Machine learning for molecular simulations: priors and predictive constraints

Tristan Bereau
Max Planck Institute for Polymer Research, Mainz, Germany
Rational design from structure-property relationships

Chemistry

Material properties

High-throughput screening
High-throughput screening
High-throughput screening
Intermolecular interactions across chemical space

High-throughput thermodynamics
Predicting the intermolecular energy landscape
Predicting the intermolecular energy landscape

\[
\{Z_i, \mathbf{r}_i\} \mapsto U(\{Z_i, \mathbf{r}_i\})
\]
Predicting the intermolecular energy landscape

\[ \{Z_i, r_i\} \rightarrow U(\{Z_i, r_i\}) \]

Nuclear charges
Coordinates
Potential energy
Transferability challenges

\[ \{Z_i, r_i\} \mapsto U(\{Z_i, r_i\}) \]

Phase (e.g., gas vs. condensed)

Conformational (e.g., cis vs. trans)

Compositional
AMOEBA polarizable force field

\[ U = U_{\text{bond}} + U_{\text{angle}} + U_{\theta} + U_{\text{oop}} + U_{\text{torsion}} + U_{\text{vdW}} + U_{\text{perm}} + U_{\text{ind}} \]

Water-sulfate anion clusters

Small molecule solvation free energies

Phase and conformational transferability
Strategies for energy landscape prediction

\[ \{Z_i, r_i\} \rightarrow U(\{Z_i, r_i\}) \]

1. Physics-based (force field)

2. Data-driven (machine learning)
Strategies for energy landscape prediction

\[ \{Z_i, r_i\} \mapsto U(\{Z_i, r_i\}) \]

1. Physics-based (force field)
   - More parameters

Transferability

2. Data-driven (machine learning)
   - Extrapolation
Extrapolation in machine learning

\[ U_{\text{LJ}}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] \]
Extrapolation in machine learning

\[ U_{\text{LJ}}(r) = 4\epsilon \left( \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right) \]
Extrapolation in machine learning

\[ U_{LJ}(r) = 4\epsilon \left( \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right) \]

Uniform prior

Training points

Prediction
Machine learning: encoding physics

\[ \hat{f}_F(x) = \sum_{i}^{M} \sum_{l}^{3N} \sum_{q}^{S} (P_q \alpha_i)_l \frac{\partial}{\partial x_l} \nabla \kappa(x, P_q x_i) \]

local symmetries
energy conservation


“Transform the configuration, and the prediction transforms with it”

\[ K(S \rho, S' \rho') = SK(\rho, \rho')S'^T \]

Machine learning: encoding physics

\[ \hat{f}_F(x) = \sum_i^{M} \sum_l^{3N} \sum_q^{S} (\mathbf{P}_q \alpha_i)_l \frac{\partial}{\partial x_l} \nabla \kappa(x, \mathbf{P}_q x_i) \]

**local symmetries**  
**energy conservation**

**Wishlist**
- long-range interactions
- chemical-space transferability

**Size of the interpolation-space?**

"Transform the configuration, and the prediction transforms with it"

\[ K(S \rho, S' \rho') = SK(\rho, \rho')S'^T \]

Intermolecular interactions across chemical space
Intermolecular interactions across chemical space

\[ \{Z_i, r_i\} \rightarrow U(\{Z_i, r_i\}) \]

Nuclear charges
Coordinates
Potential energy
Intermolecular interactions across chemical space

Any small molecule made of H, C, O, N neutral compounds
Intermolecular interactions across chemical space

\[ \{Z_i, r_i\} \mapsto U(\{Z_i, r_i\}) \]

Nuclear charges → Coordinates → Potential energy

**Physics-based models**
- Encode laws, symmetries
- Little chemical information

**Machine learning**
- Need to *learn* laws, symmetries
- Interpolate across chemistry

Any small molecule made of H, C, O, N neutral compounds
Physics-based aspect

Long-ranged

• Static electrostatics
• Many-body dispersion
• Polarization

Short-ranged

• Charge penetration
• Repulsion
• (Charge transfer)

Long-ranged

**Perturbation theory**

- Static electrostatics
- Many-body dispersion
- Polarization

\[ V_{\text{Coulomb}}(r) = \frac{1}{4\pi \varepsilon_0} \frac{q_i q_j}{r} \]

Short-ranged

**Overlap models**

- Charge penetration
- Repulsion
- (Charge transfer)

\[ S_{ij} = \int d^3r \rho_i(r) \rho_j(r) \]

Physics-based aspect

Long-ranged

• Static electrostatics
• Many-body dispersion
• Polarization

Short-ranged

• Charge penetration
• Repulsion
• (Charge transfer)

Use ML to predict atoms-in-molecules properties

• Multipole moments
• Hirshfeld ratios
• Atomic density widths/populations

Physics-based aspect

Long-ranged

- Static electrostatics
- Many-body dispersion
- Polarization

Short-ranged

- Charge penetration
- Repulsion
- (Charge transfer)

Use ML to predict atoms-in-molecules properties

- Multipole moments
- Hirshfeld ratios
- Atomic density widths/populations

Static multipole electrostatics

\[ V_{\text{Coulomb}}(r) = \frac{1}{4\pi \varepsilon_0} \frac{q_i q_j}{r} \]

(a)
Static multipole electrostatics

\[ V_{\text{Coulomb}}(r) = \frac{1}{4\pi \varepsilon_0} \frac{q_i q_j}{r} \]

\[ 4\pi \varepsilon_0 \Phi(r) = \frac{q}{R} + \frac{\mu_\alpha R_\alpha}{R^3} + \frac{1}{3} \Theta_{\alpha\beta} \frac{3R_\alpha R_\beta - R^2 \delta_{\alpha\beta}}{R^5} + \ldots \]
Static multipole electrostatics

\[ V_{\text{Coulomb}}(r) = \frac{1}{4\pi \varepsilon_0} \frac{q_i q_j}{r} \]

\[ 4\pi \varepsilon_0 \Phi(r) = \frac{q}{R} + \frac{\mu_\alpha R_\alpha}{R^3} + \frac{1}{3} \Theta_{\alpha\beta} \frac{3R_\alpha R_\beta - R^2 \delta_{\alpha\beta}}{R^5} + \ldots \]
Static multipole electrostatics

\[ V_{\text{Coulomb}}(r) = \frac{1}{4\pi \varepsilon_0} \frac{q_i q_j}{r} \]

\[ 4\pi \varepsilon_0 \Phi(r) = \frac{q}{R} + \frac{\mu_\alpha R_\alpha}{R^3} + \frac{1}{3} \Theta_{\alpha\beta} \frac{3R_\alpha R_\beta - R^2 \delta_{\alpha\beta}}{R^5} + \ldots \]
Multipoles: Learning curves

kernel-ridge regression

$K\alpha = p$

Descriptor: aSLATM

$M^I = \{Z_I, \rho_{IJ}(R), \rho_{IJK}(\theta)\}$

Huang and von Lilienfeld, arXiv:1707.04146, 2017

Easier to learn H,O than C,N
Multipoles: Correlation curves

- Monopole
- Dipole
- Quadrupole

C, N have more complex valencies
Pairwise

London dispersion:

\[ E_{AB} = -\frac{C_{6AB}}{R^6} \]

Casimir-Polder:

\[ C_{6AB} = \frac{3}{\pi} \int_0^\infty d\omega \alpha_p(i\omega)\alpha_q(i\omega) \]

\[ U_L(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \]
Many-body dispersion

Pairwise

London dispersion:

\[ E_{AB} = -\frac{C_{6AB}}{R^6} \]

Casimir-Polder:

\[ C_{6AB} = \frac{3}{\pi} \int_0^\infty d\omega \alpha_p(i\omega) \alpha_q(i\omega) \]

Need atomic polarizabilities as input parameters

\[ U_{L,J}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \]
Many-body dispersion

Pairwise

London dispersion:

\[ E_{AB} = -\frac{C_{6AB}}{R^6} \]

Casimir-Polder:

\[ C_{6AB} = \frac{3}{\pi} \int_0^\infty d\omega \alpha_p(i\omega)\alpha_q(i\omega) \]

Need atomic polarizabilities as input parameters

\[ U_{LJ}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \]

Many-body dispersion

Coupled fluctuating dipole model:
- Quantum harmonic oscillators


Atomic polarizabilities
- Atom-in-molecule: Hirshfeld ratios

Many-body dispersion

Pairwise

London dispersion:

\[ E_{AB} = -\frac{C_{6AB}}{R^6} \]

Casimir-Polder:

\[ C_{6AB} = \frac{3}{\pi} \int_0^\infty d\omega \alpha_p(i\omega)\alpha_q(i\omega) \]

Need atomic polarizabilities as input parameters

Many-body dispersion

Coupled fluctuating dipole model:
- Quantum harmonic oscillators


Atomic polarizabilities
- Atom-in-molecule: Hirshfeld ratios

Learning Hirshfeld ratios

\[ \alpha^0_p \approx \frac{V^\text{eff}}{V_p^\text{free}} \alpha^\text{free}_p = \frac{\int d\mathbf{r} r^3 w_p(\mathbf{r}) n(\mathbf{r})}{\int d\mathbf{r} r^3 n_p^\text{free}} \alpha^\text{free}_p \]

Training: 12.3k atoms in 1k molecules

MAE: 0.006

\[ C_{ij} = \begin{cases} 
\frac{1}{2} Z_i^{2.4} & \forall \ i = j \\
\frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \forall \ i \neq j 
\end{cases} \]
Polarizability

**Thole model**

*N* point dipole polarizabilities placed in a homogeneous field $\mathbf{F}$

Induced dipole moment at point $p$:

$$\mu_p = \alpha_p \left( \mathbf{F}_p - \sum_{q \neq p}^{N} T_{pq} \mu_q \right)$$

$T_{pq}$: dipole field tensor

$\mathbf{F}_p$: given by static electrostatic interactions at site $p$

Polarizability

**Thole model**

*N* point dipole polarizabilities placed in a homogeneous field $\mathbf{F}$

Induced dipole moment at point $p$:

$$
\mu_p = \alpha_p \left( \mathbf{F}_p - \sum_{q \neq p}^N T_{pq} \mu_q \right)
$$

$T_{pq}$: dipole field tensor

$\mathbf{F}_p$: given by static electrostatic interactions at site $p$

Need atomic polarizabilities and multipoles as input parameters

Molecular polarizabilities

(a)

\[ \alpha_{\text{iso}}^{\text{exp}} \ [\text{Bohr}^3] \]

\[ \alpha_{\text{iso}}^{\text{model}} \ [\text{Bohr}^3] \]

(b)

\[ F A_{\text{exp}} \]

\[ F A_{\text{model}} \]

Tkatchenko and Scheffler, Phys Rev Lett (2009)
Few global parameters

7 parameters to cover chemical space of CHON intermolecular energies

Polarization 1

Many-body dispersion 2

Repulsion C; H; O; N
Validation: S66a8

Error: 0.5 kcal/mol

Angular-displaced nonequilibrium geometries

Validation: SSI

Ref: CCSD(T)/CBS

Error: 0.4 kcal/mol


2,200+ amino-acid pairs
Validation: JSCH-2005

Error: 1.4 kcal/mol

DNA base pairs, amino-acid pairs

Toward the condensed phase

Figure 8: Correlation plots for the total intermolecular energy between reference and present calculations for (a) the water-clusters dataset and (b) the host-guest complexes in the S12L database. The colors in (a) indicate the number of molecules involved in the cluster: from two (red) to 10 (blue) molecules.

Figure 9: Comparison of the intermolecular energy as a function of dimer distance for the benzene dimer in the (a) parallel-displaced—stacked—and (b) T-shaped conformations. (c) Cohesive binding energy of the benzene crystal as a function of the scaling factor of the unit cell.
Intermolecular interactions across chemical space


**Lessons learned**
ML force fields can be systematically improved
- physical effects (perturbation theory at short range)
- chemical groups
- provide data + ML model
Outline
Outline

Intermolecular interactions across chemical space

High-throughput thermodynamics
Drug permeability

Flux of drug permeation across a lipid membrane
Drug permeability

Flux of drug permeation across a lipid membrane
Drug permeability

Flux of drug permeation across a lipid membrane

\[ P^{-1} = \int dz \frac{\exp(\frac{\Delta G(z)}{k_B T})}{D_z(z)} \]


Fokker-Planck; Smoluchowski
Drug permeability

Flux of drug permeation across a lipid membrane

Potential of mean force

\[ \Delta G(z) \]

Permeability coefficient

\[
P^{-1} = \int dz \frac{\exp(\Delta G(z)/k_B T)}{D_z(z)}
\]

Fokker-Planck; Smoluchowski

Drug permeability

Flux of drug permeation across a lipid membrane

Potential of mean force

\[ \Delta G(z) \]

Permeability coefficient

\[ P^{-1} = \int dz \frac{\exp(\Delta G(z)/k_B T)}{D_z(z)} \]

Fokker-Planck; Smoluchowski

Sample the potential of mean force from computer simulations

Drug-membrane thermodynamics in silico

Drug-membrane thermodynamics *in silico*

- Manual force-field parametrization
- Sampling: 100,000+ CPU-hours

Drug-membrane thermodynamics *in silico*

- Manual force-field parametrization
- Sampling: 100,000+ CPU-hours

Carpenter *et al.*, *Biophys J* (2014)
Automated coarse-graining: Martini
Automated coarse-graining: Martini

Chemical group:
- net charge
- hydrogen bond
- water/octanol partitioning
Automated coarse-graining: Martini

Chemical group:
- net charge
- hydrogen bond
- water/octanol partitioning

18 bead types: chemical fragments
Periole, Marrink, Biomolecular Simulations (2013)
Automated coarse-graining: Martini

Chemical group:
- net charge
- hydrogen bond
- water/octanol partitioning

Heuristic optimization of mapping

18 bead types: chemical fragments
Periole, Marrink, Biomolecular Simulations (2013)

neural network trained on experimental data

Automated parametrization for small molecules
Solute insertion in the membrane
Solute insertion in the membrane

Atomistic simulations: >100,000 CPU-hours
Coarse-grained simulations: ~100 CPU-hours

Solute insertion in the membrane

Atomistic simulations: >100,000
Coarse-grained simulations: ~100

Construct reduced chemical space

We don’t focus on specific compounds, instead explore chemical diversity

Molecular weight:

1

2

\[ \text{... (x14)} \]

\[ \text{... (x105)} \]

\[ \sim 30-160 \text{ Da} \]
Construct reduced chemical space

We don’t focus on specific compounds, instead explore chemical diversity

Molecular weight:

1  \[\ldots\] (x14)

2  \[\ldots\] (x105)

\[\sim 30\text{-}160\text{ Da}\]

CG: combinatorial explosion is reduced
Identifying simple thermodynamic relations

\[ G(z) \]

Identifying simple thermodynamic relations

\[ G(z) \]

\[ \Delta G_{W \rightarrow O} \]

\[ \Delta G_{I \rightarrow O} \]

Water

Interface

Identifying simple thermodynamic relations

\[ \Delta G_{W \to I} + \Delta G_{I \to O} - \Delta G_{W \to O} = 0 \]

\[ G(z) \]

Midplane

\[ \Delta G_{W \to O} \]

\[ \Delta G_{I \to O} \]

\[ \Delta G_{W \to I} \]

Water

Interface

Identifying simple thermodynamic relations

\[ \Delta G_{W\rightarrow I} + \Delta G_{I\rightarrow O} - \Delta G_{W\rightarrow O} = 0 \]

\[ G(z) \]

Midplane

\[ \Delta G_{W\rightarrow o} \]

\[ \Delta G_{I\rightarrow O} \]

Water

Interface

Identifying simple thermodynamic relations

\[ \Delta G_{W \rightarrow I} + \Delta G_{I \rightarrow O} - \Delta G_{W \rightarrow O} = 0 \]

\( G(z) \)

Midplane

\( \Delta G_{W \rightarrow O} \)

\( \Delta G_{I \rightarrow O} \)

Water

Interface

Identifying simple thermodynamic relations

\[ \Delta G_{W \to I} + \Delta G_{I \to O} - \Delta G_{W \to O} = 0 \]

\[ G(z) \]

Only 1 independent variable: estimated from water/octanol partitioning

PMF prediction from experimental value

PMF prediction from experimental value

PMF prediction from experimental value

Error across chemical space:
linear relations: 1.8 kcal/mol
coarse-grained: 1.4 kcal/mol

Coarse-graining reduces chemical space
Coarse-graining reduces chemical space

Reduction due to limited number of bead types

18 bead types: chemical fragments
High-throughput coarse-graining scheme

Reduction of chemical space according to size and hydrophobicity

Δ\(G_{w\rightarrow o1}\) Hydrophobicity

High-throughput coarse-graining scheme

Reduction of chemical space according to size and hydrophobicity

Insertion of neutral and charged compounds across the membrane

ΔG_{W→OIL}  

pK_a  

z

pK_a

High-throughput coarse-graining scheme

Reduction of chemical space according to size and hydrophobicity

Insertion of neutral and charged compounds across the membrane

Potential of mean force of neutral and charged species

High-throughput coarse-graining scheme

Reduction of chemical space according to size and hydrophobicity

Insertion of neutral and charged compounds across the membrane

Potential of mean force of neutral and charged species

Permeability surface across chemical space along two molecular descriptors

ΔGW→OL

Hydrophobicity

ΔG(z)

ΔGW→M

pKa

ΔG(z)

ΔG(z)

log₁₀ P

ΔGW→M

Menichetti, Kanekal, Bereau, ACS Cent. Sci. 5 (2019)
High-throughput coarse-graining scheme

**Strategy**
- Coarse-graining prior to high-throughput simulations
- Fewer simulations
- Suggests low-dimensional representation

Reduction of chemical space according to size and hydrophobicity

Insertion of neutral and charged compounds across the membrane

Potential of mean force of neutral and charged species

Permeability surface across chemical space along two molecular descriptors

Chemical-space coverage

What type of chemistry covers the surface?

What type of chemistry covers the surface?

Analysis of **500,000+** compounds

Between compounds that reduce to CG molecules made of a single bead ("unimers") from those made of two beads ("dimers") amounts to a segregation between molecular weights. We populate the permeability surfaces with these compounds—projecting them onto the two molecular descriptors: $pK_a$ and partitioning free energy. By coarse-graining every single compound, we establish a map between chemical structure and its CG thermodynamic property.

![Figure 3: Chemical-space coverage of GDB projected onto $pK_a$ and water/octanol partitioning free energies, $G_{W\rightarrow O}$. Acidic and basic $pK_a$ are shown in panels (a,b) and (c,d), respectively. Panels (a,c) and (b,d) describe the coverage corresponding to coarse-grained unimers and dimers, respectively. Regions highlighted in light blue display several representative chemical groups. Substitutions denoted by "?" correspond to H, alkyl, or aryl groups, while "?*" only corresponds to alkyl groups.

Impact of chemical group on permeability

What type of chemistry covers the surface?

Analysis of 500,000+ compounds
Navigating chemical space

FIG. 1. (a) Importance sampling across coarse-grained compounds via a Markov chain Monte Carlo scheme. Only the dark-blue region is sampled. (b) Background: Simulation setup of a solute (yellow) partitioning between water (not shown) and the lipid membrane. Foreground: Potential of mean force along the normal of the bilayer, \( G(z) \), and definition of the three transfer free energies of interest between the three state points (red circles): bilayer midplane (“M”), membrane-water interface (“I”), and bulk water (“W”). (c) The MC-sampled free energies (dark-blue region) form the training set for a machine learning model, used to predict a larger subset of compounds (light-blue region). (d) Each coarse-grained compound represents a large number of small molecules.

compounds that were not sampled by using machine learning (ML; see Fig. 1c) \([15]\). Despite known limited capabilities to extrapolate beyond the training set, we observe remarkable accuracy for the predicted compounds. This excellent transferability can be associated to a simplified learning procedure at the CG resolution: structure-property relationships are easier to establish \([13]\) and compound similarity is compressed due to the reduction of chemical space. The range of reliable predictions is made clear by means of the ML model satisfying linear thermodynamic relations across compounds \([12]\)—a more robust confidence metric compared to the predictive variance. The CG results are then systematically backmapped (Fig. 1d) to yield an unprecedentedly-large database of free energies.

II. RESULTS

We consider the insertion of a small molecule across a single-component phospholipid membrane made of 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) solvated in water. The insertion of a drug is monitored along the collective variable, \( z \), normal distance to the bilayer midplane (Fig. 1b). We focus on three thermodynamic state points of the small molecule: the bilayer midplane (“M”), the membrane-water interface (“I”), and bulk water (“W”). We link these quantities in terms of transfer free energies, e.g., \( \Delta G_{W!M} \) denotes the transfer free energy of the small molecule from water to the bilayer midplane.

A. Importance sampling

We ran MC simulations across CG linear trimers and tetramers (results for tetramers are shown in the SI), randomly changing a bead type, calculating the relative free energy difference between old and new compound in the three different environments, and accepting the trial compound using a Metropolis criterion on the water/interface transfer free energy \( G_{W!I} \) (Fig. 1a). This criterion aimed at selecting compounds that favor partitioning in the membrane. The MC algorithm yielded an acceptance ratio of 0.2. While initially most trial compounds contributed to expand the database, the sampling scheme quickly reached a stable regime where roughly half of the compounds had already been previously visited. Because each free-energy calculation is expensive, we avoid recalculating identical alchemical transformations to help efficiently converge the protocol.

A short MC sequence of accepted compounds is shown in Fig. S1. We display the sequence within the network of sampled compounds, each node being represented by the set of Martini bead types involved. Interestingly we find a large number of closed paths within this network. Since the free energy is a state function, the closed path represents a thermodynamic cycle—it must sum up to zero. We found negligible changes in the free energies regardless of whether or not this condition was enforced on the closed path, meaning that our chain of transformations does not compound significant statistical error.

Boosting the database with machine learning

Kernel-based machine learning

\[
\Delta G(x) = \sum_i \alpha_i K(x^*_i, x)
\]
Boosting the database with machine learning

Kernel-based machine learning

$$\Delta G(x) = \sum \alpha_i K(x_i^*, x)$$

Predicted free energy

$$K(x_i, x_j) = \exp \left( -\frac{|x_i - x_j|_1}{\sigma} \right)$$

Training data

Training of machine learning model on Monte Carlo dataset

Query sample

Training points

Predictions

Hoffmann, Menichetti, Kanekal, Bereau, Phys Rev E 100 (2019)
Boosting the database with machine learning

Kernel-based machine learning

\[ \Delta G(x) = \sum_{i} \alpha_i K(x^*_i, x) \]

Predicted free energy

Training points

Query sample

\[ K(x_i, x_j) = \exp \left( -\frac{|x_i - x_j|_1}{\sigma} \right) \]

Training data

\( x = \begin{pmatrix} \Delta G_{W\rightarrow O1}^{(1)} \\ \Delta G_{W\rightarrow O1}^{(2)} \\ \Delta G_{W\rightarrow O1}^{(3)} \end{pmatrix} \)

Predictions

~hydrophobic

~polar

Extrapolation?

Train Points

Hoffmann, Menichetti, Kanekal, Bereau, Phys Rev E 100 (2019)
Physics offers controlled exploration of chemical space

Existing data            Machine learning

![Graph showing distributions](image)

**FIG. 2.** Distribution functions of confidence intervals, with trimers. We observe larger deviations in the training set. As a measure of homogeneity between training and validation sets, Fig. S3. Remarkably, the prediction of 0.2 kcal/mol for the ML and linear fit, respectively, across out-of-sample mean absolute errors (MAE) as low as trained on most of the MC-sampled data, we obtained of their salient properties. [See Methods]. When backmapped (Fig. 4B-ML, on par with the 0.12%

**FIG. 3.** (a) Transfer free energies from water to interface denoted by the transfer free energy, $G_{W \rightarrow I}$, and $G_{W \rightarrow M}$, where $W$ is the bilayer midplane, $M$ is the membrane-water interface, and $I$ is the membrane-water interface. We link these quantities in terms of transfer free energies of interest between the three state points (red circles): bilayer midplane (“M”), membrane-water interface (“I”), and membrane (“W”). (b) Background: Simulation setup of a solute (yellow) partitioning between water (not shown) and the lipid bilayer (blue). The lipid bilayer is composed of dioleoyl-phosphatidylcholine (DOPC) solvated with water. The solute interacts with the lipid bilayer through a potential of mean force (PMF) along the normal distance to the bilayer, which is sampled. Foreground: PMF along the normal of the bilayer.

**RESULTS**

We systematically backmapped (Fig. 4B-ML, on par with the 0.12% accuracy compared to Auto-Martini). The linear behavior displayed across both samplings is excellent, with a correlation coefficient above 0.97. For all three models, we observe excellent performance.

We ran MC simulations across CG linear trimers and tetramers (results for tetramers are shown in the SI), whereas the MC algorithm yielded an acceptance ratio of 0.2. We found negligible changes in the free energies regardless of the sampled compound, with the randomness of changing a bead type. We then randomly changing a bead type, calculating the relative partitioning between water and the bilayer midplane, such that $G_{W \rightarrow M}$.

We also used importance sampling (IS) to yield an acceptance ratio of 0.2 using IS. The IS criterion aimed at selecting compounds that favor partitioning free energy, $G_{W \rightarrow M}$, to the database. The database is provided as supporting information. The linear behavior displayed across both samplings is excellent, with a correlation coefficient above 0.97. For all three models, we observe excellent performance.

We used importance sampling (IS) to yield an acceptance ratio of 0.2 using IS. The IS criterion aimed at selecting compounds that favor partitioning free energy, $G_{W \rightarrow M}$, to the database. The database is provided as supporting information. The linear behavior displayed across both samplings is excellent, with a correlation coefficient above 0.97. For all three models, we observe excellent performance.

**DISCUSSION**

The unprecedentedly large database of free energies obtained with the Auto-Martini in conjunction with IS was performed to test our model satisfying linear thermodynamic relations across different characteristics, e.g., more polar compounds that were not sampled by using machine learning (ML; see Fig. 1). This excellent transferability can be associated to the robustness of their salient properties [see Methods]. When backmapped (Fig. 4B-ML, on par with the 0.12% accuracy compared to Auto-Martini). The linear behavior displayed across both samplings is excellent, with a correlation coefficient above 0.97. For all three models, we observe excellent performance.

**CONCLUSION**

In conclusion, we systematically backmapped (Fig. 4B-ML, on par with the 0.12% accuracy compared to Auto-Martini). The linear behavior displayed across both samplings is excellent, with a correlation coefficient above 0.97. For all three models, we observe excellent performance.

Hoffmann, Menichetti, Kanekal, Bereau, Phys Rev E 100 (2019)
Physics offers controlled exploration of chemical space

Existing data

Machine learning


Physics offers controlled exploration of chemical space

- Thermodynamics: global constraint
- CG representation: better ML transferability
- Boosted database: 1.36M compounds


Conclusions
Conclusions

Force fields across chemical space
- Physics encode long-range interactions
- Symmetries reduce the interpolation space
- Learning from QM (atomistic) or physicochemical properties (CG)

High-throughput thermodynamics
- Coarse-graining reduces the size of chemical space: fewer simulations
- Structure-property relationships: CG suggests low-dimensional representation
Acknowledgments

Group members

- Marius Bause
- Yasemin Bozkurt
- Alexia Centi
- Arghya Dutta
- Martin Girard
- Christian Hoffmann
- Kiran Kanekal
- Chan Liu

- Roberto Menichetti
- Bernadette Mohr
- René Scheid
- Clemens Rauer
- Joseph F. Rudzinski
- Marc Stieffenhofer
- Timon Wittenstein
- Svenja Wörner

Collaborators

- Robert DiStasio Jr. (Cornell)
- Anatole von Lilienfeld (Basel)
- Alexandre Tkatchenko (Luxembourg)
- Dirk Schneider (JGU Mainz)
- Michael Wand (JGU Mainz)

- Kurt Kremer
- Denis Andrienko
- Sapun Parekh
- Tanja Weil

CARL ZEISS STIFTUNG
CHEMAXON
MAX PLANCK INSTITUTE FOR POLYMER RESEARCH
MAX PLANCK RESEARCH NETWORK ON BIG-DATA-DRIVEN MATERIALS SCIENCE
TRR 146
EMMY NOETHER-PROGRAMM