JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

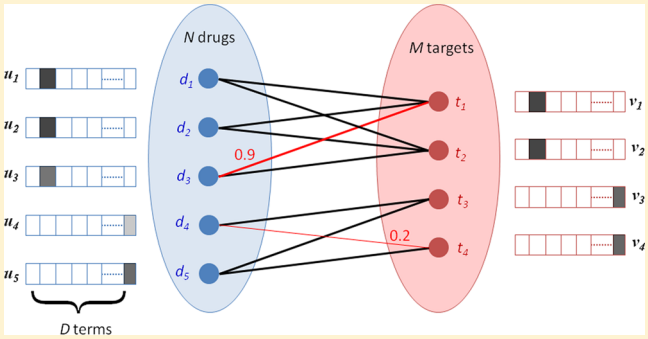
# Predicting Drug–Target Interactions Using Probabilistic Matrix Factorization

Murat Can Cobanoglu,[†] Chang Liu,[†,§] Feizhuo Hu,[†,⊥,§] Zoltán N. Oltvai,[‡] and Ivet Bahar*[,†]

[†]Department of Computational & Systems Biology and [‡]Department of Pathology, School of Medicine, University of Pittsburgh, Pennsylvania 15213, United States

Supporting Information

**ABSTRACT:** Quantitative analysis of known drug–target interactions emerged in recent years as a useful approach for drug repurposing and assessing side effects. In the present study, we present a method that uses probabilistic matrix factorization (PMF) for this purpose, which is particularly useful for analyzing large interaction networks. DrugBank drugs clustered based on PMF latent variables show phenotypic similarity even in the absence of 3D shape similarity. Benchmarking computations show that the method outperforms those recently introduced provided that the input data set of known interactions is sufficiently large—which is the case for enzymes and ion channels, but not for G-protein coupled receptors (GPCRs) and nuclear receptors. Runs performed on DrugBank after hiding 70% of known interactions show that, on average, 88 of the top 100 predictions hit the hidden interactions. De novo predictions permit us to identify new potential interactions. Drug–target pairs implicated in neurobiological disorders are overrepresented among de novo predictions.




## 1. INTRODUCTION

Drug discovery and development has become increasingly challenging in recent years, evidenced by the estimated cost of around $1.8 billion for the development of a novel molecular entity with suitable pharmacological properties.[1] This cost increase partly originates from the failure of many drug candidates in phase II or III clinical trials due to their toxicity or lack of efficacy.[2] The efficiency of drug discovery and development might be improved by adopting a systemic approach that takes into consideration the interaction of existing drugs and candidate compounds with the entire network of target proteins and other biomolecules in a cell.[3] Indeed, the "one gene, one drug, one disease" paradigm is widely recognized to fail in describing experimental observations.[4] Many drugs act on multiple targets, and many targets are themselves involved in multiple pathways. For example, β-lactam antibiotics and most antipsychotic drugs exert their effect through interactions with multiple proteins.[5,6] Biological networks are highly robust to single-gene knockouts, as recently shown for yeast where 80% of the gene knockouts did not affect cell survival.[7] Similarly, 81% of the 1500 genes knocked out in mice did not cause embryonic lethality, further corroborating the robustness of biological networks against single target perturbagens.[8] These results suggest that quantitative systems pharmacology strategies that take account of target (and drug) promiscuities can present attractive alternative routes to drug discovery.

Recent years have seen many network-based models adopted to reduce the complexity of, and efficiently explore, drug–target interaction systems.[2,5,6,9] In particular, the development of computational methods that can efficiently assess potential new interactions became an important goal. In this regard, the important role that machine learning approaches such as active learning (AL) can play has been recently been highlighted.[10] Computational approaches used to predict unknown drug–target interactions can be divided into roughly four categories: chemical-similarity-based methods,[11−13] target-similarity-based methods,[14−16] integrative (both target- and chemical-similarity-based) methods,[17−23] and holistic approaches.[24−29] The first two posit that if two entities are chemically or structurally similar they will share interactions. The integrative approaches combine the chemical- and target-similarity methods. While the intuition behind these approaches is very reasonable, their performance has been observed to be tied to the underlying similarity computation method. We also note that the utility of different methods may depend on the size of the data set being analyzed, e.g., computing chemical–chemical and target–target similarity matrices can be problematic for large databases like STITCH[30] (that contains information on the interactions between more than 2.6 million proteins and 300 000 chemicals). To overcome these limitations, holistic methods have been introduced, which utilize a number of different data sources such as gene expression perturbation[25,26] or high-throughput screening.[28]

**Received:** April 10, 2013
**Published:** December 1, 2013

   dx.doi.org/10.1021/ci400219z | J. Chem. Inf. Model. 2013, 53, 3399−3409

In this study, we propose a novel approach by using a collaborative filtering algorithm to predict interactions without reliance on chemical/target similarity or external data collection. We validate the utility of probabilistic matrix factorization (PMF) for predicting unknown drug−target interactions with the help of a detailed investigation of its performance. The method is shown to group drugs according to their therapeutic effects, irrespective of their three-dimensional (3D) shape similarity. Benchmarking computations show that the method outperforms recent methods[17,20,22] when applied to large data sets of protein−drug associations, such as those of enzyme− and ion channel−drug pairs; whereas the performance falls short of these methods with decreasing size of the examined data set (e.g., GPCR- and nuclear receptor-drug data sets). The ability of the method to efficiently analyze and make inferences from *large* data sets of protein−drug interactions suggests that, with growing sizes of those data sets, the utility (and accuracy) of the method will further improve.

Application of the same benchmarking procedure to DrugBank[31] confirms its ability to disclose hidden data: 88 out of the top 100 predictions (or 587 out of 1000) are found to hit known (but hidden) interactions, when only 30% of the entire data is used for training. Finally, when the method is trained on the entire data set of drug−target interactions compiled in DrugBank, de novo predictions for drug repurposing can be made along with the corresponding confidence levels. Top de novo predictions include many drugs indicated for neurodegenerative diseases or neurobiological disorders, including drug−target pairs apparently supported by previous experiments (not reported in DrugBank), e.g., ergotamine−serotonin receptor 1A ($5HT_{1A}$),[32] amoxapine-5-$HT_{2A}$,[33] and verapamil−calmodulin.[34] In conclusion, the newly introduced computational method provides an efficient approach for identifying potential drug−target association between chemicals and targets and formulating new hypotheses for repurposable drugs or side effects, thus complementing those deduced from chemical−chemical or target−target similarities.

## 2. MATERIALS AND METHODS

**2.1. Problem Definition.** The drug−target interaction network is a bipartite graph with two types of nodes: drugs and targets (Supporting Information Figure S1). Each edge represents an interaction between a drug and a target. The drug−target interaction identification problem is to determine the missing edges that are likely to exist given all nodes and some of the edges in the network.

**2.2. Data Set.** We used DrugBank (version of September 20, 2011) as the database.[31] All drugs annotated therein as approved, along with their annotated targets, are included in our data set (i.e., we excluded compounds annotated as withdrawn or nutraceutical), resulting in $N = 1413$ drugs and $M = 1050$ targets with 4731 interactions among them. The interaction network displays small-world characteristics: many nodes have low degree and a few, very high degree, as illustrated in panels b and c of Supporting Information Figure S1, in line with previous studies on drug−target networks.[35] On average, there are 3.35 interactions per drug, and 4.50 interactions per target.

**2.3. Probabilistic Matrix Factorization (PMF).** PMF is a member of the collaborative filtering family of machine learning algorithms that decomposes the connectivity matrix, $\mathbf{R}_{N\times M}$, of a bipartite graph of $N$ drugs and $M$ targets as a product of two matrices of latent variables (LVs).[36,37] $\mathbf{R}_{N\times M}$ is defined as

$$R_{ij} = \begin{cases} 1 & \text{if drug } i \text{ interacts with target } j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The matrix $\mathbf{R}_{N\times M}$ is modeled as the product of two matrices $\mathbf{U}^{\mathrm{T}}{}_{N\times D}$ and $\mathbf{V}_{D\times M}$, that express each drug/target in terms of $D$ LVs. Our objective is to find the best approximation for LVs, while avoiding overfitting. The predicted $\hat{\mathbf{R}}_{N\times M}$ is then expressed as

$$\hat{\mathbf{R}}_{N\times M} = \mathbf{U}^{\mathrm{T}}{}_{N\times D}\,\mathbf{V}_{D\times M} \quad (2)$$

where $\mathbf{U}^{\mathrm{T}}$ and $\mathbf{V}$ are composed of $N$ rows $\mathbf{u}_i^{\mathrm{T}}$ and $M$ columns $\mathbf{v}_j$, respectively, each being $D$-dimensional. The PMF adopts a probabilistic linear model with Gaussian noise to model the interaction. Therefore, the conditional probability over observed interactions is represented as

$$p(\mathbf{R}|\mathbf{U}, \mathbf{V}, \sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M} [f(R_{ij}|\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j, \sigma^2)]^{I_{ij}} \quad (3)$$

where $f(x|\mu, \sigma^2)$ is the Gaussianly distributed probability density function for $x$, with mean $\mu$ and variance $\sigma$, and $I_{ij}$ is the indicator function equal to 1 if the entry $R_{ij}$ is known and 0 otherwise. Therefore, $p(\mathbf{R}|\mathbf{U}, \mathbf{V}, \sigma^2)$ gives us a probabilistic representation of the connectivity matrix, $\mathbf{R}$.[37] Using zero-mean, spherical Gaussian priors on LVs, we can write

$$p(\mathbf{U}|{\sigma_U}^2) = \prod_{i=1}^{N} f(\mathbf{u}_i|\mathbf{0}, {\sigma_U}^2\mathbf{I}) \quad (4)$$

$$p(\mathbf{V}|{\sigma_V}^2) = \prod_{j=1}^{M} f(\mathbf{v}_j|0, {\sigma_V}^2\mathbf{I}) \quad (5)$$

which leads to the log-likelihood of $\mathbf{U}$ and $\mathbf{V}$ given by

$$\begin{aligned} &\ln(p(\mathbf{U}, \mathbf{V}|\mathbf{R}, \sigma^2, {\sigma_U}^2, {\sigma_V}^2)) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij} - \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j)^2 - \frac{1}{2{\sigma_U}^2}\sum_{i=1}^{N}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i \\ &\quad - \frac{1}{2{\sigma_V}^2}\sum_{j=1}^{M}\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j + C \end{aligned} \quad (6)$$

Here $C$ is a term that does not depend on LVs; the first term on the right-hand side is the squared error function to be minimized; and the two summations over the square magnitudes of $\mathbf{u}_i$ and $\mathbf{v}_j$ are regularization terms that favor simpler solutions and penalize overfitting. The above log-likelihood directly follows from the Bayes' rule where $\mathbf{R}$ stands for data, and $\mathbf{U}$ and $\mathbf{V}$ represent the model (see the Supporting Information for details). To learn an optimal model means to find the $\mathbf{U}$ and $\mathbf{V}$ matrices, or the $D$-dimensional LV vectors, $\mathbf{u}_i$ $(1 \le i \le N)$ and $\mathbf{v}_j$ $(1 \le j \le M)$, that maximize the log-likelihood function.

The PMF method yields the optimal $\mathbf{u}_i$ and $\mathbf{v}_j$ vectors corresponding to each drug, $d_i$, and each target $t_j$, respectively. The basic idea is that the model is forced toward making a "no-interaction" prediction by the regularization—i.e., there is a penalty associated with any nonzero value in the LV matrices. However, there is also a penalty for failing to capture known interactions—i.e., if the dot product of the LV vectors

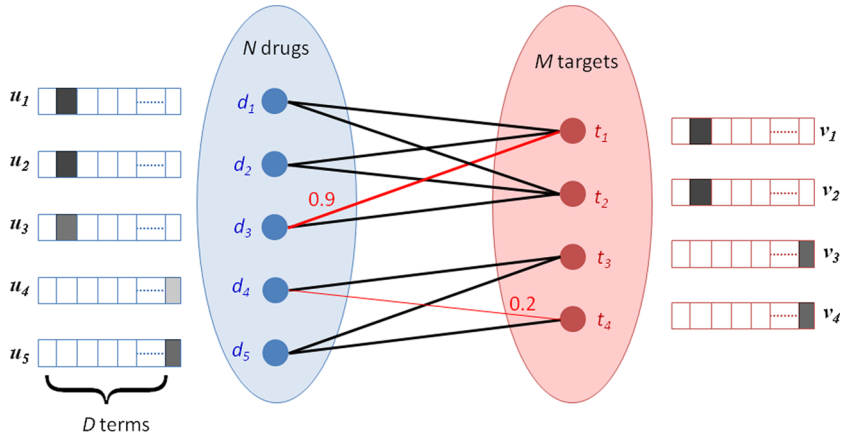



**Figure 1.** Qualitative illustration of the method for identifying drug–target interactions. The known interactions between drugs and targets (indicated by the *black* lines) are used to learn the LV vectors (shown adjacent to each node) that describe each drug and target. The dot product $\mathbf{u}_i^T\mathbf{v}_j$ of the LVs for each pair of drug $d_i$ and target $t_j$ defines the predicted statistical weight $\omega_{ij}$ of corresponding connection. Example predictions are shown in *red*.

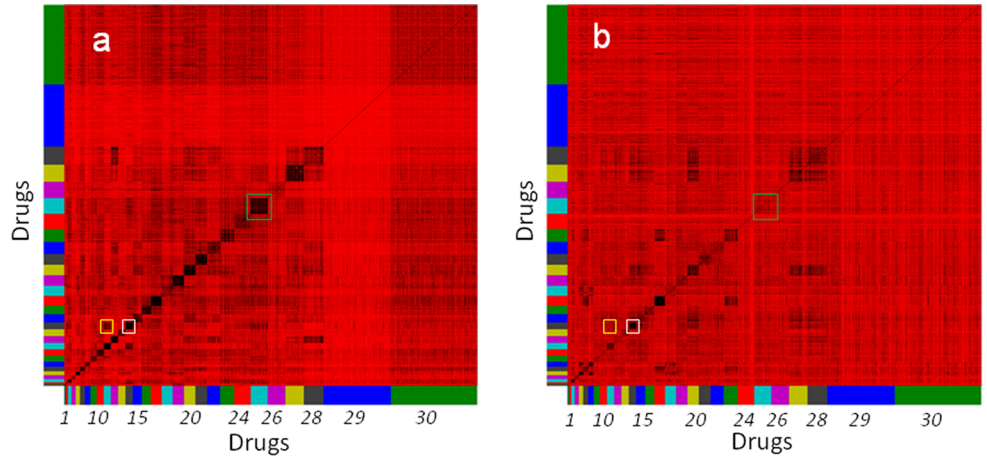



**Figure 2.** Comparison of pairwise similarities of drugs, based on their (a) therapeutic targets compiled in DrugBank and (b) 3D structure. (a) 30 clusters of drugs (color-coded along the axes; see Supporting Information Table S1 for their dominant therapeutic indication) deduced from the PMF of 1413 approved drugs and corresponding 1050 targets compiled in DrugBank. By definition, drugs belonging to a given cluster share similar interaction patterns with respect to targets. (b) 3D similarities, with the drugs being ordered as in panel a. *Dark* regions indicate high similarity based on LVs (panel a) or 3D similarities (panel b). Comparison of the panels shows that close proximity in LV space (which indicates functional similarity) does not necessarily imply 3D-structure similarity. LV distances were distributed in the range [0, 1]; with the distribution of values also skewed in different ways. To render the two sets comparable, we performed rank normalization on both the LV similarities and 3D similarities. Selected boxes are enlarged in Figures 3 (*white*) and 4 (*yellow*), and Supporting Information Figure S2 (*green*).

corresponding to an interacting drug–target pair is close to zero. Therefore the learning of a model means to optimally balance out two objectives: developing a sufficiently complex model to describe the known interactions, but not overly complex to end up in overfitting. In this study, we use gradient descent for optimization. The adoption of higher $D$ values usually yields more accurate results, although beyond a certain limit the increase in complexity and decrease in efficiency may not warrant the marginal improvement, if any, in prediction accuracy. $D$ = 50 is adopted here as an optimal dimensionality for prediction runs. The method is highly efficient: a 50-dimensional model is trained on the entire DrugBank in approximately 2 s using a 2.00 GHz AMD Opteron processor. Moreover, the computing time to learn a PMF model scales linearly with the number of interactions, and as such, the method can be advantageously used for much larger data sets.

**2.4. Active Learning (AL) Using Probabilistic Matrix Factorization.** The AL strategy adopted in the present study is, in part, motivated by the success reported by Warmuth et al.[38] who demonstrated that hit maximization is a viable AL strategy applicable to predicting drug–target interactions. The AL strategy adopted here also prioritizes the discovery of unknown interactions. Our method differs in that we aim at capturing the interactions between all drugs and targets, as opposed to predicting activity against a single target.

The procedure is the following: We begin with the set of $N$ drugs and $M$ targets, and known associations, schematically shown in Figure 1 by *black* connectors. The purpose is to identify new associations, indicated by *red* connectors. For each candidate interaction, say the possible interaction between $d_i$ and $t_j$, we compute the model's estimate, $f(R_{ij}|\mathbf{u}_i^T\mathbf{v}_j, \sigma^2)$ (eq 3). The dot product $\mathbf{u}_i^T\mathbf{v}_j$ serves as a weight $\omega_{ij}$ for the edge/connector between $d_i$ and $t_j$. Clearly, $\omega_{ij}$, or the likelihood of association between $d_i$ and $t_j$, is high when $\mathbf{u}_i$ and $\mathbf{v}_j$ have both large values of the same sign at the same dimension(s). For example, a relatively large weight may originate from the


dx.doi.org/10.1021/ci400219z | J. Chem. Inf. Model. 2013, 53, 3399–3409

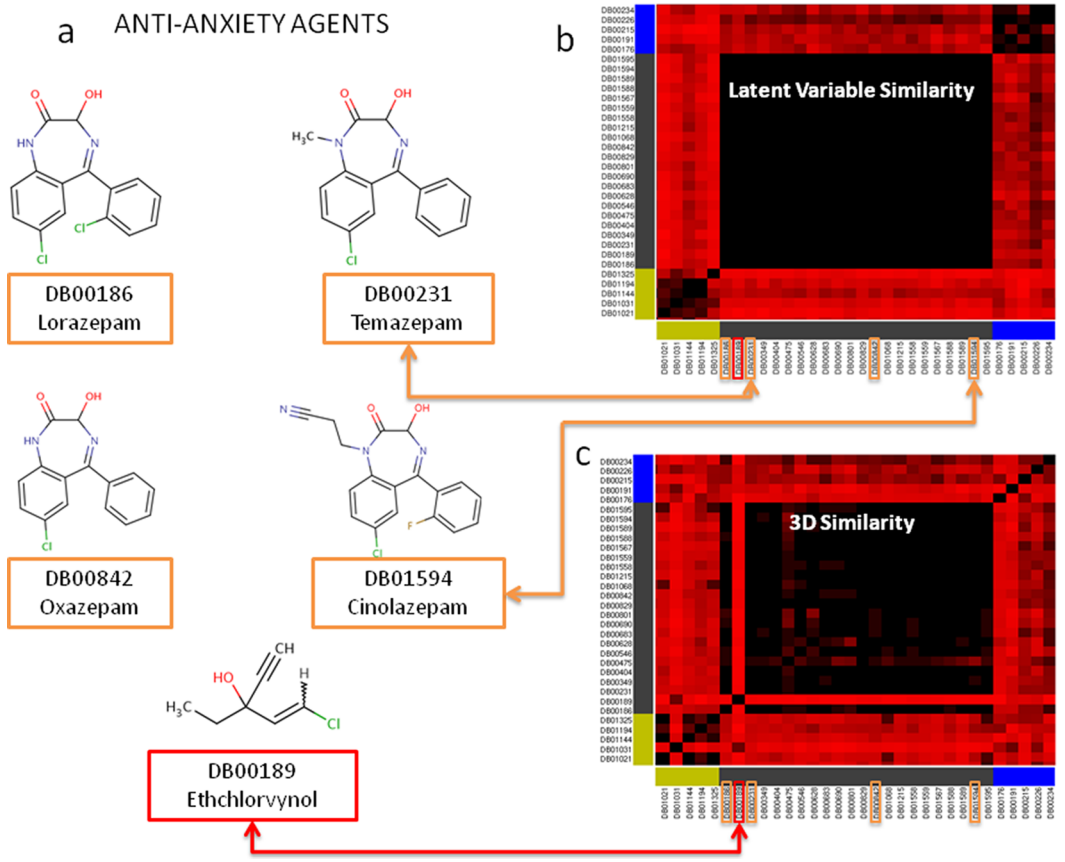



**Figure 3.** Latent variables capturing therapeutic action similarities when 3D similarity metrics cannot. Closer examination of the similarities between the members of cluster *14* in Figure 2 (enclosed in white boxes in Figure 2, enlarged in panels b and c here) shows that the cluster contains a series of antianxiety drugs. A few members of this cluster (indicated by orange boxes along the abscissa of panels b and c) are displayed in panel a, to illustrate their shared structural features, also indicated by panel c that reflects their 3D similarities. The same cluster however contains ethchlorvynol, also used as a sedative, which would have been missed if we had used exclusively used 3D similarity to identify functionally similar drugs.

second component of both $\mathbf{u}_i$ and $\mathbf{v}_j$, which means that the predicted association is mainly due to latent variable 2. We evaluated the statistical weights $\omega_{ij}(d_i, t_j)$ for the $N \times M$ pairs of drug targets for two purposes: (i) benchmarking the methodology via an iterative AL scheme and (ii) making de novo predictions. In the former case, the method is benchmarked by hiding 70% of known interactions and examining whether the top-ranking prediction is a "hit", i.e., whether it corresponds to a known (but hidden) interaction. The outcome from this test is fed back to the model, to repeat the calculation for the next prediction. Therefore, the AL model is updated at each iteration using the newly acquired "hit" or "miss" data until a predetermined number ($m$) of predictions are made. The passive learner (PL) makes the $m$ predictions simultaneously without updating its model.

In the case of de novo predictions, all DrugBank data were used as input. De novo predictions also lend themselves to an AL scheme provided that the top-ranking prediction is experimentally tested and then the new hit or miss data are incorporated in the model to perform a new prediction, and so on, until the experimentation budget is exhausted.

## 3. RESULTS AND DISCUSSION

**3.1. PMF Cluster Drugs with Therapeutic Similarities, Irrespective of Their Chemical−Structural Similarities.** To assess whether the LVs provide us with a pharmacologically meaningful metric, we examined the clustering of drugs in the $D$-dimensional space of the latent vectors. The clustering was performed for $D = 30$—the value that gave the lowest Akaike information criterion,[39] using as basis the drug−drug distance $L_1(d_i, d_j) = \sum_k |u_{ik} - u_{jk}|$ where $u_{ik}$ designates the $k$th component of $\mathbf{u}_i$, and the summation is performed over $D$ components.

Inasmuch as our method evaluates drugs based on their interaction profiles with targets, which in turn refer to specific therapeutic or phenotypic actions, the similarity of a pair of drugs should be high when their therapeutic effects are comparable and vice versa. Thus, the method will tend to cluster drugs that exhibit similar patterns of interactions (with target proteins), which we term as *functionally similar drugs*.

The heat map in Figure 2a displays the resulting organization of drugs in 30 clusters (indicated by different colors and indices along the axes). Supporting Information Table 1 lists the dominant therapeutic action associated with each cluster. The dark regions on the map indicate high functional similarity. The dark blocks along the diagonal show that most clusters include highly similar members, except for two (clusters *29* and *30*), which apparently combine the outliers.

Given that (promiscuous) proteins present more than one site for ligand-binding, different functionalities may be modulated by chemical-structurally different drugs, depending on the binding site on the target (e.g., catalysis, substrate recognition, or allosteric signaling). Furthermore, a shared phenotype may arise from the targeting of different proteins


dx.doi.org/10.1021/ci400219z | J. Chem. Inf. Model. 2013, 53, 3399−3409

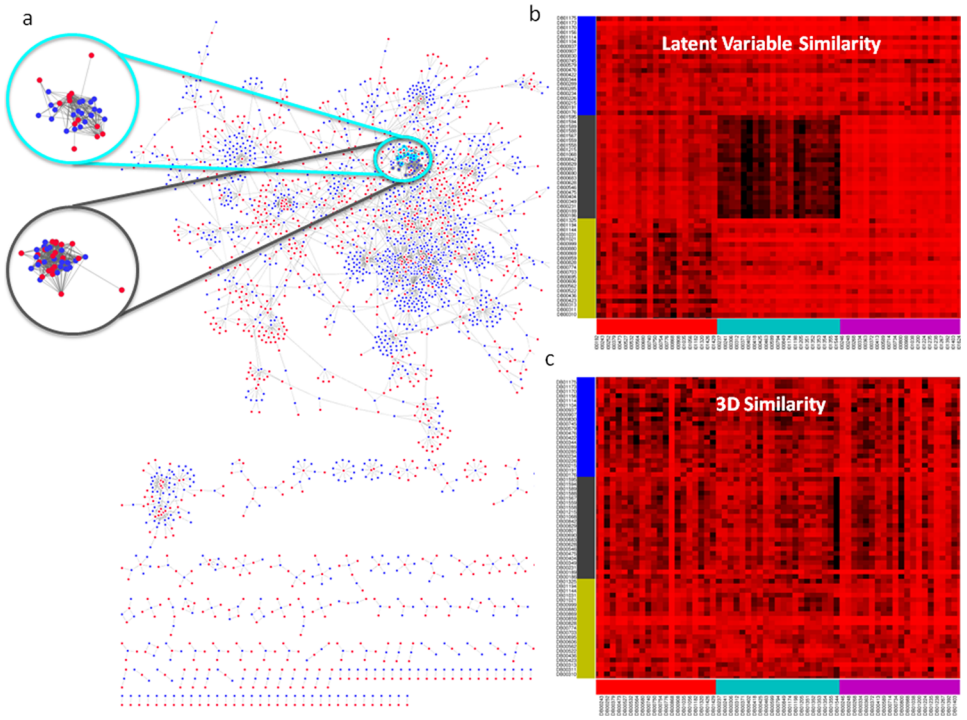



**Figure 4.** Strong cross-correlations between different clusters of drugs, consistent with their similar therapeutic functions. Cluster *11*, color-coded *cyan*, is essentially composed of hypnotics and sedatives. Cluster *14* (*dark gray*) contains antianxiety drugs. The drugs in these two clusters are located very closely on the drug−target interaction network, as shown in panel a, consistent with their similar actions. The LV-derived heat maps capture the functional similarity between these two clusters (as indicated by strong signals, or the *dark* region, in panel b); the maps based on 3D similarity (panel c) do not. In panel a drugs are shown in *blue*, protein targets in *red*. Most drugs and targets are part of a single connected component. Data are retrieved from DrugBank.[48] Cytoscape is used for visualization.[56]

along a given pathway. In order to make a better assessment of the properties of drugs grouped in those clusters, we examined their 3D structural similarities. High similarities would suggest that they bind similar epitopes, if not similar (or identical) structural domains or proteins. If, on the contrary, they are structurally dissimilar, this might indicate a different site on the same protein, a different target on the same pathway, or other indirect effect due to drug−target network connectivity.

The extent of 3D structure similarity between pairs of drugs was computed using the OpenEye Scientific software (http://www.eyesopen.com/). 3D similarity was reported to be a better predictor than 2D methods for off-target interactions and to perform equally well in on-target interactions,[40] although 3D methods may suffer from more noise due to the conformational flexibility of the small molecule. We generated for each drug all possible stereoisomers using OpenEye FLIPPER[41] and up to 200 conformers per stereoisomer using OpenEye OMEGA.[41] All combinations of conformers accessible to the examined pair of drugs were examined using OpenEye Shape[23] toolkit; and the best matching pair was adopted to assign a 3D similarity score. This computationally expensive task led to the heat map presented in panel b of Figure 2. The drugs (along the axes) are ordered as in panel a to enable visual comparison.

The comparison of Figure 2 shows that some clusters of functionally similar drugs (panel a) also exhibit some 3D similarities (panel b), whereas others display little structural similarity. We examined more closely the individual clusters to see if shared therapeutic functions were captured even when 3D similarities were absent. Figure 3 illustrates the results for cluster *14*. This cluster essentially consists of antianxiety drugs, the majority of which are both functionally (panel b) and structurally (panel c) similar. However, the cluster also includes a structurally dissimilar drug, ethchlorvynol (panel a), which shares the same type of phenotypic action (as a sedative) as the majority of the cluster membership (mostly targeting GABA receptors). The present approach thus detects chemically or structurally distinctive drugs that share common activities, which would have been missed by methods based on ligand fingerprint similarities.

Another interesting observation concerns the cross-correlations between different clusters (i.e., the off-diagonal regions of the heat maps). We note for example that cluster *11* also contains a set of sedatives. LVs are able to capture the commonality between the clusters *11* and *14* as may be seen by the strong signal (dark region) at the off-diagonal region enlarged in Figure 4b. The 3D similarity, on the other hand, cannot recognize the functional similarity and potential interference/side effects between these drugs in these two clusters (Figure 4c). See Supporting Information Figure S2, which illustrates the same behavior for another cluster, whose members are mostly antineoplastic agents, albeit with various 3D structures. The LVs thus provide information on drug groups that potentially share pathways or exhibit similar activity patterns despite their distinct physicochemical properties.

**3.2. Benchmarking Computations Support the Utility of the Method for Analyzing Large Data Sets.** To evaluate the performance of the method in comparison to previous work, we considered three important studies in this area, one recently published by Gonen[22] and two by Yamanishi et al.[17,20] Gonen used a Kernel-based matrix factorization (KBMF) with chemical and genomic similarities to predict multiple targets. Yamanishi et al., on the other hand, integrated chemical,


dx.doi.org/10.1021/ci400219z | J. Chem. Inf. Model. 2013, 53, 3399−3409

**Table 1. Properties of the Examined Space of Proteins−Drugs and Performance of the Present Method in Comparison to Others (*)**

| target type | no. of known interactions | no. of drugs ($N$) | no. of targets ($M$) | size of interaction space ($N \times M$) | percent occupancy of the space | Yamanishi (pred pharmacol effects) 2008[a] | Yamanishi (pred pharmacol effects) 2010[a] | Gonen, 2012[a] | present method ($D$ = 50) |
|---|---|---|---|---|---|---|---|---|---|
| enzymes | 1515 | 212 | 478 | 101336 | 1.50% | 0.821 | 0.845 | 0.832 | **0.861 ± 0.02** |
| ion channels | 776 | 99 | 146 | 14454 | 5.37% | 0.692 | 0.731 | 0.799 | **0.904 ± 0.02** |
| GPCRs | 314 | 105 | 84 | 8820 | 3.56% | 0.811 | 0.812 | **0.857** | 0.771 ± 0.04 |
| nuclear receptors | 44 | 27 | 22 | 594 | 7.41% | 0.814 | **0.830** | 0.824 | 0.650 ± 0.11 |
| all[b] | 2649 | 443 | 730 | | | 0.782 | 0.807 | 0.825 | **0.859** ± 0.03 |
| DrugBank | 4731 | 1413 | 1050 | 1483650 | 0.32% | | | | **0.794 ± 0.01** |

[a]The last four columns present the comparison with the results of Yamanishi[17,20] and Gonen[22] for the same data set. [b]Weighted-average mean and covariances, evaluated using the number of interactions as weights.

genomic, and pharmacological data to map all drugs and targets to the same unified feature space where each protein−compound pair closer than a predefined threshold was predicted to interact. Our approach differs from both studies, in that PMF assumes an independent LV for each row and column with Gaussian priors; whereas KBMF employs LVs spanning all rows and columns with Gaussian process priors, and Yamanishi et al. project drugs and targets into a pharmacological space based on the eigenvalue decomposition of the graph-based similarity matrix.

The benchmarking procedure that we adopted is a 5-fold cross-validation of drugs on four target classes: enzymes, ion channels, G-protein coupled receptors (GPCRs), and nuclear receptors. In order to achieve comparable results, we used the same protocol as that adopted earlier, i.e., we divided our data set into five subsets; each was used as a test set, and the others, as training sets. Due to the randomness involved in the selection of subsets, we repeated the cross-validation experiments 100 times with randomly selected subsets and evaluated the average AUC (area under the receiver operating curve) for each subset. The first four rows in Table 1 compare the results (columns 6−10) for the four classes, and the fifth row lists the average performances weighted by the size of the interaction space. Our method performs best when applied to large data sets (e.g., enzymes and ion channels); whereas Gonen's performs best in the case of GPCRs, and Yamanishi et al.[20] exhibits the highest performance for nuclear receptors, where the present method yields a relatively low (0.642) AUC value. Examination of the statistical significance of our results (Supporting Information Figure S3a) indicates that the mean AUC values obtained for all four sets are highly robust. Their covariances vary from 2% (enzymes and ion channels) to 11% (nuclear receptors). Finally, the application of the same benchmarking protocol to DrugBank yielded an accuracy rate of 79.4 ± 0.01% (Table 1, last row, and Figure S3), supporting the utility of the method when applied to large data sets.

In principle, it might be intrinsically harder to make accurate predictions for larger data sets as the size of the potential interaction space $N \times M$ grows quadratically when the number of drugs and targets grow linearly, particularly if the number of known interactions is small. The occupancy of the $N \times M$ interaction matrix is only 1.5% in the enzyme class, which could make it difficult to learn an informative model. The present PMF technique, however, successfully learned an informative model and handled the complexity of interactions in this space of interactions, apparently due to the availability of a sufficiently large (absolute) number of known interactions (Supporting Information Figure S3b).

The ion channel drug class has the second largest number of known interactions among the four. Although the size of interaction space is 1 order of magnitude smaller than the enzyme class, there are 776 known interactions leading to a percent occupancy of 5.37% of all possible ion channel-drug associations. The success of our method in this case may be attributed to both the relatively large number of known interactions and the rich annotation of that class of interactions.

The two other classes, GPCRs and nuclear receptors, are significantly smaller in terms of their interaction space and/or occupancy of that space. Nuclear receptors comprise only 27 drugs, 22 targets, and 44 interactions. A method that relies solely on connectivity, like ours, cannot presumably formulate an informative model when the set of "edges" to construct the network connectivity matrix is incomplete. In those cases, the data that come from other sources, e.g. chemical similarity and genomic patterns, amend this lack of information. Consequently, methods that incorporate such features[20,22] outperform ours.

To further examine the effect of scarcity of known interactions on the performance of the present method, we performed additional tests by varying the fraction of hidden interactions. The results are presented in Supporting Information Figure S4. Panels a−d show the performance on ion channels, enzymes, GPCRs, and nuclear receptors, respectively. These results show that the performance depends on the fraction of known interactions. To put the results into perspective, we indicated by a vertical dashed line in each panel the fraction of data (80%) used in previous studies[20,22] for training purposes. Consistent with the above findings, ion channels yield the best result: previous AUC values[22] (of 0.799; Table 1) are matched with about only 35% of the data. In the enzyme group, we match the performance of Yamanishi et al.[20] (AUC of 0.845) with roughly 70% of the data used for training. GPCRs and nuclear receptors yield AUC values lower than those previously attained,[20,22] irrespective of the fraction of hidden interactions.

In summary, the method is particularly suitable for screening and inferring repurposable drugs or potential side effects from large data sets where computational assessment of structure similarity kernels become prohibitively expensive. In cases where the data set of known interactions is too small, on the other hand, 2D or 3D similarity metrics provide more accurate assessments.

**3.3. Absolute Number of Known Interactions Overrides the Scarcity of the Data in Determining the Accuracy Rate of PMF Active Learner.** As a more stringent test, 3318 (70%) of the known 4731 interactions in DrugBank were randomly hidden, reducing the average number of interactions per drug from 3.35 to 1. The resulting "incomplete" interaction matrix was then used to predict the hidden interactions, one at a time (rank-ordered by statistical weights $\omega_{ij}(d_i, t_j)$) as described in Materials and Methods. The outcome was checked in a simulated experiment to assess whether the predicted interaction is a true positive (TP) or a false positive (FP). If the prediction is an existing, but hidden, interaction, the result is considered a TP (or hit), otherwise a FP (or miss). Then the model is updated in line with our AL scheme, and this loop is repeated until the completion of $m$ = 1000 predictions. At that point, the simulation is halted and the overall performance of the model, or the hit ratio, is evaluated. Note that this method gives us a lower bound for hit ratio because the predictions are labeled as hits only if they are annotated in DrugBank, although they can be true but not yet observed experimentally or annotated in DrugBank.

The results are presented in Figure 5. The figure displays the number of hits as a function of the number of predictions,

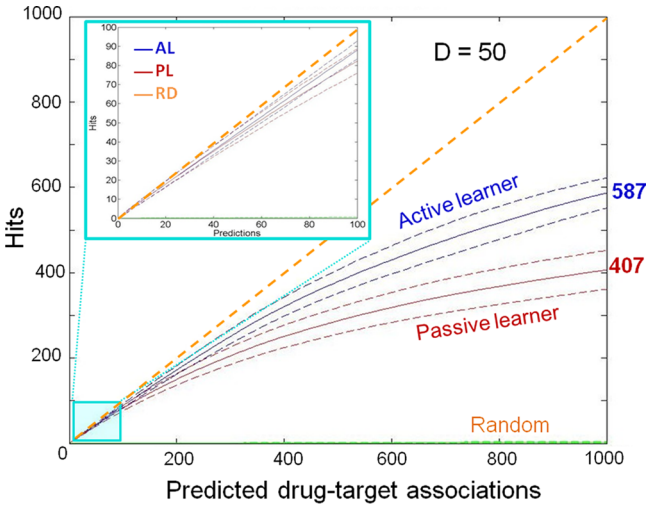



**Figure 5.** Ability of the method to recapitulate hidden drug–target interactions. The number of drug–target interactions per drug was reduced from 3.35 (average) to 1 by hiding 70% of known interactions, selected randomly. Simulations were repeated $n$ = 96 times for each of the $1 < m < 1000$ predictions (abscissa) and the number of hits (correctly identified hidden interactions) is plotted for each run, along the ordinate. The *dark blue and dark red* solid curves refer to the average performance obtained by active learning and passive learning protocols, respectively, using $D$ = 50, $\epsilon$ = 3, $\lambda$ = 0.01, and $\mu$ = 0.9 in the adopted PMF algorithm. Dashed curves show the corresponding variances (by one standard deviation) above and below the mean value. The *green* curves (practically overlapping with the abscissa) refer to results from random predictions. The inset shows a close-up of the first 100 predictions. AL reaches an accuracy rate (hit ratio) of 88.0 ± 4.7% and 58.7 ± 3.5% in the respective cases of $m$ = 100 and 1000 predictions.

obtained with three approaches: active learning (*dark blue* curves), passive learning (*dark red* curves), and random (*green*). The approach is able to achieve, on average, 587 hits out of 1000 predictions via AL, 407 hits, via PL; and the corresponding variances (indicated by the dashed curves) are 35 and 46, respectively. Compared to the random probability of 2.23 hits per 1000 predictions, the AL result is a 263-fold improvement over random. The improvement of AL over PL is 1.44 fold. The AL improvement over random was reported to be up to 3.19-fold in a previous SVM-based study for predicting the activity of 1316 drugs against a single target.[38] The same study also reported 1.59-fold improvement between passive and active learners. Closer examination of the results from the top 100 predictions (enlarged in the inset) further shows that hit ratios of 88.0 ± 4.7% and 82.2 ± 6.4% are obtained by the respective AL and PL protocols. The results are obtained with $D$ = 50, which yields optimal results, as can be seen from Supporting Information Figures S5 and 6 display the dependence of the results on $D$, in support of the choice of $D$ = 50.

These results permit us to draw two conclusions. First, a hit ratio of 88% is attainable in the top 100 predictions (and 59% in top 1000) upon adopting a PMF-based AL strategy for identifying hidden/unknown interactions in a sparse (0.32% occupancy) data set of about 1.5 million potential interactions. Second, the AL method outperforms random by 2 orders of magnitude and PL by a ratio of 1.5 approximately, in support of AL strategy for predicting new interactions.

**3.3. De novo Predictions of Drug–Target Interactions.** We used our method to predict new (potential) drug–target interactions after training our model on the latest available version of DrugBank (September 10, 2013) comprised of 5041 interactions between 1502 approved drugs and 1138 targets. The highest confidence pairs obtained for twenty distinct drugs are presented in Table 2. The pairs therein were observed to lie frequently among the top-ranking 10 pairs (in the space of $N \times M = 1.7 \times 10^6$ potential interactions), deduced from $10^4$ independent runs initiated with different random numbers.

We note that the list of de novo predictions in Table 2 is dominated by drugs used for neurological or psychiatric disorders, consistent with the known pharmacological promiscuity of this of drugs. The last column in Table 2 lists the experimental support from the literature, if any, for the possible interaction of the drug–target pair in each row. Among pairs supported by previous experiments, we note ergotamine–serotonin receptor 1A ($5HT_{1A}$),[32] amoxapine-5-$HT_{2A}$,[33] verapamil–calmodulin,[34] paliperidone-5-$HT_c$,[42] meperidine–sodium channels,[43] and cinnarizine–calmodulin.[44] We also note that chronic treatment with paroxetine has been recently reported to increase the mRNA levels of histamine receptor $H_1$, indicating an association between paroxetine and Histamine receptor H1.[45] Although paroxetine is a potent serotonin reuptake inhibitor, its weight gain side effect has been attributed to its medication action on histamine receptors.[46]

Many others (indicated as "indirect") are interactions known to occur either with subtypes of the listed targets or proteins implicated in the same phenotype (e.g., citalopram induces norepinephrine receptor hypoactivity,[47] which may relate to norepinephrine transport by NET). Yet, the validity of these predictions need to be established by experiments. Here, for exploratory purposes, we examined the potential binding pose and energetics of verapamil–calmodulin. Verapamil is known as a $Ca^{2+}$ channel entry blocker. Its interaction with calmodulin is supported by the fluorescence experiments[32] and by the inhibition of calmodulin-stimulated ($Ca^{2+}$ + $Mg^{2+}$) ATPase activity.[48] We used as template the Protein DataBank (PDB)[49] structure of the cocrystal between calmodulin and trifluoperazine (PDB identifier 1CTR)[50] and examined the binding


dx.doi.org/10.1021/ci400219z | J. Chem. Inf. Model. 2013, 53, 3399–3409

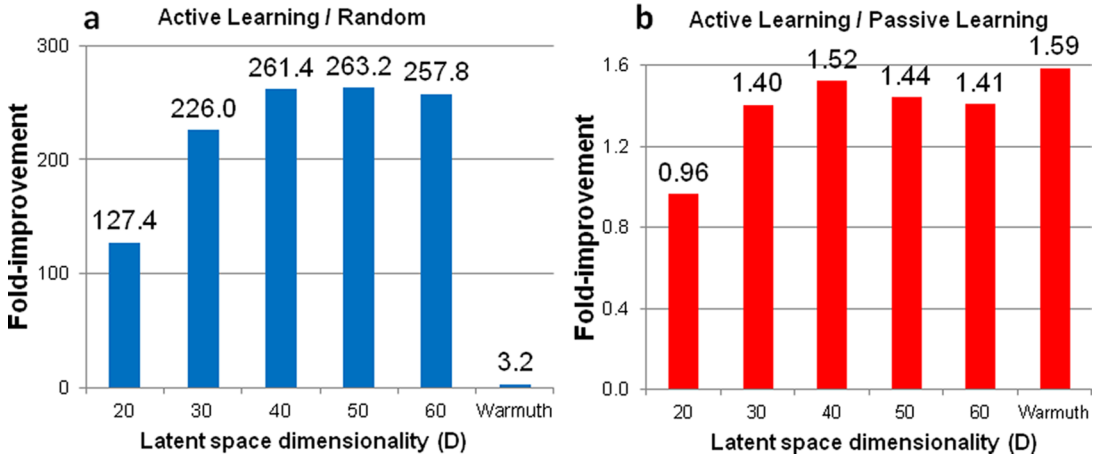



**Figure 6.** Improvement in prediction accuracy by AL over random (a) and over PL (b), as a function of the latent space dimensionality. Fold improvement is based on hit ratios obtained at the end of 1000 predictions, using same parameters as Figure 5. The AL performance levels off at about $D$ = 50 in panel a. The last bar in each panel refers to the work of Warmuth et al.[38]

**Table 2. De novo Predictions, Rank-Ordered Based on Confidence**

| drug | target | support from previous experiments (ref) |
|---|---|---|
| Ergotamine | serotonin receptor 1A (5-$HT_{1A}$[a]) | direct[32] |
| Amoxapine | serotonin receptor 2A (5-$HT_{2A}$) | direct[33] |
| Minaprine | histamine receptor $H_1$ | |
| Trimipramine | $\alpha_{2A}$ adrenergic receptor | indirect[b] |
| Amitriptyline | serotonin receptor 2C (5-$HT_{2c}$) | indirect[b] |
| Tramadol | serotonin receptor 2C (5-$HT_{2c}$) | indirect[b] |
| Clozapine | D(1B) dopamine receptor | indirect[b] |
| Doxepin | D(1A) dopamine receptor | indirect[b] |
| Nicardipine | histamine receptor $H_1$ | |
| Flunitrazepam | GABA[a]$\alpha_1$ | indirect[b] |
| Paliperidone | serotonin receptor 7 (5-$HT_7$) | direct[42] |
| Iloperidone | $\alpha_{2B}$ adrenergic receptor | indirect[b] |
| Propericiazine | $\alpha_{1A}$ adrenergic receptor (ADRA1A[a]) | indirect[b] |
| Asenapine | $\alpha_{1B}$ adrenergic receptor | indirect[b] |
| Verapamil | calmodulin | direct[34,47] |
| Meperidine | Na[a] channel 10, type $\alpha$ | direct[43] |
| Cinnarizine | calmodulin | direct[44] |
| Paroxetine (paxil) | histamine receptor $H_1$ | direct[45,46] |
| Orphenadrine | DAT[a] | indirect[c] |
| Citalopram | Na[a]-dependent NET[a] | indirect[47] |

[a]Abbreviations: GABA $\gamma$-aminobutyric-acid receptor; 5-HT 5-hydroxytryptamine (or serotonin) receptor; NET norepinephrine (or noradrenaline) transporter; ADRA1A adrenoreceptor $\alpha$1 A; DAT dopamine transporter. [b]The cases listed as "indirect" refer to interactions of the drugs with different subtypes of the identified target. [c]Orphenadrine inhibits norepinephrine reuptake thus potentiating the effect of norepinephrine There are several drugs acting as norepinephrine-dopamine reuptake inhibitors, and ophenadrine might exhibit the same behavior.

affinity of verapamil relative to that of trifluoperazine, using the module SMINA,[51] based on AutoDock VINA.[52] The software yielded an attractive energy of −4.8 kcal/mol for trifluoperazine binding to calmodulin, which after energy minimization, becomes −6.8 kcal/mol. Docking of verapamil to the same location and energy minimization led to −5.8 kcal/mol (Figure 7). While docking software with fixed target often fails to provide a quantitative assessment of binding affinity, and more elaborate simulation methods have been developed for druggability assessment (see, e.g., our recent work[53]), these results support, qualitatively, the shape complementarity between calmodulin and verapamil, as well as their favorable electrostatic interaction. Establishment of this and other predicted interactions, however, awaits experimental verifications by essays designed to test the specific drug−target interactions.

## 4. CONCLUDING REMARKS

Over the last couple of years, there have been a number of computational studies performed to identify targets of existing drugs and drug candidates other than those originally known/proposed to be targeted. A pioneering study is that of Roth, Shoichet, and co-workers[11,13] based on compound chemical similarities. Dudley et al. focused on inverse correlations between gene expression profiles in the presence of a drug and in a disease state.[25] Yamanishi and his colleagues represented drugs and targets in an integrated "pharmacological space".[17,20] Gonen used a KBMF method where chemical and genomic similarities were integrated.[22] We proposed a PMF-based AL methodology that can be advantageously used for large data sets.

The applicability of the method to large data sets is worth further attention, given that we will increasingly have access to bigger data (e.g., STITCH Database[30]), which will be exploited for repurposable drug identification. The software developed here, made accessible in http://www.csb.pitt.edu/Faculty/bahar/files/, is readily scalable. For very large data sets, which typically have more known interactions, the PMF is able to construct a better model using the plethora of available data; whereas when the number of known interactions is limited, the use of chemical and genomic kernels allows KBMF to outperform PMF. The application of KBMF to large data sets may, however, become challenging, For example, STITCH contains on the order of $10^6$ proteins and $10^5$ compounds, implying that $10^{12}$ sequence and $10^{10}$ chemical similarity comparisons are needed to make predictions. However, the PMF method is independent of chemical, structural or other

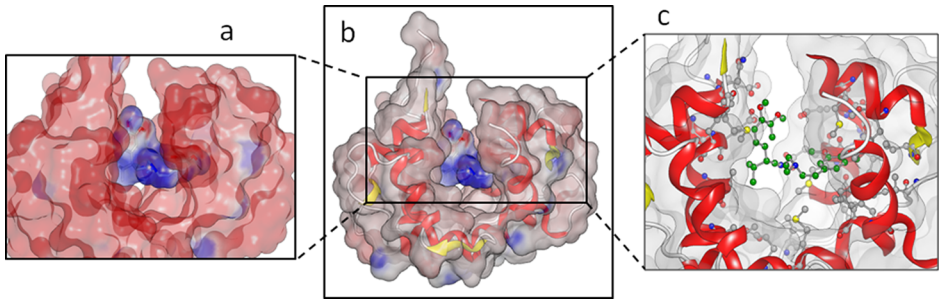



**Figure 7.** Structural model for possible binding pose of verapamil onto calmodulin. The molecular surfaces of both the drug and calmodulin are colored by electrostatic potential (a and b) and a close-up view is shown in panel c. Calmodulin surface, especially at the ligand-binding cleft, is negatively (partial) charged, favorably interacting with the verapamil surface that has partial positive charge. Note the shape between the verapamil and calmodulin. All the visualizations have been performed with OpenEye VIDA software.

similarity metrics, and its computation time scales linearly with the number of known interactions; and it proves to perform well on large data sets. The data sets reporting drug–target interactions are constantly improving in quality and quantity and, therefore, expected to give even better results when analyzed by an efficient tool. Finally, the extension of the method to analyzing big data (with millions of nodes) is foreseeable in the near future. The recently introduced GraphChi tool[54] can be used for optimized and parallelized model learning for further performance improvements.

The fact that the PMF is independent of 2D/3D shape comparison methods commonly employed in drug–target pair inferences implies that the derived LVs capture similarities based on the interaction patterns of drugs at the cellular level, even if their molecular structures are dissimilar (see Figures 2 and 3). As such, the method may be advantageously used for lead hopping, thus complementing those (e.g., SVM classification algorithms) used in conjunction with 2D or 3D pharmacophoric fingerprints (see the work of Saeh et al.[55]). Inasmuch as the currently proposed method does not require structural data for proteins but knowledge of drug–target interactions, it can be advantageously applied to membrane proteins (major drug targets) for which structural data still remain sparse. It can also be used to make predictions across major drug or target classification boundaries. One implication is that the de novo predictions are not restricted to major drug or target classification boundaries.

A major utility of the developed tool is the ability to deliver testable hypotheses with regard to repurposable drugs, thus significantly reducing the search space for identifying potent applications of existing drugs (that proved to meet ADMET requirements). The number of experiments that can be efficiently conducted is usually limited, e.g. of the order of $10^2$ if not $10^1$ for high-confidence assays as opposed to the complete space of ∼1.5 million combinations for the data set used in this study. The fact that the top-ranking predictions exhibit a hit ratio of 59% (for the top 1000 predictions; or 88% for top 100 predictions) suggest that de novo predictions made by the presently introduced method of approach applied to increasingly large data sets are likely to provide useful guidance for experimentally testing, streamlining or prioritizing existing or investigational drugs or new compounds. Another important byproduct is the probabilistic assessments on potential side effects, a topic that will become increasingly important with advances in personalized medicine.

## ■ ASSOCIATED CONTENT

### Supporting Information

Details on the methods, five figures, S1–S5, and Table S1. The figures illustrate the following: the drug–target interaction network and histograms denoting degree distribution for drug and target nodes (Figure S1), the structural heterogeneities of drugs despite their functional similarity (S2), running average of AUC over the cross-validation steps, and the correlation between AUC and number of known interactions (S3), the dependence of performance on data set size and fraction of known interactions (S4), and a comparison of the predictions from AL and PL methods for various dimensionalities of LV space (S5). Table S1 lists the therapeutic function characteristic of each of the 30 PMF-predicted clusters of drugs. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: bahar@pitt.edu.

### Author Contributions

§C.L. and F.H. provided equal contribution.

### Notes

The authors declare no competing financial interest.

⊥Visiting scholar from Tsinghua University.

## ■ ACKNOWLEDGMENTS

Support from NIH grant nos. U19 AI068021 and PO1 DK096990 is gratefully acknowledged by I.B. M.C.C. and I.B. are thankful to Dr. Ahmet Bakan for useful ideas, and they benefited from insightful discussions with Prof. Lans Taylor. A scholarship award to F.H. from the China Scholarship Council is gratefully acknowledged. We are grateful to Professor Olaf G. Wiest for careful reading of the manuscript and for making valuable comments and suggestions, especially concerning the medicinal chemistry aspects of the study.

## ■ REFERENCES

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203–214.

(2) Berg, J. M.; Rogers, M. E.; Lyster, P. M. Systems biology and pharmacology. *Clin. Pharmacol. Ther.* **2010**, *88* (1), 17–19.

(3) Csermely, P.; Korcsmaros, T.; Kiss, H. J. M.; London, G.; Nussinov, R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol. Ther.* **2013**, *138*, 333–408.

(4) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16* (1), 127−136.
(5) Sorger, P. K.; Allerheiligen, S. R. B.; Abernethy, D. R.; Altman, R. B.; Brouwer, K. L. R.; Califano, A.; David, Z.; Argenio, D.; Iyengar, R.; Jusko, W. J.; Lalonde, R.; Lauffenburger, D. A.; Shoichet, B.; Stevens, J. L.; Subramaniam, S.; Graaf, P. V. D.; Ward, R. *Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms*; NIH White Paper, 2011; pp 1−47.
(6) Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4* (11), 682−690.
(7) Hillenmeyer, M. E.; Fung, E.; Wildenhain, J.; Pierce, S. E.; Hoon, S.; Lee, W.; Proctor, M.; St Onge, R. P.; Tyers, M.; Koller, D.; Altman, R. B.; Davis, R. W.; Nislow, C.; Giaever, G. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **2008**, *320* (5874), 362−365.
(8) Zambrowicz, B. P.; Sands, A. T. Modeling drug action in the mouse with knockouts and RNA interference. *Drug Discovery Today* **2004**, *3* (5), 198−207.
(9) Berger, S. I.; Iyengar, R. Network analyses in systems pharmacology. *Bioinformatics* **2009**, *25* (19), 2466−2472.
(10) Murphy, R. F. An active role for machine learning in drug development. *Nat. Chem. Biol.* **2011**, *7* (6), 327−330.
(11) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197−206.
(12) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486* (7403), 361−367.
(13) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462* (7270), 175−181.
(14) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Comput. Biol.* **2009**, *5* (7), e1000423.
(15) Li, Y. Y.; An, J.; Jones, S. J. M. A computational approach to finding novel targets for existing drugs. *PLoS Comput Biol* **2011**, *7* (9), e1002139.
(16) Xie, L.; Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (14), 5441−5446.
(17) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug−target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232−i240.
(18) van Laarhoven, T.; Nabuurs, S. B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug−target interaction. *Bioinformatics* **2011**, *27* (21), 3036−3043.
(19) Bleakley, K.; Yamanishi, Y. Supervised prediction of drug−target interactions using bipartite local models. *Bioinformatics* **2009**, *25* (18), 2397−2403.
(20) Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **2010**, *26* (12), i246−i254.
(21) Perlman, L.; Gottlieb, A.; Atias, N.; Ruppin, E.; Sharan, R. Combining drug and gene similarity measures for drug−target elucidation. *J. Comput. Biol.* **2011**, *18* (2), 133−145.
(22) Gonen, M. Predicting drug−target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* **2012**, *28* (18), 2304−2310.
(23) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A unified, probabilistic framework for structure- and ligand-based virtual screening. *J. Med. Chem.* **2011**, *54* (5), 1223−1232.
(24) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J. P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S.; Haggarty, S. J.; Clemons, P.; Wei, R.; Carr, S.; Lander, E. S.; Golub, T. R. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313* (5795), 1929−1935.
(25) Dudley, J. T.; Sirota, M.; Shenoy, M.; Pai, R. K.; Roedder, S.; Chiang, A. P.; Morgan, A. A.; Sarwal, M. M.; Pasricha, P. J.; Butte, A. J. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Sci. Transl. Med.* **2011**, *3* (96), 96ra76.
(26) Sirota, M.; Dudley, J. T.; Kim, J.; Chiang, A. P.; Morgan, A. A.; Sweet-Cordero, A.; Sage, J.; Butte, A. J. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. *Sci. Transl. Med.* **2011**, *3* (96), 96ra77.
(27) Chiang, A. P.; Butte, A. J. Systematic Evaluation of Drug−Disease Relationships to Identify Leads for Novel Drug Uses. *Clin. Pharmacol. Ther.* **2009**, *86* (5), 507−510.
(28) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining. *J. Chem. Inf. Model.* **2011**, *51* (9), 2440−2448.
(29) Gottlieb, A.; Stein, G. Y.; Ruppin, E.; Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011**, *7*, 1−9.
(30) Kuhn, M.; Szklarczyk, D.; Franceschini, A.; von Mering, C.; Jensen, L. J.; Bork, P. STITCH 3: zooming in on protein-chemical interactions. *Nucleic. Acids. Res.* **2012**, *40* (Database issue), D876−D880.
(31) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic. Acids. Res.* **2011**, *39* (Database issue), D1035−D1041.
(32) Tfelt-Hansen, P.; Saxena, P. R.; Dahlöf, C.; Pascual, J.; Láinez, M.; Henry, P.; Diener, H.; Schoenen, J.; Ferrari, M. D.; Goadsby, P. J. Ergotamine in the acute treatment of migraine: a review and European consensus. *Brain* **2000**, *123*, 9−18.
(33) Pälvimäki, E. P.; Roth, B. L.; Majasuo, H.; Laakso, A.; Kuoppamäki, M.; Syvälahti, E.; Hietala, J. Interactions of selective serotonin reuptake inhibitors with the serotonin 5-HT2c receptor. *Psychopharmacology* **1996**, *126* (3), 234−40.
(34) Epstein, P. M.; Fiss, K.; Hachisu, R.; Andrenyak, D. M. Interaction of calcium antagonists with cyclic AMP phosphodiesterases and calmodulin. *Biochem. Biophys. Res. Commun.* **1982**, *105* (3), 1142−1149.
(35) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug−target network. *Nat. Biotechnol.* **2007**, *25* (10), 1119−1126.
(36) Salakhutdinov, R.; Mnih, A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine learning*, Helsinki, Finland, July 5−9; ACM, 2008; pp 880−887.
(37) Salakhutdinov, R.; Mnih, A. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*; NIPS, 2007; Vol. *20*, pp 1257−1264.
(38) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 667−673.
(39) Akaike, H. A New Look at the Statistical Model Identification. *IEEE. T. Automat. Contr.* **1974**, *19* (6), 716−723.
(40) Yera, E. R.; Cleves, A. E.; Jain, A. N. Chemical Structural Novelty: On-Targets and Off-Targets. *J. Med. Chem.* **2011**, *54* (19), 6771−6785.
(41) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50* (4), 572−584.

(42) Kast, R. E. Glioblastoma chemotherapy adjunct via potent serotonin receptor-7 inhibition using currently marketed high-affinity antipsychotic medicines. *Br. J. Pharmacol.* **2010**, *161* (3), 481−487.
(43) Wagner, L. E.; Eaton, M.; Sabnis, S. S.; Gingrich, K. J. Meperidine and Lidocaine Block of Recombinant Voltage-Dependent Na+ Channels. *Anesthesiology* **1999**, *91* (5), 1481−1490.
(44) Zimmer, M.; Hoffmann, F. Differentiation of the drug-binding sites of calmodulin. *Eur. J. Biochem.* **1987**, *164*, 411−420.
(45) Rahmadi, M.; Narita, M.; Yamashita, A.; Imai, S.; Kuzumaki, N.; Suzuki, T. Sleep disturbance associated with an enhanced orexinergic system induced by chronic treatment with paroxetine and milnacipran. *Synapse.* **2011**, *65* (7), 652−657.
(46) Fava, M. Weight gain and antidepressants. *J. Clin. Psychiatry* **2001**, *61* (suppl 11), 37−41.
(47) Petersen, B.; Mørk, A. Chronic treatment with citalopram induces noradrenaline receptor hypoactivity. A microdialysis study. *Eur. J. Pharmacol.* **1996**, *300*, 67−70.
(48) Raess, B. U.; Gersten, M. H. Calmodulin-stimulated plasma membrane ($Ca2^{+}$ + $Mg^{2+}$)-ATPase: inhibition by calcium channel entry blockers. *Biochem. Pharmacol.* **1987**, *36* (15), 2455−9.
(49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(50) Cook, W. J.; Walter, L. J.; Walter, M. R. Drug binding by calmodulin: crystal structure of a calmodulin-trifluoperazine complex. *Biochemistry* **1994**, *33* (51), 15259−15265.
(51) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53* (8), 1893−1904.
(52) Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31* (2), 455−461.
(53) Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2435−2447.
(54) Kyrola, A.; Blelloch, G.; Guestrin, C. GraphChi: Large-scale graph computation on just a PC. In *Proceedings of the 10th Symposium on Operating Systems Design & Implementation*, Hollywood, CA, Oct 8−10, 2012.
(55) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model.* **2005**, *45* (4), 1122−1133.
(56) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011**, *27* (3), 431−432.