Predicting protein-ligand affinity with a random matrix framework

Alpha A. Lee[a], **Michael P. Brenner**[a], **and Lucy J. Colwell**[b]

[a]School of Engineering and Applied Sciences and Kavli Institute of Bionano Science and Technology, Harvard University, Cambridge, MA 02138, USA; [b]Department of Chemistry, University of Cambridge, CB2 1EW, Cambridge, UK

This manuscript was compiled on October 18, 2016

Rapid determination of whether a candidate compound will bind to a particular target receptor remains a stumbling block in drug discovery. We use an approach inspired by random matrix theory to decompose the known ligand set of a target in terms of orthogonal "signals" of salient chemical features, and distinguish these from the much larger set of ligand chemical features that are not relevant for binding to that particular target receptor. After removing the noise caused by finite sampling, we show that the similarity of an unknown ligand to the remaining, cleaned chemical features is a robust predictor of ligand-target affinity, performing as well or better than any algorithm in the published literature. We interpret our algorithm as deriving a model for the binding energy between a target receptor and the set of known ligands, where the underlying binding energy model is related to the classic Ising model in statistical physics.

Finding new ligands that bind to a given target is both a crucial step and a major stumbling block in modern drug discovery. Numerous attempts have been made to develop computational algorithms to predict the binding affinity of a ligand to a given receptor, allowing potential compounds to be screened *in silico*, reducing costs and saving time. In particular, in response to the wealth of experimental data that exists both within pharmaceutical companies, and also in freely accessible online databases such as ChEMBL [1], approaches that attempt to 'learn' from this data are increasingly gaining attention [2].

An intuitive data-driven approach builds on the hypothesis that chemical commonalities amongst the known ligand set reveal salient features of the binding site. A corollary is that ligands with similar chemical functionality are expected to share similar binding affinity towards a particular receptor [3, 4]. This suggests that the known ligand set of a given target can be used to learn criteria that predict whether a novel ligand will bind to the target. This ligand-based approach is a powerful paradigm that does not require structural information about the receptor, which is potentially arduous to obtain, unlike other more atomistic methods such as docking or molecular dynamics.

Any ligand-based method requires a way to quantify the chemical functionalities of a ligand, and various chemical descriptors have been proposed. Examples include a vector of measured or predicted physical properties [5–8], a vector enumerating the presence or absence of known functional groups on the ligand [9, 10], a vectorial representation of connectivities in the molecular graph [11, 12] (known also as molecular fingerprints) and simply the three dimensional shape of the ligand [13–16]. Existing approaches then take the descriptor associated with each ligand and compare ligands with each other, for example through the Tanimoto coefficient [17, 18].

Nonetheless, regardless of how ligand chemical functionalities are quantified, without fortuitously knowing *a priori* which ligand features determine binding, most of the chemical features describing the ligand are likely irrelevant. While some of the features in the descriptor determine binding to the receptor interest, others are not and simply add background noise. Moreover, for any particular receptor, the known set of ligands that bind to it is often smaller, or of the same order of magnitude as the number of potentially relevant chemical features. As such, the problem of ligand-based binding prediction can be recast as a problem in signal processing – can we identify those chemical ligand features that determine binding (i.e. the "signal") amid many irrelevant ones (the "noise") in the regime where the amount of data is not significantly larger than the amount of potentially relevant information?

Random Matrix Theory (RMT) provides a natural mathematical framework for addressing this issue. Physical applications of RMT includes Wigner's study of the spectra of heavy atoms [19]. In the context of data analysis, RMT gives a null model for the similarity between samples (ligands) that can be expected by chance due to finite sampling [20]. Powerful analytical tools from RMT define a precise threshold that distinguishes the similarity that can be expected by chance from that which is caused by signal. These tools enable an effective and simple denoising algorithm, which allows us to recover the statistically significant signals. This denoising algorithm has been used in different fields, ranging from finance [21–23] to face recognition [24, 25].

Significance Statement

Developing computational methods to screen ligands against protein targets is a major challenge for drug discovery. We present a robust mathematical framework, inspired by random matrix theory, which predicts ligand binding to a target given the known ligand set of that target. Our method considers binding prediction as a denoising problem, recognizing that only some of the chemically important features associated with each ligand contribute to binding to a particular receptor. We use correlations amongst chemical features in the known ligand set, combined with random matrix theory, to eliminate statistically insignificant correlations. Our method outperforms existing algorithms in the literature. We show that our algorithm has the physical interpretation of estimating the ligand–target binding energy.

Author contributions: A.A.L., M.P.B. and L.J.C.. designed research, performed research, analyzed data and wrote the paper.

The authors declare no conflict of interest.

[1] E-mails: ljc37@cam.ac.uk (LJC), alphalee@g.harvard.edu (AAL)

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

This manuscript contains three major results: First we show that for a randomly chosen set of molecules, the eigenvalue distribution of the covariance matrix of chemical descriptors agrees with the canonical Marčenko-Pastur (MP) distribution [26] of RMT, expected in the absence of any significant signal. Second, if we consider descriptors of pharmacologically similar molecules, i.e. those that bind to the same protein receptor, then part of the eigenvalue spectrum agrees with the MP distribution, but crucially there are eigenvalues that deviate significantly from it. These eigenvalues, and their corresponding eigenvectors, describe the statistically significant signals. The most common substructure of these eigenvectors corresponds to pharmacophores. Using these two results, we can predict with higher accuracy than known methods when an unknown ligand will bind to a receptor, constructing a unique model for each protein receptor. Finally, we provide a physical interpretation of the success of the algorithm – namely, that it is effectively inferring a model of the ligand-protein binding energy from the covariance structure of fingerprints that bind to a target protein. The underlying mathematical model is closely related to the classic Ising Model in statistical physics.

Random Matrix Theory Framework

To motivate the random matrix theory framework, we focus on a popular set of descriptors that are often used in cheminformatics. Molecular fingerprints are typically constructed by first representing a ligand as a two dimensional molecular graph, and then considering all possible bond paths within the molecule [11, 12]. The set of bond paths that characterize each molecule is unique, so that only identical molecules share exactly the same bond paths; similar molecules share most bond paths. Since the set of all possible bond paths is vast, typically fingerprints are defined by first considering bond paths that are below some threshold length (i.e. within some radius of every atom of the structure) and then mapping these bond paths to a bit-string of defined length (a molecular “fingerprint” [27]) through a hash function.

The fundamental aim is to detect similarity among a set of binary strings of the same length, where each bit represents the presence or absence of a molecular feature. There is significant noise in these bit strings, because only some of the bits are truly informative - for any particular receptor, not all bond paths are equally relevant to ligand-target binding. If the individual bits of the binary strings were chosen randomly, with no information about ligand-target binding, then Random Matrix Theory predicts that the eigenvalue distribution of the covariance matrix of the bit strings obeys a specific analytical function known as the Marčenko-Pastur distribution. Therefore a highly accurate test for detecting the presence of non-random commonalities among a set of strings is to compare the eigenvalue spectrum of their covariance matrix to the MP distribution. Any deviation necessarily reflects the presence of a signal in the data, which in this case are sets of molecular features that characterise the chosen ligand-target interaction.

Mathematically, we represent the k^{th} ligand associated with the chosen receptor as a row vector of bits $\mathbf{f}_k$ using the Morgan fingerprint algorithm with radius 3, implemented using the package `rdKit` [28]. The ensemble of N ligands that bind to the chosen receptor can be arranged as a data matrix $A = [\mathbf{f}_1; \mathbf{f}_2 \cdots \mathbf{f}_N] \in \mathbb{R}^{N \times p}$, where the value of N will vary between receptors. We then remove repeated columns of the data matrix, which correspond to redundant information, and convert the data matrix to z-scores by subtracting the column mean and normalising each column to have unit variance. This allows us to construct the N by N correlation matrix $C = A^T A/N$. In general, for well-sampled data, large entries in C would indicate relationships between specific molecular features, suggesting that these features do not occur independently of one another in this dataset.

A fundamental result from random matrix theory describes the eigenvalue distribution of the correlation matrix C analytically — under certain weak assumptions, if entries in A are drawn from a Gaussian distribution with zero mean and unit variance, the probability of A having an eigenvalue λ is given by the Marčenko-Pastur distribution [26]

$$\rho(\lambda) = \frac{\sqrt{\left[\left(1+\sqrt{\gamma}\right)^2 - \lambda\right]_+ \left[\lambda - \left(1-\sqrt{\gamma}\right)^2\right]_+}}{2\pi\gamma\lambda} \qquad [1]$$

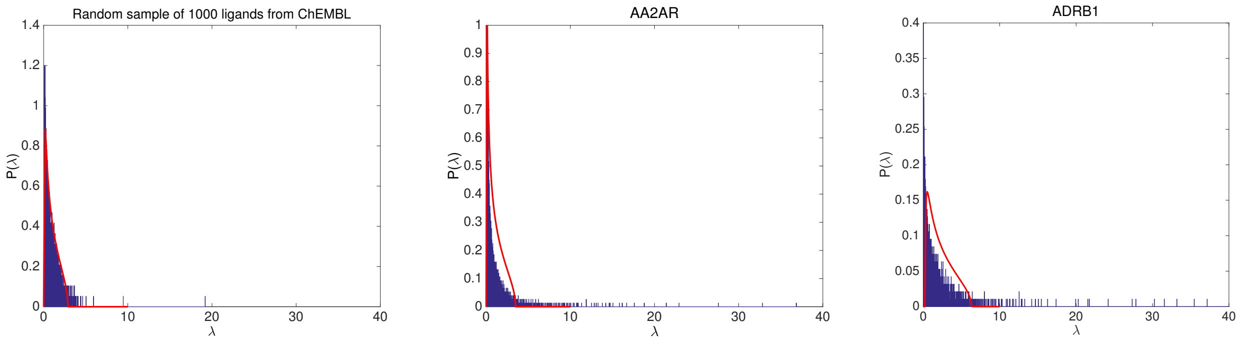
where $\gamma = p/N$ describes how well-sampled the dataset is. The probability that a random matrix has eigenvalues larger than $\left(1+\sqrt{\gamma}\right)^2$ in the absence of any signal is vanishingly small. Thus the key insight gained from equation (1) is that only eigenvalues above $\left(1+\sqrt{\gamma}\right)^2$ correspond to statistically significant signals.

Figure 1a shows that the eigenvalue distribution of the correlation matrix of 1000 ligands drawn randomly from ChEMBL [1] agrees quantitatively with the MP distribution. However, if instead we choose the ligands non-randomly, by choosing the ligand sets associated with a particular protein receptor, we find a significant number of eigenvalues above the MP threshold. As examples, Fig. 1B, C shows the eigenvalue distribution from ligand sets from ChEMBL associated with two G-protein coupled receptors, the adenosine A2a receptor (AA2AR) and the β_1 adrenergic receptor (ADRB1).

The Marčenko-Pastur distribution thus suggests an intuitive denoising algorithm for ligands binding to a particular receptor; only eigenvectors with eigenvalues larger than the MP upper bound correspond to statistically significant features of the receptor; the other eigenvectors simply reflect random noise caused by finite sampling. The set of statistically significant features, represented as orthonormal eigenvectors, are thus *orthogonal chemical features* relevant for ligand binding. In other words, if there are m eigenvalues greater than the MP upper bound, then the linear space spanned by the m associated eigenvectors, $\mathbf{V} = \mathrm{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_m)$ is the subspace of chemical feature space that facilitates binding to that particular receptor.

Classification of unknown ligands

Intuitively, if an unknown ligand is sufficiently similar to the set of known ligands binding to a receptor, the unknown ligand will likely also bind to the receptor. The random matrix framework gives a precise mathematical statement for this intuition: an unknown ligand is predicted to bind to a receptor if the bitstring vector corresponding to the unknown ligand (after transformation to z-score by subtracting the sample mean and normalising by sample variance) lies close to the subspace $\mathbf{V}$.



(a)

(b)

(c)

Fig. 1. The Marčenko-Pastur distribution (red curve) provides the null hypothesis for ligand-ligand correlations expected in the absence of signal. The eigenvalue distribution is plotted for the correlation matrix of (a) a random sample of 1000 ligands from ChEMBL, and the ligand set of the (b) adenosine A2a receptor and (c) β_1 adrenergic receptor

Let $\mathbf{u}$ by the vector of z-scores corresponding to the unknown ligand. The projection of $\mathbf{u}$ onto $\mathbf{V}$ is given by

$$\mathbf{u}_p = \sum_{i=1}^{m} (\mathbf{v}_i \cdot \mathbf{u})\mathbf{v}_i. \qquad [2]$$

$\mathbf{u}$ lies in the subspace $\mathbf{V}$ if and only if $\mathbf{u} = \mathbf{u}_p$. The distance between $\mathbf{u}$ and $\mathbf{u}_p$ is thus a quantitative metric of similarity between the unknown ligand and the set of ligands binding to the receptor in question. The ligand is predicted to bind if and only if

$$||\mathbf{u} - \mathbf{u}_p|| < \epsilon, \qquad [3]$$

where $||\cdot||$ is the Euclidean norm, and ϵ is a threshold parameter. Equation (3) has the chemical interpretation that one can be confident a ligand will bind to the receptor if it contains pharmacophores found in known ligands, and is minimally decorated with other functional groups. A pharmacophore is typically a small fragment (c.f. Figure 4), and the chemical properties of the resulting molecule will increasingly deviate from those of the pharmacophore as one incorporates additional functional groups. The threshold parameter ϵ allows the tolerance of the analysis to the presence of other functional groups to be controlled, and hence an appropriate false positive/false negative tradeoff selected; this is discussed in detail below.

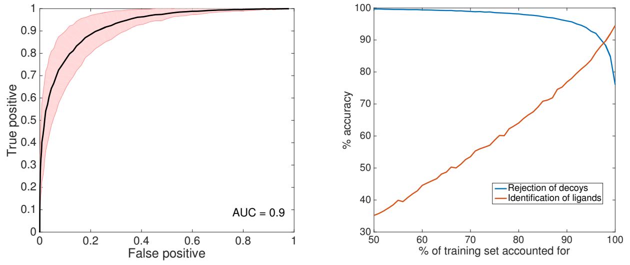
To test this, we consider human G-protein coupled receptors reported in ChEMBL. A ligand is considered to bind to a given target if their K_i, K_d, IC_{50} or EC_{50} is 1 μM or less. We consider only GPCRs with more than 120 known ligands reported in ChEMBL. We randomly sort ligands into a training set (80 %) and a verification set (20 %). To test for false positives, we need compounds that do *not* bind to the receptor. Negative results are seldom reported and the judicious selection of decoys is still a subject of intense research effort [29]. In our analysis, we use a random selection of 1000 compounds from ChEMBL as a proxy. The median number of ligands associated with each G-protein coupled receptor is ~ 400, thus even if the actual ligand set is an order of magnitude larger than those that are known, it still represents a negligible proportion of the 1,583,897 compounds in ChEMBL. Therefore, a random selection of 1000 ligands from ChEMBL is unlikely to contain any ligand that binds to a particular GPCR.

The receiver operating characteristic (ROC) curve plots the accuracy of identifying ligands (true positives) as function of false positive predictions. This characteristic is commonly used to quantify the performance of classification algorithms. In particular, the area under the ROC (the so called AUC) is the crucial figure of merit: the closer the AUC is to 1, the better the classifier. Figure 2a shows that our algorithm has a mean AUC of 0.9, surpassing methods commonly used in the literature, which have a mean AUC of $0.7 - 0.8$ [30]. As such, our algorithm comfortably outperforms the prior art.

The ROC curve is plotted by varying ϵ, the threshold parameter in Equation (3). Figure 2b shows the effect of varying ϵ, represented as the percent of the training set accounted for by each choice of ϵ. A stringent choice of ϵ corresponds to a large portion of the training set being rejected by the threshold (3), resulting in a low false positive rate but a high false negative rate. Vice versa, an ϵ value that accounts for a larger portion of the training set has higher false positive rate but lower false negative rate. In the remainder of this paper, we choose ϵ so that 95% of the training set lies within the threshold (3). With this heuristic choice, the algorithm picks out 84% of the verification set as ligands with a 7% false positive rate (i.e. it rejects 93% of randomly selected ligands from ChEMBL).

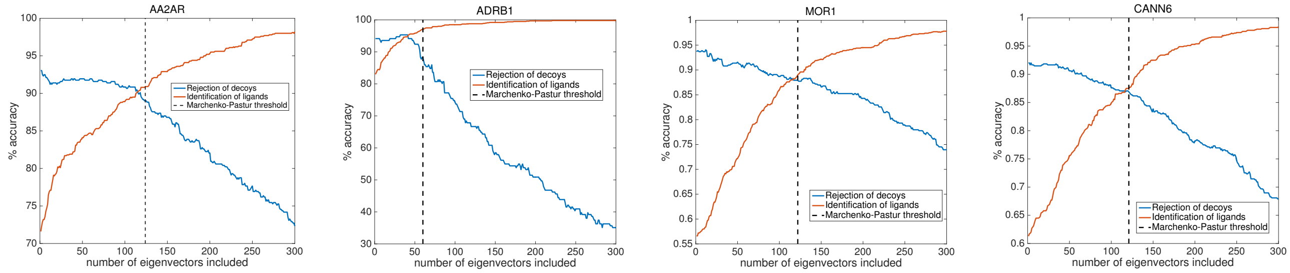
The random matrix distribution (Equation (1)) is crucial to the success of our algorithm. Figure 3 shows that including too many eigenvectors into $\mathbf{V}$ increases the false positive rate, whereas including too few eigenvectors decreases the success rate of picking out ligands from the verification set. The balance between overfitting and underfitting is achieved close to the MP bound (as the bound is probabilistic, slight sample-to-sample deviation is expected). Although Figure 3 only shows the results for the adenosine A2a receptor, the β_1 adrenergic receptor, the μ_1 opioid receptor and the cannabinoid CB1 receptor, the near optimality of the MP bound is general.

We also report that the statistically significant eigenvectors picked out by our algorithm represent pharmacophores. Formally, a fingerprint cannot be inverted directly to give a unique chemical structure because multiple structures can lead to the same fingerprint. Nonetheless we can infer the structural motif that an eigenvector represents by the common substructure amongst those ligands that lie closest to that eigenvector. Figure 4 shows the structural motif corresponding to the top two eigenvalues of the adenosine A2a receptor and the β_1 adrenergic receptor. Strikingly, the first eigenvector



(a) (b)

Fig. 2. Our RMT-inspired algorithm classifies ligands with high accuracy. (a)The receiver operating characteristic curve of our algorithm. The area under the curve (AUC) of the mean ROC curve is 0.9. The shaded region shows one standard deviation in the true positive, corresponding to AUC=0.86-0.95. (b) Accuracy at identifying ligands and rejecting decoys plotted as a function of percent of the training set rejected by the choice of the threshold ϵ.



(a) (b) (c) (d)

Fig. 3. The Marčenko-Pastur bound strikes a balance between overfitting and underfitting. The % accuracy in identifying ligands from the verification set and rejecting ligands randomly selected from ChEMBL is shown as a function of the number of eigenvectors included in **V** for (a) adenosine A2a receptor, (b) β_1 adrenergic receptor, (c) μ_1 opioid receptor and (d) cannabinoid CB1 receptor.

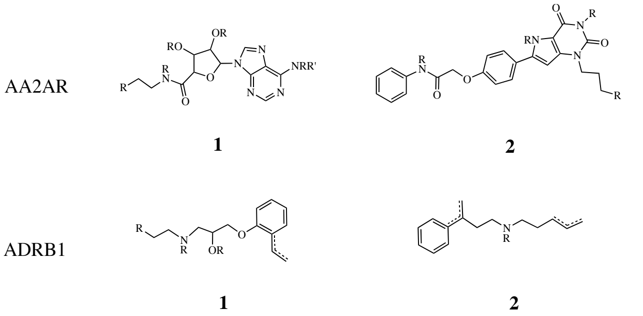


Fig. 4. The chemical motif corresponding to the first and second eigenvector of the adenosine A2a receptor and β_1 adrenergic receptor. The motif is obtained computing the common structure amongst the top 20 ligands ordered by the magnitude of the dot product between its fingerprint and the eigenvector.

of the adenosine A2a receptor is precisely the adenine motif. The second eigenvector contains a thymine motif fused to a more complex scaffold. For the β_1 adrenergic receptor, the top eigenvector is the structural motif of β blockers (e.g. propranolol), a class of successful antagonists which are used e.g. to treat hypertension.

A Physical Model

Before concluding, we address the question of why this algorithm might prove effective. What is the physics encoded in those eigenvalues larger than the Marčenko-Pastur threshold, and their associated eigenvectors?

The clearest way of determining which ligands bind to a given protein would be to accurately predict the binding energy of every possible ligand to the protein. The ligand set of the protein is then given by the set of ligands with a binding affinity greater than some threshold. Accurate determination of this binding energy is extremely computationally intensive. Nonetheless, even without a first principles determination of the ligand binding energy, we might still hope to *parameterize* a model of protein-ligand binding, where the parameters are determined from the set of ligands that bind to a given protein target. If sufficiently accurate, such a model of the binding energy could potentially still give accurate predictions as to which ligands bind to a given target protein.

We now demonstrate that there is a natural class of models for ligand binding where our algorithm precisely picks out the set of strongly binding ligands. To begin, we note that since we are describing ligands through their fingerprints $\mathbf{f}$, the ligand binding energy is a function of the fingerprints, i.e. $E = E(\mathbf{f})$. We can expand E in powers of $\mathbf{f}$, so that to leading order

$$E(\mathbf{f}) = \sum_{i=1}^{p} w_i f_i + \sum_{i,j=1}^{p} f_i J_{ij} f_j + \dots \qquad [4]$$

Here, w_i and J_{ij} are *protein specific* quantities; they parameterize how well ligands (characterized by their fingerprints) bind to the binding pocket of the protein in question. The values of w_i, J_{ij} and p also depend on the nature of the fingerprints that we use to describe the ligands. More detailed fingerprints have a better chance of accurately modeling the binding energy between the ligand and receptor than those that do not take into account parts of the molecule that binds to the receptor. The fact that the Morgan 3 fingerprints used herein have long been shown to have predictive power for ligand-target association means that they plausibly contain sufficient information to model the binding energy. It is noteworthy that since fingerprints are binary strings of length p, the model in Eq. (4) is equivalent to the *Ising Model*, well known in statistical physics.

Can we deduce w and J from the fingerprints of those ligands that bind to a protein target? Here we take as input the correlation matrix of the fingerprints that bind to each protein target in question. Indeed, determining the Ising model interaction matrix J from the correlation matrix is a classic problem in statistical physics and biophysics [31–35]. We now argue that our random matrix based procedure effectively removes noise caused by finite sampling from this problem. The essence of our algorithm is the derivation of a protein specific binding energy model J.

We can directly compute the correlation matrix of the fingerprints that bind to a given protein target (characterized by w_i, J_{ij}) by noting that our model implies that the equilibrium probability of observing a fingerprint $\mathbf{f}$ is given by

$$P(\mathbf{f}) = \frac{e^{-\beta E(\mathbf{f})}}{Z}, \qquad [5]$$

where $\beta = (k_B T)^{-1}$ characterizes the temperature and Z is the partition function, summing $e^{-\beta E}$ over all possible fingerprints. The correlation matrix follows directly from this model via $C_{ij} = \langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle$, where $< \cdot >$ denotes an average over the probability model in Eq. (5). The correlation matrix C_{ij} is a function of temperature T: at high temperatures, where $\beta E \ll 1$, all $\mathbf{f}$ are equally probable and the nontrivial correlations disappear. At lower temperatures, the set of fingerprints that are probable will reflect the structure of the interaction matrix J in Eq. (4).

Correspondingly, ligand-protein target binding only occurs over a range of temperatures, and we assume that we are in the range of temperatures where the binding is effective. Our algorithm computes the correlation matrix C_{ij} not from taking equilibrium averages but instead by averaging over n samples, where n is the number of ligands that bind to the target in question. Critically, n is the same order of magnitude as the fingerprint length p, so our computed covariance matrix does not converge to the equilibrium expectation – it is corrupted by noise. Our procedure of extracting the eigenvalues above the MP threshold corresponds to estimating the binding energy from the data matrix.

To see this, Figure 5 shows a set of simulations of the Ising model. We consider fingerprints of length $p = 50$, drawn from the distribution of Eq. (5). We take the first order coefficients to vanish ($w_i = 0$; in the case of the fingerprints this corresponds to using the z-score) and choose $J = -\alpha \mathbf{u}_\mathbf{J}' \mathbf{u}_\mathbf{J}$, where $\alpha > 0$. This is a rank one matrix, where $\mathbf{u}_\mathbf{J}$ is the (randomly chosen) direction that by construction will minimize the energy. Figure 5A shows the spectrum of the resulting correlation matrix, formed by considering $n = 200$ samples from Eq. (5) with $\beta\alpha = 0.1$. The temperature is sufficiently high that the fingerprints are uncorrelated, so the spectrum is well fit by the MP distribution (red line). Figure 5B shows the corresponding spectrum of the correlation matrix when $\beta\alpha = 0.6$. Here the bulk spectrum agrees well with the MP distribution (red line), but there is a single eigenvalue that escapes from the bulk with $\lambda \approx 9$. Figure 5C shows that the eigenvector corresponding to this eigenvalue is extremely well correlated with $\mathbf{u}_\mathbf{J}$.

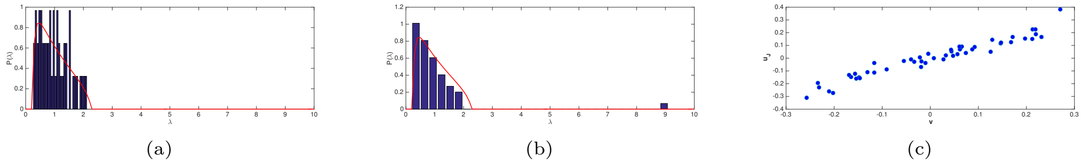
This correlation between the eigenvector and the coupling matrix J gives a physical interpretation of the projection onto the subspace of eigenvectors that escape the MP distribution in Eq. (2): We have used the data to derive a model for the binding energy of the ligand in fingerprint "coordinates", and to determine whether an arbitrary ligand binds to the target, we are simply evaluating this binding energy. The correlation structure is lost when we use a dataset of random ligands instead of those corresponding to a single protein receptor, since in this case there is no underlying energy model to learn. Although our simulations (Figure 5) use a rank 1 J for simplicity, if J is of higher rank, more eigenvectors will be pushed outside the MP distribution. Indeed, [36] showed that random matrix denoising is related to putting in a prior that the rank of J (in our case the number of independent pharmacophores) is less than the number of variables (2048 for the Morgan 3 fingerprint). We note that the Ising energy (4) provides another way to score ligands. However, the classification accuracy does not significantly improve if the energy is estimated using the leading order mean-field approximation [37].

Although interpreting our algorithm in terms of a binding energy function requires experimental verification through binding energy measurements we note that this interpretation offers several new conceptual insights. First, new candidate compounds could be uncovered by exploring the potential energy landscape of (4), and jumping between different energy minima could be related to "scaffold hopping" in drug discovery [38] as the minima would correspond to structures with pharmacophores. Investigating the topology of the energy landscape and those paths that connect distinct basins [39], as well as the statistics of energy minima could reveal properties of the binding site. Secondly, relating our algorithm to an interaction energy provides a way to extend our method to regression problems, such as predicting solubility [40].

Third, we note that chemical fingerprints may be improved by incorporating physically relevant terms such as charge and molecular volume. This is facilitated by our approach, which accounts for additional noise introduced by increasing the number of fingerprint variables. Finally, the binding energy interpretation highlights the importance of high quality negative data, i.e. which molecules do *not* bind to the desired receptor. Ref. [36] shows that including repulsive patterns could improve high dimensional inference with inverse Ising/Hopfield models. Empirically, for our system, the repulsive patterns (small eigenvalues) inferred from the data are noisy and uninformative. This can be addressed either through identification of many more ligands that bind to each protein receptor, or, perhaps more efficiently, the incorporation of negative data into this framework.

Conclusion

We have developed a classification algorithm that predicts whether a compound will bind to a particular receptor of interest, given the known ligand set of that receptor. Our algorithm decomposes signal from noise using a robust bound that is derived from random matrix theory. Applying our



(a)

(b)

(c)

Fig. 5. Eigenvalue spectrum of n = 200 fingerprints of length p = 50 sampled from $P(\lambda)$ in Eq. (5), with w = 0 and $J = -\alpha \mathbf{u}_\mathbf{J}' \mathbf{u}_\mathbf{J}$ a rank one matrix described in the text. (A) The spectrum with $\beta\alpha$ = 0.1 agrees quantitatively with the Marčenko-Pastur distribution (red line). At high temperature the covariance structure of J is irrelevant and the fingerprints are uncorrelated up to sampling noise. (B) The spectrum with $\beta\alpha$ = 0.6 has a bulk that agrees with the Marčenko-Pastur distribution (red line), but has a single eigenvalue escape from the bulk, near $\lambda \approx 9$. (C) The eigenvector **v** associated with this eigenvalue is highly correlated with $\mathbf{u}_\mathbf{J}$, the direction of J.

approach to human G-protein coupled receptors reported in ChEMBL successfully identifies 84% of known ligands with a 7% false positive rate, yielding an average AUC of 0.9. The methodology developed here complements the vast literature on optimizing fingerprint design, for example through use of high throughput screening data [7] or though application of neural networks to molecular graphs [41]. The random matrix framework described here provides a robust threshold for maximizing the information extracted from correlations between structural features, whilst avoiding overfitting the data. The algorithm has the natural interpretation as a data-driven model for the binding energy of the ligands to the target protein, in fingerprint "coordinates". This model gives a different perspective on the validity and usage chemical fingerprints for both ligand binding predictions and other purposes such as predicting ligand solubility [40] or aggregation [42], as well as revealing new insights in fingerprint design.

ACKNOWLEDGMENTS. This research was funded by a grant from Roche Pharmaceuticals. AAL acknowledges the support of a Fulbright Fellowship. LJC was supported by a Next Generation fellowship, and a Marie Curie Career Integration Grant [Evo-Couplings, Grant 631609]. MPB is an investigator of the Simons Foundation, and also acknowledges support from the National Science Foundation through DMS-1411694.

1. Bento AP et al. (2013) The chembl bioactivity database: an update. *Nucleic acids research* p. gkt1031.
2. Jacoby E (2011) Computational chemogenomics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1(1):57–67.
3. Johnson MA, Maggiora GM (1990) *Concepts and applications of molecular similarity.* (Wiley).
4. Maggiora G, Vogt M, Stumpfe D, Bajorath Jr (2013) Molecular similarity in medicinal chemistry: Miniperspective. *Journal of medicinal chemistry* 57(8):3186–3204.
5. Larsson J, Gottfries J, Muresan S, Backlund A (2007) Chemgps-np: tuned for navigation in biologically relevant chemical space. *Journal of natural products* 70(5):789–794.
6. García-Sosa AT, Oja M, Hetenyi C, Maran U (2012) Druglogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. *Journal of chemical information and modeling* 52(8):2165–2180.
7. Petrone PM et al. (2012) Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS chemical biology* 7(8):1399–1409.
8. Buonfiglio R et al. (2015) Investigating pharmacological similarity by charting chemical space. *Journal of chemical information and modeling* 55(11):2375–2390.
9. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences* 42(6):1273–1280.
10. Vilar S et al. (2014) Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature protocols* 9(9):2147–2163.
11. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50(5):742–754.
12. Sastry M, Lowrie JF, Dixon SL, Sherman W (2010) Large-scale systematic analysis of 2d fingerprint methods and parameters to improve virtual screening enrichments. *Journal of chemical information and modeling* 50(5):771–784.
13. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity–a review. *QSAR & Combinatorial Science* 22(9-10):1006–1026.
14. Putta S, Beroza P (2007) Shapes of things: computer modeling of molecular shape in drug discovery. *Current topics in medicinal chemistry* 7(15):1514–1524.
15. Kubinyi H (2008) Comparative molecular field analysis (comfa). *Handbook of Chemoinformatics: From Data To Knowledge in 4 Volumes* pp. 1555–1574.
16. Shin WH, Zhu X, Bures MG, Kihara D (2015) Three-dimensional compound comparison methods and their application in drug discovery. *Molecules* 20(7):12841–12862.
17. Todeschini R et al. (2012) Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *Journal of chemical information and modeling* 52(11):2884–2901.
18. Bausz D, Rájcz A, Héjberger K (2015) Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Chemoinformatics* 7:20.
19. Wigner EP (1955) Characteristic vectors of bordered matrices with infinite dimensions i. *Annals of Mathematics* 62(3):548–564.
20. Edelman A, Wang Y (2013) in *Advances in Applied Mathematics, Modeling, and Computational Science.* (Springer), pp. 91–116.
21. Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Physical review letters* 83(7):1467.
22. Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Stanley HE (1999) Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters* 83(7):1471.
23. Bouchaud JP, Potters M (2011) *The Oxford Handbook of Random Matrix Theory*, eds. Akemann G, Baik J, Di Francesco P. (Oxford Univ Press).
24. Turk MA, Pentland AP (1991) *Face recognition using eigenfaces.* (IEEE), pp. 586–591.
25. Turk M, Pentland A (1991) Eigenfaces for recognition. *Journal of cognitive neuroscience* 3(1):71–86.
26. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* 1(4):457.
27. Cereto-Massagué A et al. (2015) Molecular fingerprint similarity search in virtual screening. *Methods* 71:58–63.
28. Landrum G (2016) Rdkit: Open-source cheminformatics. *(Online). http://www.rdkit.org.*
29. Lagarde N, Zagury JF, Montes M (2015) Benchmarking data sets for the evaluation of virtual ligand screening methods: Review and perspectives. *Journal of chemical information and modeling* 55(7):1297–1307.
30. Unterthiner T et al. (2014) *Deep learning as an opportunity in virtual screening.*
31. Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for boltzmann machines. *Cognitive science* 9(1):147–169.
32. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087):1007–1012.
33. Cocco S, Leibler S, Monasson R (2009) Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences* 106(33):14058–14062.
34. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* 106(1):67–72.
35. Bailly-Bechet M, Braunstein A, Pagnani A, Weigt M, Zecchina R (2010) Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. *BMC bioinformatics* 11(1):1.
36. Cocco S, Monasson R, Sessak V (2011) High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Physical Review E* 83(5):051123.
37. Tanaka T (1998) Mean-field theory of boltzmann machine learning. *Physical Review E* 58(2):2302.
38. Sun H, Tawa G, Wallqvist A (2012) Classification of scaffold-hopping approaches. *Drug discovery today* 17(7):310–324.
39. Wales D (2003) *Energy landscapes: Applications to clusters, biomolecules and glasses.* (Cambridge University Press).
40. Llinas A, Glen RC, Goodman JM (2008) Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *Journal of chemical information and modeling* 48(7):1289–1303.
41. Duvenaud DK et al. (2015) *Convolutional Networks on Graphs for Learning Molecular Fingerprints.* pp. 2215–2223.
42. Irwin JJ et al. (2015) An aggregation advisor for ligand discovery. *Journal of medicinal chemistry* 58(17):7076–7087.