *p*CNHDPU (500 mg, 2.11 mmol) were dissolved in 25 mL of tetrahydrofuran (THF), and a dropping funnel charged with trimethylamine *N*-oxide dihydrate (233 mg, 2.10 mmol), THF (10 mL), and methanol (10 mL) was attached to the reaction flask, the contents of which were added dropwise to the reaction mixture over the course of one hour. The mixture was magnetically stirred for six hours at room temperature, after which the solvents were removed under reduced pressure. The residues were extracted with ethyl acetate, filtered to remove insoluble impurities, and adsorbed onto silica gel. $Mo(CO)_5$(*p*CNHDPU) was isolated by column chromatography using silica gel as the stationary phase and a mixed solvent system as eluent (5:4:1 vol. eq. CH_2Cl_2, hexanes, and ethyl acetate). The chromatographic fractions were dried under reduced pressure and the residues dissolved in *N*,*N'*-dimethylformamide (DMF). Slow titration of the DMF solution with distilled water produced colorless crys-

talline needles (606 mg, 61% yield). IR (CH_2Cl_2, cm^{-1}): ν_{CN} 2143 (m), νCO 2063 (s), 1956 (vs). ^{1}H NMR (CD_3CN, 20 °C): δ 7.05 (t, 1H, Ar-H, J = 7.41 Hz), 7.31 (t, 2H, Ar-H, J = 8.15 Hz), 7.40 (m, 3H), 7.45 (d, J = 7.61 Hz, 2H, Ar H), 7.55 (d, J = 8.59 Hz, 2H Ar H), 7.61 (br s, 1H, N-H). ^{13}C NMR (CD_3CN, 20 °C): δ 119.3, 127.2, 140.4, 151.8, isocyanide and adjacent ipso carbons not observed. MS(ESI): m/z 510 [M + $Cl^-]^-$, 537 [M + $NO3^-]^-$. Anal. Calcd for $C_{19}H_{11}MoN_3O_6$: C, 48.22; H, 2.34; N, 8.88. Found: C, 48.16; H, 2.26; N, 8.98.

H.4 Results and Discussion

The crystal structures of several *p*-substituted DPUs were acquired from the Cambridge Crystallographic Database, which represent a diverse cross-section of DPU derivatives to compare preferred molecular conformations, crystal packing configurations, and interaction energies as a function of the electron-withdrawing nature of the *p*-substituents. The systems studied were 1,3-bis(*p*-cyanophenyl)urea (*p*CyDPU), 1-(*p*-cyanophenyl)-3-(*p*-nitrophenyl)urea (*p*CyNDPU), 1,3-bis(*p*-trifluoromethylphenyl)urea (*p*CF_3DPU), 1-(*p*-cyanophenyl)-3-phenylurea (*p*CyHDPU), 1-(*p*-nitrophenyl)-3-phenylurea (*p*NHDPU), and 1-(*p*-chlorophenyl)-3-phenylurea (*p*ClHDPU) (Figure H.2). The selected structures offer both mono-and di-*para*-substituted DPUs having electron withdrawing substituents of varying magnitudes.

The implications of the *p*DPU molecular structures is that they may facilitate the dominant interactions driving crystal packing formation. Information on molecular structure can therefore provide predictive insight towards patterned crystal growth and engineering. There exists an interplay between intramolecular forces, intermolecular forces, and molecular conformational strain that govern how each DPU will behave in the gas and solid states. In the crystalline solid, any molecular DPU conformation can participate in the urea ribbon motif, but planar molecules are favored for π-stacking interactions. In instances of DPU planarity, steric hin-



R_1	R_2	Abbreviation	CCDC Code
C≡N	C≡N	*p*CyDPU	1554302
C≡N	NO_2	*p*CyNDPU	676835
CF_3	CF_3	*p*CF_3DPU	1554299
C≡N	H	*p*CyHDPU	1554301
NO_2	H	*p*NHDPU	676839
Cl	H	*p*ClHDPU	1554300

Figure H.2: Chemical structures of *p*DPUs examined in this study.

drances of adjacent rings prevent the formation of the hydrogen-bonded urea ribbon motif and rely on stabilization through π-π interactions. However, ring substituents may contribute to additional lattice stabilization by forming hydrogen bonds with the available urea moieties. In order for a DPU to achieve planarity, the electrostatics of the ring must be altered through addition of electron-withdrawing substituent groups at the *para* position. Such additions result in a coupled withdrawal of charge density at the *ortho* position, thus generating greater positive charge localized on the *ortho*-hydrogens. Although not a "true" hydrogen bond, the electrostatic interaction between the *ortho*-hydrogen on the aromatic ring and the urea oxygen may facilitate stabilization of a co-planar aryl conformation. This favorable electrostatic interaction is countered by the repulsion between the opposite *ortho*-hydrogen of the same ring and the urea hydrogen. The magnitude of the hydrogen bonding is greater than that of this H-H repulsion, which can therefore be overcome only through structural addition of

strong electron-withdrawing groups. The favored *molecular* conformation ultimately affects the *crystal* packing configuration, and the degree of DPU planarity and its ability to form stabilizing π-π interactions in the crystal phase is dependent on the magnitude of the electron-withdrawing character of the *p*-substituent group(s).

Contained in the investigated structures are the ring substituents chloro, cyano, trifluoromethyl, and nitro groups. From our single-molecule DFT calculations, Mulliken charges of the electron-withdrawing groups were determined and ranked as NO_2 > Cl > CN > CF_3 in terms of net negative group charge (averages of -0.54, -0.23, -0.13, -0.10, respectively). The propensity to form planar molecular structures, and therefore encourage π-π stacking motifs in the crystal structure, is a result of the withdrawal of electron density from the phenyl ring and shifting electron density away from the *ortho*-H, thereby increasing the strength of the *o*-H$\cdots$O$_{urea}$ stabilizing intramolecular interaction. Evident by the single-molecule structural optimizations shown in Figure H.3, the strength of the electron-withdrawing character at the *para* position has a notable effect on preferred molecular configuration. Non-substituted DPU optimizes with the phenyl groups twisted $\sim$40° from the urea N-C-N plane due to repulsion between the *o*-H and H$_{urea}$, yielding H$\cdots$H distances of 2.29 Å. These repulsive forces are balanced by the weak interactions between the adjacent *o*-H and the O$_{urea}$ at 2.40 Å. The addition of *p*-substitutions to the DPU molecular structures alters the electron density at the important *ortho* position and, hence, the ring orientations. The disubstituted *p*CyNDPU contains the strongest electron-withdrawing groups, and the resulting optimized structure is nearly planar due to the increase interaction strength between the *o*-H and the O$_{urea}$. The preferred molecular orientation has nearly equivalent *o*-H$\cdots$O$_{urea}$ and *o*-H$\cdots$H$_{urea}$ separations of 2.20 and 2.21 Å, respectively. Examination of the remaining molecule structures reveals that devi-

ation from planarity is increased. In the case of pCF$_3$DPU, which has the weakest electron withdrawing character, the resulting structure is nearly that of the non-substituted DPU structure with o-H$\cdots$O$_{urea}$ and o-H$\cdots$H$_{urea}$ distance of 2.37 and 2.28 Å, respectively.

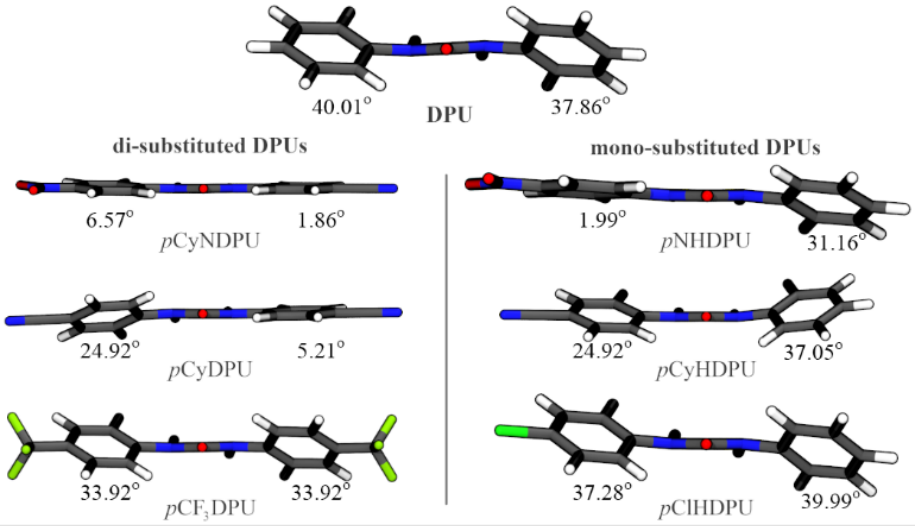


Figure H.3: Single-molecule geometry optimizations of pDPUs.

The trend in ring planarity as a function of the electron-withdrawing character of the ring substituent is clear looking at the three optimized mono-substituted DPU molecules. The o-H$\cdots$O$_{urea}$ distances with respect to the p-substituted ring were 2.19, 2.28, and 2.38 Åfor the NO_2, CN, and Cl substitutions, respectively (2.21, 2.26, and 2.30 Åfor the o-H$\cdots$H$_{urea}$). The relative orientation of the non-substituted ring was not significantly different between each of the structures, and corresponded to the structural parameters observed for the non-substituted DPU structural optimization. However, slight variations in the o-H$\cdots$O$_{urea}$ distances were calculated due to the overall molecular redistribution of electron density resulting from the single ring substitution. These distances with respect to the non-substituted ring were calculated to be 2.34, 2.39, and 2.43 Åfor the NO_2, CN, and Cl substitutions, respectively (2.27, 3.33, and 2.30 Åfor the o-H$\cdots$H$_{urea}$).

In molecular crystals systems, the relationship of packing forces and conforma-

tional strain energies often results in deviations between molecular structures observed in the solid phase and that of the optimized single-molecule 'gas phase' orientation. These structural deviations can sometimes be significant such that molecules cannot be treated as rigid building blocks in the construction of thermo-dynamically stable crystal structures. Optimization of lattice energy during crys-tal growth depends upon maximization of stabilizing intermolecular interactions while minimizing distortion of the molecules from their most stable molecular conformation. It is the relative strengths of these intra- and intermolecular forces that ultimately determine the crystalline configuration of organic molecules. As such, the single-molecule structures of the *p*DPUs were compared to those taken in their respective crystal structures in order to better understand the influence of preferred molecular conformation of crystal packing assembly. In the case of *p*DPUs, two primary intermolecular interactions are present: hydrogen bonding through the urea moiety and π-π interactions between phenyl rings. In the for-mation of a stable crystalline packing orientation, the system must either maintain the preferred molecular structure or overcome the magnitude of the intramolecu-lar interactions within the *p*DPU molecules.

The general packing assembly of the *p*DPUs are shown in Figure H.4. Although all crystal structures are composed of molecules containing the DPU motif, the crystal structures are quite different depending on the dominant supramolecular interactions. The molecular orientation of di-substituted DPUs as they are found in the crystal structure generally follow those of the single molecule optimizations. Stabilization of the planar molecular structures in the cases of *p*CyNDPU and *p*CyDPU through *o*-H$\cdots$O$_{urea}$ interactions, and inversely the lack of electrostatic stabilization through the weakly electron withdrawing *p*-substituents of *p*CF_3DPU, provides the 'rigid' building blocks for crystalline assembly. Preferentially planar molecules are likely to adhere to a π-stacking crystal orientation, and

those lacking robust *o*-H· · ·O_{urea} interactions adopt urea hydrogen-bonding rib-bon motif in the solid-state. Given the stability afforded by dominant intramolecu-lar interactions, in this case the *o*-H· · ·O_{urea} interactions, reliable predictions can be made on the molecular configurations likely to be observed in the solid phase. The intricacies of the crystalline conformation also depend on other thermodynamic factors and the ability of the substituents to form additional stabilizing interaction with neighboring molecules. Take for example the *p*CyNDPU and *p*CyDPU crys-tal systems. Even though both molecules adopt a planar molecular conformation in the crystal structure, the mechanism of stacking of these planar molecules dif-fers between the two. The *p*CyDPU molecule forms continuous chains stabilized through π - π interactions, while *p*CyNDPU forms paired molecular π - π stack-ing and lacks a continuous chain resulting from this type of interaction. The rea-son is the ability of the *p*CyNDPU to also form hydrogen bonds with neighboring molecules through adjacent -NO_2 and -CN substituents. This takes advantage of not only the spatial advantage of stacking flat molecular building blocks, but also enables strong contacts to be made between neighboring molecules. The result of this is the maximization of crystal density and minimization of total lattice energy by forming the strongest possible ensemble of favorable molecular interactions.

The situation becomes more complex for mono-substituted DPUs where only half the molecule may experience stabilizing intramolecular *o*-H· · ·O_{urea} contacts. In this case, the intricate balance of intra- and intermolecular forces guiding crystal packing configurations leans towards the dominant crystalline intermolecular in-teractions. Energetic favorability is gained through disruption of the molecular structure and subsequent optimization of molecular packing rather than main-taining the gas-phase molecular orientation. Most demonstrable in the crystal sys-tems examined is that of *p*CyHDPU in comparison to the di-substituted *p*CyDPU. Mono-substitution of the electron-withdrawing -CN group is not sufficient to pre-

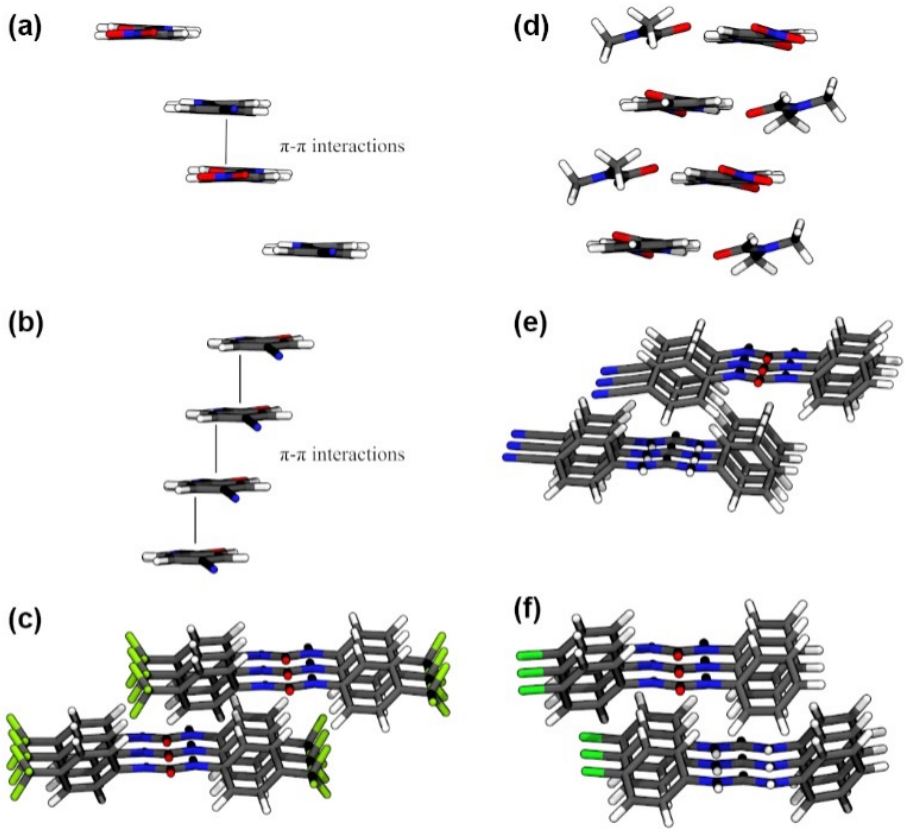


Figure H.4: Primary packing orientations within the *p*DPU crystal structures of (a) *p*CyNDPU, (b) *p*CyDPU, (c) *p*CF_3DPU, (d) *p*NHDPU, (e) *p*CyHDPU and (f) *p*ClHDPU.

serve a planar molecular orientation in the crystal phase. The resulting crystal structure contains *p*CyHDPU molecules in the preferred packing efficiency that involves the urea ribbon motif, thus overcoming the intramolecular interactions resulting in the partially planar gas-phase orientation imposed by the electron-withdrawing -CN substituent. Also a factor is the reduced ability to form stabilizing hydrogen bonding between the -CN *p*-substituents and the urea hydrogens an efficient manner due to spatial constraints and reduced molecular symmetry.

Noteworthy disparity in crystallization behavior is also discerned in *p*NHDPU. This molecule forms a solvate crystal structure in which DMF solvent molecules are intercalated into the co-crystallized system at a 1:1 ratio with *p*NHDPU. The energetic balance is addressed by allowing *p*NHDPU to maintain their preferred

planar molecular orientations afforded by the strong electron-withdrawing character of the -NO_2 substituent. Without the ability for NO_2 to form additional stabilizing intermolecular hydrogen bonds with neighboring molecules in a spatially favorable manor, it becomes advantageous for co-crystallization with DMF as opposed to a geometrically strained π-stacking configuration. Additional stabilization energy is acquired through hydrogen bonding of the DMF oxygen atom with the urea hydrogens of *p*NHDPU. The reduction in molecular symmetry between the di-substituted *p*CyNDPU and mono-substituted *p*NHDPU prevents an efficient packing configuration to enable the hydrogen bonding between *p*-substituents and urea moieties. Interestingly, there is also no π-stacking interactions stabilizing the crystal structure, which would be expected in systems containing co-planar phenyl groups. The limitations in options for hydrogen bonding within the crystal to overcome the intramolecular forces, and the associated limitations in packing configurations having one planar and one non-planar phenyl orientation, provide a more favorable environment for solvent inclusion in the thermodynamically stable *p*NHDPU crystal structure. The molecules are able to maintain their planar configuration while achieving a favorable crystal packing density through the solvent inclusion.

Solid-state DFT calculations where performed to evaluate the balance of interactions leading to the desired crystalline conformations observed for the various *p*DPUs. The lattice energy provides a measure of the driving force for the molecules to form in the solid state. In order to obtain an optimal lattice energy, the system must be able to form favorable intermolecular interactions while being counteracted by the necessity to overcome intramolecular attractions, thus creating an unfavorable molecular strain energy that is compensated by a more advantageous lattice energy. The results of energy calculations of the *p*DPU crystal systems are provided in Table H.1 and expressed in terms of lattice energy bro-

ken down into cohesive and molecular conformational strain energies. Also tabulated are the interaction energies between dimers extracted from the crystal structures that contain the desirable supramolecular interaction motif, i.e. hydrogen-bonded chain or π-stacking, exhibited in each crystal system. Although many favorable interactions for crystal assembly may be present in a particular sys-tem, only the dominant intermolecular interactions were examined. This energetic component is isolated to provide additional information that connects the single-molecule and crystalline conformations of the constituent molecules. In the case of *p*NHDPU, for which DMF solvent co-crystallizes at a 1:1 ratio, the prominent dimer energetic contribution is that of the hydrogen bonding between the DMF oxygen and the *p*NHDPU urea hydrogens. Mulliken population analysis was performed for each system to analyze partial charges on the O_{urea}, *o*-H, and *o*-C atoms of the crystalline *p*DPU structures (Table H.2). The results show that when the *p*-substituted groups are sufficiently electron-withdrawing, the hydrogens in the *ortho* position are more positively charged and can therefore interact more strongly with O_{urea}. This is best demonstrated by comparing the symmetric *p*CyDPU and *p*CF_3DPU molecules, for which the stronger electron withdrawing group, -CN, produced a net positive charge of +0.090 at the *o*-H's, whereas CF_3 substitution yielded only a charge of +0.056 on these atoms. This seemingly small difference in atomic charge is sufficient to greatly influence the intramolecular structure and the resulting primary intermolecular motif exhibited in the crystal structure.

The lattice energies are similar among the di-substituted *p*DPU structures, with a slight increase in stabilization energy as the strength of the electron-withdrawing *p*-substituent decreases (Table H.1). The same trend is observed in the dimer energies and is much more pronounced, suggesting that the hydrogen bonded urea ribbon motif is considerably more stable than the primary π-stacking interactions exhibited in the *p*CyDPU and *p*CyNDPU crystal packing (Figure H.4).

Table H.1: Calculated lattice, cohesive, strain, and dimer energies for *p*DPU crystal structures. Energies provided in units of kcal mol^{-1}.

Complex	$E_{Lattice}$	$E_{Cohesive}$	E_{Strain}	E_{Dimer}
di-substituted				
*p*CyDPU	-72.83	-94.78	21.95	-49.07
*p*CyNDPU	-74.60	-97.58	22.98	-57.17
pCF_3DPU	-76.72	-105.46	28.74	-73.64
mono-substituted				
*p*CyHDPU	-65.17	-87.39	22.22	-57.27
*p*NHDPU	-90.87	-112.70	21.83	-42.49[a]
*p*ClHDPU	-66.39	-88.55	22.15	-57.38

[a] *p*NHDPU:DMF dimer interaction energy

Table H.2: Mulliken population analysis of partial charges on the O_{urea}, *o*-H, and *o*-C atoms in *p*DPUs (reported as elementary charge units). The average charges are reported for di-substituted DPUs, and both substituted and non-substituted phenyl groups are reported for mono-substituted DPUs. The phenyl ring with the electron withdrawing group (EWG) or hydrogen (H) at the *para* position are denoted.

di-substituted	O_{urea}	***o*-H**		***o*-C**	
*p*CyDPU	-0.73	+0.09		-0.05	
*p*CyNDPU	-0.73	+0.09		-0.05	
pCF_3DPU	-0.75	+0.06		-0.00	
mono-substituted	O_{urea}	***o*-H (EWG)**	***o*-H (H)**	***o*-C (EWG)**	***o*-C (H)**
*p*CyHDPU	-0.77	+0.06	+0.04	-0.01	+0.01
*p*NHDPU	-0.74	+0.09	+0.07	-0.06	-0.04
*p*ClHDPU	-0.75	+0.05	+0.05	+0.00	-0.01

However, in order to achieve hydrogen bonding between urea moieties of adjacent molecules, the molecules must undergo large conformational changes. Accompanied with this change is a substantial increase in conformational strain energy. For instance, the *p*CyDPU and *p*CyNDPU molecules exhibited comparable conformational strain energies of 21.95 and 22.98 kcal mol^{-1}, whereas *p*CF_3DPU revealed a larger 28.74 kcal mol^{-1} due to the molecular deformation required to attain the urea ribbon motif. The strain energy of the planar *p*CyDPU and *p*CyNDPU molecules would be necessarily much higher to achieve an analogous ribbon hydrogen bonding pattern since more energy would be required to break their stronger *o*-H$\cdots$$O_{urea}$ contacts. It is therefore more energetically favorable for the planar *p*CyDPU and *p*CyNDPU configurations to maintain intramolecular *o*-H$\cdots$$O_{urea}$ contacts while minimizing conformational strain energy. The trade-off of energy contributions yields comparable lattice energies for each of the disubstituted DPU crystal structures.

Of the molecular crystals examined in this study, only *p*CyDPU and *p*CyNDPU display evidence of π-stacking (Figure H.5). The symmetrically substituted *p*CyDPU exhibits isotropic molecular stacking, forming contiguous π-stacked arrangements in the solid state, whereas the π-stacked *p*CyNDPU molecular dimers exist in orientations inverted to one another. This difference nonetheless yields comparable distances between ring centroids of 3.76 and 3.83 Å, respectively. Energy calculations show the *p*CyNDPU dimer interaction to be more favorable by 8.1 kcal mol^{-1} due to added constructive interactions primarily between the *o*-H and N_{urea} of adjacent DPU molecules. The variations in relative orientations and overall packing configurations are accounted for by the additional hydrogen bonding interactions available between *p*-substituents and the DPU urea group in the crystal structures.

The most favorable lattice energy is observed in *p*NHDPU. The dominant inter-

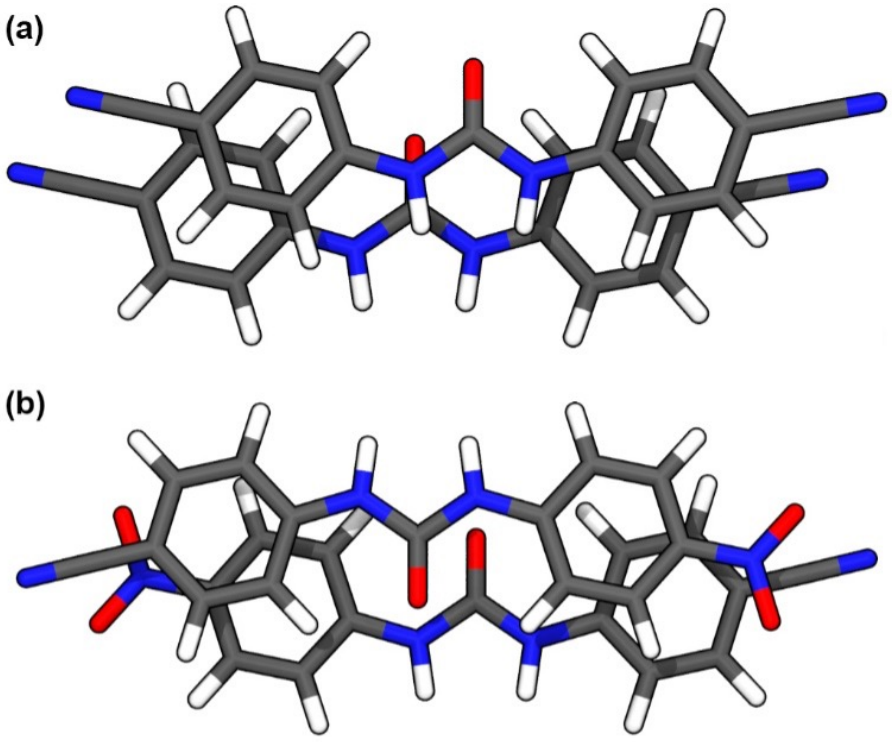


Figure H.5: (a) Top view of *p*CyDPU dimer showing isotropic slipped π-stacking motif. The distance between the ring centroids is 3.76 Å. (b) Top view of *p*CyNDPU dimer showing slipped π-stacking motif, with a distance between ring centroids of 3.83 Å in the crystal structure.

action in this crystal is hydrogen bonding between the urea hydrogens and the DMF oxygen. The molecule remains moderately planar allowing the crystal packing to adopt a herringbone stacking pattern. Although the planar conformation of *p*NHDPU has the potential for π-stacking, the distance between the ring centroids (7.39 Å) in the crystal structure is too great to expect that orbital overlap facilitates the crystal self-assembly. Despite the most favorable lattice energy, the weakest calculated primary interaction energy was between *p*NHDPU and the co-crystallized DMF solvent. This was not surprising given the relatively small size of the DMF molecule and that the hydrogen bonding interaction between *p*NHDPU and DMF is likely not as strong as those in structures with the urea ribbon motif since the DMF molecule will exhibit more thermal motion, whereas the urea ribbon in stabilized on both sides by the ribbon-like hydrogen bond net-

work. However, the additional spatial freedom afforded by the inclusion of the smaller DMF molecules allows the crystal system to adopt the more energetically favorable herringbone packing configuration.

The least favorable lattice energy is observed in the asymmetric *p*CyHDPU which crystallizes according to the urea ribbon motif. The calculated lattice energy was -65.17 kcal mol^{-1}, in contrast to that of *p*NHDPU of -90.87 kcal mol^{-1}. This raises the question as to why additional thermodynamic stability could not be achieved by *p*CyHDPU through the formation of a DMF solvate as did its *p*NHDPU counterpart. Inspection of the stabilizing partial charges on the *o*-H atoms and the calculated strain energies of the two molecules suggests that considerable energy would be required for the *p*CyHDPU molecule to adapt to a planar configura-tion in the crystalline configuration. The additional conformational strain en-ergy needed to force this configuration would outweigh the stabilization energy gained by incorporating DMF solvent molecules into the crystal structure. One must also consider the nucleation and growth processes involved in crystallization. The DMF-*p*NHDPU dimer interaction of -42.49 kcal mol^{-1} is much weaker than that of the *p*CyHDPU dimer of -57.27 kcal mol^{-1}. The stronger interaction of the *p*CyHDPU dimer likely drives the contiguous kinetic growth of the crys-tal, whereas the combination of DMF stabilization and the inclination to conserve the planar intramolecular structure of *p*NHDPU drives the thermodynamically favorable crystallization of the DMF-*p*NHDPU solvate, even though the crystal structure lacks any obvious intermolecular interactions constructively leading to the auspiciously high lattice energy.

Comparing structural similarities/disparities of those DPUs forming the urea ribbon motif, i.e. *p*CF_3DPU, *p*CyHDPU, and *p*ClHDPU, two primary ring orientations are observed (Figure H.6). In order to achieve the ribbon hydrogen bonding pattern, the rings substituents must deviate from planarity. The *p*CF_3DPU and

*p*ClHDPU have isostructural molecular conformations in their respective crystal structures. The ring substituent groups are twisted in opposite directions in relation to the urea moiety at plane angles formed between phenyl rings of 83.90° and 90.77°, respectively. The *p*CyHDPU molecule adopts a different orientation in which the rings are rotated in the same direction in relation to the urea groups, possibly driven by the stronger electron-withdrawing character of the -CN group maintaining a higher degree of the *o*-H$\cdots$O$_{urea}$ contact within the molecule. The resulting angle between phenyl planes was 142.58°. The differences observed in the DPUs are again the balance of intra- and intermolecular forces guiding the crystal formation, with variations in phenyl orientations resulting from the system pushing for optimal packing density through intermolecular interactions while preserving favorable intramolecular stabilization.

The underlying motivation to investigate the intricacies of the crystal structures of *p*-substituted DPUs comprises the interest of acquiring a broader understanding of how molecular structure, intra- and intermolecular interactions, and packing forces govern the organization of molecular solids. Progression of crystal engineering capabilities to direct the organization of molecular solids according to desired structural motifs requires the knowledge of not only the easily discernable intermolecular interactions, but also the subtle interactions that may exist within a molecule or that may be exposed following crystallization—many of which are often unexpected. This holds true for many crystal systems of related molecular species. Our interest in *p*DPUs originates from our current developmental endeavors of low- and zerovalent transition metal carbonyl complexes incorporating *p*-substituted *N,N'*-diarylurea moieties as a facilitating agent for supramolecular chemistry and crystal engineering of advanced materials. The multiple supramolecular motifs made accessible by such DPUs offers much versatility in the generation of novel organometallic structures and functional mate-

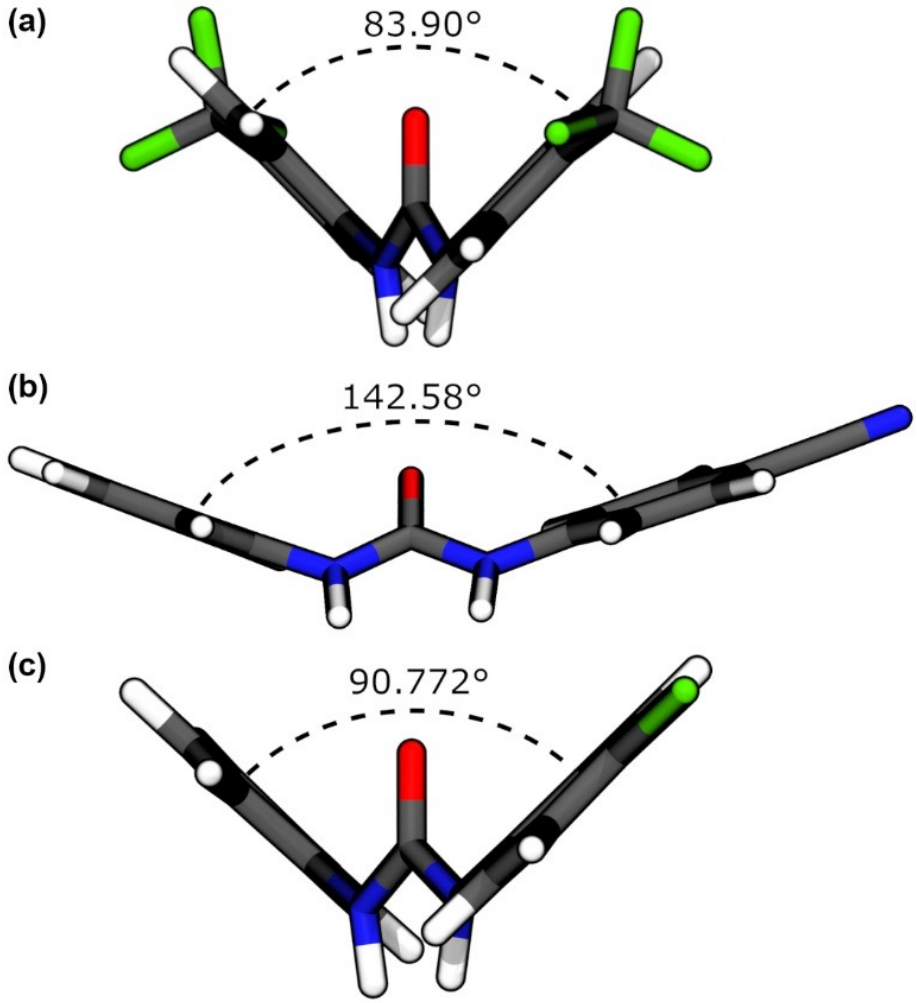


Figure H.6: Angle between ring planes in DPU structures crystallizing according to the urea ribbon motif: (a) *p*CF_3DPU, (b) *p*CyHDPU, and (c) *p*ClHDPU.

rials.

We recently detailed the multifaceted interactions found in the crystal structures of organometallic complexes containing 1,3-bis(*p*-isocyanophenyl)urea tethered to $M(CO)_5$ fragments (M = Cr, Mo, and W) via isocyanide linkages (Figure H.7). The coordination of bulky metal carbonyl fragments to the DPU scaffold introduced additional layers of complexity in that significant steric effects would expectedly influence the crystal packing configuration and the observed interaction motifs. The binuclear organometallic complexes were observed to pack into ladder-like anisotropic arrays in the solid state. Crystallographic and computational evidence

suggested that the observed packing motif could be attributed to a combination of intermolecular π-π and urea-π interactions. Intriguingly, the magnitude of the unconventional urea-π interactions was nearly equivalent to the familiar π-π stacking motif, with respective cohesive energies of urea-π and π-π dimers calculated at -111.64 and -118.43 kcal mol^{-1}. The strength of the urea-π interactions was explained through comprehensive analysis of electrostatics and favorable HOMO/LUMO overlap of the dimer complexes. Notable structural similarities of the carbonyl complexes with the *p*-substituted DPUs presented in this study are that of ring planarity resulting from strong electron-withdrawing ring substituents and the resultant slipped-stacked molecular packing configuration, as opposed to a urea ribbon hydrogen bonding motif. The supramolecular assembly of these complexes seems to be governed first by the favorable intramolecular *o*-H$\cdots$O$_{urea}$ contacts and secondly by spatial acclimations afforded by intermolecular motifs advantageous to planar DPU compounds.

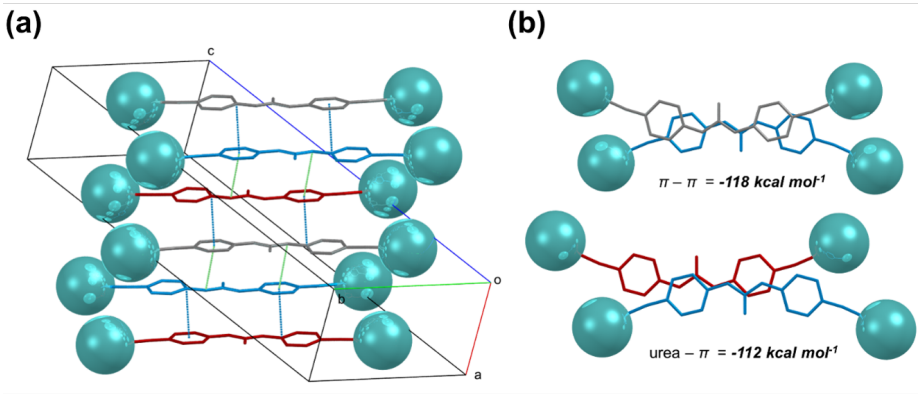


Figure H.7: (a) Solid-state packing of Mo(CO)$_5$-appended 1,3-bis(*p*-isocyanophenyl)urea showing the orientation of the molecules relative to the unit cell axes. (b) Alternative representation viewed through the *N,N′* diarylurea planes. All CO ligands and H atoms are omitted and Mo atoms are depicted at full Van der Waals radius (reproduced from CCDC 1905536) [30]

Based on the findings of the present and past works, we were inspired to extend our computational investigations to the crystal structure of an *asymmetric* organometallic complex prepared by tethering a Mo(CO)$_5$ fragment to *N*-

(*p*-isocyanophenyl)-*N*′-phenylurea (*p*Mo(CO)$_5$HDPU). Of primarily speculation, what are the principal interactions present, and are they consistent with the studied *p*DPUs and/or the dioptic metal carbonyl complexes? Could the primary structural characteristics of the thermodynamically stable crystal be predicted from the knowledge gained through computational modeling of *p*DPU crystallization behaviors? Again, the understanding of how to control the crystallization propensity of compounds synthesized comprising the versatile *p*DPU moiety will guide our exploration into novel organometallic crystal systems with unique materials applications.

Crystallographic characterization of the asymmetric *p*Mo(CO)$_5$HDPU exhibited the *p*DPU moiety crystallized in a nearly planar configuration despite the sin-gle asymmetric substitution, analogous to the *p*NHDPU crystal structure (Fig-ure H.8). This occurrence is likely due to the enhanced electron-withdrawing character of the Mo(CO)$_5$-CN$^-$ substituent as compared with the asymmetrically substituted *p*CyHDPU, which exhibits the urea ribbon motif in the solid state. The bolstered electron-withdrawing nature of the metal carbonyl fragment provides the necessary intramolecular stabilization through *o*-H$\cdots$O$_{urea}$ hydrogen bonding to maintain the planar molecular orientation in its crystalline configuration. Also akin to *p*NHPDU was the co-crystallization of DMF solvent molecules at a 1:1 ratio. The DMF forms the stable H$_{urea}\cdots$O$_{DMF}$ hydrogen bonding complex in a likewise manner, thus filling otherwise void space created between the planar DPUs and bulky Mo(CO)$_5$ substituents, and adding additional stabilizing lattice energy.

Based on the careful observations of the studied *p*DPU systems, one may have predicted the observed nature of crystallization of the *p*Mo(CO)$_5$HDPU molecule. The similar balance of intramolecular and intermolecular forces driving crystal nucleation and growth in the DPU crystal systems is met through the intricate

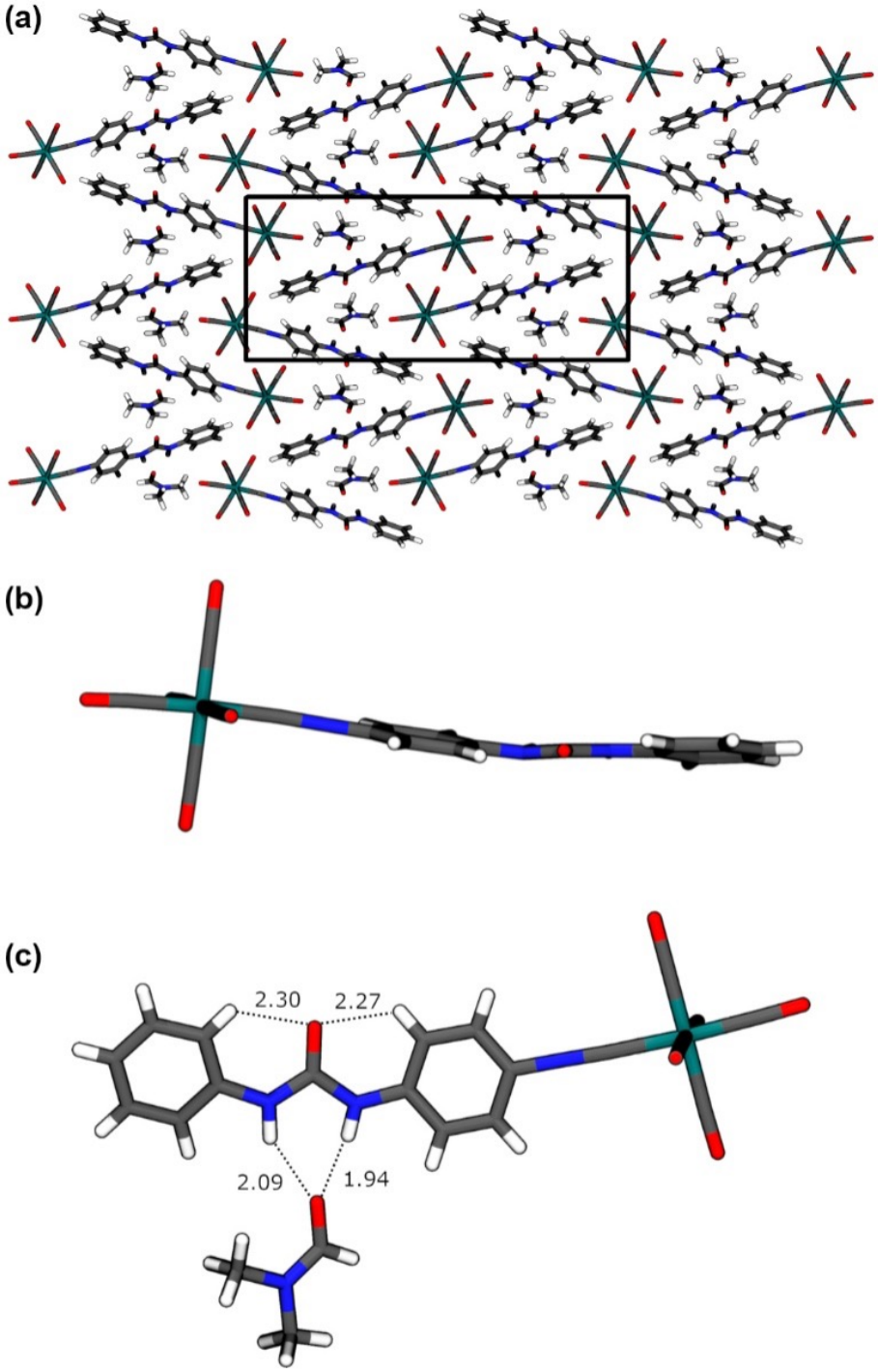


Figure H.8: (a) Molecular packing in the *p*Mo(CO)$_5$HDPU crystal system. (b) Planar configuration of the *p*DPU moiety as found in the *p*Mo(CO)$_5$HDPU crystal structure. (c) Hydrogen bonding distances of *o*-H$\cdots$O$_{urea}$ and H$_{urea}\cdots$O$_{DMF}$ contacts in the crystal structure.

interactions of the DPU moiety internally, with neighboring *p*Mo(CO)$_5$HDPU molecules, and with the abundantly available solvent molecules. The resulting

molecular conformations found in the isolated molecule and in the crystal structure are primarily consequences of the electron-withdrawing character of the DPU *p*-substituents. Solvent molecules may or may not be incorporated into the crystal system based on spatial constraints and the energies associated with disrupting a preferred planar molecular conformation.

Remarkably, the $Mo(CO)_5$ moieties in crystalline *p*$Mo(CO)_5$HDPU form an ordered array similar to that observed in the previously reported di-substituted 1,3-bis(*p*-isocyanophenyl)urea. The distances between metal centers of *p*$Mo(CO)_5$HDPU were 6.06, 6.85, and 8.18 Å, which is remarkable close to those of the di-substituted species of 6.01, 6.87, and 7.58 Å. This demonstrates the role in crystallization of the appended metal carbonyl fragments to the *p*DPU structure. While the DPU elements behave in a seemingly predictable manner, the $Mo(CO)_5$ groups may as well. The result is the concerted balance between the two prominent molecular constituents that guides the formation of the supramolecular structure. Although the current sample size is limited for this specific class of molecular species, the convoluted balance of energies steering the nucleation and crystallization of complex solid-state systems may too become more predictable as we continue to gather data on such systems as this field of study progresses.

We have established the role of *p*DPUs in facilitating the growth of organic and organometallic crystal systems, and we have also shed light on the engagement of metal carbonyl fragments in patterning crystal growth. This study relies on quantum chemical calculations, experimental crystallographic analysis, and the chemical synthesis of a new class of organometallic complexes to improve the understanding of the intricacies of complex intra- and intermolecular interactions driving the formation of tailorable crystal structures with the hopes that this knowledge can be leveraged for predictive capabilities in the design of further novel materials. Continued perseverance and comprehensive computational and ex-

perimental investigations into DPU and organometallic derivatives may provide the necessary predictive ability for the engineering of novel crystal systems with enhanced performance in processing, development, and applications as charge-transfer materials.

H.5 Conclusions

Understanding the balance of opposing packing forces and conformational strain energies is paramount for the engineering of novel organic and organometallic crystalline materials. In this study, we categorize the intricate features of energetic contributions to the supramolecular assembly of *p*-substituted *N*,*N*′-diphenylureas (*p*DPUs) and organometallic derivatives utilizing a holistic DFT approach. The results show a relationship between gas-phase and solid-state molecular conformations, and how the observed structural differences may or may not lead to deviations in expected crystallization behaviors based on the strengths of relative intra- and intermolecular interactions. Building on previous reports of DPU-based organometallic crystal systems, we also present the synthesis of an asymmetric $Mo(CO)_5$-substituted *N*-(*p*-isocyanophenyl)-*N*′-phenylurea in efforts to further relate the crystallization behaviors and interactions of organometallic DPU derivatives with those of the base *p*DPU systems. It was observed that the supramolecular assembly of the organometallic complexes display similar predictive crystallization patterns, as well as additional complexities evolving due to the bulky metal carbonyl substituents. We present a comprehensive computational classification of *p*DPU-based crystal systems in complement to experimental data in our endeavors to develop reliable design rules for patterned crystal growth of DPU and related systems for the future generation of new materials having unique solid-state structures and properties.

H.6 Acknowledgements

We would like to acknowledge high-performance computing support of the R2 compute cluster (DOI: 10.18122/B2S4^{1}H) provided by Boise State University's Research Computing Department.

H.7 Supporting Information

X-ray data collection and refinement parameters for *p*Mo(CO)$_5$HDPU. (Appendix I)

H.8 Accession Codes

CCDC 1554302, 676835, 1554299, 1554301, 676839, 1554300, 1905536, and 2035534 contain the supplementary crystallographic data for this paper. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre at www.ccdc.cam.ac.uk/data_request/cif or by emailing data_request@ccdc.cam.ac.uk.

APPENDIX I:

SUPPORTING INFORMATION FOR CATALOGUING THE ENERGETIC CONTRIBUTIONS TO THE SUPRAMOLECULAR ASSEMBLY OF *P*-SUBSTITUTED *N,N'*-DIPHENYLUREAS AND THEIR ORGANOMETALLIC DERIVATIVES IN THE SOLID STATE: A DENSITY FUNCTIONAL THEORY APPROACH

I.1 Crystal Structure Determination of $Mo(CO)_5$(*p*CNHDPU)

X-ray diffraction data were collected at 100 K on a Bruker D8 Venture using MoKα-radiation (λ = 0.71073 Å). Data have been corrected for absorption using SADABS [342] area detector absorption correction program. Using Olex2 [343], the structure was solved with the SHELXT [345] structure solution program using Direct Methods and refined with the SHELXL [344] refinement package using least squares minimization. All non-hydrogen atoms were refined with anisotropic thermal parameters. Hydrogen atoms attached to heteroatoms were found from the residual density maps, placed, and refined with isotropic thermal parameters. All other hydrogen atoms in the investigated structure were located from difference Fourier maps but finally their positions were placed in geometrically calculated positions, and refined using a riding model. Isotropic thermal parameters of the placed hydrogen atoms were fixed to 1.2 times the *U* value of the atoms they are linked to (1.5 times for methyl groups). Calculations and refinement of structures were carried out using APEX2 [347], SHELXTL [346], and Olex2 software. Graphical representations of crystallographic data were generated using the Mercury software package [348].