The Journal of Chemical Physics

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

Cite as: J. Chem. Phys. **154**, 244114 (2021); https://doi.org/10.1063/5.0053931
Submitted: 13 April 2021 • Accepted: 14 June 2021 • Published Online: 29 June 2021

Arghya Dutta, Jilles Vreeken, Luca M. Ghiringhelli, et al.

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Chemical physics software
The Journal of Chemical Physics **155**, 010401 (2021); https://doi.org/10.1063/5.0059886

Machine learning meets chemical physics
The Journal of Chemical Physics **154**, 160401 (2021); https://doi.org/10.1063/5.0051418

Perspective on integrating machine learning into computational chemistry and materials science
The Journal of Chemical Physics **154**, 230903 (2021); https://doi.org/10.1063/5.0047760

Lock-in Amplifiers
up to 600 MHz
Zurich Instruments
Watch

© 2021 Author(s).

The Journal of Chemical Physics | ARTICLE | scitation.org/journal/jcp

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

Cite as: J. Chem. Phys. 154, 244114 (2021); doi: 10.1063/5.0053931
Submitted: 13 April 2021 • Accepted: 14 June 2021 •
Published Online: 29 June 2021 • Publisher Corrected: 30 June 2021

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

AFFILIATIONS

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

[a)] **Author to whom correspondence should be addressed:** dutta@mpip-mainz.mpg.de

ABSTRACT

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 additionally consider the functional relationship between passive permeation and acidity. To discover easily interpretable equations that explain the data well, we use the recently proposed sure-independence screening and sparsifying operator (SISSO), an artificial-intelligence technique that combines symbolic regression with compressed sensing. Our study is based on a large *in silico* dataset of 0.4×10^6 small molecules extracted from coarse-grained simulations. We rationalize the equation suggested by SISSO via 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 but surprisingly contributes stepwise rather than proportionally. Our results are in line with previously observed changes in permeability, suggesting the distinction between liquid-disordered and liquid-ordered 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.

© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0053931

I. INTRODUCTION

Passive lipid-membrane permeation is and remains of great relevance for pharmaceutical applications and an improved physicochemical understanding of small-sized molecules in complex biological materials.[1,2] The technological implications of the problem have sustained the need for experiment- and simulation-free prediction of passive permeation, which are rapid, inexpensive, and accurate.[3,4] Various types of surrogate models have been proposed over the years, with 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 drawbacks:

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 tends to 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 nonlinear hierarchical functions, making them nearly impossible to conceptually grasp.[7] Quantitative structure–activity relationship (QSAR) models can build multivariate models, but

© Author(s) 2021

the combination of too many descriptors will lead to similar difficulties. Beyond predicting individual datapoints, 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. Critically, we address overfitting by relying on large datasets applied to simple models. Instead of experimental approaches, we base our study on *in silico* measurements, taking advantage of the rapid rise in high-throughput molecular dynamics simulations.[8] All-atom simulations offer a gold standard in terms of simulation-based permeability modeling that can reach exquisite correlation against experimental reference, but their overwhelming computational costs unfortunately limit studies to tens of compounds.[9–11] Here, instead, we use an approach based on particle-based coarse-grained (CG) simulations, making use of the transferable Martini model.[12–15] The Martini model sacrifices some chemical resolution but retains the essential driving force, mainly the partitioning coefficient of a chemical group between different phases. This allows the CG approach to recover excellent accuracy, 1.4 kcal/mol along a potential of mean force (PMF), translating to 1 $\log_{10}$ unit in the permeability coefficient, validated across an extensive set of structurally distinct compounds against both atomistic simulations and experimental measurements.[16] Critically, the accuracy of the CG model is accompanied by a three-orders-of-magnitude speedup compared to atomistic simulations. The high-throughput coarse-grained (HTCG) simulations offer unprecedentedly large databases of permeability coefficients: Menichetti *et al.* reported results for 511 427 compounds.[16] The ability to screen over so many compounds results mainly from the transferable nature of the coarse-grained model—many molecules map to the same set of CG beads, which effectively *reduces* the size of chemical space.[17] The benefits for an efficient computational-screening procedure outweigh the impinged degeneracy, as indicated by the above-mentioned accuracy. Overall, the database contains a nearly exhaustive subset of small organic molecules in the range of 30–160 Da, thereby ensuring a dense coverage of the chemical space in this subset. A deeper analysis of this large database is the subject of this study.

As for the data-driven model, we explicitly avoid building and using a black-box model and instead turn to learning an *equation*. In particular, we will rely on recently proposed data-driven techniques to discover equations.[18,19] 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.[20,21] Generalization capabilities are critical, because typical training datasets are minuscule relative to the size of chemical compound space, such that overfitting can easily prevail. How do we discover simple equations from the data? To this end, we follow Occam's razor and limit the complexity of the equations we consider. The combination of descriptors through various mathematical operators will lead to an overwhelming number of trial equations, many of which may fit the data similarly well. Following ideas from compressed sensing, we select *simple* equations by applying the ℓ_0 regularization. Identifying simple equations benefits both generalization aspects and also interpretability, i.e., rationalizing the derived equation.

Passive drug permeation measures the propensity of a solute to spontaneously cross a lipid membrane. In this paper, we exclude transporter-mediated uptake to focus solely on thermal diffusion.[22] 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[23] and Overton[24] 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, namely, the partitioning coefficient between water and octanol. Water/octanol partition, or more generally hydrophobicity, has long been identified as strongly correlating with membrane permeability.[25] 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 PMF, $G(z)$, as

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

where $R(z)$ and $\beta = 1/k_\mathrm{B}T$ correspond to the associated resistivity and the inverse temperature, respectively.[26] The competition between different protonation states naturally follows the sum of inverse resistivities, analogous to the total resistance in a parallel electrical circuit.[10] PMFs are shifted according to the difference between the solution's pH and the compound's acid dissociation constant, pK_a. In the following, we take the perspective of a neutral compound, which can deprotonate (acidic, apK_a) or protonate (basic, bpK_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-h per system at an atomistic resolution).[9,11]

Even at a CG resolution, Eq. (1) still requires free-energy calculations to determine the PMF. The objective of this study is to gain further insight into the key physical determinants of the permeability coefficient. Beyond the widely known effect of hydrophobicity, we focus on incorporating the role of acidity via the compound's protonation state.[27] The role of acidity is crucial, partly because of the sheer number of ionizable drugs: they make ≈62% of the World Health Organization's list of essential drugs.[28,29] It has long been hypothesized that the neutral form of an ionizable drug is the only contributor to its permeability, known as the pH–partition hypothesis.[30,31] However, this hypothesis is limiting for two reasons. First, permanently charged compounds are known to permeate lipid membranes.[32,33] Second, differences between a compound's pK_a and the surrounding pH can lead to a protonation-coupled permeation mechanism, which calls for the combined contributions of the neutral and ionized forms.[34,35] Clearly, the role of acidity is expected to couple with hydrophobicity. Establishing a *functional* relationship connecting the two quantities is the objective of this work. To this end, we rely on modern data-driven

© Author(s) 2021

techniques to discover an equation. Relating acidity to the permeability coefficient would not only help establishing rapid estimates for ionic compounds but also offer insight into the coupling of these physicochemical properties that are valid for a wide class of compounds.

Limitations in the number of candidate descriptors and correlations between features have recently been addressed by sure-independence screening and sparsifying operator (SISSO),[36,37] which is an artificial-intelligence technique that combines symbolic regression with compressed sensing.[38–40] SISSO provides a data-driven framework to discover equations—mathematical relations between input variables that best correlate with the target property. We will discuss several equations of various complexities to illustrate the balance between accuracy and interpretability. We root the simplest variant in the underlying physics by comparison 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 the 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

Our analysis is based on the passive-permeability database provided by Menichetti *et al.*[16] Reference information includes the water/octanol partitioning free energy ($\Delta G_{\mathrm{W}\to\mathrm{Ol}}$), acid dissociation constant for acids and bases (apK_a and bpK_a), and simulation-based permeability coefficient (expressed by its order of magnitude, $\log_{10} P$) for solutes through a single-component bilayer made of 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC). The water/octanol partitioning free energies were predicted using the neural network ALOGPS,[41] and acid dissociation constants pK_a were predicted from ChemAxon Marvin.[42] Filtering of the Generated DataBase (GDB)[43] for compounds that mapped to a one- or two-bead coarse-grained Martini representation, i.e., a monomer or a dimer, led to 511 427 small organic molecules (30–160 Da). Enhanced-sampling molecular dynamics simulations yielded the PMFs for both neutral and (de)protonated species, and Eq. (1) was used to compute the permeability coefficient, P. The diffusivity profile was *not* extracted from coarse-grained simulations but instead from previous atomistic studies taking advantage of its relatively small chemical dependence and its logarithmic impact on the permeability.[10] The pK_a of a chemical group can be either acidic (apK_a) or basic (bpK_a), and in that, a neutral compound can either deprotonate or protonate (see the supplementary material for definitions). While the ionization constant of conjugated acid/base pairs typically lies between 10^{-2} and 10^{16}, we considered compounds with pK_a values between −10 and 20.[44] 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).[45]

B. Learning algorithm

To learn about an interpretable model of passive permeability, we used SISSO, as implemented in Ref. 46. SISSO aims at establishing a functional relationship, $y = f(\Phi)$, between n primary features, $\Phi_0 = \{\phi_1, \phi_2, \ldots, \phi_n\}$, 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 applies 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 candidate features. Φ_q denotes the set of candidate features at each level of recursion q. The number of candidate features in Φ_q increases sharply with the increase in the recursion level q, the number of operators used, and the number of primary features n. For each q, SISSO iteratively selects subsets of candidate features that have the largest linear correlations with the target y and then with the subsequent residuals, i.e., each portion of y that is captured by the previous iterations (see Ref. 37). The number of iterations in this procedure, which equals the number of terms in the linear expansion of $f(\Phi)$, and hereby denoted as the dimension of the model, is controlled by a sparsifying ℓ_0 regularization. For each q and number of dimensions, SISSO selects the model with the smallest 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, $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$, and the acid dissociation constants apK_a and bpK_a as provided by Menichetti *et al.*[16] We thereby seek a refinement or correction to the commonly used model based on hydrophobicity alone.[25] The mean absolute errors associated with the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ and pK_a predictions (0.36 kcal/mol[41] and 0.86 units,[47] respectively) make them reliable primary descriptors. To avoid constructing features with different units, we multiply the partitioning free energy by the inverse temperature: $\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ using $T = 300$ K following Menichetti *et al.*[16] Starting with the set of primary features $\Phi_0 = \{\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}, \mathrm{ap}K_\mathrm{a}, \mathrm{bp}K_\mathrm{a}\}$, 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 2000 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 (see the supplementary material for the input script) 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.*[36] emphasized that SISSO works reliably when the number of materials in a training set, N, sufficiently exceeds the number of candidate features. In particular, SISSO requires $N \geq kd\ln(\#\Phi)$, where $k \sim 1$–10, d is the dimension, and $\#\Phi$ is the size of the feature space. 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 (~42 000 compounds), SISSO is well within a regime to provide meaningful and consistent results.

© Author(s) 2021

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 Table I. For all models, $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ takes on a central role, as expected by the performance of the baseline. The simplest SISSO model, $f^{\mathrm{1D}} = c_0^{\mathrm{1D}} + c_1^{\mathrm{1D}}(\mathrm{ap}K_{\mathrm{a}} - \mathrm{bp}K_{\mathrm{a}} - 2\beta\Delta G_{\mathrm{W}\to\mathrm{Ol}})$, is remarkably robust: it is systematically identified as the best performing 1D model across all ten training sets. Given that we trained on only 10% (~42 000 compounds) of the dataset, this highlights this model's performance compared to all other candidates (see the supplementary material for a list of candidate one-dimensional models). The stability of the model—given such a small training fraction—speaks for the robustness of the equation. Its improved accuracy compared to the baseline will be evaluated further down.

We also report more complex 2D and 3D models in Table 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 ten training sets, meaning that other models of similar complexity closely compete.

B. Model performance

We now turn to the performance of these four models. Table 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 the baseline to SISSO models of increasing complexity, the distribution does lean increasingly toward the $y = x$ correlation line. The presence of horizontal stripes in Fig. 1 results from the degenerate use of CG mappings for many molecules.[17] This artifact is most notable for f^{Hyd}, which solely relies on hydrophobicity, whereas the others have chemically specific acidity information. For the 2D and 3D models, we also point out outliers at the lowest permeability values. Figure S1 of the supplementary material shows the distribution of compounds: these low-permeability values are scarcely populated, both in algorithmically and synthesized compound databases.[16] Here, they represent only 0.07% of the dataset, and our uniform sampling of training points likely brought in only few of them. They likely result from poor extrapolation behavior of the 2D and 3D models, which notably include powers of two of several variables.

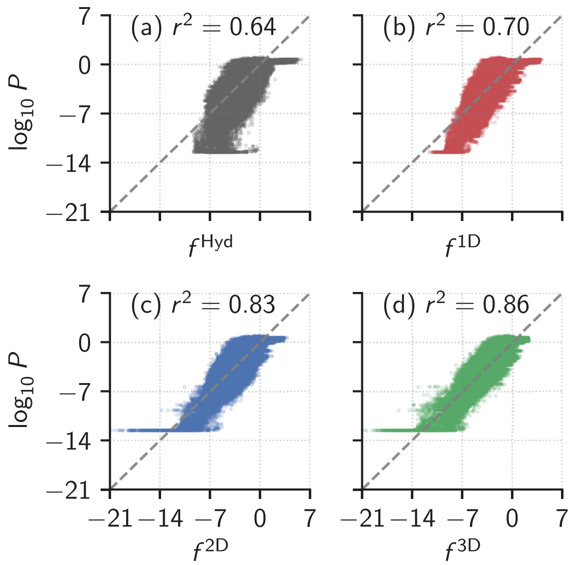

FIG. 1. Performance of the four permeability models against the reference for the dataset of 418 K small molecules: (a) baseline hydrophobicity model f^{Hyd} and (b)–(d) 1D to 3D data-driven SISSO models. See Table I for their expressions.

To better understand the performance of each model, we analyze their errors in greater detail. Figure 2(a) shows the distribution

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 three feature dimensions: f^{1D}, f^{2D}, and f^{3D}. 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 supplementary material.

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

© Author(s) 2021

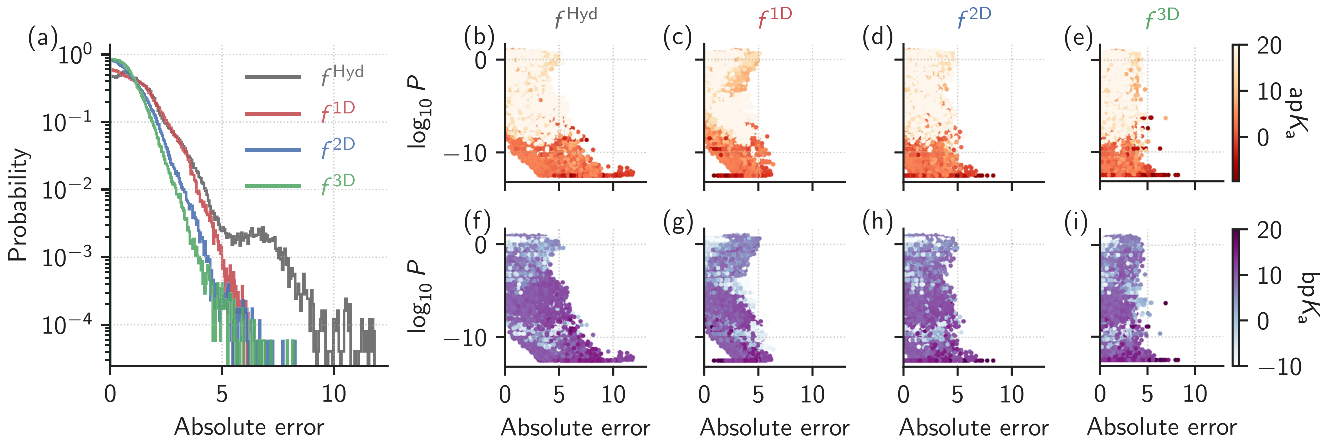

FIG. 2. Absolute error analysis. (a) Error (in $\log_{10}$ units) distribution for all models. Error decomposed as a function of (b)–(e) apK_a and (f)–(i) bpK_a. Stronger acids/bases are shown in darker colors.

of absolute error. The large dataset at our disposal allows us to evaluate more than four 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. 2(a)] 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 apK_a or bpK_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 should naturally be prone to systematic errors inherent to the training dataset. While we expect our systematic integration of the ISDM permeability coefficient [Eq. (1)] to ensure robust functional relationships, systematic errors in the parameters are likely to affect the fitting coefficients. 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.[16] Here, we compare the performance of the four permeability models against the atomistic simulations of Carpenter *et al.*[10] 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, comparable to that of real drugs, given that more than 60% of drugs have molecular weight below 300 Da.[48] On the other hand, our training HTCG database only contained compounds up to 160 Da. This thus presents a challenging test for the generalizability of the SISSO models.

Figure 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 apK_a and bpK_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 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 notable 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 in the MAE with increasing complexity in Figs. 3(b)–3(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,

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

© Author(s) 2021

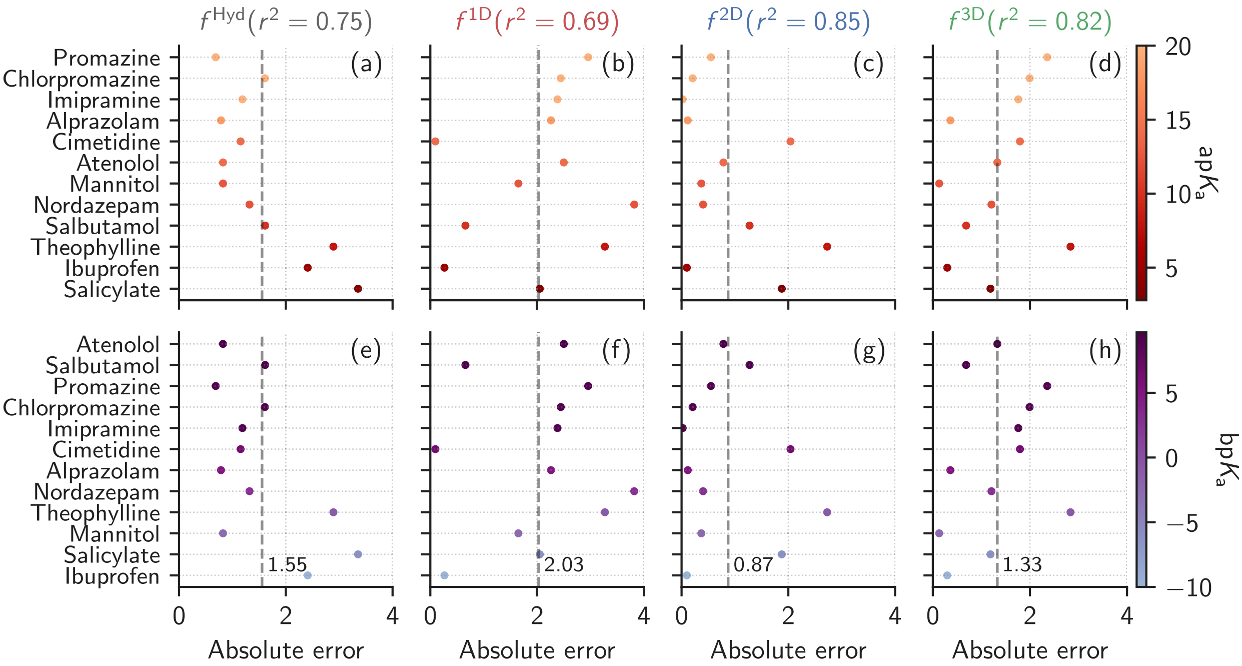

FIG. 3. Absolute error (in $\log_{10}$ units) against atomistic simulations for 12 reference small molecules.[10] The error is displayed for all four models and as a function of (a)–(d) $\mathrm{ap}K_\mathrm{a}$ and (e)–(h) $\mathrm{bp}K_\mathrm{a}$.

Figure 4 displays the permeability coefficient as a function of these two contributions. The symmetric contribution of $\beta\Delta G_{\mathrm{W}\to\mathrm{OI}}$ 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 $\mathrm{ap}K_\mathrm{a} - \beta\Delta G_{\mathrm{W}\to\mathrm{OI}} \sim 0$ and $-\mathrm{bp}K_\mathrm{a} - \beta\Delta G_{\mathrm{W}\to\mathrm{OI}} \sim -15$ and represent strong-acid and strong-base regimes. In what follows, we provide an asymptotic rationalization of the functional form of Eq. (2).

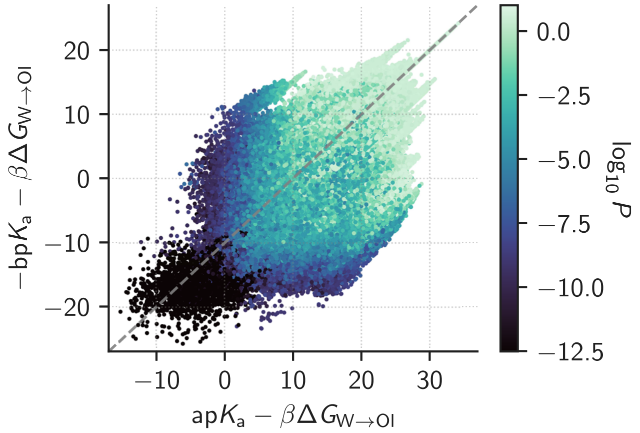

FIG. 4. SISSO 1D model of passive permeability. The decomposition of the secondary feature in two axes highlights the role of hydrophobicity (along the diagonal) compared to acidity (vertical and horizontal).

To rationalize Eq. (2), we first outline the role of our three primary descriptors in the ISDM model [Eq. (1)]. Figure 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 $G_\mathrm{N}(z)$, $G_\mathrm{B}(z)$, and $G_\mathrm{A}(z)$, respectively.

The difference between $\mathrm{ap}K_\mathrm{a}$ or $\mathrm{bp}K_\mathrm{a}$ and pH dictates the propensity for the PMFs to cross each other. The ISDM relies on a total resistivity [defined in Eq. (1)], R_T, such that $R_\mathrm{T}^{-1}(z) = R_\mathrm{N}^{-1}(z) + R_\mathrm{B}^{-1}(z) + R_\mathrm{A}^{-1}(z)$, 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 $\mathrm{ap}K_\mathrm{a}$ and $\mathrm{bp}K_\mathrm{a}$ values, as well as the pH of the environment, through the following equations:

$$\beta^{-1}(\mathrm{pH} - \mathrm{ap}K_\mathrm{a})\ln 10 = G_\mathrm{N}(\infty) - G_\mathrm{A}(\infty), \tag{3}$$

$$\beta^{-1}(\mathrm{pH} - \mathrm{bp}K_\mathrm{a})\ln 10 = G_\mathrm{B}(\infty) - G_\mathrm{N}(\infty), \tag{4}$$

where $G(\infty)$ indicates that the compound is located in bulk water. Equations (3) and (4) effectively link the difference between the pH of the environment with the $\mathrm{p}K_\mathrm{a}$ 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. Equations (3) and (4) allow us to explicitly link $\mathrm{p}K_\mathrm{a}$ information with the total resistivity

$$R_\mathrm{T}^{-1}(z) = D(z)\Big[\mathrm{e}^{-\beta(G_\mathrm{N}(z)-G_\mathrm{N}(\infty))} + \mathrm{e}^{-\beta(G_\mathrm{B}(z)-G_\mathrm{B}(\infty))} + \mathrm{e}^{-\beta(G_\mathrm{A}(z)-G_\mathrm{A}(\infty))}\Big], \tag{5}$$

© Author(s) 2021

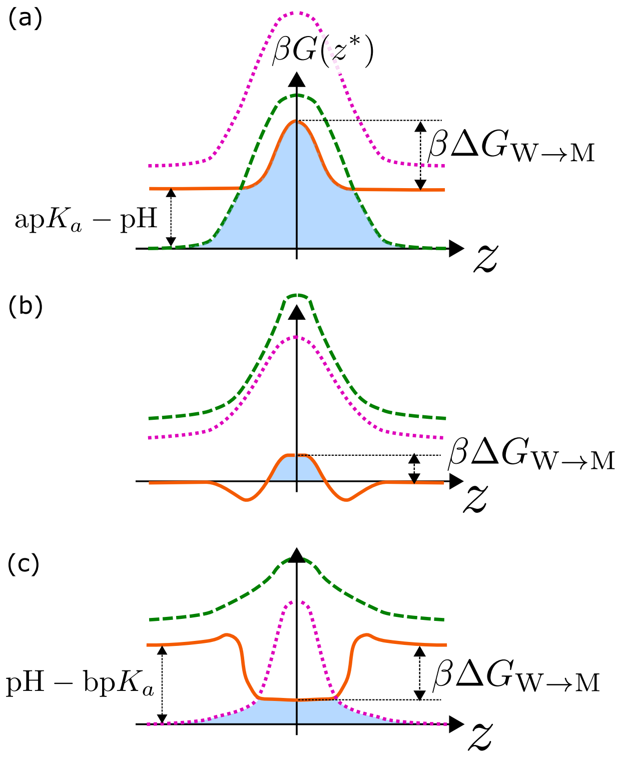
(a)
$\beta G(z^*)$
$\beta\Delta G_{\mathrm{W}\to\mathrm{M}}$
$\mathrm{ap}K_a - \mathrm{pH}$
z
(b)
$\beta\Delta G_{\mathrm{W}\to\mathrm{M}}$
z
(c)
$\mathrm{pH} - \mathrm{bp}K_a$
$\beta\Delta G_{\mathrm{W}\to\mathrm{M}}$
z

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 (orange solid curve), acidic (green dashed curve), and basic (pink dotted curve) PMFs. The blue area under the effective PMF directly links to the permeability coefficient.

where we assume that all protonation states yield identical diffusivity.[16] 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., $\mathrm{ap}K_\mathrm{a} \gg \mathrm{pH}$ and $\mathrm{bp}K_\mathrm{a} \ll \mathrm{pH}$, such that the compound is effectively unable to (de)protonate, and $G_\mathrm{N}(\infty) = 0$. Equation (5) can be simplified to $R_\mathrm{T}(z) \approx D^{-1}(z)\exp[\beta G_\mathrm{N}(z)]$.[49]
- Strong acids consist of solutes that display at least one functional group for which $\mathrm{ap}K_\mathrm{a} \ll \mathrm{pH}$. For neutral pH (≈7), this would correspond to $\mathrm{ap}K_\mathrm{a} < 4$. We set $G_\mathrm{A}(z \to \infty) = 0$. In this regime, the third exponential in Eq. (5) would dominate the other two, leading to $R_\mathrm{T}(z) \approx D^{-1}(z)\exp[\beta G_\mathrm{N}(z) + (\mathrm{pH} - \mathrm{ap}K_\mathrm{a})\ln 10]$.
- 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 $R_\mathrm{T}(z) \approx D^{-1}(z)\exp[\beta G_\mathrm{N}(z) - (\mathrm{pH} - \mathrm{bp}K_\mathrm{a})\ln 10]$.

The total resistivities still require integration over the reaction coordinate z, which we simplify to the largest contribution of the PMF.[11] 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 $z^* = \underset{z}{\arg\max}\, G_\mathrm{eff}(z)$. 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:

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

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 shifts the permeability coefficient across the chemical space of compounds considered. Figure 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): $\mathrm{ap}K_\mathrm{a} - \beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ and $-\mathrm{bp}K_\mathrm{a} - \beta\Delta G_{\mathrm{W}\to\mathrm{Ol}}$. We first note that $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ 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 is hydrophobic compounds for which the highest point in the PMF is in water [Fig. 5(c)]. Furthermore, $G(z^* = 0)$, which corresponds to the transfer free energy from water to the membrane midplane, has been shown to linearly relate to the water/octanol partitioning free energy, $\Delta G_{\mathrm{W}\to\mathrm{M}} \propto \Delta G_{\mathrm{W}\to\mathrm{Ol}}$.[50] This connects $G_\mathrm{N}(z^*)$ present in Eq. (6) to $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ 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 $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ by the compound's $G_\mathrm{N}(z^*)$ in the relevant among the two contributions of Eq. (2), we indeed recover one of the asymptotes. As

© Author(s) 2021

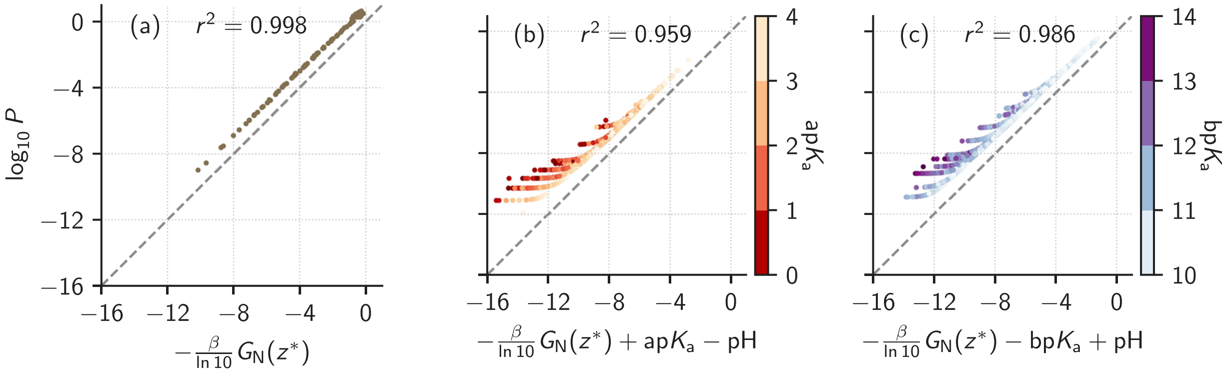

FIG. 6. Simple asymptotic estimates [Eq. (6)] against reference permeability coefficients at neutral pH. (a) Neutral compounds, (b) strong acids, and (c) strong bases.

for neutral compounds, $G_N(z^*)$ is the only relevant quantity while estimating permeability.

E. Drug–membrane permeability across membranes

The following paragraphs analyze 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.[45] We specifically analyze the change in permeability when turning to phosphocholine (PC) membranes made of different lipids, varying in both the tail length and level of unsaturation.

Figure 7 shows the relation of $-\beta G_N(z^*)$—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 ~1.4.

To better understand these results, we first recall that $G_N(z^*)$ corresponds to the highest value of the neutral PMF. We can safely ignore contributions of the charged PMFs such that $G_N(z^*)$ denotes the highest point of the *effective* PMF. The excellent agreement between DLPC and DPPC indicates that the tail length (3 and 4 beads long, respectively) does not impact the permeability. This is expected, given that the tail length is only expected to change the length, but not the height, of the hydrophobic plateau. On the other hand, the further agreement between them and POPC and DOPC indicates 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, previous Martini studies of ternary membranes with DPPC and cholesterol have shown that DIPC and DAPC strongly phase separate into a liquid-disordered (Ld) phase.[51–53] Inconsistent conclusions were drawn from different studies with DOPC, pointing to a thermodynamic drive that is weak at best.[52,54,55] As for POPC, there is no sign of phase transition.[52] The results on Fig. 7 mirror these trends: we find a clear shift of $-\beta G_N(z^*)$, depending on the ability of the membrane to form an Ld domain. The Ld-domain formation of course hinges on the presence of DPPC and cholesterol, which are notably absent from our reference simulations.[45] The trends are surprising in that they show a dependence on lipid-tail unsaturation that is *stepwise* rather than proportional. While we defer a more detailed study to future work, we suggest the role of the entropic character of the lipid-tail fluctuations.

Other studies before us have reported a clear change in permeability between Ld and Lo domains: Ghysels *et al.* used both

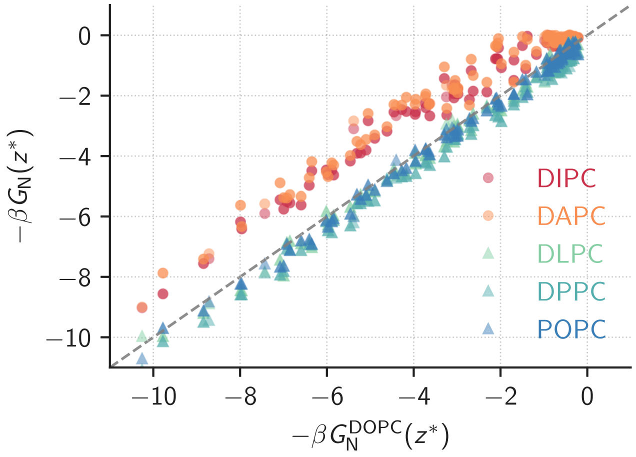

FIG. 7. Variation of $-\beta G_N(z^*)$ between the original membrane, DOPC, and others. Note the shift between (i) DLPC, DPPC, and POPC and (ii) DIPC and DAPC.

© Author(s) 2021

atomistic simulations and electron paramagnetic resonance spectroscopy experiments on the permeation of water and oxygen and found a permeability ratio of $P(\text{Ld})/P(\text{Lo}) \approx 3$.[56] Here, the CG simulations yield a shift in $-\beta G_N(z^*)$, which translates into a ratio of the permeability coefficients of ≈ 25. Our estimates are thus within one $\log_{10}$ unit of the results of Ghysels *et al.* for their specific compounds. The mechanism remains to be clearly identified, although this could be consistent with the proposed role of local membrane surface density (i.e., its propensity to form transient holes).[57]

To summarize, our use of single-component lipid membranes only allows us to speculate the shift in Fig. 7. The link to Lo/Ld-domain formation is in line with prior atomistic simulations and experiments for specific compounds.[56] In the broader context, our results could help generalize their results across the chemical space of drugs. The results also suggest simple *additive* corrections to our effective equation when considering different membranes.

IV. CONCLUSIONS

We propose to learn the functional relationship between hydrophobicity and acidity as a simple surrogate for passive membrane permeability. Our approach, combining symbolic regression and compressed sensing, is data-driven *and* interpretable and based on large databases of high-throughput coarse-grained simulations. 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 $f^{1\text{D}} = c_0^{1\text{D}} + c_1^{1\text{D}}(\text{ap}K_\text{a} - \text{bp}K_\text{a} - 2\beta\Delta G_{\text{W}\to\text{Ol}})$, where $\text{ap}K_\text{a}$ and $\text{bp}K_\text{a}$ characterize acidity, $\Delta G_{\text{W}\to\text{Ol}}$ is the water/octanol partitioning coefficient, and $c_0^{1\text{D}}$ and $c_1^{1\text{D}}$ 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, implicitly testifying to the accuracy of the underlying HTCG resolution. Broad agreement numerically validates the use of a single bulk hydrophobicity measure to effectively replace the potential of mean force, which has been exploited by others before.[11] Importantly, the interplay of hydrophobicity together with acidity leads to a significant improvement in the model accuracy of the SISSO 1D equation for ionizable groups. The SISSO equations show improvements over a challenging set of compounds that are much larger than the training set. Critically, our work refines the common role of hydrophobicity in passive permeation to relate it *functionally* with acidity.

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 observe a surprising change in permeability coefficient: lipid-tail unsaturation contributes stepwise rather than proportionally. Our findings are in line with recent atomistic simulations and electron paramagnetic resonance spectroscopy experiments, highlighting the distinction between lipids primarily involved in liquid-disordered (Ld) and liquid-ordered (Lo) domains. The approach offers a data-driven, interpretable analysis of drug–membrane passive permeability across both drugs and membranes.

SUPPLEMENTARY MATERIAL

See the supplementary material for definitions of dissociation constants ($\text{ap}K_\text{a}$ and $\text{bp}K_\text{a}$), the distribution of compounds across permeability, Table I along with the error values, the list of best one-dimensional descriptors, and the SISSO input script.

ACKNOWLEDGMENTS

The authors thank Oleksandra Kukharenko, Roberto Menichetti, and Yasemin Bozkurt Varolgüneş for critical reading of the manuscript and Kiran Kanekal and Martin Girard for insightful discussions. Data analysis relied extensively on the open source packages Numpy,[58] Matplotlib,[59] and Pandas.[60,61] We acknowledge support from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science. T.B. was partially supported by the Emmy Noether Programme of the Deutsche Forschungsgemeinschaft (DFG).

DATA AVAILABILITY

In this study, we used the database provided by Menichetti *et al.*[16] It is openly available at https://doi.org/10.1021/acscentsci.8b00718.

REFERENCES

[1] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," Adv. Drug Delivery Rev. **46**, 3–26 (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, "Predicting passive permeability of drug-like molecules from chemical structure: Where are we?," Mol. Pharm. **13**, 4199–4208 (2016).

[5] N. Bodor, A. Harget, and M. J. Huang, "Neural network studies. 1. Estimation of the aqueous solubility of organic compounds," J. Am. Chem. Soc. **113**, 9480–9483 (1991).

[6] C. M. Dobson, "Chemical space and biology," Nature **432**, 824–828 (2004).

[7] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, 2018).

[8] T. Bereau, "Computational compound screening of biomolecules and soft materials by molecular simulations," Modell. Simul. Mater. Sci. Eng. **29**, 023001 (2021).

[9] M. Orsi and J. W. Essex, "Passive permeation across lipid bilayers: A literature review," in *Molecular Simulations and Biomembranes: From Biophysics to Function* (The Royal Society of Chemistry, 2010), pp. 76–90.

[10] T. S. Carpenter, D. A. Kirshner, E. Y. Lau, S. E. Wong, J. P. Nilmeier, and F. C. Lightstone, "A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations," Biophys. J. **107**, 630–641 (2014).

[11] 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,

© Author(s) 2021

"Simulation-based approaches for determining membrane permeability of small compounds," J. Chem. Inf. Model. **56**, 721–733 (2016).

[12] S. J. Marrink, A. H. de Vries, and A. E. Mark, "Coarse grained model for semiquantitative lipid simulations," J. Phys. Chem. B **108**, 750–760 (2004).

[13] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The Martini force field: Coarse grained model for biomolecular simulations," J. Phys. Chem. B **111**, 7812–7824 (2007).

[14] S. J. Marrink and D. P. Tieleman, "Perspective on the Martini model," Chem. Soc. Rev. **42**, 6801–6822 (2013).

[15] R. Alessandri, F. Grünewald, and S. J. Marrink, "The martini model in materials science," Adv. Mater. **33**, 2008635 (2021).

[16] R. Menichetti, K. H. Kanekal, and T. Bereau, "Drug–membrane permeability across chemical space," ACS Cent. Sci. **5**, 290–298 (2019).

[17] K. H. Kanekal and T. Bereau, "Resolution limit of data-driven coarse-grained models spanning chemical space," J. Chem. Phys. **151**, 164106 (2019).

[18] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Phys. Rev. Lett. **114**, 105503 (2015).

[19] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proc. Natl. Acad. Sci. U. S. A. **113**, 3932–3937 (2016).

[20] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science **324**, 81–85 (2009).

[21] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," Sci. Adv. **6**, eaay2631 (2020).

[22] D. B. Kell and S. G. Oliver, "How drugs get into cells: Tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion," Front. Pharmacol. **5**, 231 (2014).

[23] H. Meyer, "Zur theorie der alkoholnarkose," Arch. Exp. Pathol. Pharmakol. **42**, 109–118 (1899).

[24] C. E. Overton, *Studien Über die Narkose: Zugleich ein Beitrag zur Allgemeinen Pharmakologie* (G. Fischer, 1901).

[25] V. A. Levin, "Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability," J. Med. Chem. **23**, 682–684 (1980).

[26] J. M. Diamond and Y. Katz, "Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water," J. Membr. Biol. **17**, 121–154 (1974).

[27] B. K. Radak, C. Chipot, D. Suh, S. Jo, W. Jiang, J. C. Phillips, K. Schulten, and B. Roux, "Constant-pH molecular dynamics simulations for large biomolecular systems," J. Chem. Theory Comput. **13**, 5933–5944 (2017).

[28] D. T. Manallack, "The pK_a distribution of drugs: Application to drug discovery," Perspect. Med. Chem. **1**, 1177391X0700100 (2007).

[29] D. T. Manallack, R. J. Prankerd, E. Yuriev, T. I. Oprea, and D. K. Chalmers, "The significance of acid/base properties in drug discovery," Chem. Soc. Rev. **42**, 485–496 (2013).

[30] P. A. Shore, B. B. Brodie, and C. A. M. Hogben, "The gastric secretion of drugs: A pH partition hypothesis," J. Pharmacol. Exp. Ther. **119**, 361–369 (1957).

[31] S. M. Saparov, Y. N. Antonenko, and P. Pohl, "A new model of weak acid permeation through membranes revisited: Does Overton still rule?," Biophys. J. **90**, L86–L88 (2006).

[32] P. Langguth, A. Kubis, G. Krumbiegel, W. Lang, H. P. Merkle, W. Wächter, H. Spahn-Langguth, and R. Weyhenmeyer, "Intestinal absorption of the quaternary trospium chloride: Permeability-lowering factors and bioavailabilities for oral dosage forms," Eur. J. Pharm. Biopharm. **43**, 265–272 (1997).

[33] H. Fischer, M. Kansy, A. Avdeef, and F. Senner, "Permeation of permanently positive charged molecules through artificial membranes—Influence of physicochemical properties," Eur. J. Pharm. Sci. **31**, 32–42 (2007).

[34] K. Palm, K. Luthman, J. Ros, J. Gråsjö, and P. Artursson, "Effect of molecular charge on intestinal epithelial drug transport: pH-dependent transport of cationic drugs," J. Pharmacol. Exp. Ther. **291**, 435–443 (1999).

[35] Z. Yue, C. Li, G. A. Voth, and J. M. J. Swanson, "Dynamic protonation dramatically affects the membrane permeability of drug-like molecules," J. Am. Chem. Soc. **141**, 13421–13433 (2019).

[36] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, "SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates," Phys. Rev. Mater. **2**, 083802 (2018).

[37] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, and L. M. Ghiringhelli, "Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO," J. Phys.: Mater. **2**, 024002 (2019).

[38] D. L. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory **52**, 1289–1306 (2006).

[39] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory **52**, 489–509 (2006).

[40] R. G. Baraniuk, "Compressive sensing," IEEE Signal Process. Mag. **24**, 118–121 (2007).

[41] I. V. Tetko and V. Y. Tanchuk, "Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program," J. Chem. Inf. Comput. Sci. **42**, 1136–1145 (2002).

[42] See https://chemaxon.com/products/marvin for calculator plugin of chemaxon marvin; accessed 12 January 2021.

[43] T. Fink, H. Bruggesser, and J.-L. Reymond, "Virtual exploration of the small-molecule chemical universe below 160 daltons," Angew. Chem., Int. Ed. **44**, 1504–1508 (2005).

[44] S. L. Dixon and P. C. Jurs, "Estimation of pK_a for organic oxyacids using calculated atomic charges," J. Comput. Chem. **14**, 1460–1467 (1993).

[45] C. Hoffmann, A. Centi, R. Menichetti, and T. Bereau, "Molecular dynamics trajectories for 630 coarse-grained drug-membrane permeations," Sci. Data **7**, 51 (2020).

[46] R. Ouyang, SISSO, https://github.com/rouyang2017/SISSO, 2017.

[47] C. Liao and M. C. Nicklaus, "Comparison of nine programs predicting pK_a values of pharmaceutical substances," J. Chem. Inf. Model. **49**, 2801–2812 (2009).

[48] M. Feher and J. M. Schmidt, "Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry," J. Chem. Inf. Comput. Sci. **43**, 218–227 (2003).

[49] R. Menichetti and T. Bereau, "Revisiting the Meyer-Overton rule for drug-membrane permeabilities," Mol. Phys. **117**, 2900–2909 (2019).

[50] R. Menichetti, K. H. Kanekal, K. Kremer, and T. Bereau, "*In silico* screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force," J. Chem. Phys. **147**, 125101 (2017).

[51] C. Rosetti and C. Pastorino, "Comparison of ternary bilayer mixtures with asymmetric or symmetric unsaturated phosphatidylcholine lipids by coarse grained molecular dynamics simulations," J. Phys. Chem. B **116**, 3525–3537 (2012).

[52] R. S. Davis, P. B. Sunil Kumar, M. M. Sperotto, and M. Laradji, "Predictions of phase separation in three-component lipid membranes by the MARTINI force field," J. Phys. Chem. B **117**, 4072–4080 (2013).

[53] G. A. Pantelopulos and J. E. Straub, "Regimes of complex lipid bilayer phases induced by cholesterol concentration in MD simulation," Biophys. J. **115**, 2167–2178 (2018).

[54] S. Baoukina, E. Mendez-Villuendas, and D. P. Tieleman, "Molecular view of phase coexistence in lipid monolayers," J. Am. Chem. Soc. **134**, 17543–17553 (2012).

[55] M. Girard and T. Bereau, "Regulating lipid composition rationalizes acyl tail saturation homeostasis in ectotherms," Biophys. J. **119**, 892–899 (2020).

[56] A. Ghysels, A. Krämer, R. M. Venable, W. E. Teague, E. Lyman, K. Gawrisch, and R. W. Pastor, "Permeability of membranes in the liquid ordered and liquid disordered phases," Nat. Commun. **10**, 5616 (2019).

[57] C. Liu, P. Elvati, S. Majumder, Y. Wang, A. P. Liu, and A. Violi, "Predicting the time of entry of nanoparticles in lipid membranes," ACS Nano **13**, 10221–10232 (2019).

[58] 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, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson,

© Author(s) 2021

P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," Nature **585**, 357–362 (2020).

[59] J. D. Hunter, "Matplotlib: A 2D graphics environment," Comput. Sci. Eng. **9**, 90–95 (2007).

[60] J. Reback, W. McKinney, jbrockmendel, J. V. den Bossche, T. Augspurger, P. Cloud, gfyoung, Sinhrks, S. Hawkins, M. Roeschke, A. Klein, T. Petersen, J. Tratner, C. She, W. Ayd, S. Naveh, M. Garcia, J. Schendel, A. Hayden, D. Saxton, V. Jancauskas, A. McMaster, P. Battiston, S. Seabold, chris b1, h-vetinari, K. Dong, S. Hoyer, W. Overmeire, and M. Gorelli (2020). "pandas-dev/pandas: Pandas 1.1.4," Zenodo. http://doi.org/10.5281/zenodo.4161697.

[61] W. McKinney, "Data structures for statistical computing in Python," in Proceedings of the 9th Python in Science Conference, edited by S. van der Walt and J. Millman (2010), pp. 56–61.

© Author(s) 2021