Data-driven equation for drug–membrane permeability across drugs and membranes

Arghya Dutta,[1, a)] Jilles Vreeken,[2] Luca M. Ghiringhelli,[3] and Tristan Bereau[4, 1]

[1)] *Max Planck Institute for Polymer Research, Mainz, Germany*
[2)] *CISPA Helmholtz Center for Information Security, Saarbrücken, Germany*
[3)] *NOMAD Laboratory, Fritz Haber Institute of the Max Planck Society, Berlin, Germany*
[4)] *Van 't Hoff Institute for Molecular Sciences and Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands*

(Dated: 4 December 2020)

Drug efficacy depends on its capacity to permeate across the cell membrane. We consider the prediction of passive drug–membrane permeability coefficients. Beyond the widely recognized correlation with hydrophobicity, we apply sure-independence screening and sparsifying operator (SISSO), a data-driven compressed-sensing technique, to a large (0.4 million compounds) database of coarse-grained computer simulations as a way to also incorporate the role of acidity. We rationalize our derived equation by means of an analysis of the inhomogeneous solubility-diffusion model in several asymptotic acidity regimes. We further extend our analysis to the dependence on lipid-membrane composition. Lipid-tail unsaturation plays a key role: we report a permeability ratio between liquid-disordered (Ld) and liquid-ordered (Lo) domains of roughly 25, largely independent of the chemistry of the drug. They confirm the role of membrane surface-density fluctuations in passive permeation. Together, compressed sensing with analytically derived asymptotes establish and validate an accurate, broadly applicable, and interpretable equation for passive permeability across both drug and lipid-tail chemistry.

I. INTRODUCTION

Passive lipid-membrane permeation remains of great relevance for pharmaceutical applications and an improved physico-chemical understanding of small-sized molecules in complex biological materials.[1,2] The technological implications of the problem has sustained the need for experiment- and simulation-free prediction of passive permeation that are rapid, inexpensive, and accurate.[3,4] Various types of surrogate models have been proposed over the years, the field having adopted machine learning early on.[5] While modern deep-learning approaches take advantage of unchallenged model expressivity to offer unprecedented prediction accuracy, they suffer from two important aspects:

1. Overfitting: The size of chemical space of drug-like molecules is overwhelmingly large ($\sim 10^{60}$ compounds).[6] Deep surrogate models need large numbers of parameters to establish complex relationships. Unfortunately the body of experimentally-available data is minuscule compared to the size of chemical space. This can lead to surrogate models that shift dangerously upon addition/removal of small numbers of compounds in the training set. The problem is aggravated by databases that are often proprietary, preventing broad availability and reproducibility. Relying on different measurement batches may also accentuate systematic errors.

2. Lack of interpretability: Surrogate models are oftentimes black-box techniques that typically cloud *why* a certain prediction has been made. Deep neural networks exhibit an overwhelming number of parameters and rely on highly non-linear hierarchical functions, making them nearly impossible to conceptually grasp.[7] Quantitative structure–activity relationship (QSAR) models can build multivariate models, but the combination of too many descriptors will lead to similar difficulties. Beyond predicting individual data points, we seek to gain further insight. Insight is essential, for instance as a stepping stone to solving the inverse problem, thereby establishing structure–property relationships and enabling compound design.

In this work we address these points using a combination of approaches: First, we consider a nearly exhaustive database of permeability coefficients for a subset of the chemical space of small organic molecules. These coefficients are extracted from physics-based, coarse-grained computer simulations. The resolution of the coarse-grained model offers an accurate reproduction of the system's thermodynamics—the key driving force—combined with a 3-order-of-magnitude speedup compared to atomistic simulations. The high-throughput coarse-grained (HTCG) simulations yielded a large database of permeability coefficients, together with potentials of mean force and acidity constants, for 511 427 compounds.[8] The database contains a nearly exhaustive subset of small organic molecules in the range 30–160 Da, thereby ensuring a dense coverage of the chemical space in this subset.

As for the data-driven model, we explicitly avoid build-

a)Electronic mail: dutta@mpip-mainz.mpg.de

ing and using a black box, and instead turn to learning an *equation.* In particular, we will rely on recently proposed data-driven techniques to discover equations.[9,10] Equations relevant to physical problems often display simplifying properties, such as symmetries or separability easing both their data-driven discovery and generalization beyond the training set. Several studies have demonstrated the ability to (re)discover physics equations.[11,12] Generalization capabilities are critical: typical training datasets are minuscule relative to the size of chemical compound space, such that overfitting can easily prevail. To this end, we follow Occam's razor and limit the complexity of the equations we consider. Beyond the generalization aspects, it also helps interpretability: rationalizing the derived equation.

Passive drug permeation measures the propensity of a solute to spontaneously cross a lipid membrane. Upon doing so, the solute interacts with a great variety of physicochemical environments—from an aqueous phase to a hydrophobic membrane core. Permeation is quantified by means of its coefficient, P, as the steady state flux of the solute across the soft interface. Early on, Meyer[13] and Overton[14] modeled passive permeation as diffusion across a homogeneous slab via $P = KD/\sigma$, where K and D are the water-membrane partitioning coefficient and diffusivity of the compound, respectively, and σ is the thickness of the bilayer core. K is typically approximated by a simpler proxy: the partitioning coefficient between water and octanol. Water/octanol partition, or more generally hydrophobicity, has long been identified as strongly correlating with membrane permeability.[15] Notable refinements to the homogeneous Meyer-Overton rule include the inhomogeneous solubility–diffusion model (ISDM), estimating the permeability coefficient via an integral over the membrane extension, z, of its potential of mean force (PMF), $G(z)$, as

$$P^{-1} = \int \mathrm{d}z \, R(z) = \int \mathrm{d}z \, \frac{\exp[\beta G(z)]}{D(z)}, \qquad (1)$$

where $R(z)$ and $\beta = 1/k_\mathrm{B}T$ correspond to an associated resistivity and the inverse temperature, respectively.[16] The competition between different protonation states naturally follows the sum of inverse resistivities, analogous to the total resistance in a parallel electrical circuit.[17] PMFs are shifted according to the difference between the solution's pH and the compound's acid dissociation constant, $\mathrm{p}K_\mathrm{a}$. In the following we take the perspective of a neutral compound, which can deprotonate (acidic, $\mathrm{ap}K_\mathrm{a}$) or protonate (basic, $\mathrm{bp}K_\mathrm{a}$). Knowledge of the PMF(s) and the diffusivity thereby fully determines the permeability coefficient. Unfortunately these (i) are so far only available by computational techniques and (ii) typically require extensive calculations ($\sim 10^5$ CPU-hour per system at an atomistic resolution).[18,19]

We apply SISSO,[20,21] which combines symbolic regression with compressed sensing,[22–24] to discover a (possibly approximated) equation of Eq. 1 that considers the effect of hydrophobicity, but also explicitly incorporates acidity via the compound's protonation state. We will discuss several equations of various complexities to illustrate the balance between accuracy and interpretability. Remarkably, the simplest variant can be validated against analytical acid-base asymptotic regimes. This simple model incorporating both hydrophobicity and acidity allows us to easily extend our analysis to different lipid membranes starting from limited information. From knowledge of neutral species alone, we predict the change in passive permeability in various lipid membranes. We finally discuss the change in permeability in the context of membrane phase behavior.

II. METHODS

A. Database of drug–membrane thermodynamics

We rely on the passive-permeability database provided by Menichetti *et al.*[8] It consists of the potential of mean force, acid dissociation constant for acids and bases ($\mathrm{ap}K_\mathrm{a}$ and $\mathrm{bp}K_\mathrm{a}$), and the order of magnitude of the permeability coefficient ($\log_{10} P$) for $511\,427$ small drug molecules through a single-component bilayer made of 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC). Small molecules in the range 30–160 Da were mapped to a one- or two-bead coarse-grained representation, i.e., a monomer or a dimer, of the coarse-grained Martini model.[25–28] Enhanced-sampling molecular dynamics simulations yielded the PMFs for both neutral and (de)protonated species, ultimately leading to the permeability coefficient, P. The $\mathrm{p}K_\mathrm{a}$ of a chemical group can be either acidic ($\mathrm{ap}K_\mathrm{a}$) or basic ($\mathrm{bp}K_\mathrm{a}$), in that a neutral compound can either deprotonate or protonate (see SI for definitions). While the ionization constant of conjugated acid/base pairs typically lie between 10^{-2} and 10^{16}, we considered compounds with $\mathrm{p}K_\mathrm{a}$ values between -10 and 20.[29] This led to a dataset of $418\,828$ compounds used as part of this work. A follow up work to Menichetti *et al.* provided PMFs for the same set of neutral Martini small molecules in different phosphocholine (PC) lipid membranes: 1,2-diarachidonoyl-*sn*-glycero-3-PC (DAPC); 1,2-dilinoleoyl-*sn*-glycero-3-PC (DIPC); 1,2-dilauroyl-*sn*-glycero-3-PC (DLPC); 1,2-dioleoyl-*sn*-glycero-3-PC (DOPC); 1,2-dipalmitoyl-*sn*-glycero-3-PC (DPPC); and 1-palmitoyl-2-oleoyl-*sn*-glycero-3-PC (POPC).[30]

B. Learning algorithm

To learn an interpretable model of passive permeability, we used the recently proposed sure-independence screening and sparsifying operator (SISSO).[20,21] We relied on the open-source implementation of Ouyang.[31] SISSO aims at establishing a functional relationship,

$y = f(\Phi)$, between n primary features, $\Phi_0 = \{\phi_1, \phi_2, \cdots, \phi_n\}$, and a target property, y, based on N training compounds. SISSO assumes that y can be reliably expressed as a linear combination of non-linear, but closed-form, functions of primary features. To construct these non-linear functions, SISSO recursively operates a set of user-defined unary and binary operators (we used $\{+, -, \times, \div, \exp, \log, ()^{-1}, ()^2, ()^3, \sqrt[3]{()}, \sqrt{()}\}$) on the primary features and builds up sets of secondary features. Φ_q denotes the set of secondary features at each level of recursion q. The number of secondary features in Φ_q increases sharply with increase in the recursion level q, the number of operators used, and the number of primary features n. At each q, SISSO selects a subset of features that features the strongest linear correlations with the target y (see Ouyang *et al.*[21]). To keep the number of terms low, a linear combination of secondary features is combined with a sparsifying ℓ_0 regularization. The number of constituent secondary features is herein denoted as its dimension. At each level, SISSO proposes a large number of trial functions, which we rank according to their root mean-squared error (RMSE). We also quantify model performance using the maximum absolute error (MaxAE) and the square of the Pearson correlation coefficient, r^2.

C. Feature construction and training

We apply SISSO to three easily accessible and interpretable primary features: the water-octanol partitioning free energy, $\Delta G_{W \to Ol}$, and the acid dissociation constants apK_a and bpK_a. We thereby seek a refinement or correction to the commonly used model based on hydrophobicity alone.[15] To avoid constructing features with different units, we multiply the partitioning free energy by the inverse temperature: $\beta \Delta G_{W \to Ol}$, using $T = 300\,\mathrm{K}$ following Menichetti *et al.*[8] Starting with the set of primary features $\Phi_0 = \{\beta \Delta G_{W \to Ol}, apK_a, bpK_a\}$, we consider the construction of secondary features for up to two iterations (i.e., $q = 2$), where Φ_1 and Φ_2 consist of roughly 30 and 2 000 features, respectively. We limit the SISSO screening size to 500 and consider up to three-dimensional descriptors. We train on 10% of the available data, and use the remaining 90% for hold-out evaluation. We draw these train/test sets uniformly at random, without replacement. To reduce variance, we report the average performance over ten independently drawn train/test sets. Ouyang *et al.*[20] emphasized that SISSO works reliably when the number of materials in a training set, N, sufficiently exceeds the number of features. In particular, SISSO requires $N \geq kq \ln(\#\Phi_q)$, where $k \sim 1 - 10$ and $\#\Phi_q$ is the size of the feature space at recursion level q. For Φ_2, the relevant feature space used to train our model, the relation requires $N \geq 10 \cdot 2 \cdot \ln(2 \cdot 10^3) \simeq 150$. By training our models with only 10% of the dataset ($\sim 42\,000$ compounds), SISSO is well within a regime to provide meaningful and consistent results.

III. RESULTS AND DISCUSSIONS

A. Learning permeability models

Table I contains the four models considered in this work: (*i*) f^{Hyd} is a baseline hydrophobicity model, which linearly correlates with water/octanol partitioning free energy; and (*ii-iv*) f^{1D} to f^{3D} linearly correlate with one to three secondary feature(s) identified by SISSO. For each model, c_i^{m} correspond to non-zero coefficients for model m and index i, reported in Tab. I. For all models, $\Delta G_{W \to Ol}$ takes on a central role, as expected by the performance of the baseline. The simplest SISSO model, $f^{\mathrm{1D}} = c_0^{\mathrm{1D}} + c_1^{\mathrm{1D}}(apK_a - bpK_a - 2\beta \Delta G_{W \to Ol})$, is remarkably robust: it is systematically identified as the best performing 1D model across all 10 training sets. Given that we trained on only 10% ($\sim 42\,000$ compounds) of the dataset, this highlights this model's performance compared to all other candidates (see SI for a list of candidate one-dimensional models). Its improved accuracy compared to the baseline will be evaluated further down.

We also report more complex 2D and 3D models in Tab. I. While we will show below that these yield even better accuracy compared to f^{1D}, they are specifically tailored to the training set used: f^{2D} and f^{3D} are ranked as the best model in eight and five out of the 10 training sets, meaning that other models of similar complexity closely compete.

B. Model performance

We now turn to the performance of these four models. Tab. I reports their RMSE, MaxAE, and squared Pearson correlation coefficient, r^2, averaged over the test sets. Going from the baseline to more complex SISSO models, the systematic decrease in the RMSE is accompanied by an increase in the correlation coefficient. On the other hand, the maximum absolute error does show a clear minimum for f^{1D}. This offers a first hint at the appealing balance between generalization and interpretability of the 1D SISSO model. The performance of these four models is depicted in Fig. 1 for the entire dataset, where we report each model against reference values. Going from baseline to SISSO models of increasing complexity, the distribution does lean increasingly toward the $y = x$ correlation line. For the 2D and 3D models, we also point out outliers at the lowest permeability values. These points represent only 0.07% of the data, which likely was not part of the training. This apparent overfitting aspect is made clear by our small training-set fraction, and is more likely to appear for complex models. The presence of horizontal stripes in Fig. 1 results from the use of the transferable coarse-grained Martini model to calculate permeability. A large number of chemically similar compounds map down to a single coarse-grained representation, thus yielding the

TABLE I. Permeability models, descriptor coefficients, and model performance: RMSE, MaxAE (both in $\log_{10}$ units of the permeability coefficient), and r^2. The models considered are: baseline hydrophobicity model, f^{Hyd}, as well as SISSO with up to 3 feature dimensions: $f^{1\mathrm{D}}$, $f^{2\mathrm{D}}$, and $f^{3\mathrm{D}}$. Compared to the baseline, SISSO systematically leads to more accurate models. Descriptor coefficients and performance metrics are averaged over training and test sets, respectively. All standard errors are small and reported in the SI.

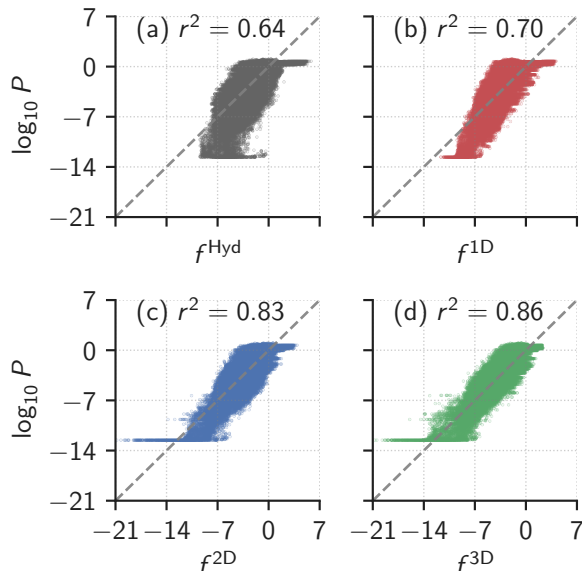| Model		c_0	c_1	c_2	c_3	RMSE	MaxAE	r^2
f^{Hyd}	$= c_0^{\mathrm{Hyd}} + c_1^{\mathrm{Hyd}}\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}$	-3.444	-0.648			1.53	11.82	0.64
$f^{1\mathrm{D}}$	$= c_0^{1\mathrm{D}} + c_1^{1\mathrm{D}}(\mathrm{ap}K_a - \mathrm{bp}K_a - 2\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}})$	-5.419	0.163			1.40	6.35	0.70
$f^{2\mathrm{D}}$	$= c_0^{2\mathrm{D}} + c_1^{2\mathrm{D}}(\sqrt[3]{\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}} + \beta\Delta G_{\mathrm{W}\to\mathrm{Ol}} - \mathrm{ap}K_a)$ $+ c_2^{2\mathrm{D}}(\mathrm{ap}K_a^2 + \mathrm{bp}K_a^2)$	-5.753	-0.487	-0.017		1.06	8.28	0.83
$f^{3\mathrm{D}}$	$= c_0^{3\mathrm{D}} + c_1^{3\mathrm{D}}(\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}} - \mathrm{ap}K_a)$ $+ c_2^{3\mathrm{D}}(\mathrm{bp}K_a^2(\mathrm{ap}K_a + \mathrm{bp}K_a))$ $+ c_3^{3\mathrm{D}}(\mathrm{ap}K_a^2 + (\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}})^2)$	-7.101	-0.614	-0.001	-0.018	0.94	8.19	0.86

same permeability coefficient.[28]

FIG. 1. Performance of the four permeability models against the reference for the dataset of 418k small molecules: (a) baseline hydrophobicity model f^{Hyd}; (b-d) 1D to 3D data-driven SISSO models. See Tab. I for their expressions.

To better understand the performance of each model, we analyze their errors in greater detail. Fig. 2a shows the distribution of absolute error. The large dataset at our disposal allows us to evaluate more than 4 orders of magnitude of this distribution. The comparison between f^{Hyd} and $f^{1\mathrm{D}}$ proves insightful: while they are remarkably close up to errors of 5 $\log_{10}$ units, the baseline then displays a significant hump, while the SISSO 1D model keeps decaying monotonously. Both models rely on $\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}$, which explains the remarkable agreement early on, while the stark difference between the two curves is entirely due to the effect of acidity. This is confirmed by a further decomposition of the error as a function of acidity, showing that f^{Hyd} leads to larger errors for stronger acids and bases (Panels b and f), while SISSO 1D significantly reduces the error in this regime (Panels c and g).

In comparison, the more complex SISSO 2D and 3D display a shift in the error distribution toward lower errors (Fig. 2a) compared to the 1D model. At low probability however, we observe a significant change in the slope of the decay, indicating worse performance for a small number of outliers. This is also illustrated when decomposing the error in terms of acidity in Panels (d-i): while the overall performance improves, we identify more outliers. These outliers mostly lie at low permeability values (reminiscent of Fig. 1), and for strong $\mathrm{ap}K_a$ or $\mathrm{bp}K_a$. Poor performance at large acidity values could take place if these were absent of the small training fraction.

C. Validation against atomistic simulations

The SISSO models we learn may naturally be prone to systematic errors inherent to the training dataset. Reference permeability values were extracted from computationally efficient coarse-grained computer simulations, at the cost of force-field accuracy. Still, a comparison of the coarse-grained simulations against atomistic computer simulations had indicated an excellent agreement of 1 $\log_{10}$ unit across a limited set of small molecules[8]. Here we compare the performance of the four permeability models against the atomistic simulations of Carpenter et al.[17] This set of 12 organic compounds covers a range of molecular weights that goes significantly beyond our training set: an average of 243 Da and up to 319 Da, while the database contains information only up to 160 Da. This thus presents a challenging test for the generalizability of the SISSO models.

Fig. 3 shows the absolute error against atomistic simulations across all 12 small molecules and for our four models. For each model we display the error as a function of both $\mathrm{ap}K_a$ and $\mathrm{bp}K_a$. The baseline model yields absolute errors between 1 and almost 4 $\log_{10}$ units. While larger errors correlate with strong acids, they do not seem
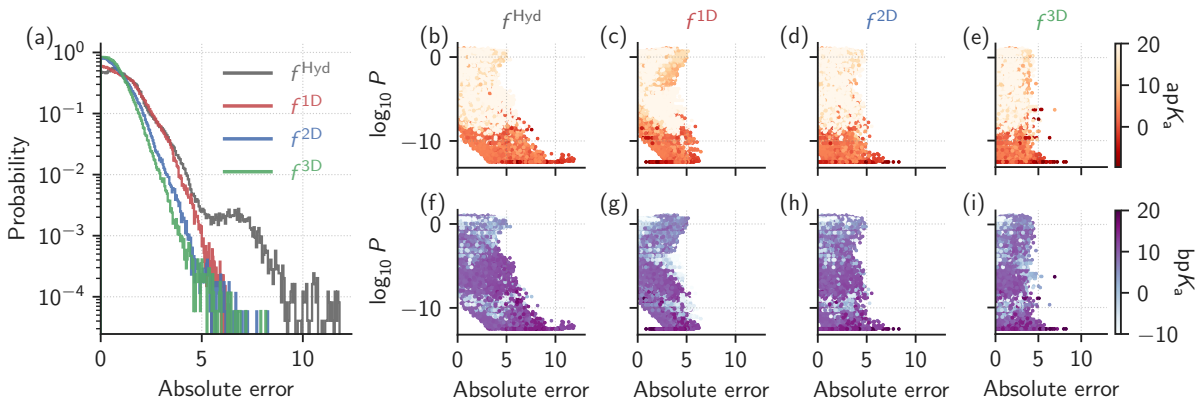
FIG. 2. Absolute error analysis. (a) Error (in $\log_{10}$ units) distribution for all models; Error decomposed as a function of (b-e) apK_a and (f-i) bpK_a. Stronger acids/bases are shown in darker colors.
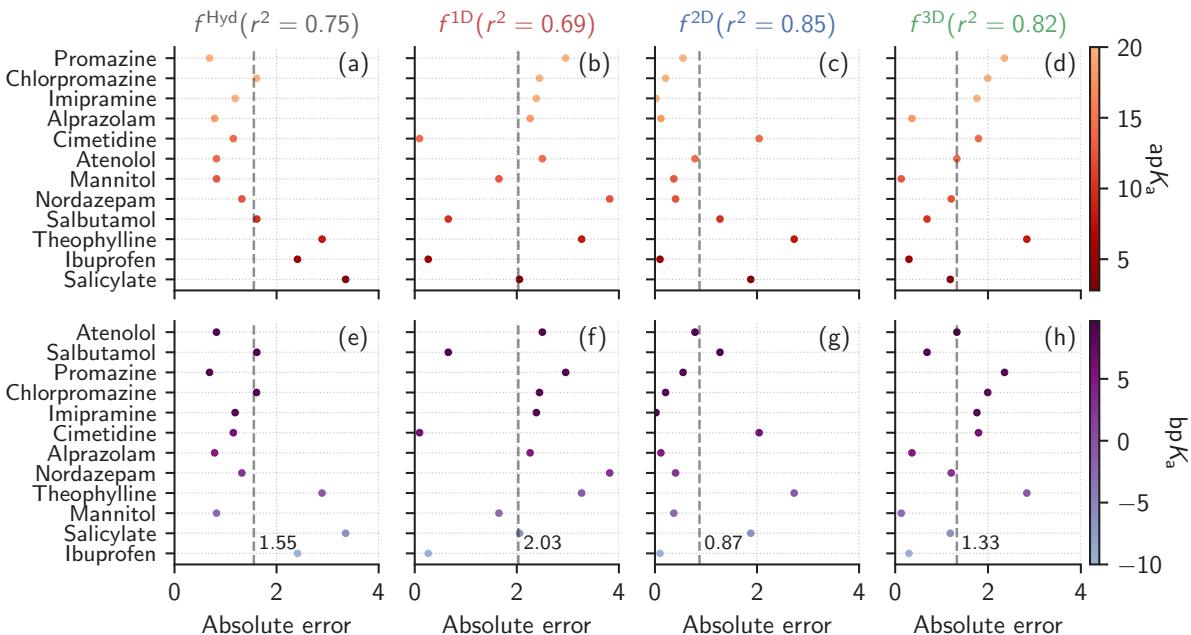

FIG. 3. Absolute error against atomistic simulations for 12 reference small molecules.[17] The error is displayed for all four models and as a function of (a-d) apK_a and (e-h) bpK_a.

to correlate with larger bases. Unlike Fig. 2, the minuscule set of atomistic compounds prevents us from drawing conclusions that would hold across any significant subset of chemical space. Turning to SISSO 1D, we observe a small but noticeable decrease in performance, where the mean absolute error (MAE) increases from 1.55 to 2.03 $\log_{10}$ units. The MAE, however, decreases against the baseline when considering the more complex SISSO 2D and 3D: 0.87 and 1.33 $\log_{10}$ units, respectively. The non-monotonic decrease of the MAE with increasing complexity in Fig. 3b–d suggests the role of the small validation dataset considered. Overall though, the incorporation of acidity does lead to an improved reproduction of the permeability coefficient. It validates the SISSO-derived

equations on permeability coefficients derived using independent methods and outside the chemical space of the training data.

D. Acid-base asymptotes

The analysis so far highlights how model complexity impacts accuracy. Missing from the analysis so far is the consideration of interpretability. The two one-dimensional models—the baseline and SISSO 1D—highlight a simple mechanism as to the functional dependence of the permeability coefficient on both hydrophobicity and acidity. Focusing on SISSO 1D specifically, we

rewrite the model in terms of two contributions

$$f^{1D} = c_0^{1D} + c_1^{1D}\big[\,(\mathrm{ap}K_a - \beta\Delta G_{W\to Ol})$$
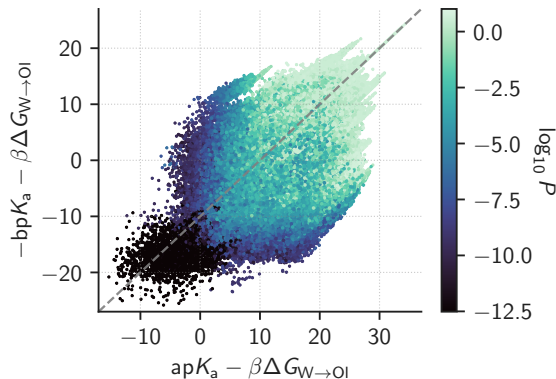$$+ (-\mathrm{bp}K_a - \beta\Delta G_{W\to Ol})\,\big]. \quad (2)$$

FIG. 4. SISSO 1D model of passive permeability. The decomposition of the secondary feature in two axes highlights the role of hydrophobicity (along the diagonal) compared to acidity (vertical and horizontal).

Fig. 4 displays the permeability coefficient as a function of these two contributions. The symmetric contribution of $\beta\Delta G_{W\to Ol}$ in the two terms of Eq. 2 indicates that the baseline hydrophobicity model manifests itself along the diagonal. Notably missing from the diagonal behavior are the dark vertical and horizontal basins. They localize at $\mathrm{ap}K_a - \beta\Delta G_{W\to Ol} \sim 0$ and $-\mathrm{bp}K_a - \beta\Delta G_{W\to Ol} \sim -15$, and represent strong-acid and strong-base regimes. In what follows we provide an asymptotic rationalization of the functional form of Eq. 2.

To rationalize Eq. 2, we first outline the role of our three primary descriptors in the ISDM model (Eq. 1). Fig. 5 illustrates the well known interplay between PMF and acidity, in particular how the latter shifts the PMFs of the neutral and (de)protonated species. In the following we will denote the PMFs of the neutral, protonated, and deprotonated species as $G_N(z)$, $G_B(z)$, and $G_A(z)$, respectively.

The difference between $\mathrm{ap}K_a$ or $\mathrm{bp}K_a$ and pH dictate the propensity for the PMFs to cross each other. The ISDM relies on a total resistivity (defined in Eq. 1), R_T, such that $R_T^{-1}(z) = R_N^{-1}(z) + R_B^{-1}(z) + R_A^{-1}(z)$, analogous to the total resistance in a parallel electric circuit.

The PMFs of the neutral, protonated, and deprotonated species can be linked in water thanks to their $\mathrm{ap}K_a$ and $\mathrm{bp}K_a$ values, as well as the pH of the environment, through the following equations

$$\beta^{-1}(\mathrm{pH} - \mathrm{ap}K_a)\ln 10 = G_N(\infty) - G_A(\infty), \quad (3)$$
$$\beta^{-1}(\mathrm{pH} - \mathrm{bp}K_a)\ln 10 = G_B(\infty) - G_N(\infty), \quad (4)$$

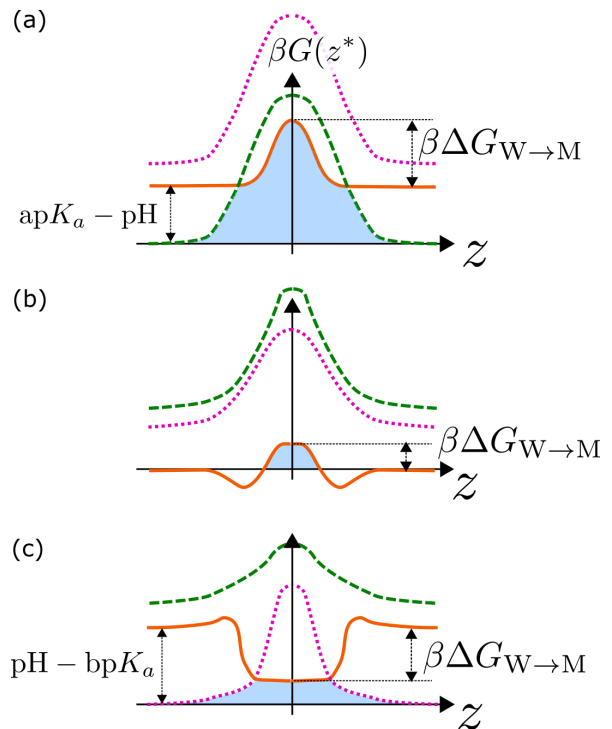where $G(\infty)$ indicates that the compound is located in

FIG. 5. Sketch of the permeation mechanism in three regimes: (a) strong acid, (b) neutral compound, and (c) strong base. The curves display the neutral (solid orange), acidic (green dashed), and basic (pink dotted) PMFs. The blue area under the effective PMF directly links to the permeability coefficient.

bulk water. Eqs. 3 and 4 effectively link the difference between the pH of the environment with the pK_a of the compound to a shift in the PMFs. Without loss of generality, we will shift all free energies such that zero corresponds to the most favorable compound in bulk water. Eqs. 3 and 4 allow us to explicitly link pK_a information with the total resistivity

$$R_T^{-1}(z) = D(z)[e^{-\beta(G_N(z)-G_N(\infty))}$$
$$+ e^{-\beta(G_B(z)-G_B(\infty))}$$
$$+ e^{-\beta(G_A(z)-G_A(\infty))}], \quad (5)$$

where we assume that all protonation states yield identical diffusivity.[8] Because Eq. 5 takes on a relatively complex form, we will consider only asymptotic regimes:

- Neutral compounds entail no strong acid or base characteristic, i.e., $\mathrm{ap}K_a \gg \mathrm{pH}$ and $\mathrm{bp}K_a \ll \mathrm{pH}$, such that the compound is effectively unable to (de)protonate, and $G_N(\infty) = 0$. Eq. 5 can be simplified to $R_T(z) \approx D^{-1}(z)\exp[\beta G_N(z)]$.[32]

- Strong acids consist of solutes that display at least one functional group for which $\mathrm{ap}K_a \ll \mathrm{pH}$. For neutral pH ($\approx 7$), this would correspond to $\mathrm{ap}K_a < 4$. We set $G_A(z \to \infty) = 0$. In this regime the third exponential in Eq. 5

would dominate the other two, leading to $R_\mathrm{T}(z) \approx D^{-1}(z) \exp\left[\beta G_\mathrm{N}(z) + (\mathrm{pH} - \mathrm{ap}K_\mathrm{a}) \ln 10\right]$.

- Strong bases would display at least one functional group where $\mathrm{bp}K_\mathrm{a} \gg \mathrm{pH}$. For neutral pH, this would correspond to $\mathrm{bp}K_\mathrm{a} > 10$. We set $G_\mathrm{B}(z \to \infty) = 0$. Using a similar argument, Eq. 5 would be dominated by the second exponential, leading to $R_\mathrm{T}(z) \approx D^{-1}(z) \exp\left[\beta G_\mathrm{N}(z) - (\mathrm{pH} - \mathrm{bp}K_\mathrm{a}) \ln 10\right]$.

The total resistivities still require integration over the reaction coordinate z, which we simplify to the largest contribution of the PMF.[19] The effective resistivity model is equivalent to choosing the lowest PMF at any value of z: $G_\mathrm{eff}(z) = \min_i G_i(z)$, where i runs over the neutral, protonated, and deprotonated species. In addition, the dominating contribution of the effective permeability will come from its maximum value, corresponding to a position $z^* = \arg\max_z G_\mathrm{eff}(z)$. Assuming that the largest contribution of the permeability arises from the total resistivity at z^*, we obtain $P \approx R_\mathrm{T}^{-1}(z^*)$. This yields the following acid-base asymptotic regimes

$$\log_{10} P \approx \begin{cases} \log_{10} D(z^*) - \frac{\beta}{\ln 10} G_\mathrm{N}(z^*) - \mathrm{pH} + \mathrm{ap}K_\mathrm{a} & \text{if } \mathrm{ap}K_\mathrm{a} \ll \mathrm{pH}, \\ \log_{10} D(z^*) - \frac{\beta}{\ln 10} G_\mathrm{N}(z^*) + \mathrm{pH} - \mathrm{bp}K_\mathrm{a} & \text{if } \mathrm{bp}K_\mathrm{a} \gg \mathrm{pH}, \\ \log_{10} D(z^*) - \frac{\beta}{\ln 10} G_\mathrm{N}(z^*) & \text{if } \mathrm{ap}K_\mathrm{a} \gg \mathrm{pH} \text{ and } \mathrm{bp}K_\mathrm{a} \ll \mathrm{pH}. \end{cases} \quad (6)$$

To numerically test Eq. 6, we identify datapoints corresponding to the three asymptotic regimes: the neutral compounds ($\mathrm{ap}K_\mathrm{a} > 10$ and $\mathrm{bp}K_\mathrm{a} < 4$), strong acids ($0 < \mathrm{ap}K_\mathrm{a} < 4$ and $\mathrm{bp}K_\mathrm{a} < 4$), and strong bases ($10 < \mathrm{bp}K_\mathrm{a} < 14$ and $\mathrm{ap}K_\mathrm{a} > 10$). For simplicity, we only considered non-zwitterionic compounds. We assume that $\log_{10} D(z)$ yields no significant, chemically specific effect, and uniformly shift the permeability coefficient across the chemical space of compounds considered. Fig. 6 shows the agreement between Eq. 6 and the reference permeability coefficients. All show strong linear correlation: for neutral compounds ($r^2 = 0.998$), strong acids ($r^2 = 0.959$), and strong bases ($r^2 = 0.986$). These results numerically validate the asymptotes of Eq. 6.

More importantly, the asymptotes provide a physically motivated rationale for the two contributions of Eq. 2: $\mathrm{ap}K_\mathrm{a} - \beta \Delta G_\mathrm{W \to Ol}$ and $-\mathrm{bp}K_\mathrm{a} - \beta \Delta G_\mathrm{W \to Ol}$. We first note that $\Delta G_\mathrm{W \to Ol}$ is related to $G_\mathrm{N}(z^*)$. The depth at which the effective PMF is the highest, z^*, will almost always be close to the membrane midplane: $z^* \approx 0$. The main exception to this are hydrophobic compounds, for which the highest point in the PMF is in water (Fig. 5c). Furthermore, $G(z^* = 0)$, which corresponds to the transfer free energy from water to membrane midplane, has been shown to linearly relate to the water-octanol partitioning free energy, $\Delta G_\mathrm{W \to M} \propto \Delta G_\mathrm{W \to Ol}$.[33] This connects $G_\mathrm{N}(z^*)$ present in Eq. 6 to $\Delta G_\mathrm{W \to Ol}$ in Eq. 2. We then observe that the asymptotes and Eq. 2 have the same signs and exponents of $\mathrm{ap}K_\mathrm{a}$ and $\mathrm{bp}K_\mathrm{a}$. For an acidic or a basic compound, if we consider the compound-specific $\mathrm{p}K_\mathrm{a}$ and substitute $\Delta G_\mathrm{W \to Ol}$ by the compound's $G_\mathrm{N}(z^*)$ in the relevant among the two contributions of Eq. 2, we indeed recover one the asymptotes. As for neutral compounds, $G_\mathrm{N}(z^*)$ is the only relevant quantity while estimating permeability.

E. Drug–membrane permeability across membranes

The following analyzes how drug–membrane permeability changes according to membrane composition. We hypothesize that the functional form of SISSO 1D is applicable to other lipid membranes, and use it as a starting point. We take advantage of the above-mentioned asymptotic regimes to *limit* the amount of information needed from new membranes. The regime of neutral compounds described in Eq. 6 can be used advantageously because it only requires information on neutral PMFs. We rely on the dataset of Hoffmann *et al.*, which precisely contains PMF information—but no permeability—in various membranes, and only for neutral compounds.[30] We specifically analyze the change in permeability when turning to phosphocholine (PC) membranes made of different lipids, varying in both tail length and level of unsaturation.

Fig. 7 shows the relation of $-\beta G_\mathrm{N}(z^*)$—the dominant term for the permeability of neutral compounds (Eq. 6)—between the original membrane used in this work, DOPC, and others. All curves follow a line, indicating that the asymptotic regime for neutral compounds of Eq. 6 holds for all membranes. We find two families of lines with different intercepts: DLPC, DPPC, and POPC show an intercept with DOPC that is roughly 0, while DIPC and DAPC have an intercept that is approximately 1.4.

To better understand these results, we first recall that $G_\mathrm{N}(z^*)$ corresponds to the highest value of the neutral PMF. We can safely ignore contributions of the charged PMFs, such that $G_\mathrm{N}(z^*)$ denotes the highest point of the *effective* PMF. The excellent agreement between DLPC and DPPC indicates that tail length (3 and 4 beads long, respectively) does not impact the permeability. This is expected, given that tail length is only expected to change the length, but not the height, of the hydropho-
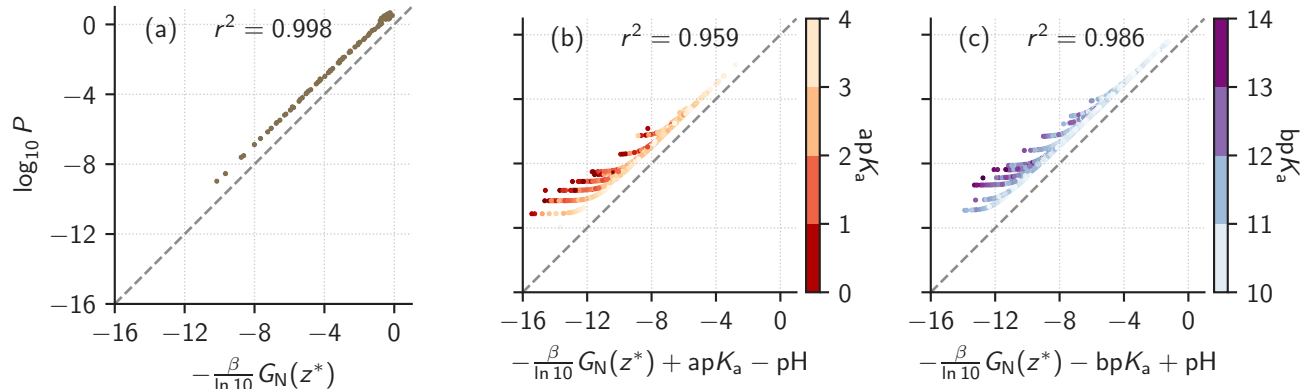
FIG. 6. Simple asymptotic estimates (Eq. 6) against reference permeability coefficients at neutral pH. (a) Neutral compounds; (b) Strong acids; and (c) Strong bases.
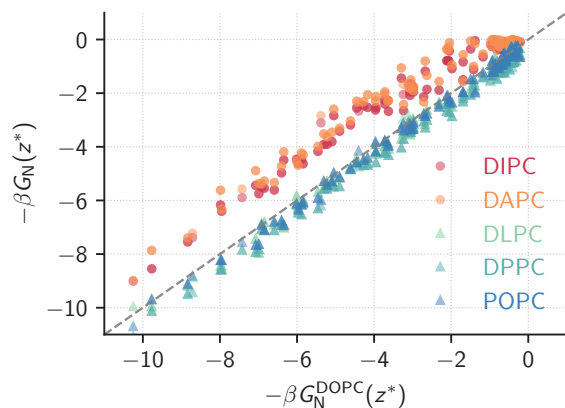

FIG. 7. Variation of $-\beta G_N(z^*)$ between the original membrane, DOPC, and others. Note the shift between (i) DLPC, DPPC, and POPC and (ii) DIPC and DAPC.

bic plateau. On the other hand, the further agreement between them and POPC and DOPC indicate a lack of dependence on tail saturation for these lipids (1 and 2 unsaturated beads, respectively). Remarkably, the shift in the intercept appears only for DIPC and DAPC—lipids that exhibit more unsaturations: 4 and 8, respectively. Notably, they display the *same* shift in $-\beta G_N(z^*)$. As such, the level of unsaturation is a determining factor, but does not gradually impact $-\beta G_N(z^*)$.

Interestingly, DIPC and DAPC are the only lipids in Martini capable of phase separating into a liquid-disordered (Ld) phase, when combined with DPPC and cholesterol.[34] On the other hand, the force field fails to phase separate when using POPC or even DOPC. The results on Fig. 7 thereby indicate a clear shift of $-\beta G_N(z^*)$ depending on the ability of the membrane to form an Ld domain. Drug–membrane permeability is thereby a direct function of the entropic character of the lipid-tail fluctuations.

While the force field incorrectly assigns lipids capable of stabilizing an Ld domain, the shift in permeabilities between liquid-ordered- (Lo) and Ld-forming membranes is realistic. The results shown in Fig. 7 translate to a ratio of the permeability coefficients between Ld and Lo roughly, $P(\text{Ld})/P(\text{Lo}) \approx 25$. Recent atomistic simulations and electron paramagnetic resonance (EPR) spectroscopy experiments probing the permeation of water and oxygen reported a ratio of 3.[35] Our results are thus within one $\log_{10}$ unit for this specific compound, and extend the trend across the chemical space of small drugs. Mechanistically this would emphasize the link between local membrane surface density (i.e., its propensity to form transient holes) and passive permeation.[36]

The shift observed in Fig. 7 suggests how to generalize the permeability equation (Eq. 2) across membranes. It only needs a simple additive constant that corrects the free-energy term to describe Lo-forming versus Ld-forming membranes. The correlation tapers off for $-\beta G_N(z^*) \approx 0$, but these correspond to compounds that are largely hydrophobic, for which solubility is likely an issue.

IV. CONCLUSIONS

Using a data-driven compressed-sensing analysis, we propose refinements to the common hydrophobicity model of passive permeability. Sure-Independence Screening and Sparsifying Operator (SISSO) builds a hierarchy of models of increasing complexity. Models prove increasingly accurate, yet more complex models are more prone to discrepancies for a few outliers. Our SISSO 1D model offers improved accuracy compared to the hydrophobicity baseline, and yet excellent interpretability. We identify the simple and interpretable equation

$f^{1D} = c_0^{1D} + c_1^{1D}(\mathrm{ap}K_\mathrm{a} - \mathrm{bp}K_\mathrm{a} - 2\beta\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}})$, where $\mathrm{ap}K_\mathrm{a}$ and $\mathrm{bp}K_\mathrm{a}$ characterize acidity, $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ is the water/octanol partitioning coefficient, and c_0^{1D} and c_1^{1D} are the only two fitting parameters. We rationalize the model by an analysis of the asymptotic regimes of the inhomogeneous solubility–diffusion model (ISDM). The asymptotes are validated numerically and confirm the SISSO 1D equation.

The separation of the ISDM in asymptotic regimes allows us to build drug–permeability models across membranes with limited information only. Using only the potential of mean force of neutral solutes, we infer the change in permeability for membranes with lipids of varying tail length and level of unsaturation. We report a change in permeability coefficient depending on the membrane's ability to form a liquid-disordered domain, linking to the entropic character of the lipid-tail fluctuations. The results are in excellent agreement with recent atomistic simulations and EPR experiments, and further extend to large subsets of chemical space. The approach offers a data-driven, interpretable analysis of drug–membrane passive permeability across both drugs and membranes.

[1] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, Advanced Drug Delivery Reviews 46, 3 (2001).

[2] A. Avdeef, *Absorption and drug development: solubility, permeability, and charge state* (John Wiley & Sons, 2012).

[3] L. Di and E. H. Kerns, *Drug-like properties: concepts, structure design and methods from ADME to toxicity optimization* (Academic press, 2015).

[4] F. Broccatelli, L. Salphati, E. Plise, J. Cheong, A. Gobbi, M.-L. Lee, and I. Aliagas, Molecular Pharmaceutics 13, 4199 (2016).

[5] N. Bodor, A. Harget, and M. J. Huang, Journal of the American Chemical Society 113, 9480 (1991).

[6] C. M. Dobson, Nature 432, 824 (2004).

[7] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, 2018).

[8] R. Menichetti, K. H. Kanekal, and T. Bereau, ACS Central Science 5, 290 (2019).

[9] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Physical Review Letters 114 (2015), 10.1103/PhysRevLett.114.105503.

[10] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Proceedings of the National Academy of Sciences 113, 3932 (2016).

[11] M. Schmidt and H. Lipson, Science 324, 81 (2009).

[12] S.-M. Udrescu and M. Tegmark, Science Advances 6, eaay2631 (2020).

[13] H. Meyer, Archiv für experimentelle Pathologie und Pharmakologie 42, 109 (1899).

[14] C. E. Overton, *Studien über die Narkose: zugleich ein Beitrag zur allgemeinen Pharmakologie* (G. Fischer, 1901).

[15] V. A. Levin, Journal of Medicinal Chemistry 23, 682 (1980).

[16] J. M. Diamond and Y. Katz, The Journal of Membrane Biology 17, 121 (1974).

[17] T. Carpenter, D. Kirshner, E. Lau, S. Wong, J. Nilmeier, and F. Lightstone, Biophysical Journal 107, 630 (2014).

[18] M. Orsi and J. W. Essex, Molecular Simulations and Biomembranes , 76 (2010).

[19] C. T. Lee, J. Comer, C. Herndon, N. Leung, A. Pavlova, R. V. Swift, C. Tung, C. N. Rowley, R. E. Amaro, C. Chipot, Y. Wang, and J. C. Gumbart, Journal of Chemical Information and Modeling 56, 721 (2016).

[20] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, Physical Review Materials 2, 083802 (2018).

[21] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, and L. M. Ghiringhelli, Journal of Physics: Materials 2, 024002 (2019).

[22] D. L. Donoho, IEEE Transactions on Information Theory 52, 1289 (2006).

[23] E. J. Candes, J. Romberg, and T. Tao, IEEE Transactions on Information Theory 52, 489 (2006).

[24] R. G. Baraniuk, IEEE Signal Processing Magazine 24, 118 (2007).

[25] S. J. Marrink, A. H. de Vries, and A. E. Mark, The Journal of Physical Chemistry B 108, 750 (2004).

[26] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, The Journal of Physical Chemistry B 111, 7812 (2007).

[27] S. J. Marrink and D. P. Tieleman, Chemical Society Reviews 42, 6801 (2013).

[28] T. Bereau and K. Kremer, Journal of chemical theory and computation 11, 2783 (2015).

[29] S. L. Dixon and P. C. Jurs, Journal of Computational Chemistry 14, 1460 (1993).

[30] C. Hoffmann, A. Centi, R. Menichetti, and T. Bereau, Scientific Data 7 (2020), 10.1038/s41597-020-0391-0.

[31] R. Ouyang, "SISSO," https://github.com/rouyang2017/SISSO (2017).

[32] R. Menichetti and T. Bereau, Molecular Physics 117, 2900 (2019).

[33] R. Menichetti, K. H. Kanekal, K. Kremer, and T. Bereau, The Journal of Chemical Physics 147, 125101 (2017).

[34] R. S. Davis, P. B. Sunil Kumar, M. M. Sperotto, and M. Laradji, Journal of Physical Chemistry B 117, 4072 (2013).

[35] A. Ghysels, A. Krämer, R. M. Venable, W. E. Teague, E. Lyman, K. Gawrisch, and R. W. Pastor, Nature Communications 10, 1 (2019).

[36] C. Liu, P. Elvati, S. Majumder, Y. Wang, A. P. Liu, and A. Violi, ACS Nano 13, 10221 (2019).

[37] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, Nature 585, 357 (2020).

[38] J. D. Hunter, Computing in Science & Engineering 9, 90 (2007).

[39] The pandas development team, "pandas-dev/pandas: Pandas 1.1.3," (2020).

[40] Wes McKinney, in *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman (2010) pp. 56–61.