Ligand biological activity predicted by cleaning positive and negative chemical correlations

Alpha A. Lee[a,1], Qingyi Yang[b], Asser Bassyouni[c], Christopher R. Butler[b], Xinjun Hou[b], Stephen Jenkinson[c], and David A. Price[b]

[a]Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom; [b]Medicine Design, Pfizer Inc., Cambridge, MA 02139; and [c]Drug Safety Research and Development, Pfizer Inc., La Jolla, CA 92121

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved November 27, 2018 (received for review July 15, 2018)

Predicting ligand biological activity is a key challenge in drug discovery. Ligand-based statistical approaches are often hampered by noise due to undersampling: The number of molecules known to be active or inactive is vastly less than the number of possible chemical features that might determine binding. We derive a statistical framework inspired by random matrix theory and combine the framework with high-quality negative data to discover important chemical differences between active and inactive molecules by disentangling undersampling noise. Our model outperforms standard benchmarks when tested against a set of challenging retrospective tests. We prospectively apply our model to the human muscarinic acetylcholine receptor M1, finding four experimentally confirmed agonists that are chemically dissimilar to all known ligands. The hit rate of our model is significantly higher than the state of the art. Our model can be interpreted and visualized to offer chemical insights about the molecular motifs that are synergistic or antagonistic to M1 agonism, which we have prospectively experimentally verified.

random matrix theory | ligand-based drug discovery | bioactivity prediction | machine learning | chemoinformatics

Finding novel hits to a target receptor is an important initial step in the long process of drug discovery. Although biochemical assays are increasingly high throughput, an experiment-only strategy that attempts to screen chemical space exhaustively remains intractable. To accelerate drug discovery, researchers have developed computer-aided virtual screening strategies in the literature over the last decades (1–5). Structure-based approaches require knowledge of the receptor structure and the binding site, and they predict the binding free energy by modeling protein–ligand interactions (6–8). However, determining the receptor structure and parameterizing protein–ligand interactions are often challenging, and notwithstanding those challenges, it is still computationally intensive to compute the protein–ligand binding free energy (9, 10).

Ligand-based methods sidestep the challenges of structure-based approaches and only require a set of molecules that are known to be active against a particular receptor or trigger a particular phenotype (11, 12). Those methods are built on the hypothesis that the receptor binding site recognizes a specific set of chemical motifs in a molecule, and those motifs can be uncovered from chemical motifs that are shared between the known active molecules. Therefore, an unknown molecule is likely to be active if it also contains those common chemical motifs and inactive otherwise. Those chemical motifs characterize the protein binding site, and proteins can be related based on the chemical similarity of their ligands (13, 14).

However, regardless of how one defines chemical motifs—common strategies include using pharmacological intuitions (15–17), enumerating all linear or circular fragments below a certain size around every atom (17, 18), or unsupervised learning (19)—a molecule has many motifs, but only a few are important for biological activity of a particular target. Unless one fortuitously knows a priori which motifs are important for a specific receptor and eliminates the other "nuisance" motifs, with a finite amount of data the nuisance motifs could drown out the motifs that are actually important for binding simply by chance. This problem is all of the more challenging, as it is often the confluence of different motifs rather than a single motif that drives binding, but correlations are known to be especially sensitive to noise due to undersampling (20–22). Pioneering advances in machine learning that infer the optimal representation of molecules directly from data (23–26) do not resolve this undersampling problem, as the available data are usually significantly less than the number of parameters in the model.

In this paper, we show that removing the noise arising from statistical undersampling—not having enough samples compared with the number of motifs—is needed to reveal chemical differences between active (positive) and inactive (negative) molecules and identify important chemical motifs that determine activity. We develop a statistical method based on random matrix theory and use our model to prospectively discover experimentally confirmed agonists of human muscarinic acetylcholine receptor M1 that are chemically dissimilar to known ligands. Our model also compares favorably with the prior art on retrospective benchmark tests. Importantly, we can interpret the model to offer pharmacological insights about the roles of different chemical motifs in determining activity.

Our work significantly extends a random matrix framework developed by some of us (27) by having an explicit statistical

Significance

Predicting ligand biological activity is a key challenge in drug discovery. Although there is an increasing amount of activity data, a data-driven approach needs to overcome the challenge that the number of molecules known to be active or inactive is vastly less than the number of possible chemical features that might determine binding. We develop a framework using random matrix theory that discovers important chemical features by disentangling undersampling noise. This method is used to prospectively discover four experimentally confirmed agonists of the human muscarinic acetylcholine receptor M1, a target for diseases such as Alzheimer's disease and schizophrenia. Crucially, our method is interpretable and yields prospectively validated chemical insights on the binding modes of the M1 receptor.

Author contributions: A.A.L., Q.Y., C.R.B., X.H., and D.A.P. designed research; A.A.L., Q.Y., A.B., C.R.B., X.H., S.J., and D.A.P. performed research; A.A.L., Q.Y., A.B., C.R.B., X.H., S.J., and D.A.P. analyzed data; and A.A.L. and Q.Y. wrote the paper.

Conflict of interest statement: Q.Y., A.B., C.R.B., X.H., S.J., and D.A.P. are current employees of Pfizer. The structure highlighted in red in Fig. 3*A* is exemplified in Patent WO/2013/072705.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: aal44@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810847116/-/DCSupplemental.

Published online February 11, 2019.

model for the inactive set, prospective experiments on a therapeutically relevant receptor, and robust retrospective tests with confirmed inactives as well as offering biochemical hypotheses through model visualization and interpretation. We present a prospective validation of the random matrix methodology applied to drug discovery. Our work also demonstrates the importance of high-quality negative data and methodologies that clean and exploit negative correlations.

Random Matrix Theory and Chemical Correlations

We focus on a popular set of descriptors used in chemoinformatics. Molecular fingerprints are typically constructed by first representing a molecule as a 2D molecular graph and then considering all possible bond paths (contiguous atoms connected by chemical bonds) within the molecule. Only identical molecules would share the same bond paths, and similar molecules share most bond paths; thus, comparing bond paths is a reasonable way to quantify chemical similarity. As the set of all possible bond paths is vast, typically fingerprints are defined by first considering bond paths that are within some radius of every atom in the molecule and then mapping these bond paths to a bit string of defined length through a hash function. Throughout this paper, we will use the 1,024-bit Morgan fingerprint of radius 3 (18) generated using the open source package `rdkit` (28).

To determine which bond paths are important for binding, we need to determine which bond paths are correlated in their presence/absence in the set of active molecules relative to the set of inactive molecules. Principal component analysis provides way to do that. Mathematically, each molecule can be characterized as a binary vector of length p, $\mathbf{f}_i \in \mathbb{R}^p$. The ensemble of N_+ active molecules can be arranged as a data matrix $R_+ = [\mathbf{f}_1; \mathbf{f}_2 \cdots \mathbf{f}_{N_+}] \in \mathbb{R}^{N_+ \times p}$, and similarly, R_- can be constructed from the N_- inactive molecules. We rescale the features such that the columns of $R_\pm$ have zero mean and unit variance; columns with zero variance are removed. Persistent correlations in bond paths can be identified from the eigendecomposition of each sample covariance matrix

$$C_\pm = \frac{1}{N_\pm} R_\pm^T R_\pm = \sum_{i=1}^{p} \lambda_i^\pm \mathbf{v}_i^\pm \otimes \mathbf{v}_i^\pm, \qquad [1]$$

where $\{\lambda_i^\pm\}$ are the eigenvalues and $\{\mathbf{v}_i^\pm\}$ are the eigenvectors. Each eigenvector $\mathbf{v}_i^\pm$ identifies a particular combination of the p bond paths, which explains a fraction $\lambda_i / \sum_i \lambda_i$ of the variance.

However, not all eigenvectors are equally important. The question of discriminating signal from noise due to the undersampling in the context of molecular fingerprints has been discussed in the literature (27, 29, 30). Under certain weak assumptions, if entries in $R_\pm$ are random and drawn from a Gaussian distribution with zero mean and unit variance, the probability of $C_\pm$ having an eigenvalue λ is given by the Marchenko–Pastur distribution (31)

$$\rho_\pm(\lambda) = \frac{\sqrt{\left[(1+\sqrt{\gamma_\pm})^2 - \lambda\right]_+^2 \left[\lambda - (1-\sqrt{\gamma_\pm})^2\right]_+^2}}{2\pi\gamma_\pm\lambda}, \qquad [2]$$

where $\gamma_\pm = p/N_\pm$ describes how well sampled the dataset of active or inactive molecules is and $(\cdot)_+ = \max\{\cdot, 0\}$. Eq. **2** provides a suitable null distribution—only eigenvectors with eigenvalues outside the Marchenko–Pastur distribution are statistically significant. In practice, as the Marchenko–Pastur distribution is nonzero only in the region $\lambda \in [(1-\sqrt{\gamma_\pm})^2, (1+\sqrt{\gamma_\pm})^2]$, only eigenvectors with an eigenvalue greater than $(1+\sqrt{\gamma_\pm})^2$ are significant. In other words, for a less well-sampled dataset ($\gamma_\pm$ large), an eigenvector needs to have a large eigenvalue (i.e., explains a lot of the variance in the data) before one can believe that it is significant. The statistically significant orthonormal eigenvectors are orthogonal chemical features that are relevant for binding. If there are $m_\pm$ significant eigenvalues, then the linear space spanned by those $m_\pm$ associated eigenvectors, $\mathbf{V}_\pm = \mathrm{span}\{\mathbf{v}_1^\pm, \mathbf{v}_2^\pm, \cdots \mathbf{v}_{m_\pm}^\pm\}$, is the subspace of chemical feature space that causes binding/nonbinding to that particular receptor.

Intuitively, a molecule is likely to be active if it is chemically similar to the set of known active molecules and dissimilar to the set of known inactive molecules. We can capture this intuition by requiring the molecule, represented as the Morgan fingerprint $\mathbf{u}$, to be close to the linear subspace $\mathbf{V}_+$ but far from $\mathbf{V}_-$. Therefore, we should classify a molecule $\mathbf{u}$ as active if

$$\mathcal{D}(\mathbf{u}, \mathbf{V}_+) < \mathcal{D}(\mathbf{u}, \mathbf{V}_-) + \epsilon, \qquad [3]$$

where ϵ is a tolerance parameter that captures the trade-off between false positive and false negative and $\mathcal{D}(\mathbf{u}, \mathbf{V}_\pm) = \left|\left|\mathbf{u} - \sum_{i=1}^{m_\pm} \left[\mathbf{v}_i^\pm \cdot (\mathbf{u} - \boldsymbol{\mu}_\pm)/\boldsymbol{\sigma}_\pm\right]\mathbf{v}_i^\pm\right|\right|_2$ is the distance between the vector $\mathbf{u}$, translated by the mean of the active/inactive set $\boldsymbol{\mu}_\pm$ and scaled by the variance $\boldsymbol{\sigma}_\pm$, and the linear subspace $\mathbf{V}_\pm$.

Eq. **3** is the central result of this paper. It extends the result of our previous work (27) by explicitly accounting for the set of inactive molecules. This is important, as the set of molecules tested is often clustered around scaffolds, and those scaffolds will appear as statistically significant eigenvectors regardless of whether they are important for binding. Therefore, one should focus only on correlations that are present in the active set but not those in the inactive set, which is the interpretation of Eq. **3**. Henceforth, we will refer to Eq. **3** as the random matrix discriminant (RMD).

Retrospective Benchmarks

We first validate the random matrix distribution as a suitable null hypothesis. Fig. 1 shows that the eigenvalue distribution of 200 random molecules drawn from ChEMBL (32) indeed agrees with the random matrix null distribution (Eq. **2**). The Kolmogorov–Smirnov test statistic is $D = 0.087$; thus, the hypothesis that the eigenvalue distribution follows Eq. **2** cannot be rejected at the 0.95 confidence level. A small number of eigenvalues are outside

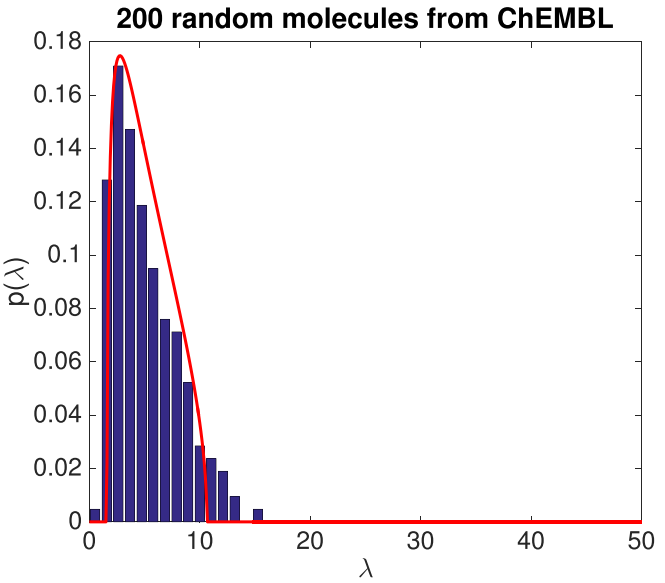


Fig. 1. The eigenvalue distribution of random molecules drawn from ChEMBL follows the random matrix distribution. The histogram shows the eigenvalue distribution of 200 random molecules drawn from ChEMBL, and the red curve is the random matrix distribution (Eq. **2**) for $p = 1024$ and $N = 200$.

the threshold predicted by Eq. **2** because of two reasons. First, Eq. **2** is derived asymptotically in the limit $p, n \to \infty$ with p/n fixed; thus, there are finite size corrections that we neglected. Second, Eq. **2** describes the typical behavior of random matrices rather than extreme value distributions (33). Nonetheless, by analyzing 10 batches of 200 random molecules drawn from ChEMBL, we find that more than 95% of the total number of eigenvalues are within the bound predicted by Eq. **2**; thus, Eq. **2** is a suitable null hypothesis.

Following our previous study (27), we benchmark RMD using the challenge of identifying ligands of human G protein-coupled receptors (GPCRs). We consider GPCRs where there are more than 500 known active molecules in ChEMBL and more than 150 inactive molecules from the internal Pfizer database; we consider only structures that are already in the public domain so that the dataset can be fully disclosed in *SI Appendix*. A ligand is considered active against a given target if its K_i, K_d, IC_{50}, or EC_{50} is 1 μM or less and inactive otherwise. Our previous study (27) only considered active molecules, because confirmed inactives from a "pharmacologically plausible" chemical space are difficult to obtain from the literature. This study importantly benefits from high-quality negative data from proprietary historic high-throughput screening campaigns. All in all, four receptors—μ opioid receptor (MOR1), serotonin receptor 2B (5-HT2B), α-2A adrenergic receptor (ADRA2A), and histamine H1 receptor—have sufficient numbers of disclosable confirmed inactives.

Table 1 shows that RMD outperforms the benchmark (27) as well as the naive Bayes method and classification based on the mean Tanimoto coefficient against active molecules in the training set. The latter two methods are common chemoinformatics methods used in the industry. Our method takes into account the mean and pairwise correlations of chemical features in a noise-robust manner; thus, a natural question to ask is whether our method outperforms nonlinear methods in the literature that use higher-order correlations. To make this comparison, we focus on the support vector machine (SVM) with a cubic kernel (which accounts for third-order correlations) and the Graph Convolutional Neural Network fingerprint (25). Table 1 also shows that RMD outperforms SVM as well as the Graph Convolutions [implemented in the open source package `DeepChem` (34)]. In all tests, the active and inactive sets are randomly split into a training set (90%) and a test set (10%). The figure of merit that we consider is the area under the curve (AUC) of the receiver operating characteristic. The mean AUC and standard error of the mean are estimated by analyzing 10 random partitions. In *SI Appendix*, we show that our method also outperforms other common machine learning methods for the Avalon fingerprint, another descriptor based on a handcrafted set of chemical features (16), showing that the importance of cleaning undersampling noise is independent of descriptor choice. The AUCs of the two fingerprints with RMD are comparable, and we use the Morgan 3 fingerprint for ease of directly interpreting the model in terms of correlations between chemical fragments (see Fig. 4 for visual interpretation).

One looming question is whether RMD performs well because the chemical space probed by Pfizer is different to the published literature on ChEMBL. To answer this question, we search our database for a receptor system where there are more than 500 active molecules and 150 inactive molecules. The receptor system that fulfills those criteria is the chemokine receptor type 5 (CCR5). Table 1 also shows that RMD works equally well when the chemical space of both the active and inactive molecules originates from the same source.

Thus far, we have only considered GPCRs. However, RMD is agnostic as to whether the target is a GPCR or even a protein receptor. To illustrate this point, Table 1 shows that RMD can accurately predict binding to human ether-à-go-go–related gene (hERG), a potassium ion channel that is an "antitarget" for drug discovery, as binding can cause cardiac arrest (35, 36). Both the active and inactive data come from internal screens. Following commonly used thresholds for hERG activity (35, 36), we classify compounds with an $IC_{50} < 9$ μM as active and those with an $IC_{50} > 29.9$ μM as inactive.

Prospective Discovery of Human M1 Agonists

We select the human muscarinic acetylcholine receptor M1 as a target to prospectively deploy our algorithm. The muscarinic acetylcholine receptors are members of the rhodopsin-like GPCRs and regulate the functions of the central and peripheral nervous systems. Currently, drugs that target muscarinic receptors have been approved for the treatment of chronic obstructive pulmonary disease, overactive bladder, and Sjogren's syndrome (37, 38). Gene knockout studies suggest that the M1 agonists may ameliorate the symptoms of Alzheimer's disease and related cognitive disorders as well as ameliorate psychosis-like symptoms in schizophrenia (37, 38).

We use data from a historic campaign to train our model. In total, 5,445 compounds were screened for agonist activity, of which 222 were active ($EC_{50} < 1$μM). It is apparent that the number of molecules, even in an industrial campaign, is small compared with the number of chemical features; thus, removing undersampling noise is essential. Fig. 2 *A* and *B* shows that the eigenvalue distribution of the active and inactive sets follows Eq. **2**, save for a few statistically significant eigenvectors. Fig. 2*C* reassures the reader by showing the classification accuracy of RMD on the historic data; as before, the data are partitioned into training/testing (90/10%) sets.

The model is then deployed to screen the entire `eMolecules` database (https://www.emolecules.com), a publicly accessible database of 5.9 million commercially available chemical compounds. We select the top 150 molecules using RMD and the naive Bayes classifier as well as by maximum Tanimoto similarity to training set, and we perform a prospective experimental test (*Methods*). Fig. 3 shows that RMD discovered four agonists. The closest structure in the training set by Tanimoto similarity is shown in Fig. 3, *Lower Insets*. RMD has found active molecules that are structurally dissimilar to molecules in the training set and can successfully hop between chemical scaffolds. In terms of Tanimoto similarity, none of the agonists that we found have a

Table 1. The AUC of our method, RMD, outperforms the random matrix theory benchmark (27) as well as the naive Bayes, Tanimoto similarity, SVM, and graph convolutional fingerprint methods

Target	RMD (Eq. 3)	Random matrix	Naive Bayes	Tanimoto	SVM	Graph convolutions
MOR1	**0.99 ± 0.001**	0.91 ± 0.003	0.95 ± 0.003	0.91 ± 0.005	0.70 ± 0.01	0.93 ± 0.007
5-HT2B	**0.93 ± 0.007**	0.82 ± 0.005	0.85 ± 0.01	0.85 ± 0.008	0.67 ± 0.02	0.87 ± 0.01
ADRA2A	**0.90 ± 0.01**	0.75 ± 0.01	0.84 ± 0.009	0.77 ± 0.02	0.61 ± 0.03	0.90 ± 0.006
Histamine H1	**0.97 ± 0.003**	0.87 ± 0.005	0.94 ± 0.007	0.87 ± 0.008	0.65 ± 0.02	0.84 ± 0.01
CCR5	**0.92 ± 0.007**	0.89 ± 0.008	0.90 ± 0.006	0.86 ± 0.01	0.68 ± 0.02	0.91 ± 0.009
hERG	**0.83 ± 0.02**	0.51 ± 0.01	0.79 ± 0.01	0.66 ± 0.02	0.60 ± 0.02	0.71 ± 0.01

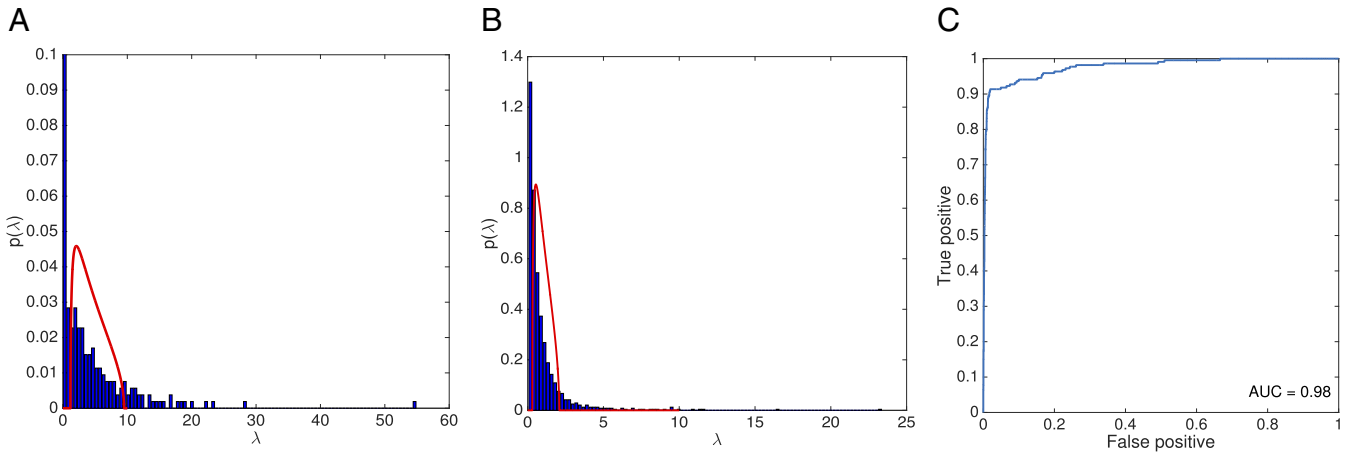


Fig. 2. Our random matrix model captures the statistics of M1 agonists and confirmed inactives from a historic campaign. The random matrix distribution (red curve) agrees with the histograms of eigenvalues of the (*A*) active agonists and (*B*) confirmed inactives. (*C*) A classification model built using the statistically significant eigenvectors achieves an accuracy of 98%.

Tanimoto coefficient greater than 0.41 to any molecule in the training set—this level of (dis-)similarity is the same as two molecules randomly drawn from ZINC, a large database often used in virtual screening (39).

The hit rate of RMD is also greater than common chemoinformatics methods. For comparison, the naive Bayes classifier only discovered two agonists, and the agonists predicted by the Tanimoto classifier are all inactive. Moreover, the hit rate of a typical high-throughput screen is ∼0.01–0.14% (40); thus, our model performance is around almost 40-fold better than the background hit rate. This finding corroborates the results from the retrospective test (Table 1).

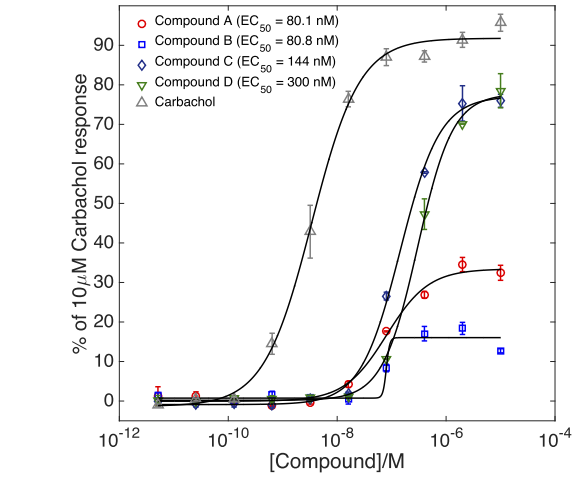
Model Interpretation and Visualization

The ability of RMD to prospectively discover M1 agonists that are chemically dissimilar to the training set motivates us to unbox the model and interpret its "reasoning." RMD classifies molecules based on their distance from the linear subspace spanned by motifs in the active set relative to the inactive set. As such, an intuitive way to interpret the model is to directly visualize the difference between the two linear subspaces. We can operationalize this difference by defining the following effective correlation matrix:

$$\mathcal{A} = \sum_{i=1}^{m_+} \mathbf{v}_i^+ \otimes \mathbf{v}_i^+ - \sum_{i=1}^{m_-} \mathbf{v}_i^- \otimes \mathbf{v}_i^- . \qquad [4]$$

A positive entry $\mathcal{A}_{ij}$ denotes that motifs i and j are jointly and positively contributing to binding, whereas a negative entry denotes that the motifs are jointly and negatively effecting binding. A similar approach has been proposed in the context of principal component analysis, where the goal is to reveal the contrast between two groups (41).

Fig. 4 shows that the motif–motif correlation matrix (4) can be visualized and interpreted as a network of chemical fragments. Each motif is an entry of the molecular fingerprint. Due to bit collision, multiple fragments can be assigned to the same motif. Interestingly, the model identifies the carbamate fragment (node 10 in Fig. 4) as distinct and positively correlated to the piperazine fragments (node 2) and puts them together to form the active molecules C and D (see Fig. 3). Crucially, the model is able to learn that a carbamate with a ternary nitrogen is the relevant fragment rather than a six-membered ring piperidine-*N*-carboxylate or seven-membered ring azepane-*N*-carboxylate present in the training set. In other words, the model interpolates between structures in the training set by learning generalizable chemical features rather than simply memorizing the training data. Another trend that the model extracts is the correlation between carbamate (node 10) and aromatic fragments (nodes



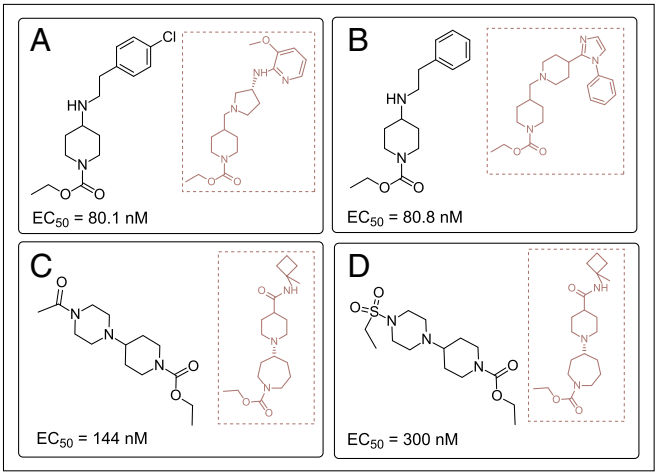


Fig. 3. The RMD model discovered four human M1 agonists, compounds *A*–*D*. (*Upper*) The measured dose–response curves for the agonists. (*Lower*) The molecular structures of the agonists (*A*–*D*); *Insets* show the closest molecule in the training set by Tanimoto coefficient.

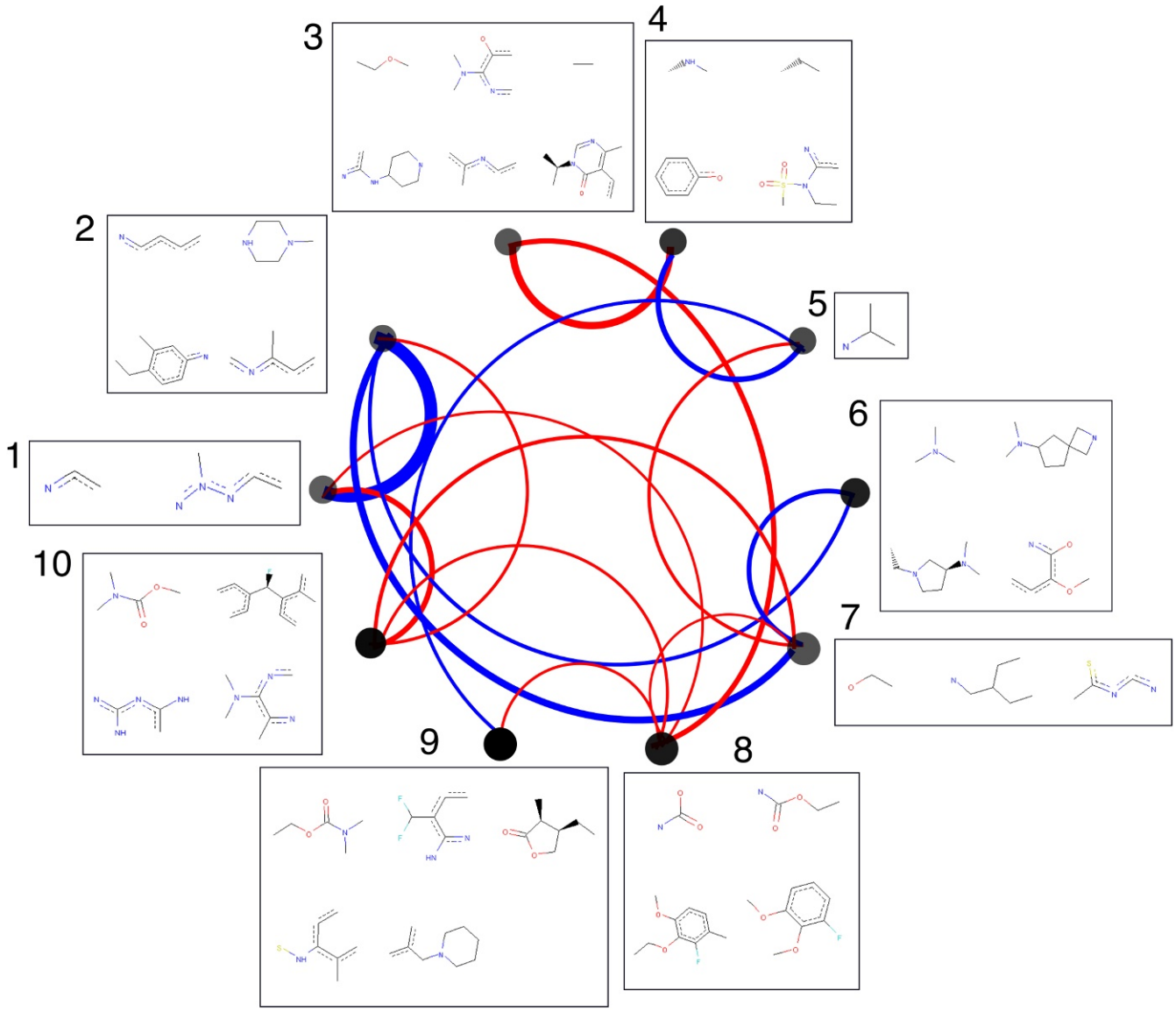


Fig. 4. Our model can be interpreted as a network of features, where each feature is an entry of the molecular fingerprint. The opacity of the nodes is proportional to the difference in the number of times that the feature is present in the active set relative to the inactive set; only the top 10 features are shown. Red (blue) edges correspond to a positive (negative) correlation, and the width of the edges is proportional to the strength of the correlation.

1 and 2), agreeing with the heuristic in the literature that M1 agonists generally have a hydrogen bond acceptor and a distal aromatic moiety (42).

The negative correlations are also chemically significant. One interesting prediction is that, although the aromatic nitrogen fragment (node 1) and the piperazine fragment (node 2) are both overrepresented in the active molecules, having both fragments is strongly detrimental to agonist activity. Intriguingly, the model does not predict a similar negative correlation between the aromatic nitrogen fragment and the cognate piperidine fragment, which is structurally identical to piperazine except that it has one fewer nitrogen atom. This subtle prediction of an activity difference between piperidine and piperazine is ripe for experimental testing.

We search our internal database for pairs of molecules, both containing an aromatic nitrogen moiety, and share the same chemcial structure except one has a piperidine ring and the other has a piperazine ring. We exclude molecules used in model training, testing, or in the `e − Molecules` database. Those "matched molecular pairs" (43) were experimentally tested for M1 agonist activity. Note that none of those matched molecular pairs are considered in the original model; thus, this is a completely independent out-of-sample validation of model prediction. Fig. 5 shows that, in all cases, swapping the piperazine motif for the

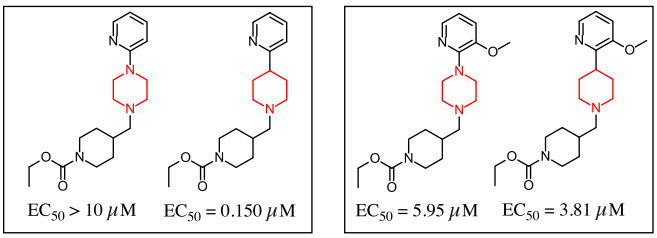


Fig. 5. Prospective matched molecular pair analysis corroborates the significant negative correlation between the piperazine and the aromatic nitrogen motif that the model predicts.

piperidine motif leads to a significant increase in activity. This confirms the strongly negative motif–motif correlation that our model has picked up as well as demonstrates how we can exploit this chemical insight to introduce a small chemical modification to the molecule that significantly increases binding affinity. The visualization in Fig. 4 combined with insights on synthetic accessibility can provide ideas for de novo design.

Conclusion

We derived a statistical framework inspired by random matrix theory for ligand biological activity prediction that discovers important correlations between chemical features by disentangling undersampling noise and subtracting the correlation structure of the active compounds from the inactive compounds. We showed that the model outperforms standard benchmarks when tested against a set of challenging retrospective tests. We prospectively applied the model to the human muscarinic acetylcholine receptor M1 and found four experimentally confirmed agonists that are chemically dissimilar to all known agonists. Moreover, we can visualize and interpret the model to yield pharmacological insights. Our method distills which combinations of chemical motifs are positively/negatively responsible for binding to the receptor; it predicts pairs of motifs in which each individual motif is overrepresented in the active molecule (thus naively expected to be important for binding), but when occurring as a pair, it is detrimental to binding. We experimentally validated the pharmacological insights predicted by the model, showing how one can exploit insights afforded by the model to make a small chemical change to a molecule to evade those negative motif pairs, drastically improving potency. A broader conclusion of our study is the power of high-quality inactive data, which allows the model to generate meaningful hypotheses about motif combinations that lead to inactivity.

Methods

M1 Agonist Assay. CHO cells stably expressing the human muscarinic 1 receptor were plated at a density of 7,500 cells per well (50 μL per well) in black-walled, clear-bottomed 384-well plates, and they were incubated overnight (20–24 h) in a 37 °C humidified incubator with 5% carbon dioxide. Agonist activity was determined by measuring compound stimulated changes in intracellular calcium levels using a calcium sensitive dye. Before the start of the experiment, the medium was removed from the plates; 80 μL of HBSS containing Hepes (20 mM), Calcium 5 dye (catalog no. R8186; Molecular Devices), and probenecid (1.25 mM) was added to each well, and the plate was returned to the incubator for 1 h to allow for dye loading. For compound assessment in the agonist format, 10 μL of the compound solution was added to each well by the FLIPR Tetra instrument (Molecular Devices), and the change in fluorescence from baseline over a 60-s period (excitation: 470–495 nM; emission: 515–575 nM) was measured.

Data and Code Availability. Active and inactive compounds for MOR1, 5-HT2B, ADRA2A, histamine H1, CCR5, M1, and hERG, as well as data from the prospective experiments on M1, are available in *Datasets S1–S7*. For CCR5 and M1, the training set compounds are reported as 1024 bit Morgan fingerprint of radius 3. The code used to perform RMD analysis is on GitHub (https://github.com/alphaleegroup/RandomMatrixDiscriminant).

ACKNOWLEDGMENTS. A.A.L. acknowledges support from the Winton Program for the Physics of Sustainability.

1. Alvarez J, Shoichet B, eds (2005) *Virtual Screening in Drug Discovery* (CRC Press, Boca Raton, FL).
2. Klebe G (2006) Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discov Today* 11:580–594.
3. Kubinyi H, et al. (2008) *Virtual Screening for Bioactive Molecules* (John Wiley & Sons, Weinheim, Germany), Vol. 10.
4. Koeppen H (2009) Virtual screening-what does it give us? *Curr Opin Drug Discov Dev* 12:397–407.
5. Schneider G (2010) Virtual screening: An endless staircase? *Nat Rev Drug Discov* 9:273–276.
6. Lyne PD (2002) Structure-based virtual screening: An overview. *Drug Discov Today* 7:1047–1055.
7. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-based virtual screening for drug discovery: A problem-centric review. *AAPS J* 14:133–141.
8. Lionta E, Spyrou G, Vassilatis D, Cournia Z (2014) Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Curr Top Med Chem* 14:1923–1938.
9. Chodera JD, et al. (2011) Alchemical free energy methods for drug discovery: Progress and challenges. *Curr Opin Struct Biol* 21:150–160.
10. Hansen N, Van Gunsteren WF (2014) Practical aspects of free-energy calculations: A review. *J Chem Theor Comput* 10:2632–2647.
11. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J Chem Inf Model* 50:205–216.
12. Ripphausen P, Nisius B, Bajorath J (2011) State-of-the-art in ligand-based virtual screening. *Drug Discov Today* 16:372–376.
13. Keiser MJ, et al. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25:197–206.
14. Keiser MJ, et al. (2009) Predicting new molecular targets for known drugs. *Nature* 462:175–181.
15. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 42:1273–1280.
16. Gedeck P, Rohde B, Bartels C (2006) QSAR- how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J Chem Inf Model* 46:1924–1936.
17. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminformatics* 5:26.
18. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742–754.
19. Schneider N, Fechner N, Landrum GA, Stiefl N (2017) Chemical topic modeling: Exploring molecular data sets using a common text-mining approach. *J Chem Inf Model* 57:1816–1831.
20. Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Phys Rev Lett* 83:1467–1470.
21. Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Stanley HE (1999) Universal and nonuniversal properties of cross correlations in financial time series. *Phys Rev Lett* 83:1471–1474.
22. Bun J, Bouchaud JP, Potters M (2017) Cleaning large correlation matrices: Tools from random matrix theory. *Phys Rep* 666:1–109.
23. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. *IEEE Trans Neural Networks* 20:61–80.
24. Lusci A, Pollastri G, Baldi P (2013) Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. *J Chem Inf Model* 53:1563–1575.
25. Duvenaud DK, et al. (2015) Convolutional networks on graphs for learning molecular fingerprints. *Adv Neural Inf Process Syst* 28:2224–2232.
26. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P (2016) Molecular graph convolutions: Moving beyond fingerprints. *J Computer-Aided Mol Des* 30:595–608.
27. Lee AA, Brenner MP, Colwell LJ (2016) Predicting protein–ligand affinity with a random matrix framework. *Proc Natl Acad Sci USA* 113:13564–13569.
28. GitHub (date unknown) RDKit: Open-Source Cheminformatics. Available at www.rdkit.org. Accessed January 6, 2019.
29. Lee AA, Brenner MP, Colwell LJ (2017) Optimal design of experiments by combining coarse and fine measurements. *Phys Rev Lett* 119:208101.
30. Cortes Cabrera A, Petrone PM (2018) Optimal HTS fingerprint definitions by using a desirability function and a genetic algorithm. *J Chem Inf Model* 58:641–646.
31. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Math USSR-Sbornik* 1:457–483.
32. Gaulton A, et al. (2016) The ChEMBL database in 2017. *Nucleic Acids Res* 45:D945–D954.
33. Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* 29:295–327.
34. Wu Z, et al. (2018) MoleculeNet: A benchmark for molecular machine learning. *Chem Sci* 9:513–530.
35. C Braga R, et al. (2014) Tuning hERG out: Antitarget QSAR models for drug development. *Curr Top Med Chem* 14:1399–1415.
36. Didziapetris R, Lanevskij K (2016) Compilation and physicochemical classification analysis of a diverse hERG inhibition database. *J Computer-Aided Mol Des* 30:1175–1188.
37. Wess J, Eglen RM, Gautam D (2007) Muscarinic acetylcholine receptors: Mutant mice provide new insights for drug development. *Nat Rev Drug Discov* 6:721–733.
38. Kruse AC, et al. (2014) Muscarinic acetylcholine receptors: Novel opportunities for drug development. *Nat Rev Drug Discov* 13:549–560.
39. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2013) Molecular similarity in medicinal chemistry: Miniperspective. *J Med Chem* 57:3186–3204.
40. Zhu T, et al. (2013) Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis: Miniperspective. *J Med Chem* 56:6560–6572.
41. Abid A, Bagaria VK, Zhang MJ, Zou J (2017) Contrastive principal component analysis. arXiv:1709.06716.
42. Bhandare RR, Canney DJ (2011) Modifications to five-substituted 3, 3-diethyl-4, 5-dihydro-2 (3h)-furanones en route to novel muscarinic receptor ligands. *Med Chem Res* 20:558–565.
43. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool: Miniperspective. *J Med Chem* 54:7739–7750.