David R Bell, PhD

Research Statement

Research Interests: The motivation for my research is the discovery and characterization of novel materials for human advancement. To achieve this goal, I use various computational techniques to model, interrogate, and predict material properties. These techniques include multiscale modeling and classical simulation techniques such as Molecular Dynamics (MD) simulations which I used in my dissertation work to study nucleic acid structure. During my postdoc at IBM, I was introduced to a wide portfolio of AI methods, which I now routinely incorporate into my materials modeling projects. Although I mostly use machine learning (ML) techniques for supervised tasks such as parameter optimization and enhancing computational efficiency, I am excited to explore using ML methods for generative and discovery purposes: to discover novel structures and material combinations.

Research Background: My research work focuses on modeling native and synthetic materials, from proteins and nucleic acids to nanotubes and quantum dots. There exists a copious amount of questions for these systems with implications to related fields such as polymer chemistry and solid-state physics. My emphasis has been to explore the interdependence of structure, dynamics, and thermodynamics present in biological and nanoscale materials. These molecules, typically interacting only through relatively weak electrostatic charge distributions and induced dipoles, can achieve robust binding interactions.

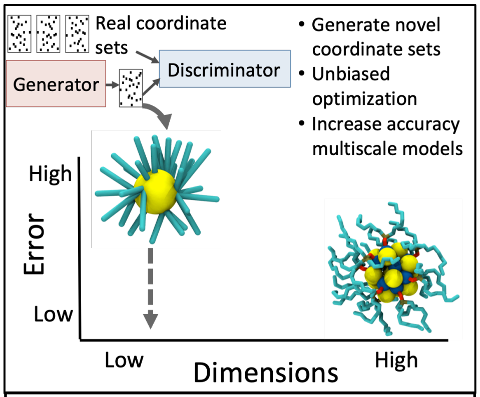
Immune complexes, in particular, I find to be vastly complex, as typified by catch bonds[1] and selection against minimum energy configurations[2]. I have studied at length the immune complex of MHC-II-TCR, which is responsible for activation of immune T-cells for a variety of different effects, including autoimmune disease. This interaction is mediated by a small peptide, which strongly binds the MHC, but binds the TCR through much weaker interactions. I have predicted MHC binding affinity of a small peptide implicated in Type-1 Diabetes using MD simulations and the Free Energy Perturbation (FEP) technique[3]. During this project, I generated a large dataset of MD simulations and FEP data that I have subsequently explored using classification techniques and an autoencoder-ML pipeline (under review). Predicting peptide-MHC binding affinity is difficult due to strong attractive dispersion forces present in the binding site. I have used ML techniques to reduce the computational cost of this prediction as well as generate predictions of novel small peptides that could be used as immunotherapy vaccines.

Nucleic acid structure, dynamics, and interactions also prove a formidable challenge to model, primarily because nucleic acids are highly charged polymers. I have built a multiscale model to predict nucleic acid structure using statistical potentials and experimental free energy information[4], with mixed results. In a more successful attempt, I predicted nucleic acid structure using bioinformatics constraints and focused on Mg^{2+} ion binding, which screens the charged backbone of the nucleic acid[5]. Under this approach, I built a 3-D RNA structure that I used to predict RNA-protein binding affinity with verification by calorimetry experiments.

Synthetic nanomaterials, in contrast, tend to exhibit less dynamic behavior than biological systems, yet hold more direct application. I studied CdSe quantum dot aggregation and interaction with a widespread protein domain[6] to understand quantum dot toxicity. I found interesting concentration-dependent quantum dot binding behavior that blocked the protein active site at high concentrations, indicating a potential mechanism of toxicity. I also studied methane and hydrogen gas adsorption onto MoS_2 nanotubes for prospective gas storage. Energy gases are difficult to store efficiently, so various nanomaterial substrates are being tested for storage capacity. Our work determined hydrogen and methane adsorption and release profiles onto MoS_2 and found favorable but sub-ideal storage efficiencies of MoS_2 substrates.

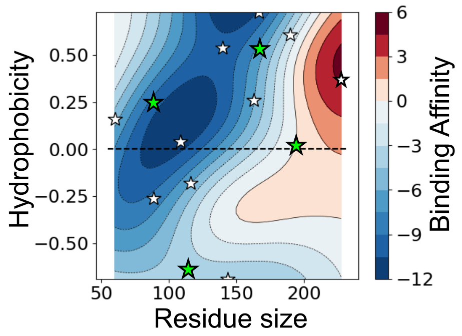
Assistant Professor: As an AP, I aim to explore and characterize novel materials and biomolecules using ML and multiscale modeling techniques. Towards this goal, I propose three starting projects: (1) Use Generative Adversarial Networks (GANs) for generative multiscale materials modeling, (2) use Gaussian Processes to predict novel material interactions, and (3) use the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) financial AI model to accelerate the prediction of material properties.

GANs for multiscale materials modeling: Multiscale chemical models are plagued by high error and/or high bias with no straightforward way to quantify their inaccuracies. Multiscale models are designed to study long time- and length-scale molecular phenomena[7-9] such as polymer aggregation[10], protein folding and amyloid formation[11], and semiconductor doping[12]. However, many models suffer from either high bias when parameters are manually tuned to achieve the desired result[13], or models suffer from high error when parameter optimization methods select inaccurate coordinate sets with sparsely sampled interaction terms[4]. I will address these problems[14-15] and increase model accuracy by using GANs to generate accurate coordinate sets that are parameterized by unbiased optimization ML methods such as regression. I will use this pipeline to generate multiscale models at varying degrees of complexity and quantify multiscale model error at each complexity by comparing to the original high-dimensional representation.



Project 1. Use GANs to generate novel multiscale model coordinate sets that are optimized by ML methods and tested to quantify multiscale accuracy.

Gaussian Processes for material interaction: Determining mutant binding affinities for applications such as immunotherapy or enzyme design is resource intensive, both computationally and experimentally. To accelerate binding affinity calculation, I will use Gaussian Processes to learn and interpolate Free Energy landscapes from a small subset of affinity values. Specifically, I will probe amino acid contributions to protein-protein binding. Based on a small subset of amino acid affinity data, I will interpolate binding affinity to the whole amino acid set, interpolating over the features size and hydrophobicity. I will determine how many amino acids are necessary to accurately interpolate binding affinity and search for a consistent amino acid mutation set that can be used for general protein design research.



- Sample residue subset (green)
- Interpolate binding affinity to remaining residues (white) using Gaussian Process

Project 2. Accelerate mutant protein design using Gaussian Process to interpolate residue binding affinity data

GARCH finance ML model for trajectory analysis: I will use GARCH, a time series ML model used in finance, to predict long-time values for several materials modeling systems. System 1) I will use DFTB+ to model CdSe quantum dot formation under experimental conditions bound to aliphatic solvents. Based on this data, I will use GARCH to predict quantum dot formation and growth to larger quantum dot sizes. I will compare growth rates and mechanisms under several conditions[16-17] to discover novel catalytic mechanisms. System 2) I will use GARCH to predict converged free energy values from short-time calculations. Currently, accurate binding free energy calculations are very computationally expensive[3] and difficult to converge. I will explore using ML methods to predict converged affinity values from short generated trajectories for efficient free energy prediction.

References

1. Pullen, R. H.; Abel, S. M., Catch Bonds at T Cell Interfaces: Impact of Surface Reorganization and Membrane Fluctuations. *Biophys. J.* **2017,** *113* (1), 120-131.
2. Levisetti, M. G.; Suri, A.; Petzold, S. J.; Unanue, E. R., The Insulin-Specific T Cells of Nonobese Diabetic Mice Recognize a Weak MHC-Binding Segment in More Than One Form. *The Journal of Immunology* **2007,** *178* (10), 6051-6057.
3. Ahmed, R.; Omidian, Z.; Giwa, A.; Cornwell, B.; Majety, N.; Bell, D. R.; Lee, S.; Zhang, H.; Michels, A.; Desiderio, S.; Sadegh-Nasseri, S.; Rabb, H.; Gritsch, S.; Suva, M. L.; Cahan, P.; Zhou, R.; Jie, C.; Donner, T.; Hamad, A. R. A., A Public BCR Present in a Unique Dual-Receptor-Expressing Lymphocyte from Type 1 Diabetes Patients Encodes a Potent T Cell Autoantigen. *Cell* **2019,** *177* (6), 1583-1599.e16.
4. Bell, D. R.; Cheng, S. Y.; Salazar, H.; Ren, P., Capturing RNA Folding Free Energy with Coarse-Grained Molecular Dynamics Simulations. *Sci Rep* **2017,** *7*, 45812.
5. Bell, D. R.; Weber, J. K.; Yin, W.; Huynh, T.; Duan, W.; Zhou, R., In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers. *Proceedings of the National Academy of Sciences* **2020,** *117* (15), 8486-8493.
6. Bell, D. R.; Kang, S.-G.; Huynh, T.; Zhou, R., Concentration-dependent binding of CdSe quantum dots on the SH3 domain. *Nanoscale* **2018,** *10* (1), 351-358.
7. Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M., Multiscale modeling of emergent materials: biological and soft matter. *Physical Chemistry Chemical Physics* **2009,** *11* (12), 1869-1892.
8. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C., The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics* **2008,** *128* (24), 244114.
9. Schlick, T.; Collepardo-Guevara, R.; Halvorsen, L. A.; Jung, S.; Xiao, X., Biomolecularmodeling and simulation: a field coming of age. *Q. Rev. Biophys.* **2011,** *44* (2), 191-228.
10. Shahidi, N.; Chazirakis, A.; Harmandaris, V.; Doxastakis, M., Coarse-graining of polyisoprene melts using inverse Monte Carlo and local density potentials. *The Journal of Chemical Physics* **2020,** *152* (12), 124902.
11. Sharp, M. E.; Vázquez, F. X.; Wagner, J. W.; Dannenhoffer-Lafage, T.; Voth, G. A., Multiconfigurational Coarse-Grained Molecular Dynamics. *Journal of Chemical Theory and Computation* **2019,** *15* (5), 3306-3315.
12. Zographos, N.; Zechner, C.; Martin-Bragado, I.; Lee, K.; Oh, Y.-S., Multiscale modeling of doping processes in advanced semiconductor devices. *Materials Science in Semiconductor Processing* **2017,** *62*, 49-61.
13. Chen, M.; Wolynes, P. G., Aggregation landscapes of Huntingtin exon 1 protein fragments and the critical repeat length for the onset of Huntington's disease. *Proceedings of the National Academy of Sciences* **2017,** *114* (17), 4406-4411.
14. Foiles, S.; McDowell, D. L.; Strachan, A., Preface for focus issue on uncertainty quantification in materials modeling. *Modelling and Simulation in Materials Science and Engineering* **2019,** *27* (8), 080301.
15. van der Giessen, E.; Schultz, P. A.; Bertin, N.; Bulatov, V. V.; Cai, W.; Csányi, G.; Foiles, S. M.; Geers, M. G. D.; González, C.; Hütter, M.; Kim, W. K.; Kochmann, D. M.; Llorca, J.;

Mattsson, A. E.; Rottler, J.; Shluger, A.; Sills, R. B.; Steinbach, I.; Strachan, A.; Tadmor, E. B., Roadmap on multiscale materials modeling. *Modelling and Simulation in Materials Science and Engineering* **2020,** *28* (4), 043001.
16. Ohta, Y.; Okamoto, Y.; Irle, S.; Morokuma, K., Rapid Growth of a Single-Walled Carbon Nanotube on an Iron Cluster: Density-Functional Tight-Binding Molecular Dynamics Simulations. *ACS Nano* **2008,** *2* (7), 1437-1444.
17. Page, A. J.; Ohta, Y.; Irle, S.; Morokuma, K., Mechanisms of Single-Walled Carbon Nanotube Nucleation, Growth, and Healing Determined Using QM/MD Methods. *Accounts Chem. Res.* **2010,** *43* (10), 1375-1385.