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, and Tristan Bereau

ARTICLES YOU MAY BE INTERESTED IN

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

Multi-body effects in a coarse-grained protein force field
The Journal of Chemical Physics 154, 164113 (2021); https://doi.org/10.1063/5.0041022
Data-driven equation for drug–membrane permeability across drugs and membranes

Arghya Dutta, Jilles Vreeken, Luca M. Ghiringhelli, and Tristan Bereau

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

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 $0.4 \times 10^6$ 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 physical-chemical understanding of small-sized molecules in complex biological materials. 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. Various types of surrogate models have been proposed over the years, with the field having adopted machine learning early on. 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). 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 non-linear hierarchical functions, making them nearly impossible to conceptually grasp. 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 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. 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. Here, instead, we use an approach based on particle-based coarse-grained (CG) simulations, making use of the transferable Martini model. 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 unit in the permeability coefficient, validated across an extensive set of structurally distinct compounds against both atomistic simulations and experimental measurements. 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. 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. 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. 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. 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 $\ell_0$ 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. Upon doing so, the solute interacts with a great variety of physico-chemical 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 and Overton 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 $\sigma$ 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 correlated with membrane permeability. 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

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

where $R(z)$ and $\beta = 1/k_B T$ correspond to the associated resistivity and the inverse temperature, respectively. The competition between different protonation states naturally follows the sum of inverse resistivities, analogous to the total resistance in a parallel electrical circuit. PMFs are shifted according to the difference between the solution’s pH and the compound’s acid dissociation constant, p$K_a$. In the following, we take the perspective of a neutral compound, which can deprotonate (acidic, p$K_a$) or protonate (basic, bp$K_a$). 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).

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. 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. 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. However, this hypothesis is limiting for two reasons. First, permanently charged compounds are known to permeate lipid membranes. Second, differences between a compound’s p$K_a$ and the surrounding pH can lead to a protonation-coupled permeation mechanism, which calls for the combined contributions of the neutral and ionized forms. 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
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-inducement screening and sparsifying operator (SISSO), which is an artificial-intelligence technique that combines symbolic regression with compressed sensing. 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. Reference information includes the water/octanol partitioning free energy (ΔG_{W→O,l}), 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, and acid dissociation constants pK_a were predicted from ChemAxon Marvin. Filtering of the Generated DataBase (GDB) for compounds that mapped to a one- or two-head 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 diffusion 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. 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 $20$. 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-diarchachidonoyl-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).

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, $\Phi_n = \{\phi_1, \phi_2, \ldots, \phi_n\}$, 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 {+,-,×,÷, exp, log, ( )^{-1}, ( )^2, ( )^3, √( )} on the primary features and builds up sets of candidate features. $\Phi_q$ denotes the set of candidate features at each level of recursion $q$. The number of candidate features in $\Phi_q$ 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 $\ell_0$ 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, ΔG_{W→O,l}, and the acid dissociation constants apK_a and bpK_a as provided by Menichetti et al. We thereby seek a refinement or correction to the commonly used model based on hydrophobicity alone. The mean absolute errors associated with the ΔG_{W→O,l} and pK_a predictions (0.36 kcal/mol and 0.86 units, respectively) make them reliable primary descriptors. To avoid constructing features with different units, we multiply the partitioning free energy by the inverse temperature: $\beta\Delta G_{W→O,l}$ using $T = 300$ K following Menichetti et al. Starting with the set of primary features $\Phi_0 = \{\beta\Delta G_{W→O,l}, \text{apK}_a, \text{bpK}_a\}$, we consider the construction of secondary features for up to two iterations (i.e., $q = 2$), where $\Phi_1$ and $\Phi_2$ 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. 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 $N \geq k d \ln(*\Phi)$, where $k \sim 1$–10, $d$ is the dimension, and *\Phi is the size of the feature space. For $\Phi_2$, the relevant feature space used to train our model, the relation requires $N \geq 10 \cdot 2 \cdot \ln(2 \cdot 10^3) \approx 150$. 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.
III. RESULTS AND DISCUSSIONS

A. Learning permeability models

Table I contains the four models considered in this work: (i) \( f^{\text{Hyd}} \) 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, \( c_i \) correspond to non-zero coefficients for model m and index i, reported in Table I. For all models, \( \Delta G_{W\rightarrow O} \) takes on a central role, as expected by the performance of the baseline. The simplest SISSO model, \( f^{1D} = c_0^{1D} + c_1^{1D}(\alpha p K_a - b p K_a - 2\beta \Delta G_{W\rightarrow O}) \), 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. This artifact is most notable for \( f^{\text{Hyd}} \), 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. 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.

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

The presence of horizontal stripes in Fig. 1 results from the degenerate use of CG mappings for many molecules. This artifact is most notable for \( f^{\text{Hyd}} \), 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. 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.

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

![FIG. 1. Performance of the four permeability models against the reference for the dataset of 418 K small molecules: (a) baseline hydrophobicity model \( f^{\text{Hyd}} \) and (b)–(d) 1D to 3D data-driven SISSO models. See Table I for their expressions.](image-url)
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 \( f^{1D} \) 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 \( \beta \Delta G_{W\rightarrow Ol} \), 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 \( \text{ap}K_a \) or \( \text{bp}K_a \). 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 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.\(^{24}\) 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 \( \text{ap}K_a \) and \( \text{bp}K_a \). 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 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,

\[
f^{1D} = c_0^{1D} \left[ (\text{ap}K_a - \beta \Delta G_{W\rightarrow Ol}) + (\text{bp}K_a - \beta \Delta G_{W\rightarrow Ol}) \right].
\]  (2)
FIG. 3. Absolute error (in log_{10} units) against atomistic simulations for 12 reference small molecules. The error is displayed for all four models and as a function of (a)–(d) apK_a and (e)–(h) bpK_a.

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 $G_N(z)$, $G_B(z)$, and $G_A(z)$, respectively.

The difference between apK_a or bpK_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

$$R_T^{-1}(z) = R_T^{-1}(N) + R_T^{-1}(B) + R_T^{-1}(A),$$

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 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. Equations (3) and (4) allow us to explicitly link pK_a information with the total resistivity.

$$R_T^{-1}(z) = D(z)\left[ e^{-\beta(G_N(z) - G_N(\infty))} + e^{-\beta(G_B(z) - G_B(\infty))} + e^{-\beta(G_A(z) - G_A(\infty))}\right].$$
The Journal of Chemical Physics

Sponding to the three asymptotic regimes: the neutral compounds show strong linear correlation for neutral compounds (uniformly shifts the permeability coefficient across the chemical rationale for the two contributions of Eq. (2): apK

To numerically test Eq. (6), we identify datapoints corresponding to the three asymptotic regimes: the neutral compounds (apK

More importantly, the asymptotes provide a physically motivated rationale for the two contributions of Eq. (2): apK

where we assume that all protonation states yield identical diffusivity. 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., apK

- Strong acids consist of solutes that display at least one functional group for which apK

- Strong bases would display at least one functional group, where bpK

The total resistivities still require integration over the reaction coordinate z, which we simplify to the largest contribution of the PMF. The effective resistivity model is equivalent to choosing the lowest PMF at any value of z: G_{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_{i} G_{i}(z). Assuming that the largest contribution of z in the permeability arises from the total resistivity at z^*, we obtain P = R_{i}^{-1}(z^*). This yields the following acid–base asymptotic regimes:

\[
\begin{align*}
\log_{10} P &= \begin{cases} 
\frac{\beta}{\ln 10} \frac{G_{N}(z^*) - \mathrm{pH} + \mathrm{apK}_{a}}{bpK_{a}} & \text{if } \mathrm{apK}_{a} < \mathrm{pH} \\
\frac{\beta}{\ln 10} \frac{G_{N}(z^*) + \mathrm{pH} - \mathrm{apK}_{a}}{bpK_{a}} & \text{if } \mathrm{apK}_{a} > \mathrm{pH} \\
\frac{\beta}{\ln 10} \frac{G_{N}(z^*)}{bpK_{a}} & \text{if } \mathrm{apK}_{a} \approx \mathrm{pH} \text{ and } \mathrm{bpK}_{a} \ll \mathrm{pH}.
\end{cases}
\end{align*}
\]
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. 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 $\sim 1.4$.

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. Inconsistent conclusions were drawn from different studies with DOPC, pointing to a thermodynamic drive that is weak at best. As for POPC, there is no sign of phase transition. As for POPC, there is no sign of phase transition. 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. 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
atomistic simulations and electron paramagnetic resonance spectroscopy experiments on the permeation of water and oxygen and found a permeability ratio of $P(\text{Ld})/P(\text{Lo}) \approx 3$. Here, the CG simulations yield a shift in $-\beta G_{\text{N}}(z^2)$, which translates into a ratio of the permeability coefficients of $\approx 25$. Our estimates are thus within one log$_{10}$ unit of the results of Ghyseels 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).

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. 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 $f^{1\text{D}} = c^{1\text{D}} + c^{1\text{D}}(apK_a - bpK_a - 2\beta W_{\text{Ol}} - \Delta G_{\text{W} \rightarrow \text{Ol}})$, where $apK_a$ and $bpK_a$ characterize acidity, $\Delta G_{\text{W} \rightarrow \text{Ol}}$ is the water/octanol partitioning coefficient, and $c^{1\text{D}}$ and $c^{1\text{D}}$ 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.

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 ($apK_a$ and $bpK_a$), 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 Kukhareno, Roberto Menichetti, and Yasemin Bozkurt Varolgunes for critical reading of the manuscript and Kian Kanekal and Martin Girard for insightful discussions. Data analysis relied extensively on the open source packages Numpy, Matplotlib, and Pandas. 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. It is openly available at https://doi.org/10.1021/acscentsci.8b00718.

REFERENCES


