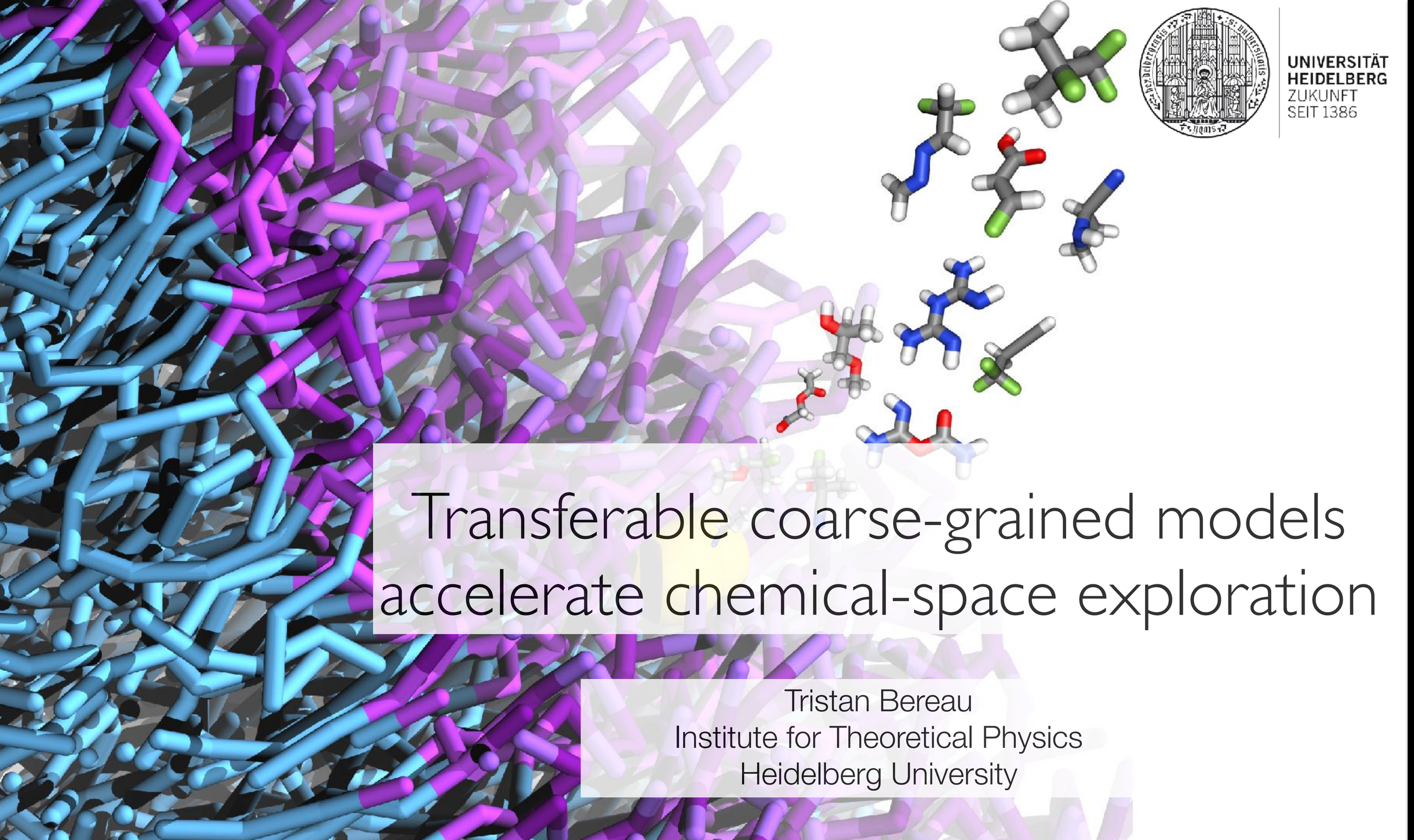


UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

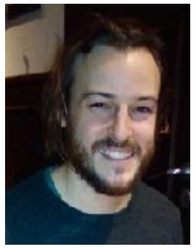
Transferable coarse-grained models accelerate chemical-space exploration

Tristan Bereau
Institute for Theoretical Physics
Heidelberg University

Acknowledgments

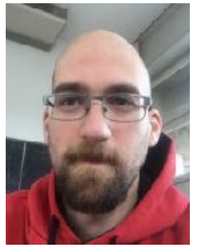
Past/present group members

Atreyee Benerjee, Marius Bause, Yasemin Bozkurt Varolgunes, Alessia Centi, Arghya Dutta, Martin Girard, Christian Hoffmann, Kiran Kanekal, Chan Liu, Roberto Menichetti, Bernadette Mohr, Clemens Rauer, Joseph Rudzinski, Marc Stieffenhofer, Svenja Wörner


Marc Stieffenhofer


Joseph Rudzinski


Kiran Kanekal


Roberto Menichetti

Max Planck Inst. Polymer Res.

Denis Andrienko, Kurt Kremer, Christoph Scherer, Tanja Weil, Christopher Synatschke


Bernadette Mohr


Kirill Schmilovich


Andrew Ferguson

University of Chicago


Isabel Kleinwächter

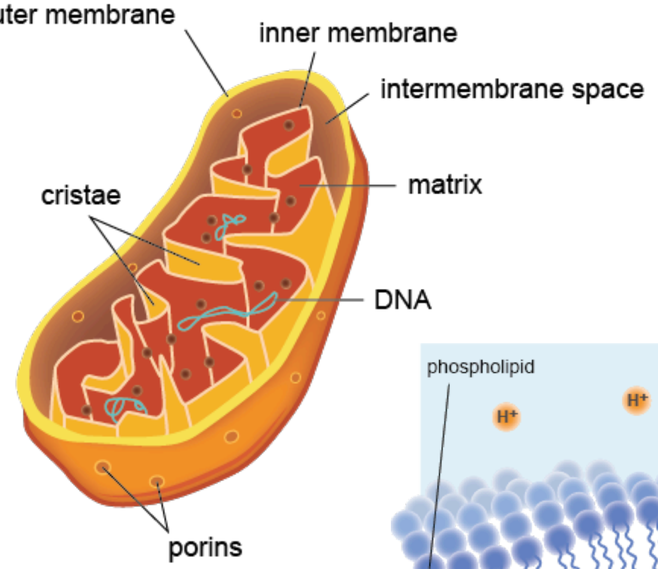

Dirk Schneider

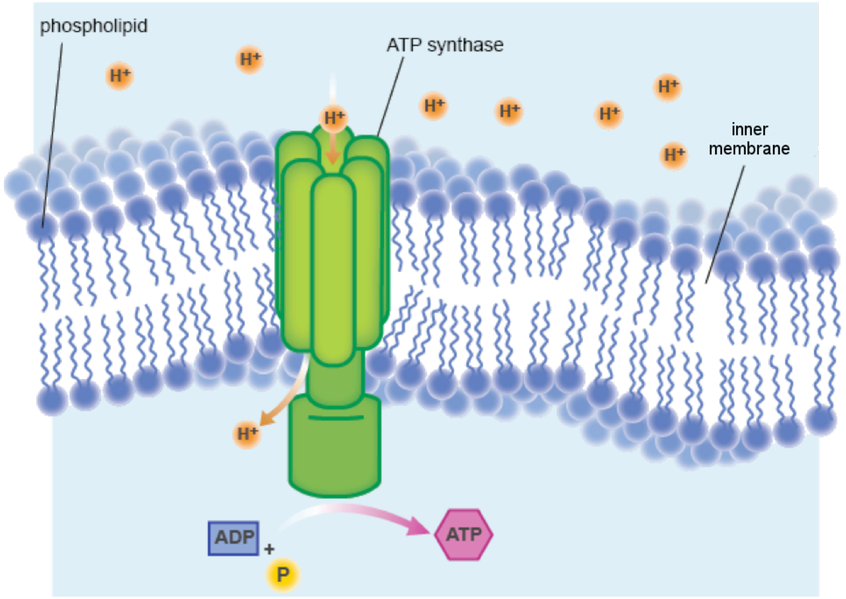
University of Mainz

Collaborators

Robert DiStasio Jr. (Cornell), Anatole von Lilienfeld (Basel), Andrew Ferguson (U Chicago), Sapun Parekh (UT Austin), Dirk Schneider (JGU Mainz), Alexandre Tkatchenko (Luxembourg), Michael Wand (JGU Mainz), Jilles Vreeken (Saarland), Luca Ghiringhelli (FHI Berlin)

Mitochondria



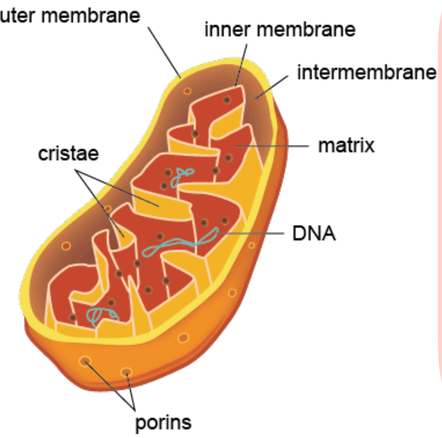


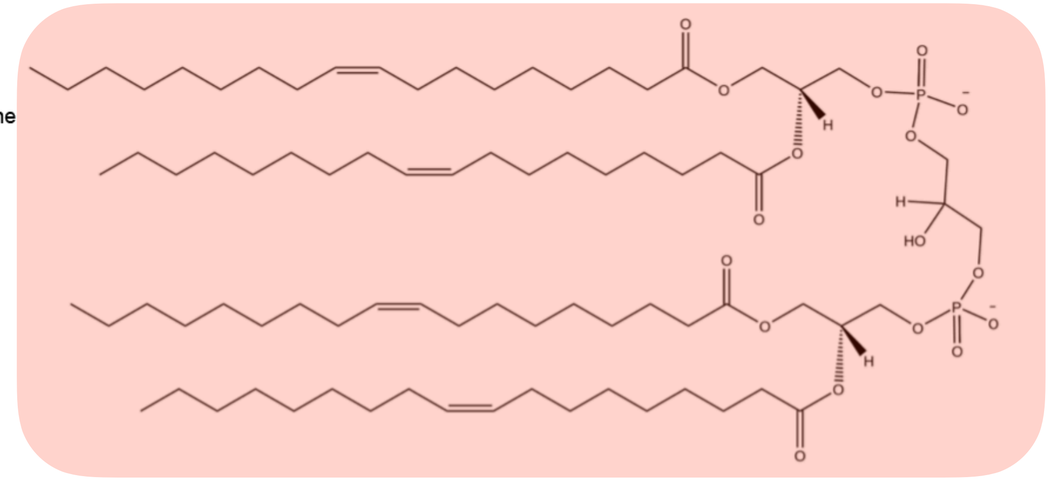
- Produce approx. 64 kg of ATP per day
- Also involved in biogenesis and metabolic cycles

J. Dudek, *Frontiers in Cell and Developmental Biology* **5** (2017)

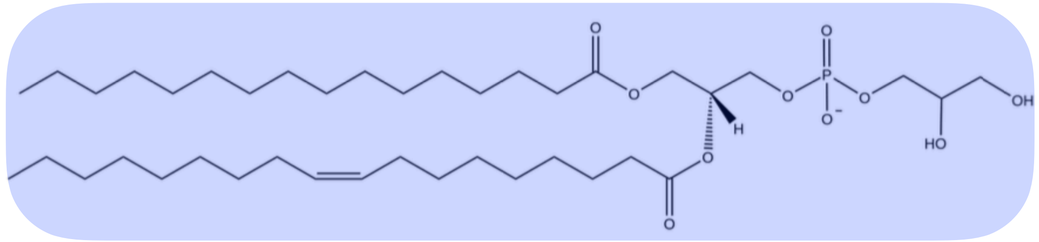
Cardiolipin-linked pathologies

- Inner membrane contains ~20% of cardiolipin (CL)
- Abnormalities in CL composition of IM are linked to Barth syndrome, Tangier disease, heart failure, and neurodegeneration

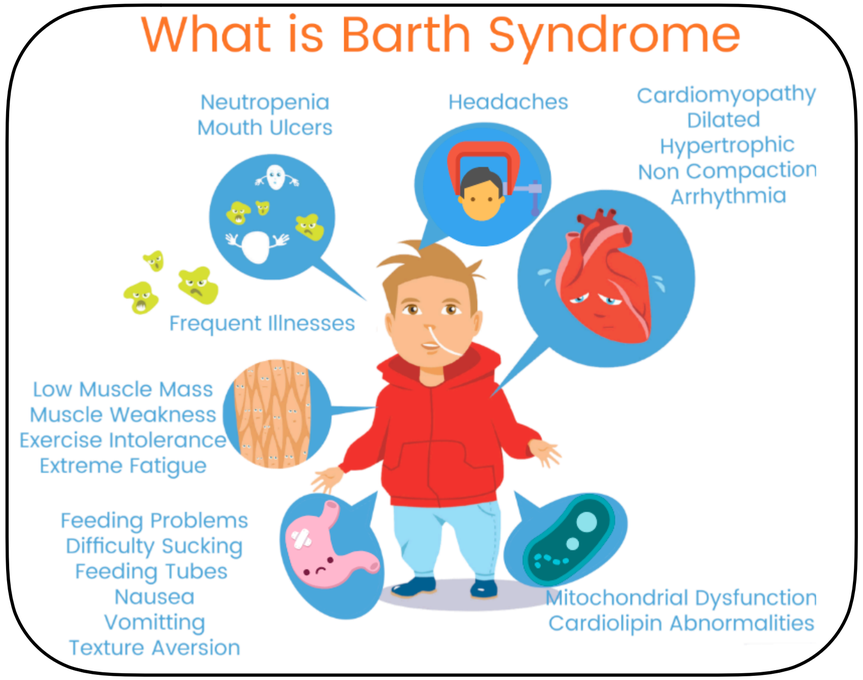




cardiolipin (CL)



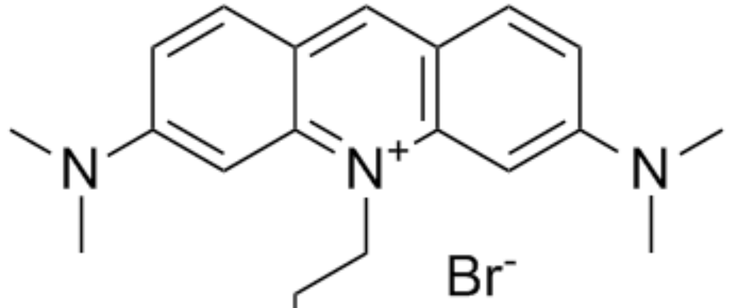
palmitoyloleoyl phosphatidylglycerol (POPG)

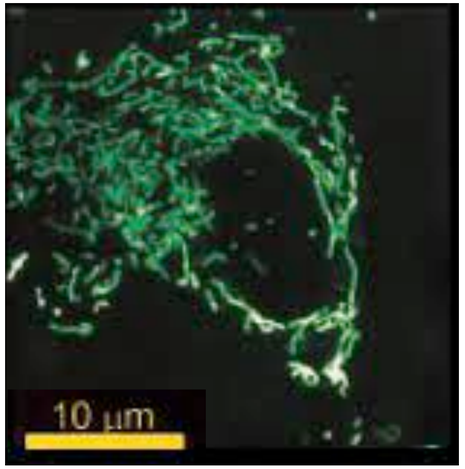


Paradies, Paradies, Ruggiero, Petrosillo, *Cells* **8** (2019)

NAO fluorescent stain

- 10-*N*-nonyl acridine orange (NAO) is a high-affinity probe
- Used as a stain for fluorescent visualization and quantification of IM CL
- Physicochemical basis for specificity is unknown





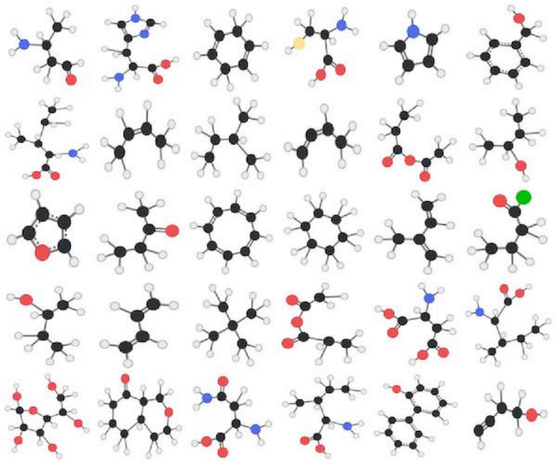
Can we:

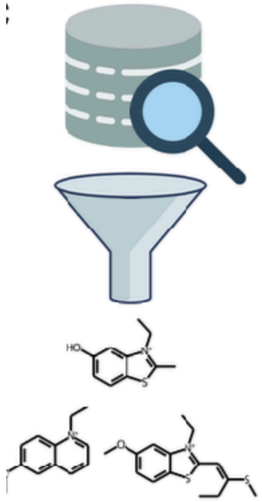
1. **Discover** small organic CL dyes with selectivity superior to NAO?
2. Extract molecular design rules for **mechanism of action**?

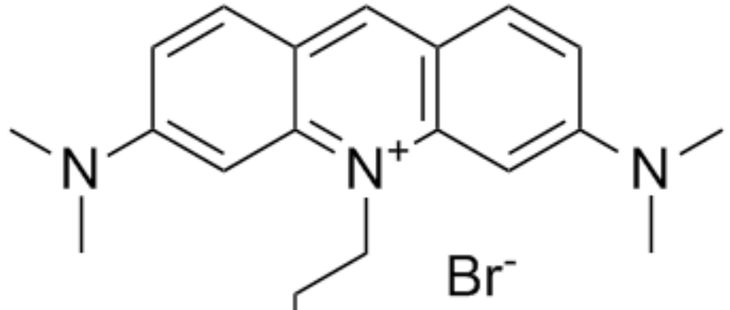
Dirk Schneider
(JGU Mainz)

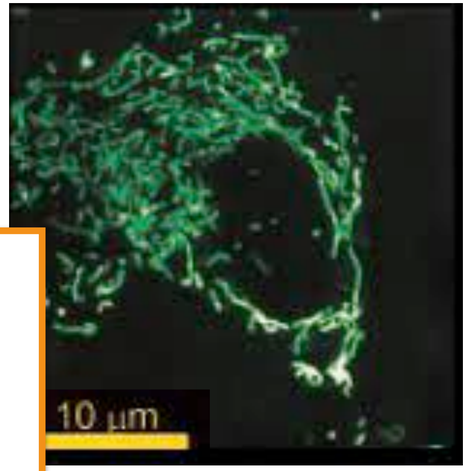
Fernandez, Ceccarelli, Muscatello, *Analytical Biochemistry* **328** (2004); Rodriguez *et al.*, *Mitochondrion* **8** (2008); Horvath, Daum, *Progress in Lipid Research* **52** (2013); Zielonka *et al.*, *Chemical Reviews* **117** (2017); Jacobson *et al.*, *Journal of Neurochemistry* **82** (2002)

Approaches









1. Fit experimental data to machine-learning model

   Little to no c

2. Run all-atom m

   Unbearably

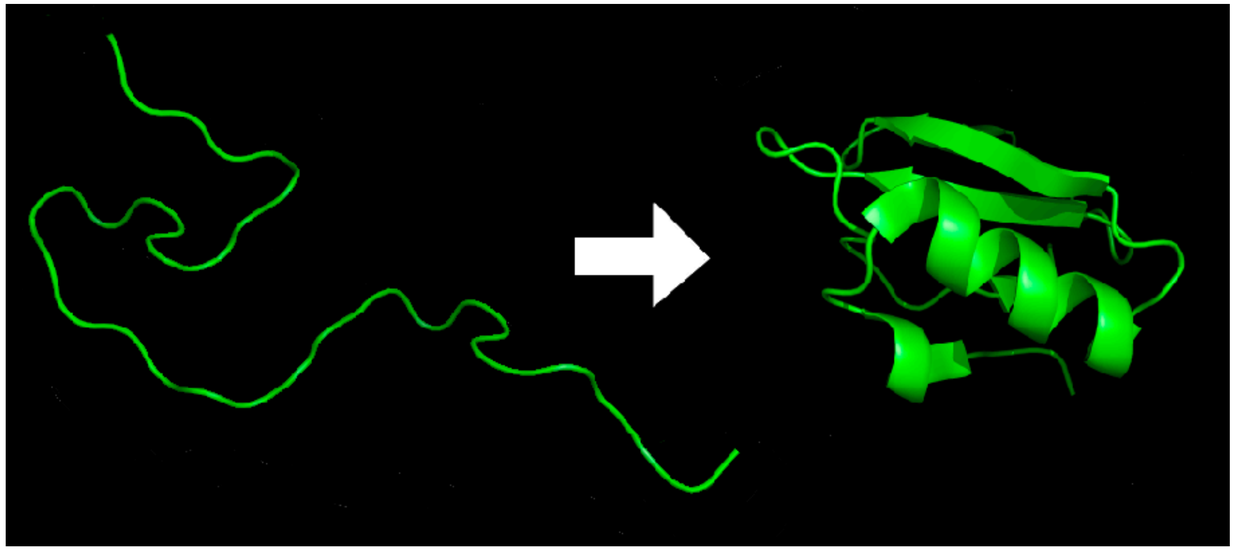
3. Run coarse-grained molecular dynamics simulations

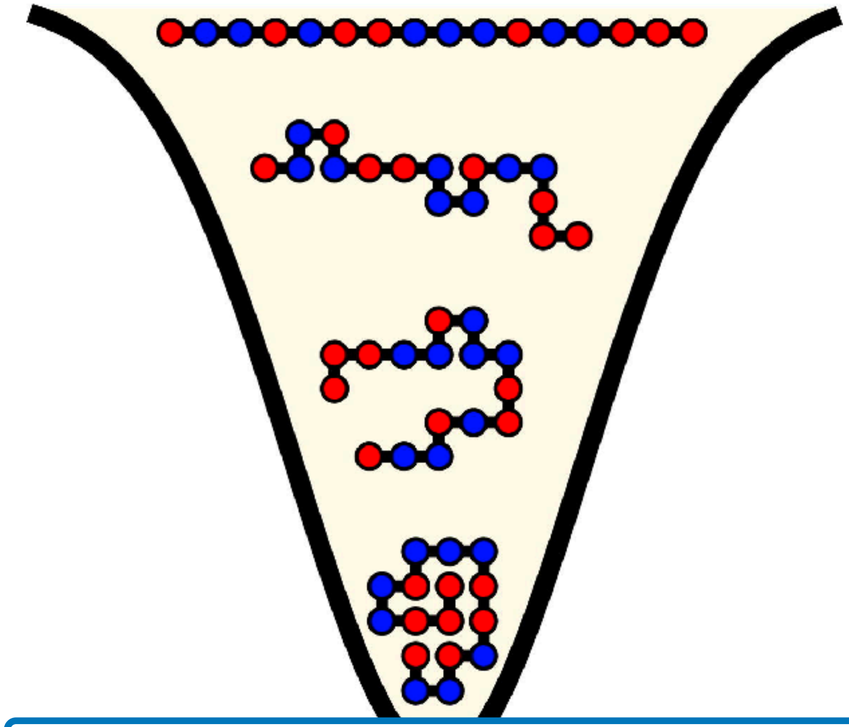
   Precision vs model simplicity?

Challenges

1. Coarse-graining too simplistic?
2. Generate ML training data from computer simulations

HP protein models





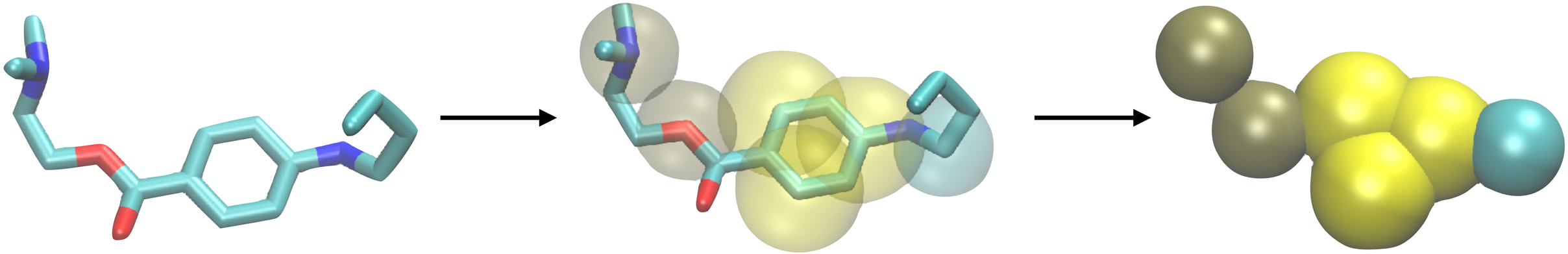
- Recast protein-folding problem as a lattice polymer model
- Discovery of funnel-shaped energy landscape theory
- Map 20 amino acids into binary code: **H**ydrophobic and **P**olar beads

Top-down model with bead types: Martini does something similar (with more chemical fidelity)

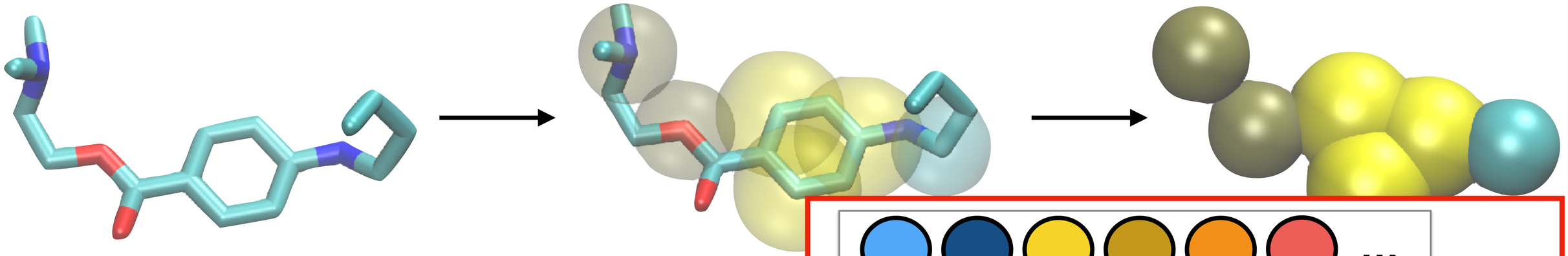
21)

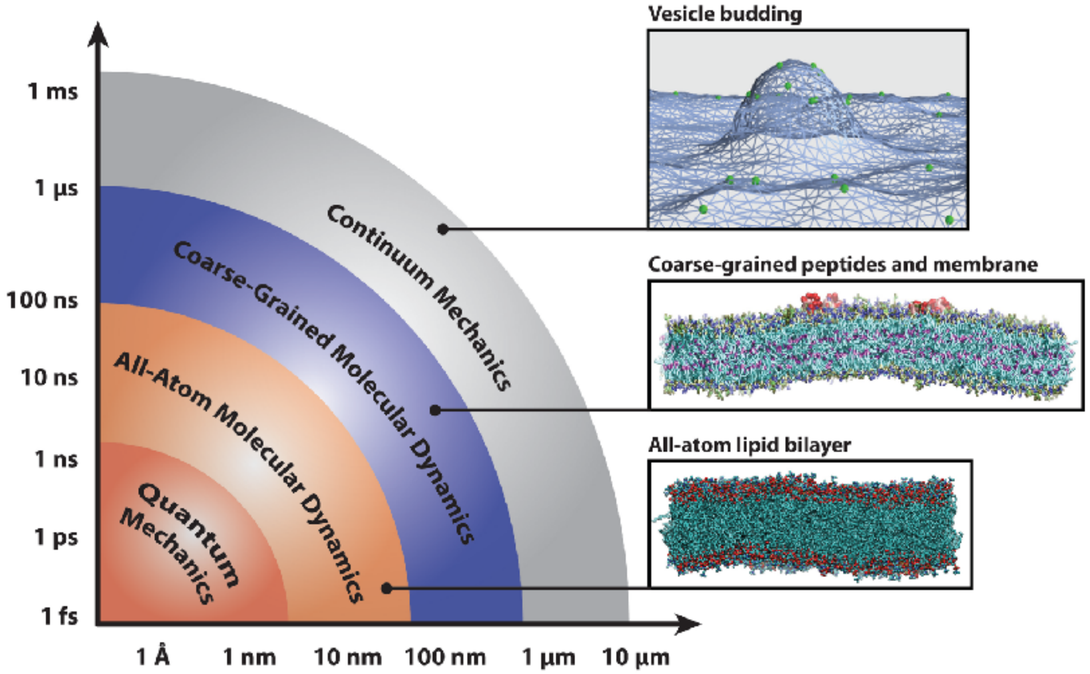
See work from Dill, MacCallum, Wolynes, Chan, Shakhnovich, and many others

Coarse-graining molecules



Short primer on coarse-graining





Bradley and Radhakrishnan, *Polymers* **5** (2013)

Bead types: *chemical fragments*

- **Top-down**: from phenomenological information / large-scale physics

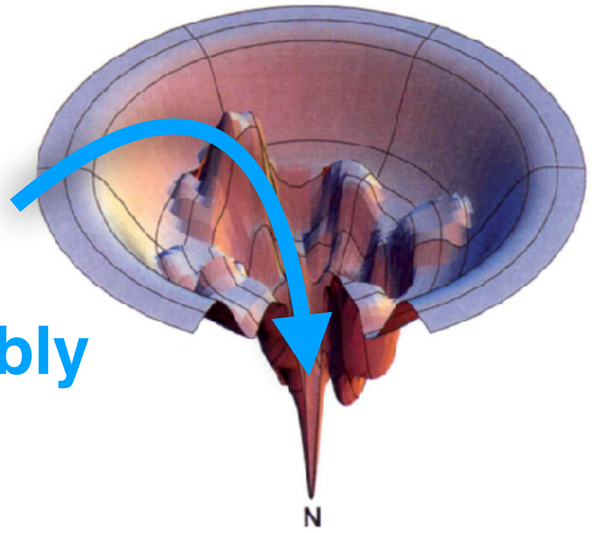
- **Bottom-up**: from microscopic information (e.g., atomistic simulations)

Noid, *J Chem Phys* **139** (2013)

Compounding challenges in chemical-space exploration

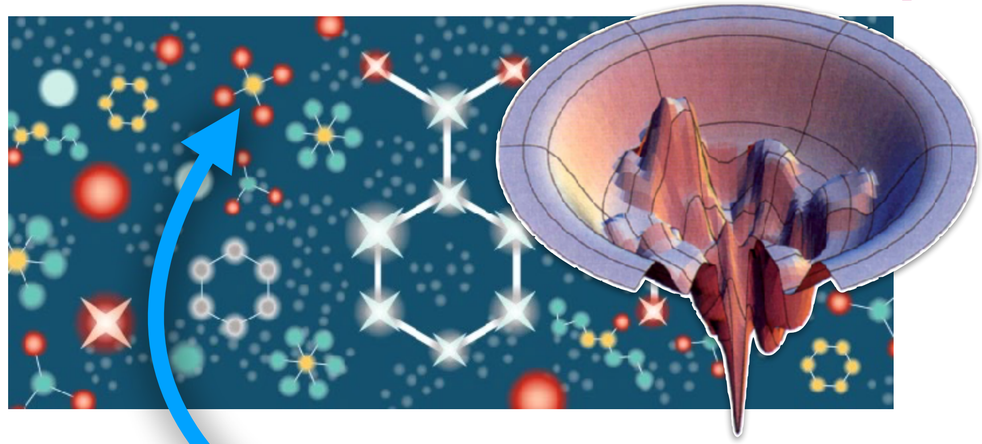
"Compositional landscape"

Free-energy landscape



Voth, *CRC Press* (2008); Noid, *J Chem Phys* **139** (2013)

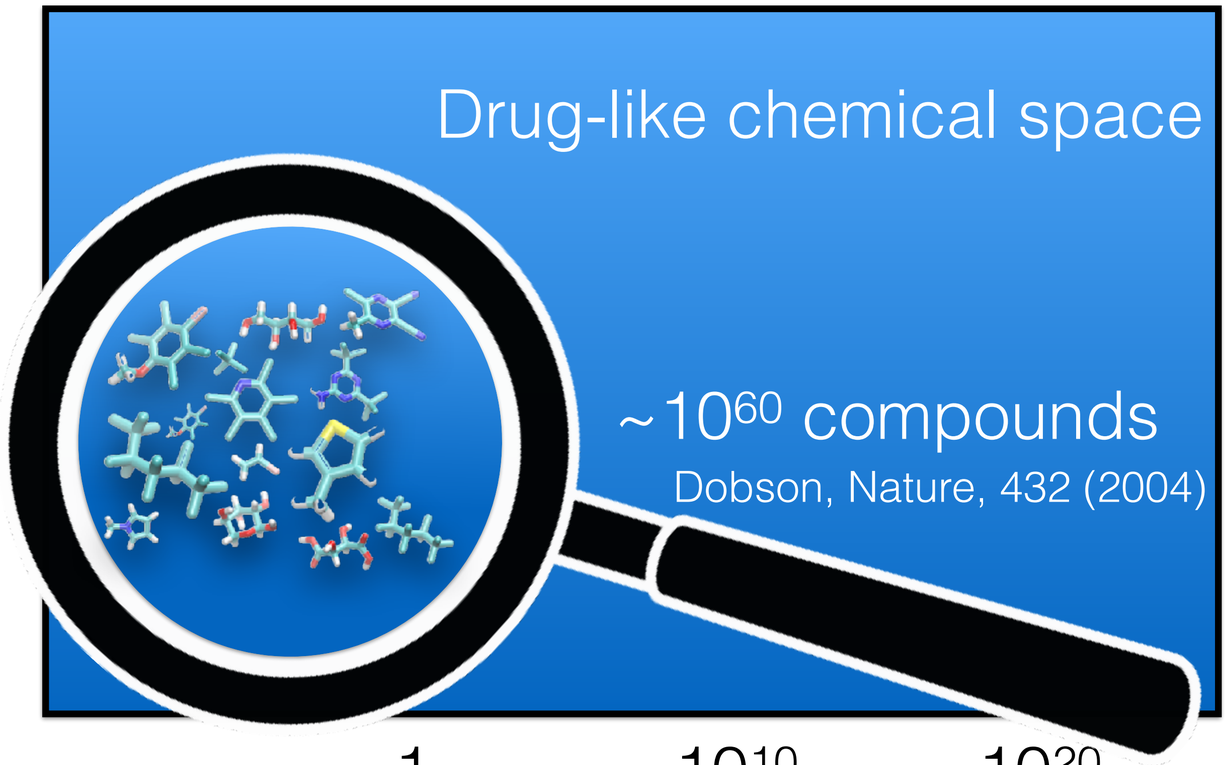
Chemical compound space



Von Lilienfeld *et al.*, *Nat Rev Chem* **4** (2020)

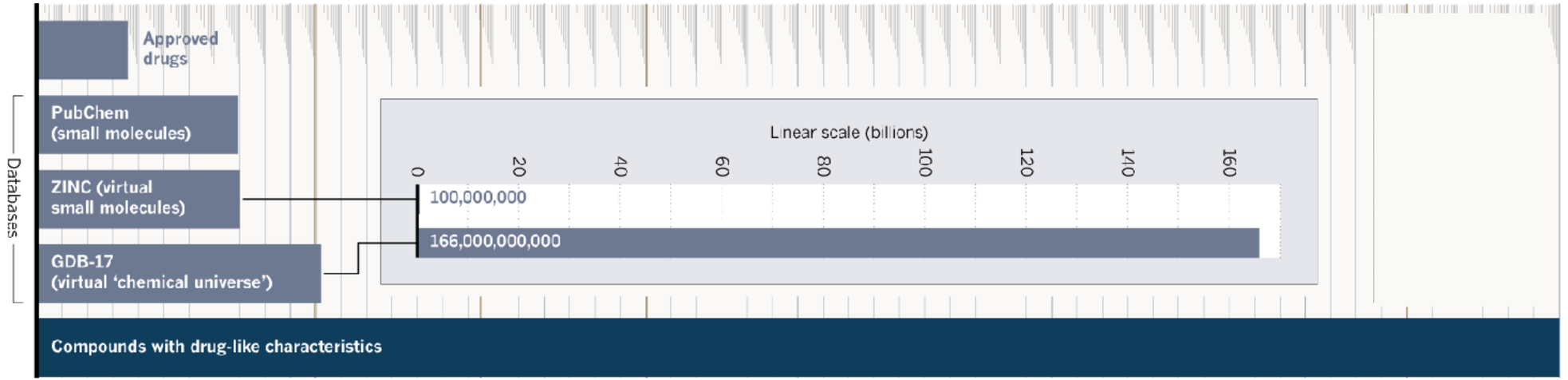
In both cases: Importance-sampling problems!

Chemical space is large



number of carbon atoms in the universe

1 10^{10} 10^{20} 10^{30} 10^{40} 10^{50} 10^{60}

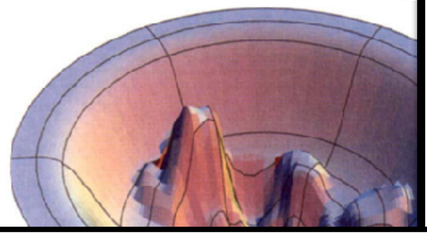


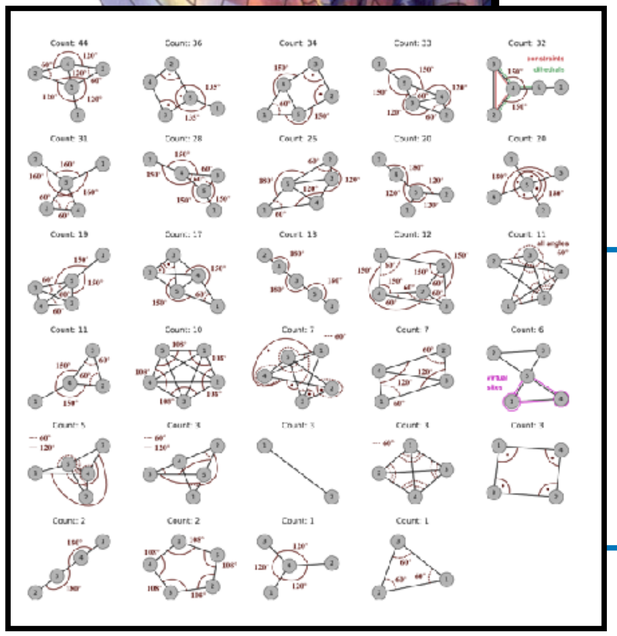
Mullard, Nature, 549 (2017)

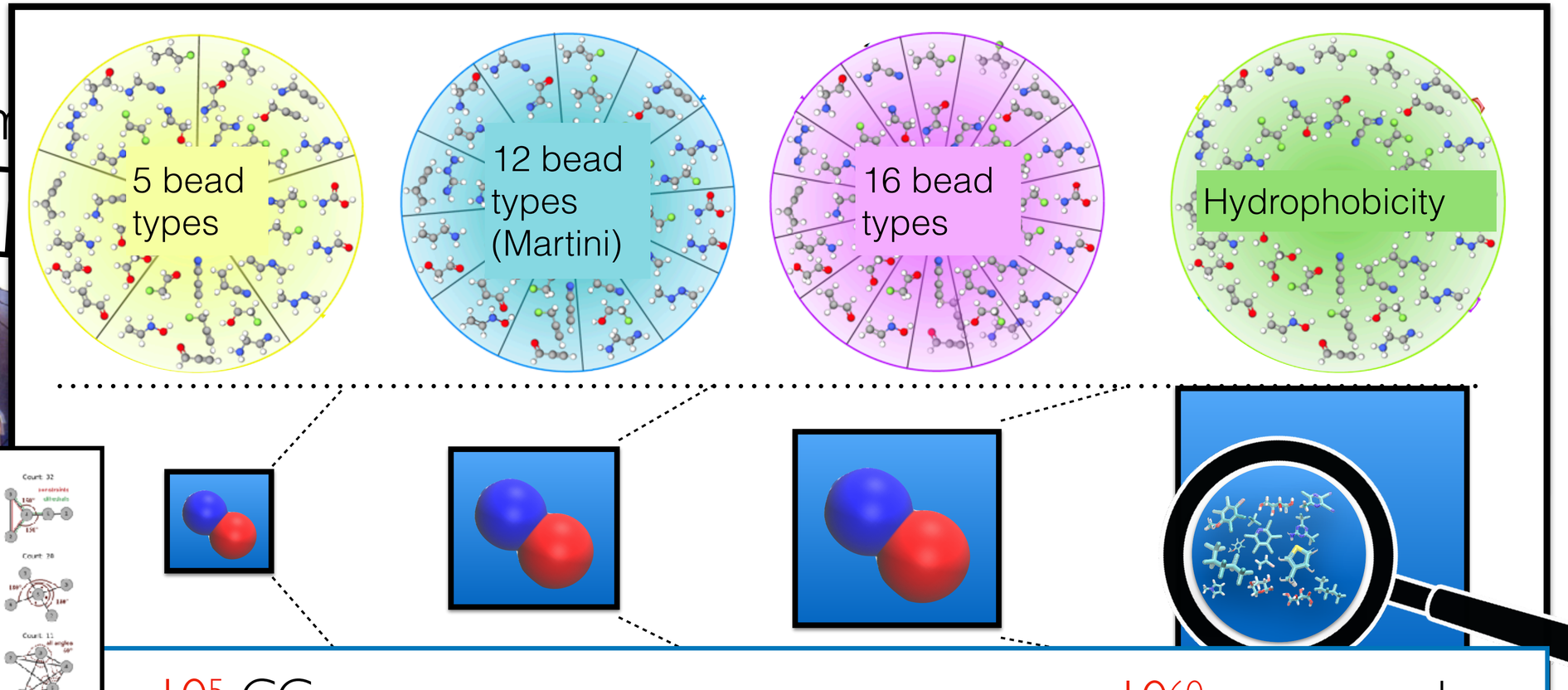
Sampling efficiency of coarse-grained models

Conform

All CG models







~10^5 CG molecules

~10^{60} compounds

Dobson, Nature, 432 (2004)

Drug-like small-molecule chemical space

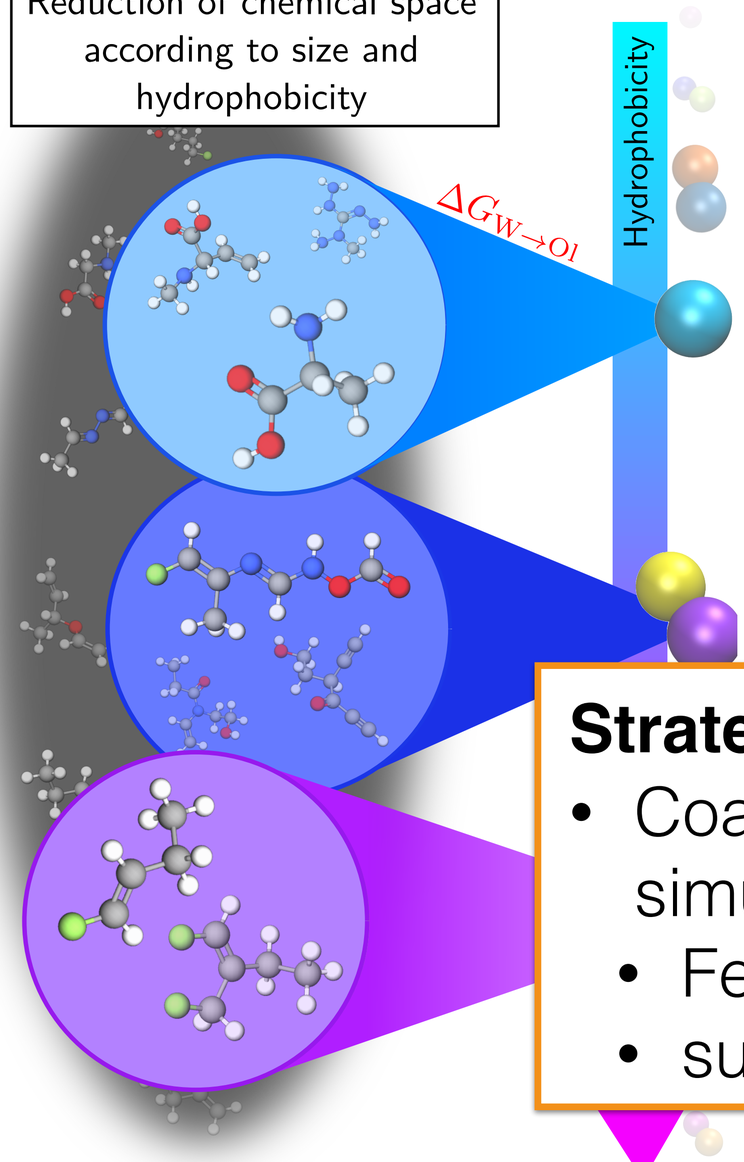
due to finite set of bead types

Work from Voth, Noid, Shell, Kremer, van der Vegt, and others

Kanekal, Bereau, *J Chem Phys* **151** (2019)

High-throughput coarse-graining scheme

Reduction of chemical space according to size and hydrophobicity



Strategy

- Coarse-graining prior to high-throughput simulations
  - Fewer simulations
  - suggests low-dimensional representation

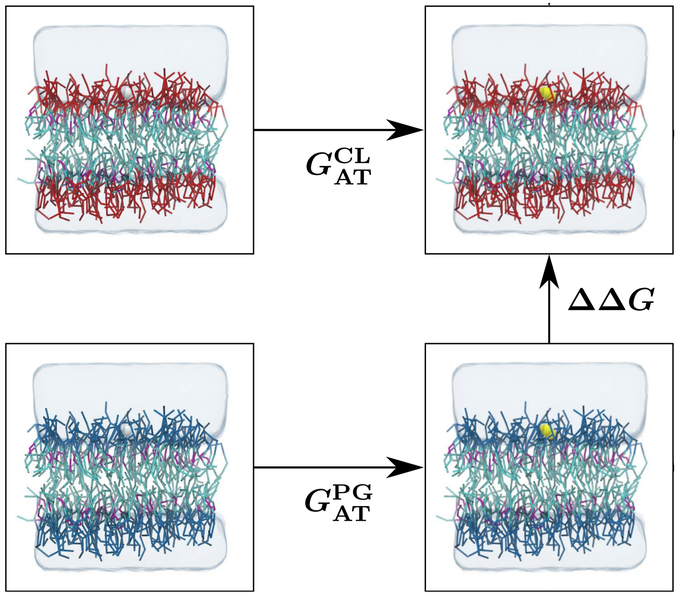
Menichetti, Kanekal, Bereau, *ACS Cent. Sci.* **5** (2019)

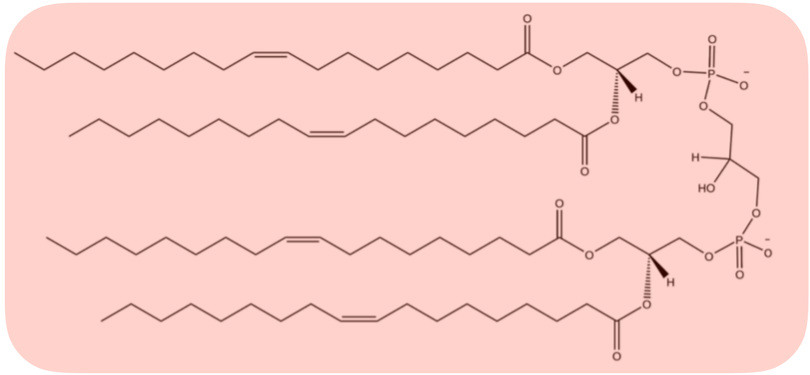
Kiran Kanekal

Roberto Menichetti

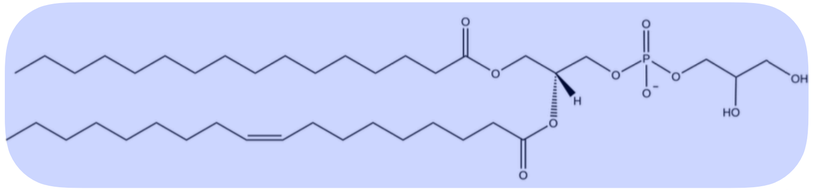
CGMD selectivity measurements

- Compute transfer free energies into CL and POPG membranes using alchemical transformations
- Objective function to maximize CL vs POPG selectivity:





cardiolipin (CL)



palmitoyloleoyl phosphatidylglycerol (POPG)

$$\Delta\Delta G_{\mathrm{POPG}\rightarrow\mathrm{CL}} = \Delta G_{\mathrm{CL}} - \Delta G_{\mathrm{POPG}}$$

Maximize the gap

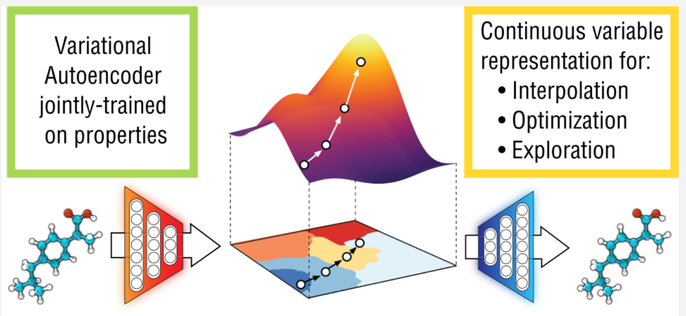
Thermodynamic preference for desired CL membrane

Thermodynamic preference for most chemically similar competitor

High-throughput virtual screening

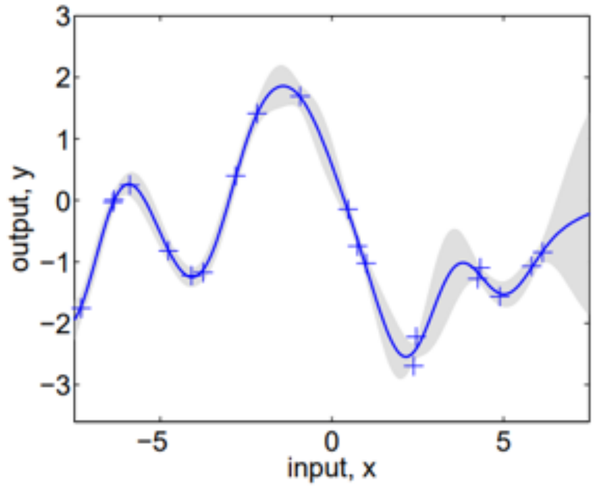
Learn featurization

Unsupervised deep representation learning of chemical space using regularized autoencoders



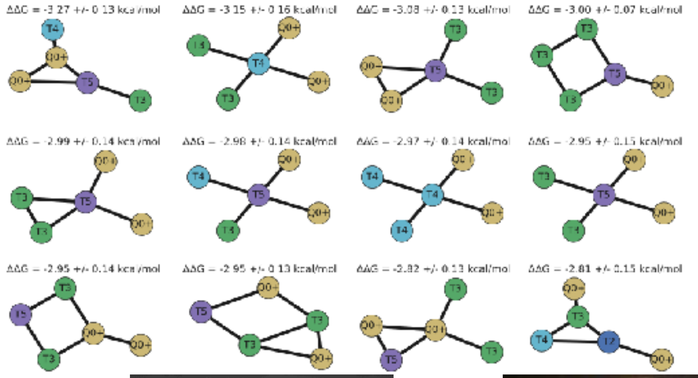
Estimate fitness

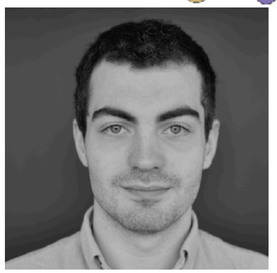
Supervised learning of sequence-selectivity relationship using Gaussian process regression



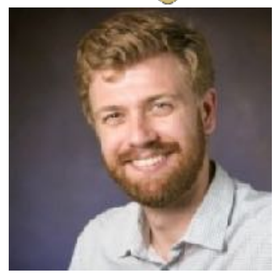
Explore chemical space

Active learning (Bayesian optimization) to optimally deploy coarse-grained molecular dynamics to explore chemical space





Kirill Schmilovich

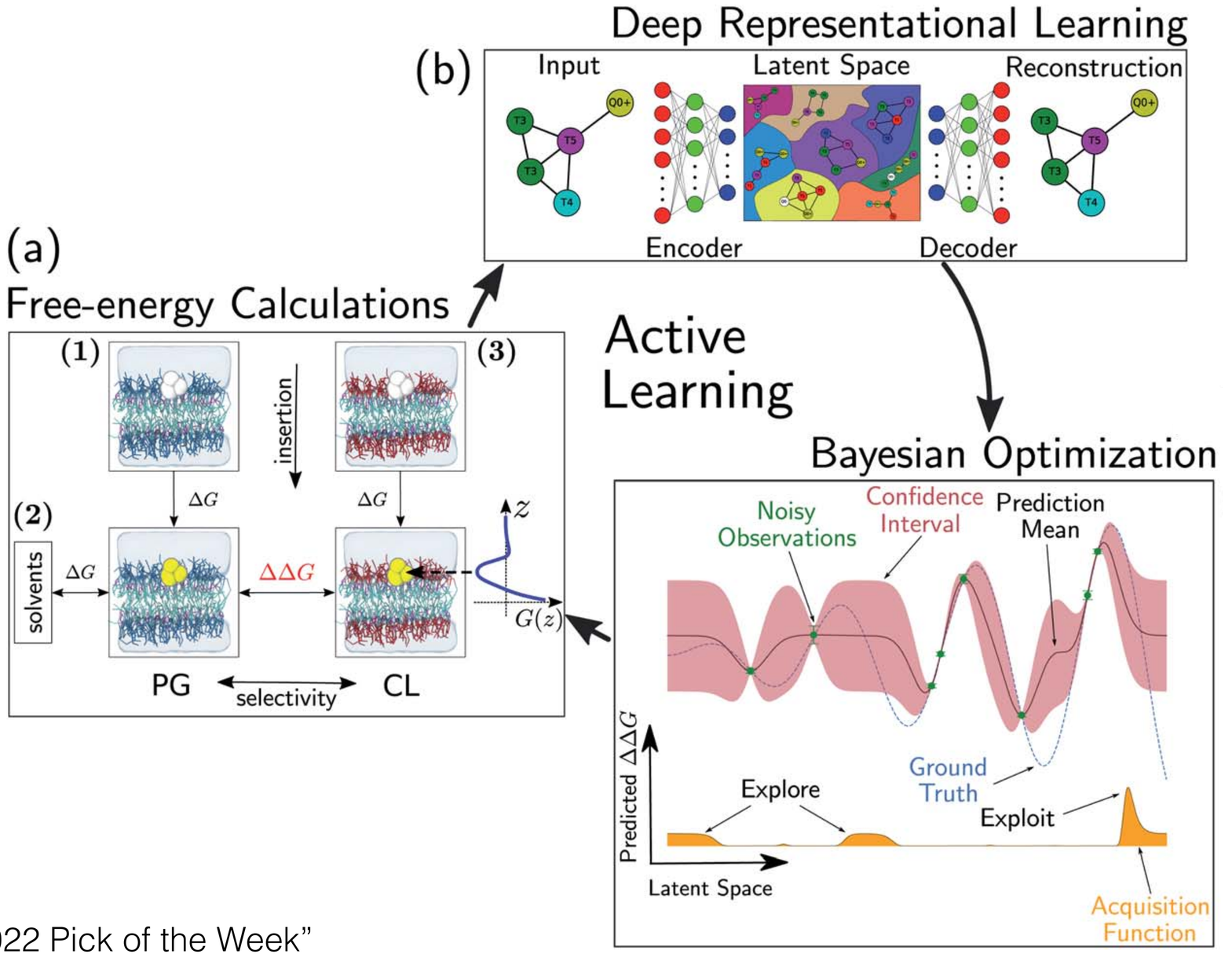


Andrew Ferguson

Gomez-Bombarelli *et al.*, *ACS Central Science* **4** (2018)
Rasmussen "Gaussian processes in machine learning" *Springer*, 2003

Workflow: Multiscale modeling & machine learning

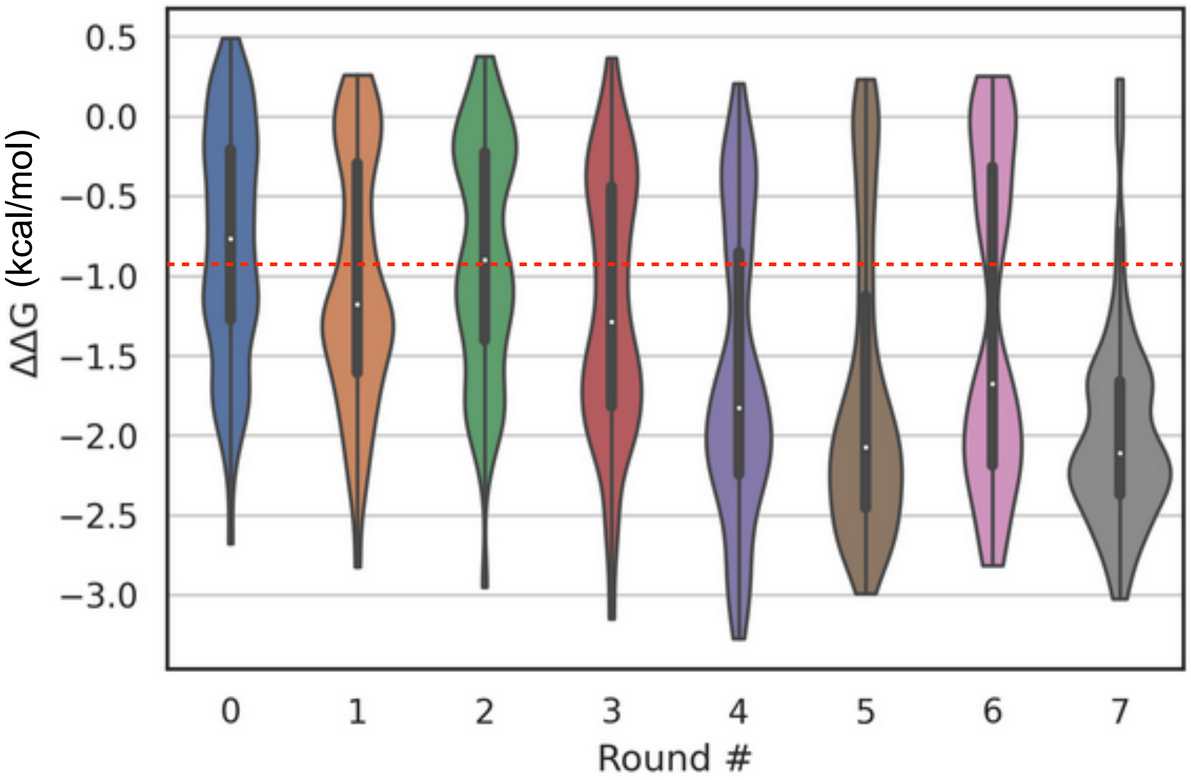
1. Run free-energy calculations for active learning cycle *n*
2. Attach new free energies to augment training of GPR inside the latent space
3. Bayesian optimization selects next compounds to simulate
4. Repeat until convergence

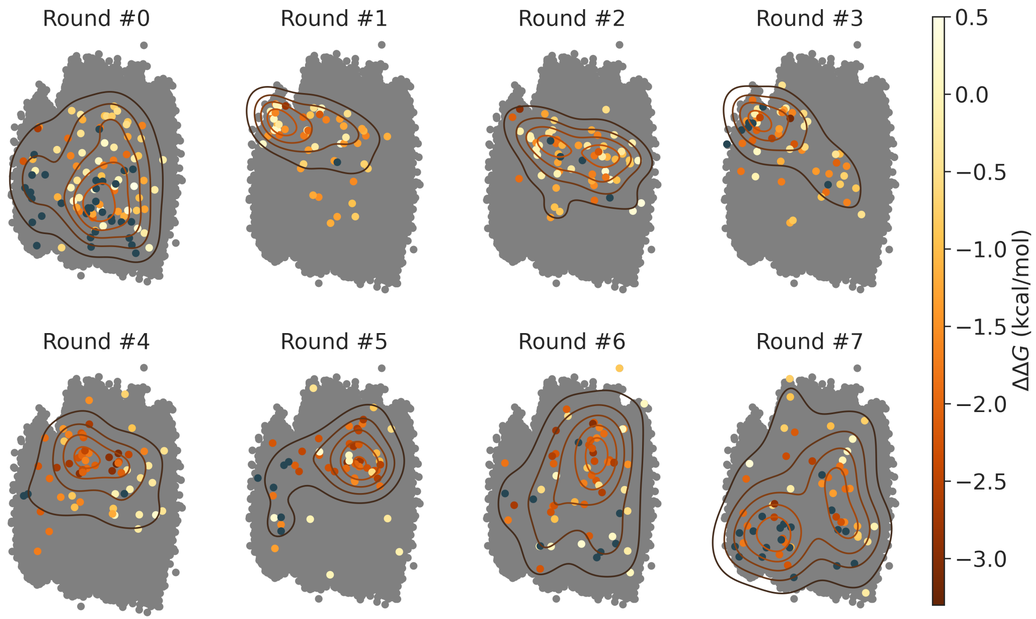


Mohr *et al.*, *Chemical Science* (2022); "2022 Pick of the Week"

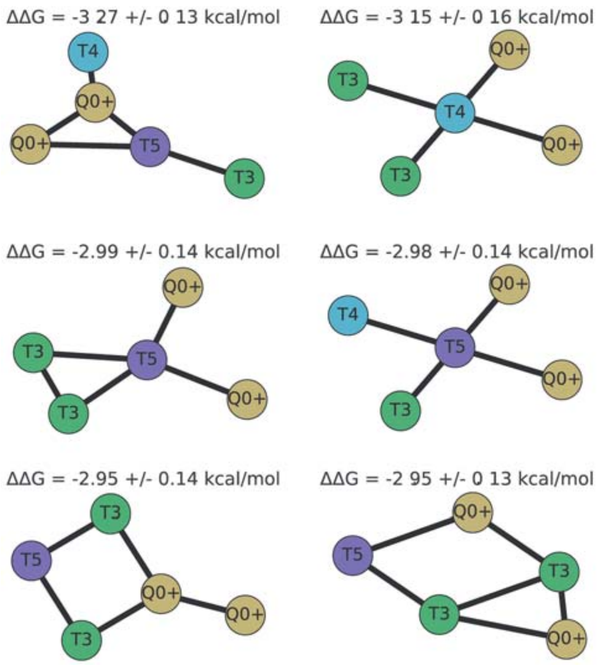
Screening results

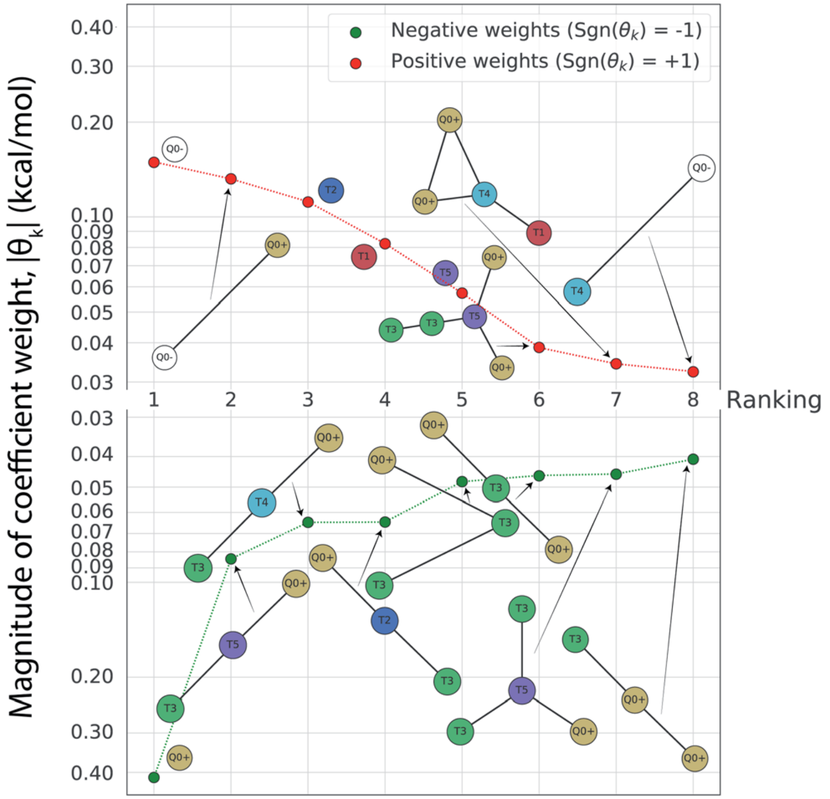
- Performed 7 rounds of screening with 60 candidates per round
- Discover optimal candidate after 4 rounds with $\Delta\Delta G = -3.27$ kcal/mol
  - Sampled only **0.42%** of molecular candidate space
  - **720 GPU-days** of computation
  - **180x selectivity increase** over NAO

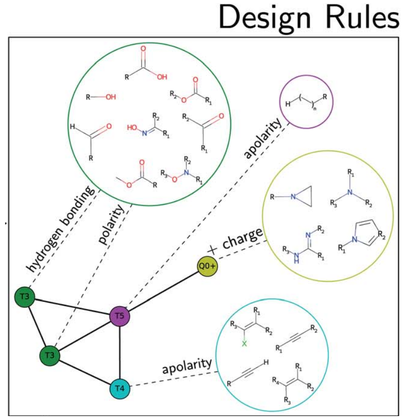




Results: From CG representations to design rules







In silico discovery of design rules

- **Positive charge** (lipid's polar head)
- (coarse-grained) Hbonds
- **Hydrophobicity** (stabilizes in membrane)

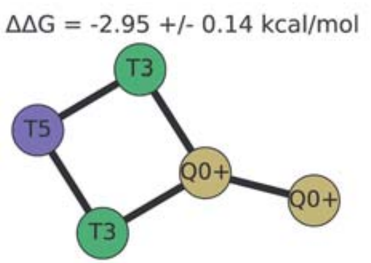
Top coarse-grained compounds

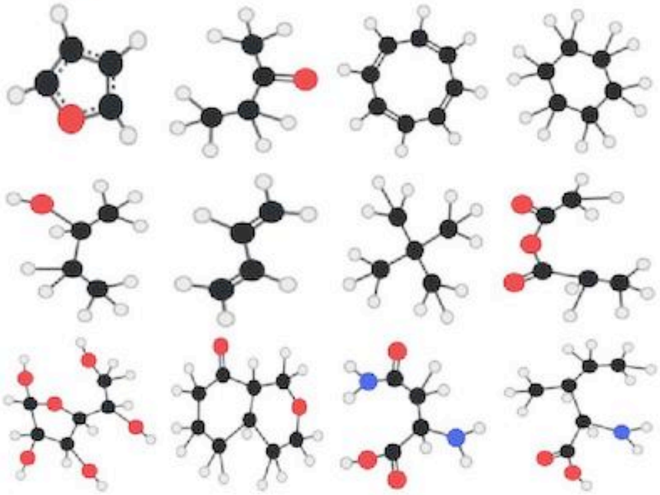
LASSO regression to identify best/worst subgraphs

Design rules

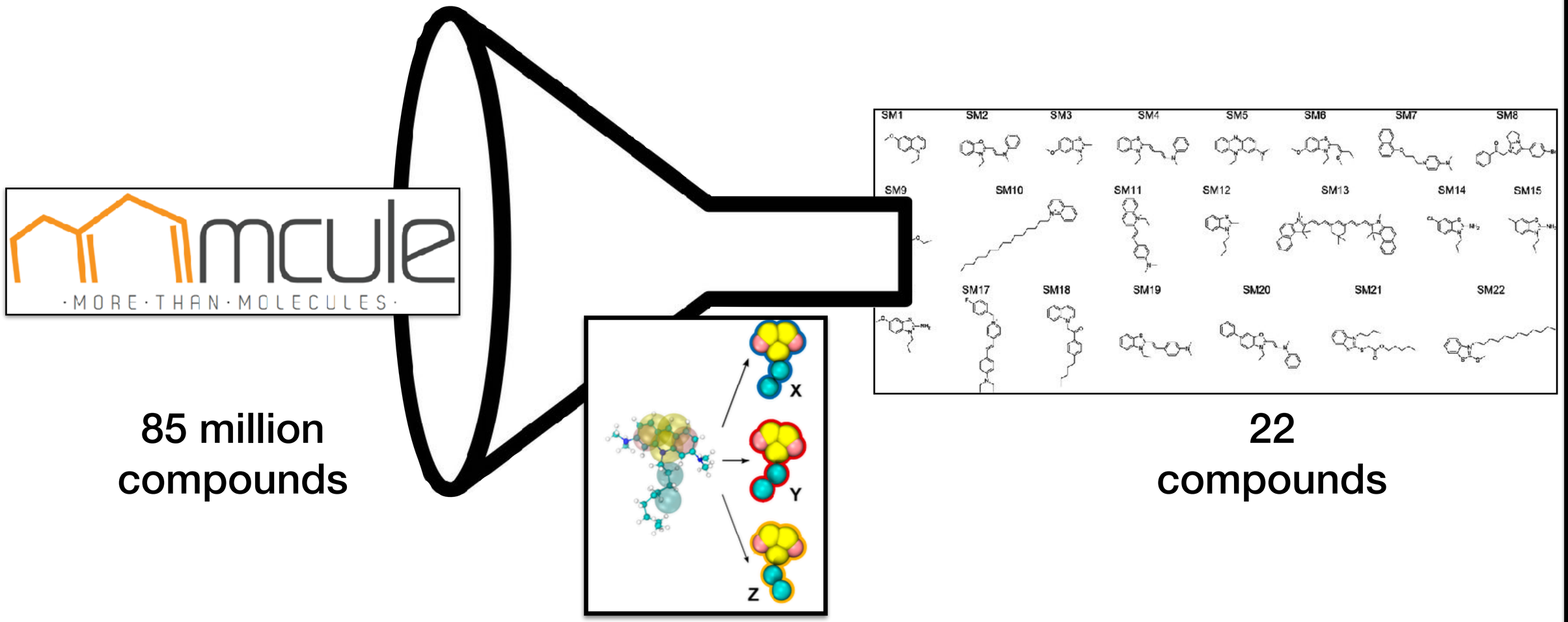
Mohr *et al.*, *Chemical Science* (2022); "2022 Pick of the Week"

How do we translate CG design rules to chemical structures?





Filter vendor databases against design rules



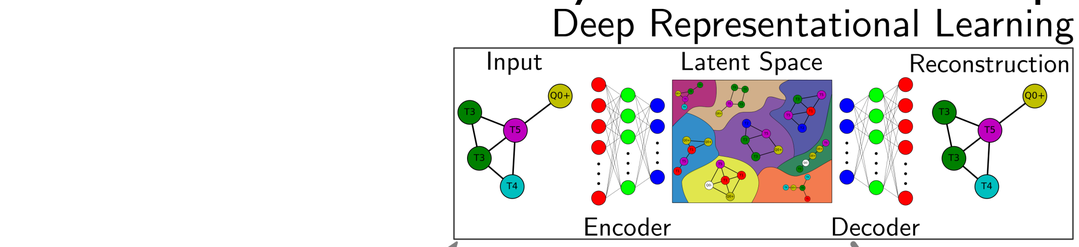
85 million compounds

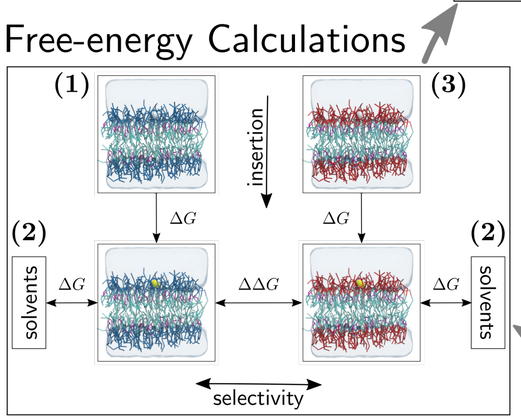
CG design rules

22 compounds

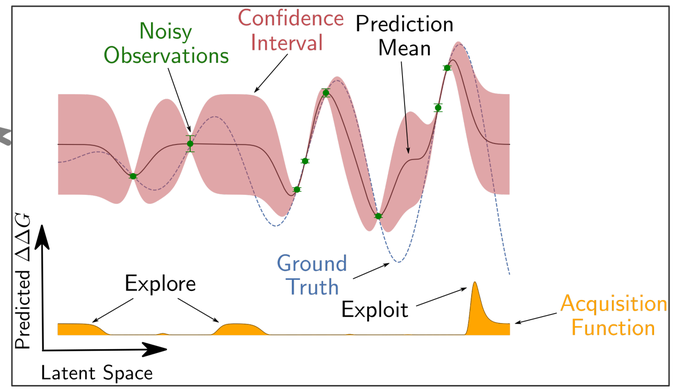
Kleinwächter *et al.*, *RSC Chemical Biology* (2022)

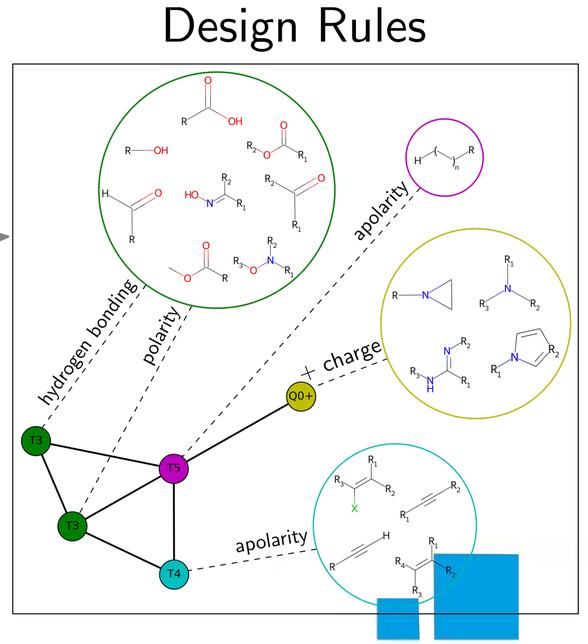
Data-driven discovery of cardiolipin-selective small molecules



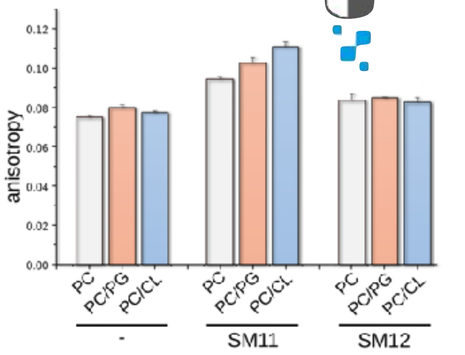


Active Learning





1. Use a *minimal set* of computer simulations to discover **design rules**
2. Suggested 20 compounds for experiments, of which 3 show selectivity *in vitro*, 1 *in vivo*



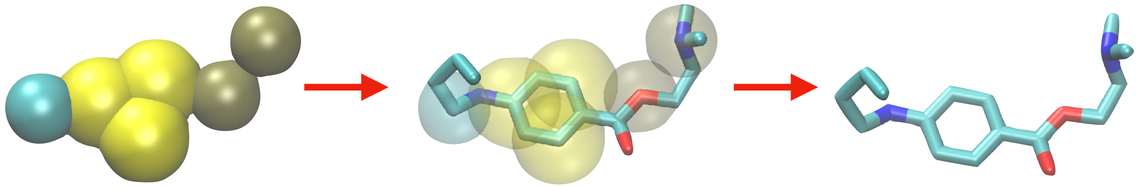
Bernadette Mohr

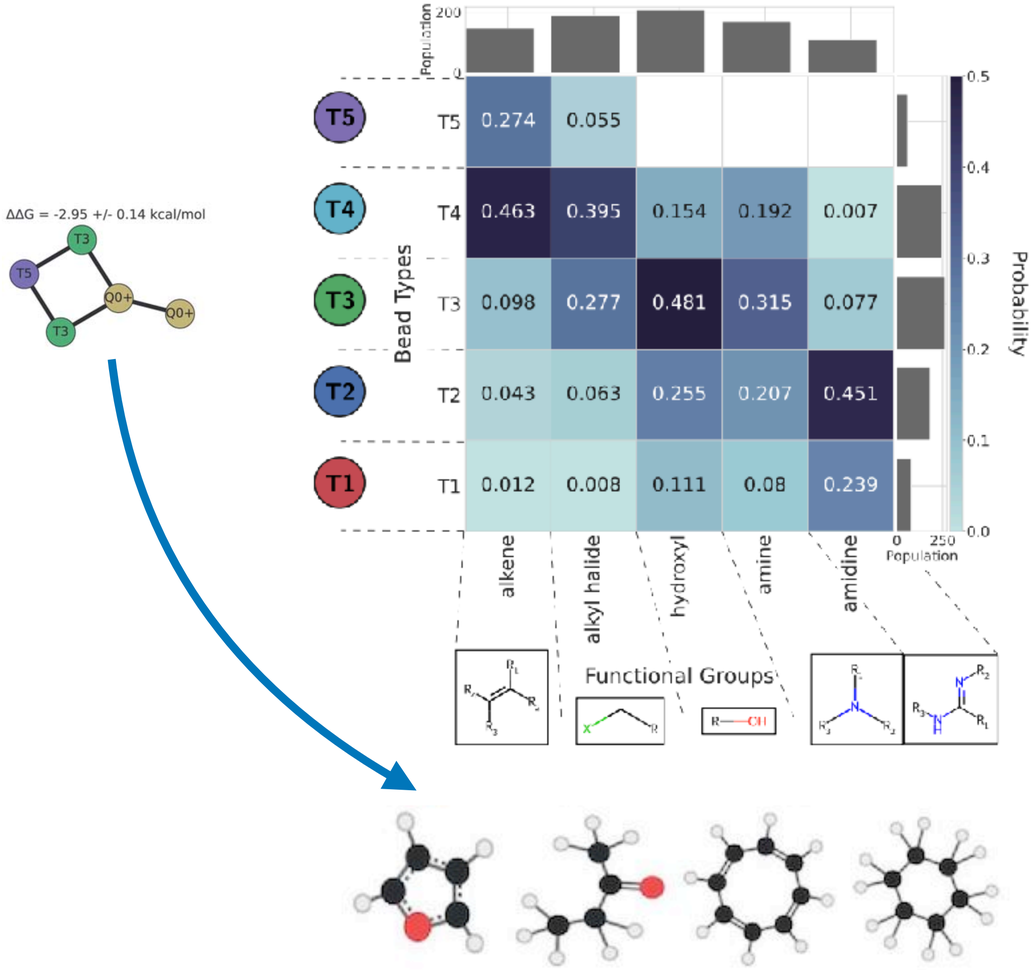
Dirk Schneider

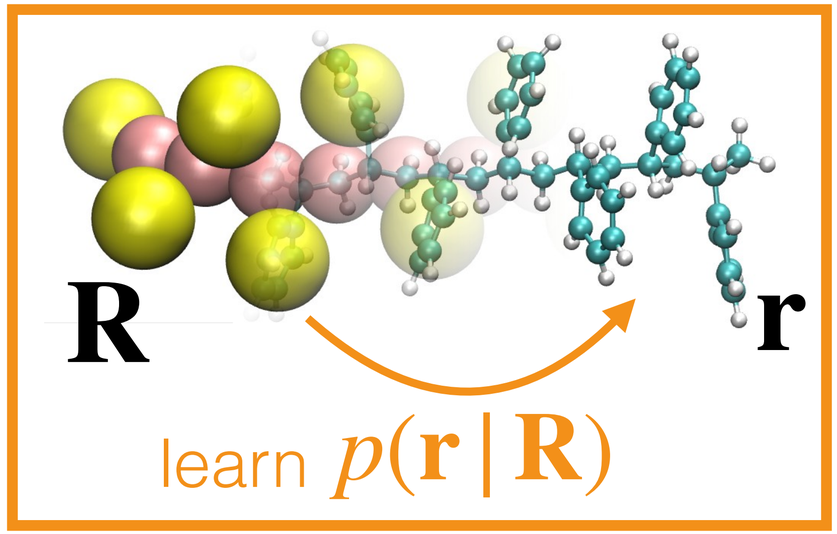
Andrew Ferguson

Mohr *et al.*, *Chemical Science* (2022); "2022 Pick of the Week"
Kleinwächter *et al.*, *RSC Chemical Biology* (2022)

Outlook: reconstructing chemical details

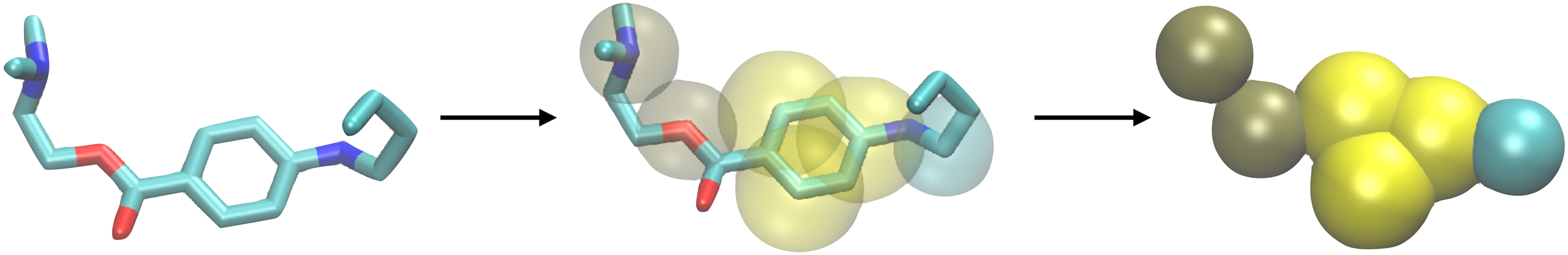






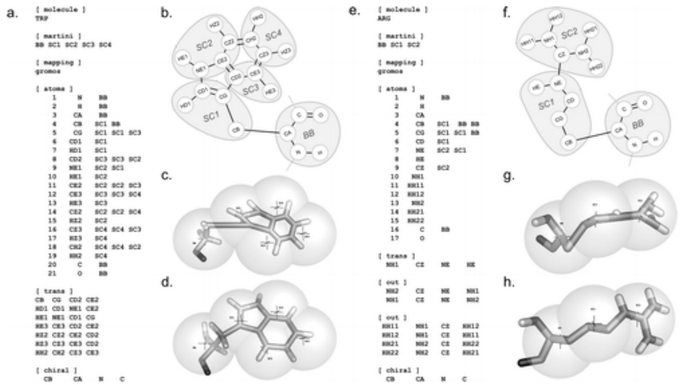
Bernadette Mohr, *PhD thesis*

Fine-graining?



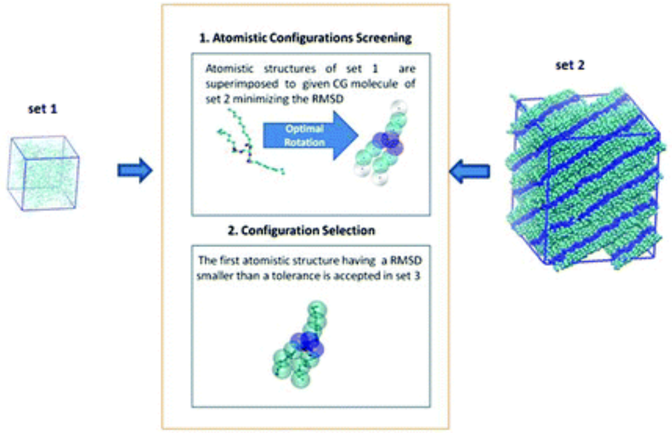
- Going right is straightforward: define a mapping
- **Going left?**

Energy minimization



Wassenaar *et al.*, *JCTC* **10** (2014)

Structural libraries

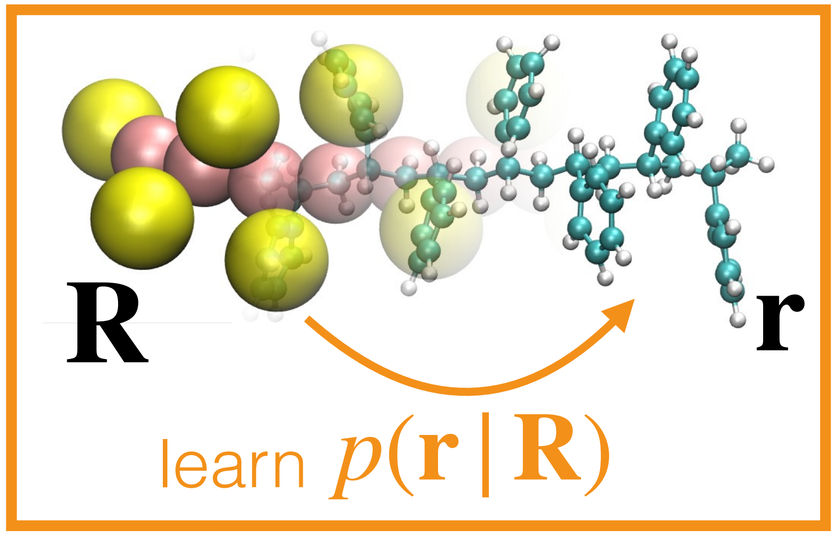


Brasiello *et al.*, *Faraday Disc* **158** (2012)

Generative models

Goodfellow *et al.* NIPS (2014);
Karras et al., *arXiv*:1812.04948 …

Generative adversarial networks: architecture



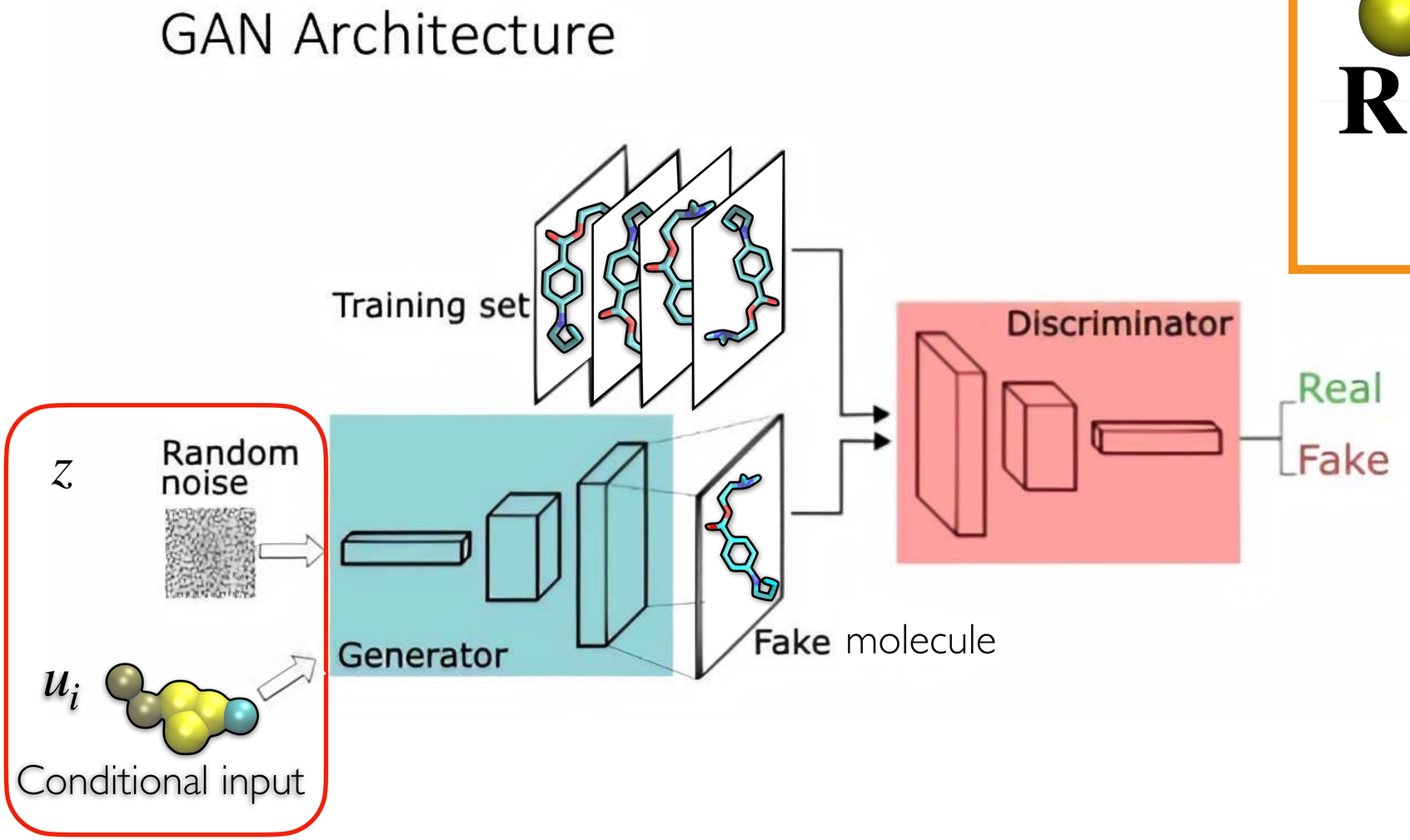
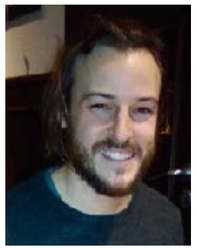
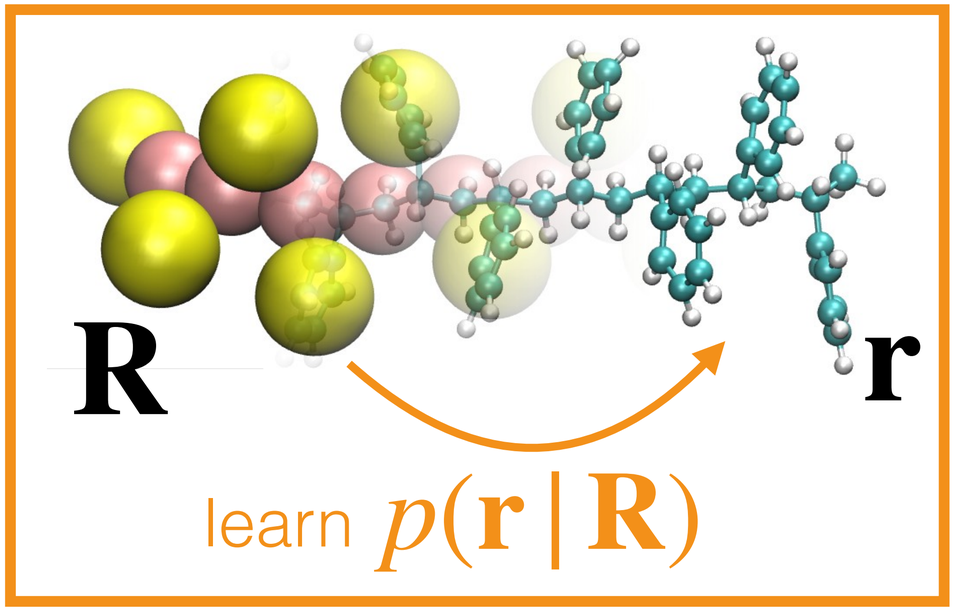


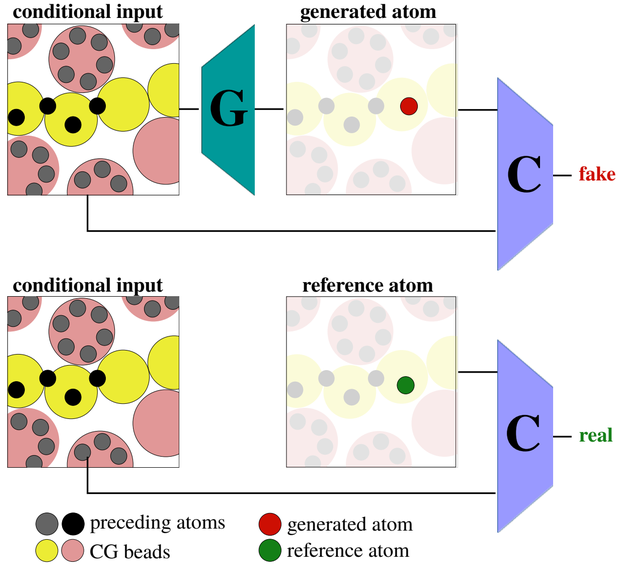
Image adapted from A. Jarda, Medium

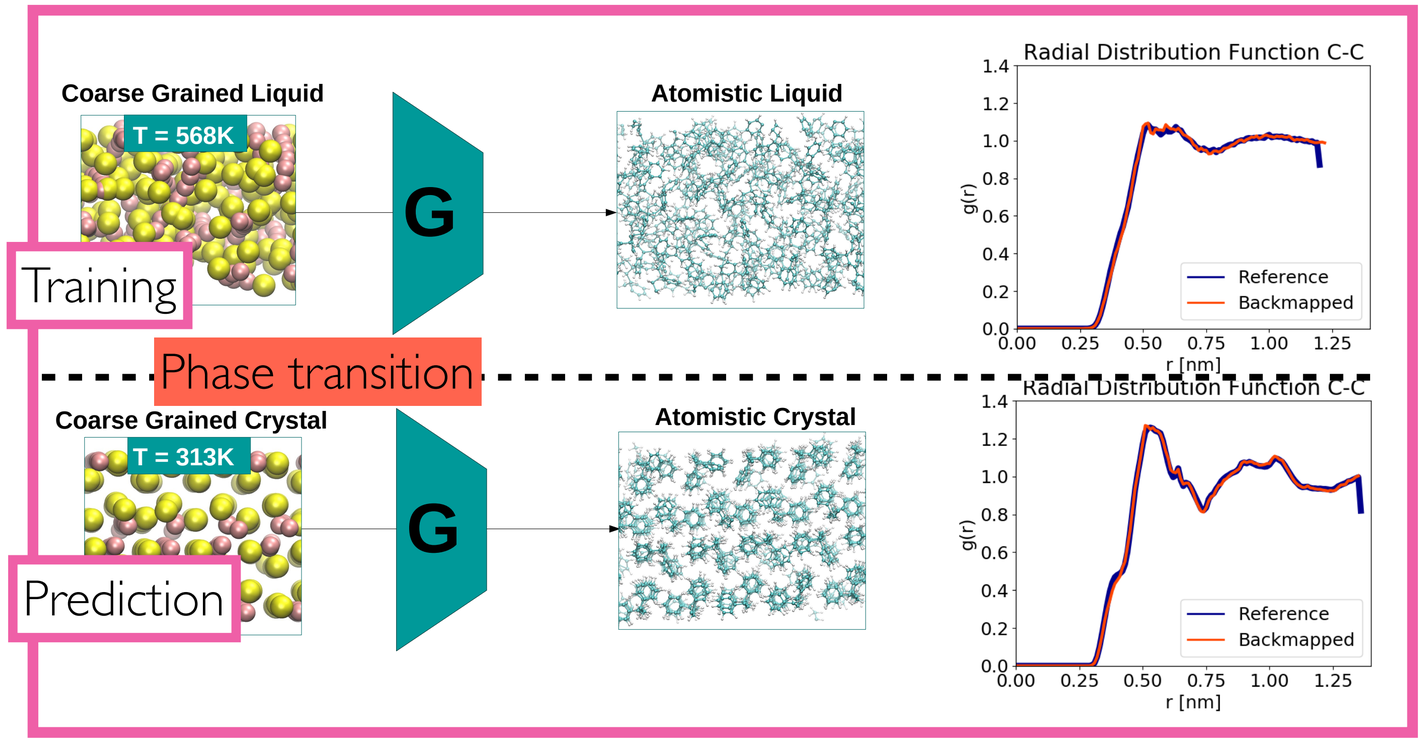


Marc Stieffenhofer

Adversarial backmapping of equilibrated molecular structures





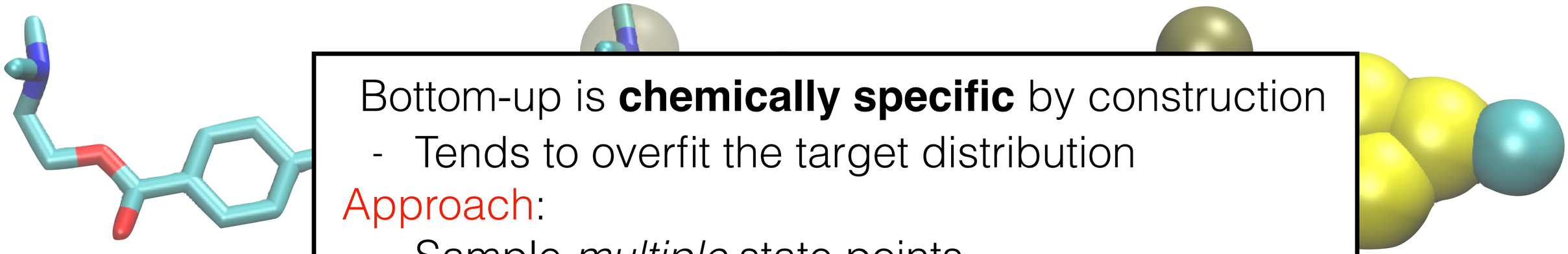


Features

- Loss function incorporates physical priors
- Learns the Boltzmann distribution beyond pairwise statistics
- Temperature transf. from crystal to melt

Stieffenhofer, Wand, Bereau, *Machine Learn Sci Tech* **1** (2020)

Short primer on coarse-graining



Bottom-up is **chemically specific** by construction

- Tends to overfit the target distribution

Approach:

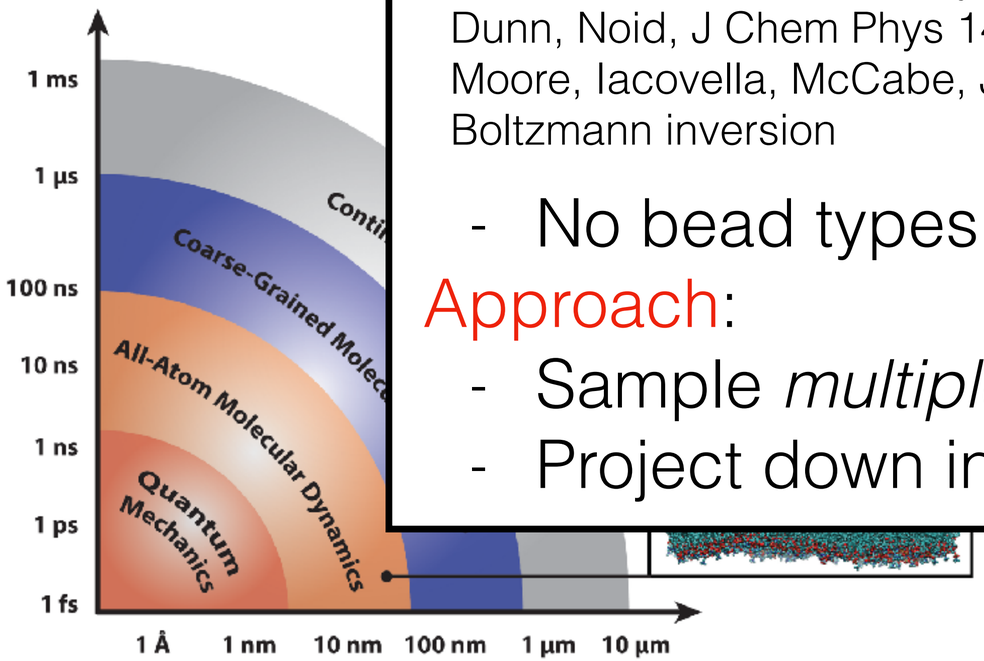
- Sample *multiple* state points

Mullinax, Noid, J Chem Phys 131 (2009): Extended ensemble force matching
Dunn, Noid, J Chem Phys 144 (2016): Ext. ensemble with pressure matching
Moore, Iacovella, McCabe, J Chem Phys 140 (2014): Multistate iterative Boltzmann inversion

- No bead types

Approach:

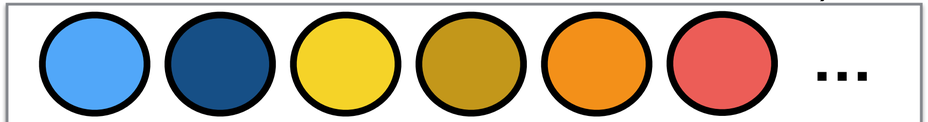
- Sample *multiple* state points + chemistries
- Project down interactions onto bead types



Bradley and Radhakrishnan, *Polymers* **5** (2013)
Noid, *J Chem Phys* **139** (2013)

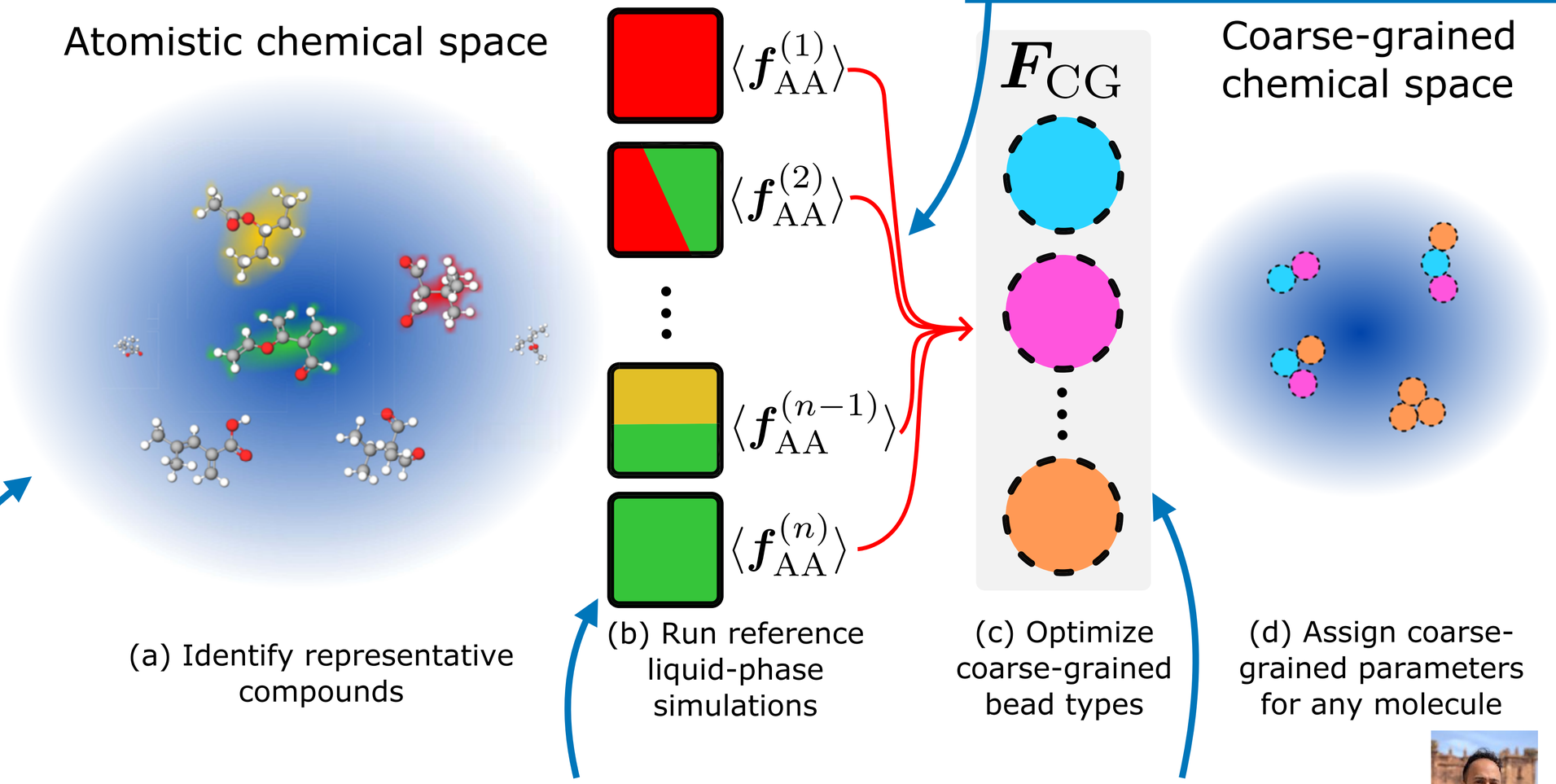
scale physics

ation (e.g., atomistic simulations)



Structure-based and chemically-transferable CG model

Extended-ensemble force matching
Mullinax, Noid, *J Chem Phys* **131** (2009)



3,441 C_7O_2 isomers

703 state points

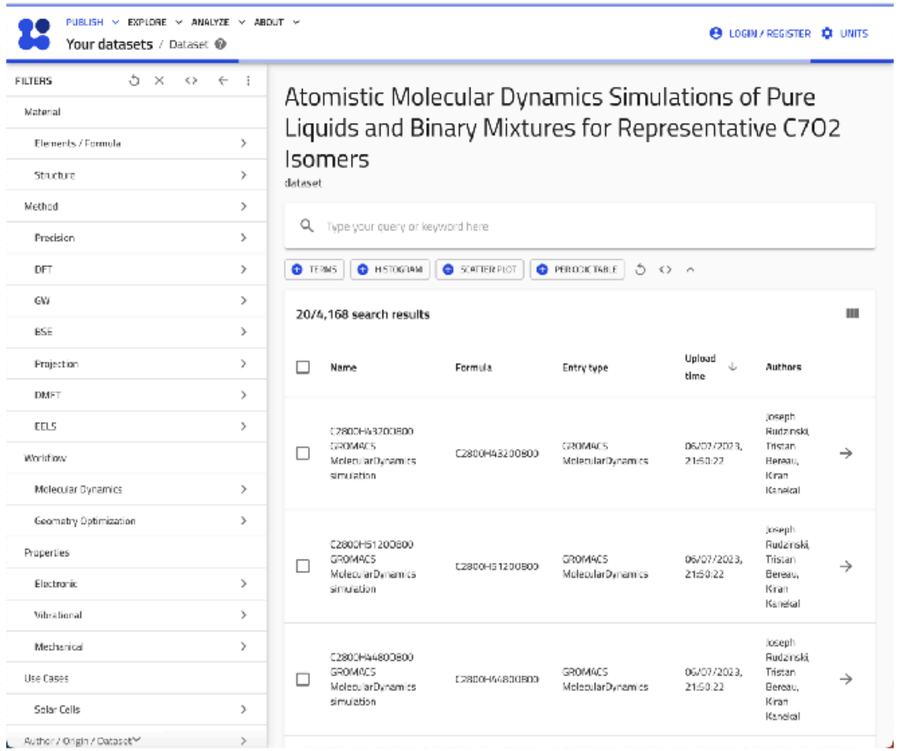
14 CG bead types

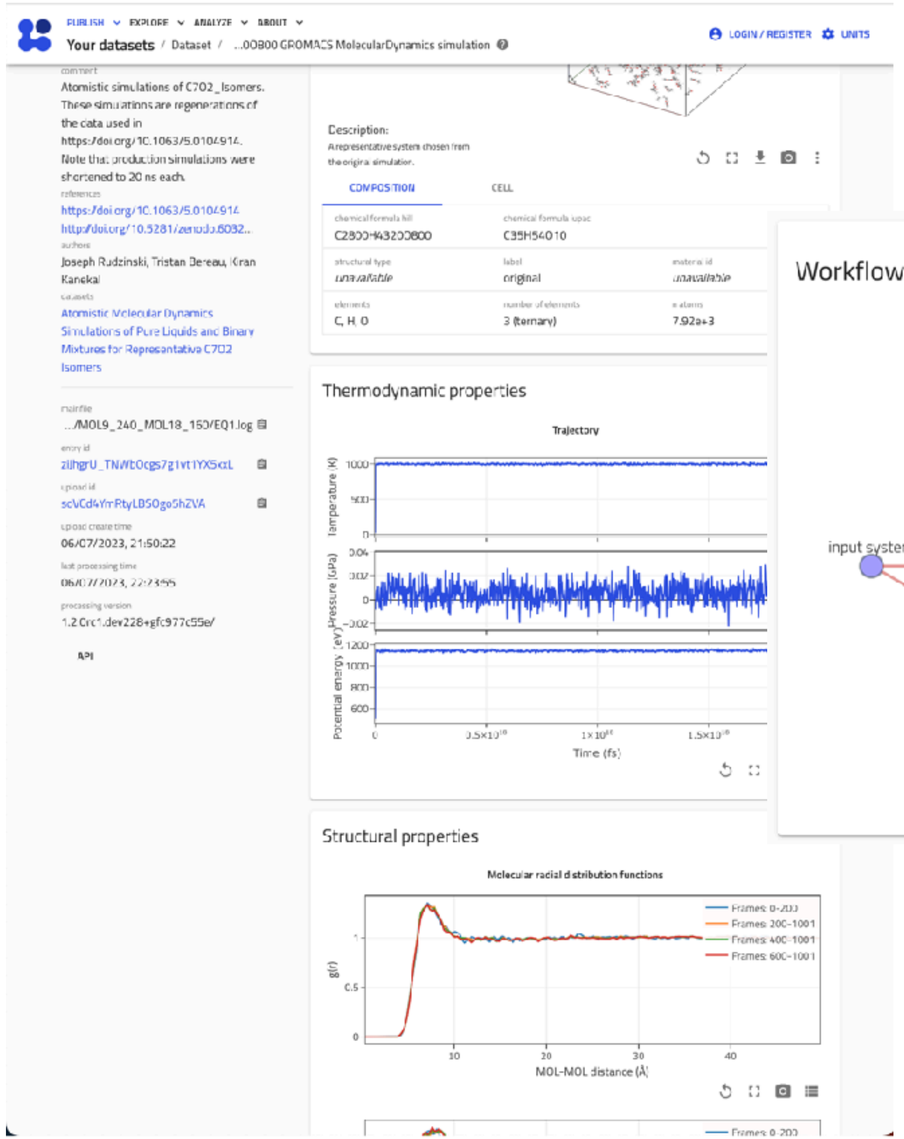
Kanekal, Rudzinski, Bereau, *J Chem Phys* **157** (2022)

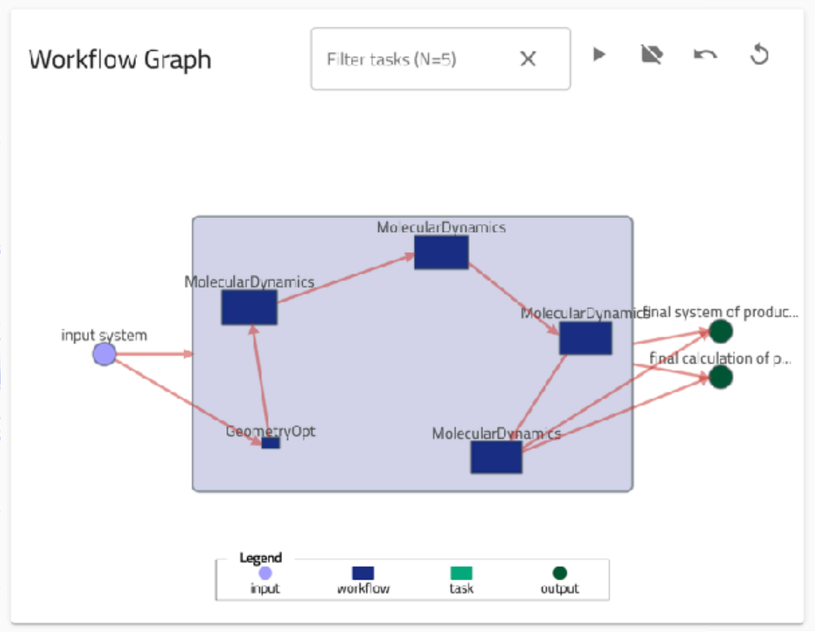

Kiran Kanekal


Joseph Rudzinski

Atomistic MD simulations available on NOMAD





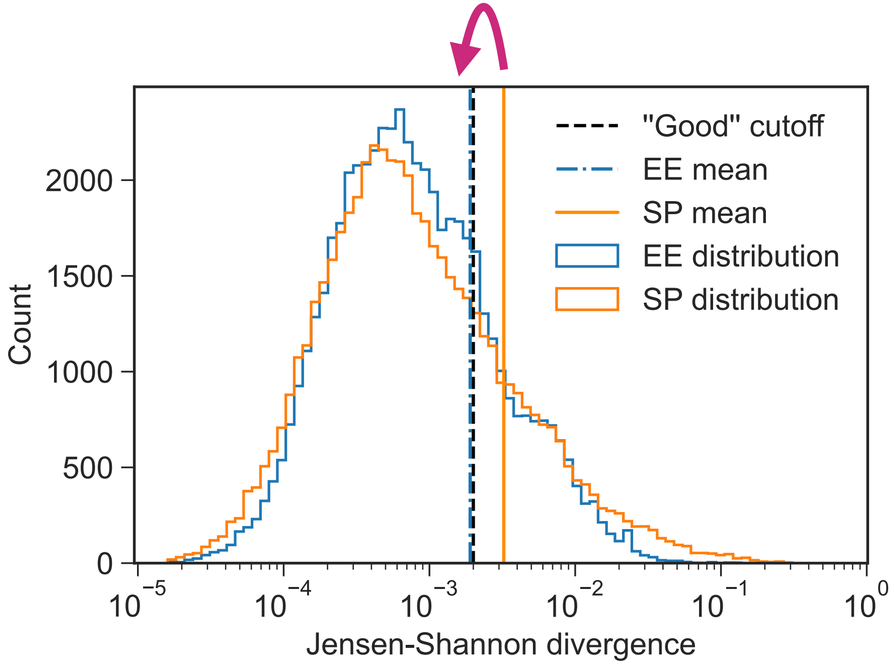


https://nomad-lab.eu/prod/v1/gui/user/datasets/dataset/id/k0FIIN93TDqcDatfaBWpXQ

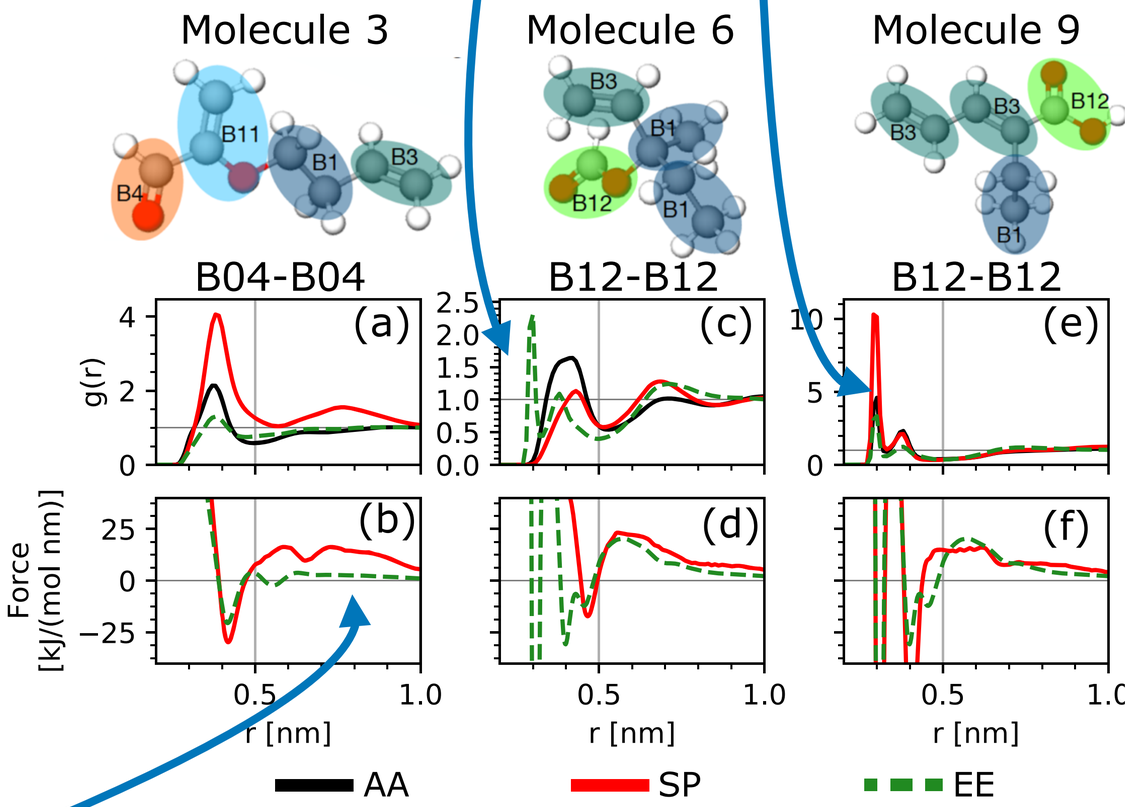
Extended ensemble more accurate than state-point specific models

Extended ensemble models are **more accurate**!

Regularization effect



Same bead type: Averaging smoothens sharp features



Smoothening removes long-range features

Dunn, Noid, *J Chem Phys* **144** (2016)

Kanekal, Rudzinski, Bereau, *J Chem Phys* **157** (2022)

Kiran Kanekal

Joseph Rudzinski

Outlook

Chemical-space perspective

New perspective: Parametrize, calculate, analyze simulations *across* chemical space

Multiscale modeling to explore chemical space

Multiscale description of chemical space exploits scale separation. Accelerates search for structure-property relationships, compound discovery