ANTENNA, a Multi-Rank, Multi-Layered Recommender System for Inferring Reliable Drug-Gene-Disease Associations: Repurposing Diazoxide as a Targeted Anti-Cancer Therapy

Annie Wang, Hansaim Lim, Shu-Yuan Cheng, and Lei Xie

Abstract—Existing drug discovery processes follow a reductionist model of "one-drug-one-gene-one-disease," which is inadequate to tackle complex diseases involving multiple malfunctioned genes. The availability of big omics data offers opportunities to transform drug discovery process into a new paradigm of systems pharmacology that focuses on designing drugs to target molecular interaction networks instead of a single gene. Here, we develop a reliable multi-rank, multi-layered recommender system, ANTENNA, to mine large-scale chemical genomics and disease association data for prediction of novel drug-gene-disease associations. ANTENNA integrates a novel tri-factorization based dual-regularized weighted and imputed One Class Collaborative Filtering (OCCF) algorithm, tREMAP, with a statistical framework based on Random Walk with Restart and assess the reliability of specific predictions. In the benchmark, tREMAP clearly outperforms the single-rank OCCF. We apply ANTENNA to a real-world problem: repurposing old drugs for new clinical indications without effective treatments. We discover that FDA-approved drug diazoxide can inhibit multiple kinase genes responsible for many diseases including cancer and kill triple negative breast cancer (TNBC) cells efficiently $(\mathrm{IC}_{50} = 0.87\,\mu\mathrm{M})$. TNBC is a deadly disease without effective targeted therapies. Our finding demonstrates the power of big data analytics in drug discovery and developing a targeted therapy for TNBC.

Index Terms—Anti-cancer targeted therapy, big data analytics, data mining, diazoxide, drug discovery, drug repurposing, machine learning, multi-layered network, tri-factorization, triple negative breast cancer, prediction reliability

1 Introduction

The cost of bringing a drug to market has risen to approximately 2.6 billion dollars (Tufts Center for the Study of Drug Development, 2015), and the failure rate is daunting: only about one-third of drugs in phase III clinical trials reach the market. The limited success of the conventional drug discovery process is largely attributed to the wide adoption of a reductionist model of "one-drug-one-gene-one-disease" [1], [2], [3]. As a matter of fact, the onset and progress of many complex diseases such as cancer is a systematic process that involves multiple interacting genes. Thus, it is necessary to design drugs that target gene interaction networks instead of a single gene. Moreover, drug repurposing that reuses existing safe drugs to treat new diseases has emerged as a new paradigm to accelerate drug discovery and development. As the safety profile of existing medicines has already been well documented, the cost of clinical trials can be significantly reduced.

Recent advances in high-throughput technologies have generated abundant chemical genomics data on drug actions and disease genes. These big, complex, heterogeneous data sets provide unprecedented opportunities for identifying genome-wide drug-gene-disease associations, thereby facilitating multi-targeted drug design and drug repurposing. However, several challenges remain in mining chemical genomics and disease association data for drug discovery. Firstly, chemical genomics data from high-throughput screening campaigns are not only extremely large but also highly noisy, biased, and incomplete. Many existing data mining algorithms cannot be directly applied to model chemical genomics data. Secondly, drug action is a complex process. It starts with drug-gene interactions at the molecular level, and manifest clinical outcomes through biological network. A single genomics data set can only capture one part of whole drug process. Thus, it is necessary to integrate multiple data sets for chemical-gene interactions, gene-disease associations, and chemical-disease associations to model the drug action on a multi-layer. Finally, one of the fundamental problems in biomedical data mining has not been fully addressed: how to

- *A. Wang is with the Bronx High School of Science, 75 W 205th St., Bronx, NY 10468. E-mail: wangan@bxscience.edu.*
- *H. Lim is with the Ph.D. Program in Biochemistry, the City University of New York, 365 5th Avenue, New York, NY 10016. E-mail: hlim1@gradcenter.cuny.edu.*
- *S. Cheng is with Department of Sciences, John Jay College, the City University of New York, 365 5th Avenue, New York, NY 10016. E-mail: shcheng@jjay.cuny.edu.*
- *L. Xie is with Department of Computer Science, Hunter College, and the Graduate Center, the City University of New York, 695 Park Ave., New York, NY 10065. E-mail: lei.xie@hunter.cuny.edu.*

Manuscript received 9 Feb. 2018; accepted 26 Feb. 2018. Date of publication 16 Mar. 2018; date of current version 6 Dec. 2018.
(Corresponding author: Lei Xie).
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2812189

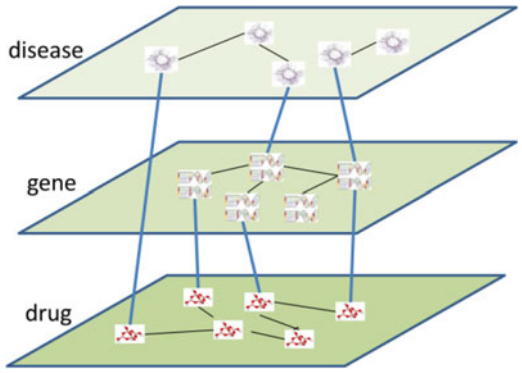


Fig. 1. Illustration of multi-layered network model (MULAN) that integrates multiple genomics data sets.

assess the individual reliability of a specific prediction from a data mining agent under a rigorous statistics framework. The reliable and unbiased assessment of the prediction quality for an individual instance is critical for cost-sensitive drug discovery process. For example, the selection of a novel chemical that is structurally different from patented drugs as a lead compound from a ranked list of candidate chemicals is a billion-dollar decision. Information on the individual predictive reliability of a novel chemical entity based on its weak chemical similarity to existing drugs in terms of bioactivity is invaluable. Most existing data mining tools can only provide an average predictive accuracy based on the population of training data, but not reliability for a specific new case. For example, in a ranking system, it is not straightforward to determine what the threshold is to select top-ranked hits. For a specific case, the top-first ranked hit could be a false positive. In another scenario, the top-N $(N > 1)$ ranked hits could all be true positives.

2 Contributions of This Work

To address challenges in the predictive modeling of drug-gene-disease associations as well as unmet needs in the treatment of complex diseases such as cancer, this work makes contributions to both methodology development and translational medicine.

On the side of methodology development, our contribution is twofold. First, we have developed a novel algorithm tREMAP based on tri-factorization to optimize matrix completion problem in which row and column have significantly different ranks. tREMAP formulates the chemical-gene predictions as a multi-rank dual-regularized weighted and imputed One Class Collaborative Filtering (OCCF) problem. Under the formulation of OCCF, negative data is not needed for the training, which is sparse and even unavailable. By using element-specific weights and imputation, tREMAP can handle noisy chemical genomics data in which the label is often uncertain. Finally, unlike conventional OCCF algorithm that applies a single rank to all layers, tREMAP assigns a different rank to a different layer. It is important since different layers can have dramatically different dimensions thus optimal ranks. For example, the dimension of a chemical layer is in the order of millions, while the dimension of a gene layer is only thousands. Our benchmark studies clearly show that tREMAP outperforms single-rank OCCF method. Second, to tailor the nature of chemical-gene-disease association data sets where observed chemical-disease associations are far sparser than known chemical-gene interactions and few three-way chemical-gene-disease associations exist, we have developed a multi-rank, multi-layered framework ANTENNA for inferring novel chemical-gene-disease associations. ANTENNA has three main components. (1) ANTENNA integrates multiple chemical genomics and disease association data set, and links them as a multi-layered network [4], as shown in Fig. 1. (2) ANTENNA uses tREMAP to infer genome-wide novel chemical-gene associations. (3) Based on the genome-wide chemical-gene association, ANTENNA applies Random Walk with Restart (RWR) and a statistics framework, Enrichment of Topological Similarity (ENTS) [5], to predict chemical-disease associations and assess their reliabilities.

Arguably, the most important contribution of this work is to discover a potentially safe and effective targeted therapy for triple negative breast cancer (TBNC). Using ANTENNA, we predicted that an FDA-approved drug diazoxide may inhibit multiple kinase genes. The malfunction of kinases is associated with many diseases such as cancer and Alzheimer's disease. Among the kinases with the highest percentage of inhibition by diazoxide, one gene TTK is specifically over-expressed in the patients with TNBC [6], [7]. Thus, we hypothesized that diazoxide may kill TNBC cells. Our predictions were supported by multiple experimental evidence. TNBC is a subgroup of breast cancers, which is associated with the most aggressive clinical behavior. No targeted therapy is currently available for the treatment of TNBC. Our finding has a great potential for developing a targeted therapy for the effective treatment of TNBC.

3 Relevant Works

In principle, tensor factorization is a powerful method to infer three-way relationships. However, observed three-way chemical-gene-disease relations are extremely sparse. Majority of observed chemical-gene pairs are not associated with any diseases. Thus, the tensor factorization may be not the best option for this work. OCCF has been applied to a bipartite graph for predicting drug-target interactions [8], but not to inferring multiple drug-gene-disease associations. Moreover, existing OCCF algorithm is mainly based on the formulization of matrix factorization that only allows a common rank for both row and column. FASCINATE is an algorithm that can jointly infer missing links from a multi-layered network model [4]. However, FASCINATE is based on the formulation of a single rank collective OCCF. Moreover, it can only rank predicted relations [4]. There is no reliability information associated with each individual prediction. This work will address the drawbacks in matrix factorization, OCCF, and FASCINATE when applied to inferring chemical-gene-disease associations.

4 Experimental and Computational Details

4.1 Overview of Computational and Experimental Procedure

Our primary purpose is to mine chemical genomics and disease association data to identify novel targeted therapies for unmet biomedical problems such as the treatment of TBNC. As shown in Fig. 2, the input of ANTENNA is the existing chemical genomics, drug, and disease databases including

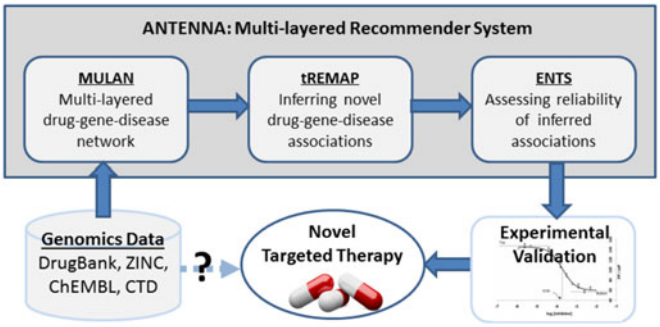


Fig. 2. Workflow of drug discovery process using ANTENNA, a multi-layered recommender system.

DrugBank [9], ZINC [10], ChEMBL [11], and CTD [12]. We first integrate these data sets into a multi-layered chemical-gene-disease network, MULAN. Then we apply tREMAP, a multi-rank dual-regularized weighted imputed OCCF algorithm, to infer novel chemical-gene associations. Next, we used ENTS to predict drug-disease association and to assess the reliability for each inferred association. The output of ANTENNA is a list of ranked drug-disease associations ranked by their statistical significance. Finally, we experimentally validate the top-ranked predictions.

4.2 Construction of Multi-Layered Chemical-Gene-Disease Network (MULAN)

We integrated heterogeneous data sets from genomics into a multi-layered network model, MULAN. In the MULAN, each node is a chemical entity (drugs and other chemicals), a biological entity (genes or proteins that it encodes), or a phenotypic entity (disease and side effect). Nodes in the same entity class are linked together by similarities (e.g., chemical-chemical similarity) or interactions (e.g., protein-protein interactions). Nodes that belong to different entity classes reside in different network layers and are linked by known associations (e.g., drug-target interactions, disease-gene associations). Integration of genomics data into a bipartite graph is of a proven value [13]. The MULAN can be considered as the unification of multiple bipartite graphs; thus, our new method is likely to be more robust than traditional approaches.

Chemical-gene associations including drug-gene associations were obtained from the ZINC [14], ChEMBL [15] and DrugBank [9] databases. To obtain reliable chemical-gene association pairs, binding assays records with IC_{50} (concentration of the chemical needed to inhibit 50 percent of the activity of the target protein) information were extracted from the databases, and the cutoff IC_{50} value of $10\,\mu M$ was used where applicable. Chemical-gene pairs were considered associated if $IC_{50} \leq 10\,\mu M$ (active pairs), unassociated if $IC_{50} > 10\,\mu M$ (inactive pairs), ambiguous if records exist in both ranges (ambiguous pairs), and unobserved otherwise (unknown pairs). A total of 198,712 unique chemicals and 3,549 unique genes were obtained from the combination of ChEMBL and ZINC with 228,725 unique chemical-gene active pairs, 76,643 inactive pairs, and 4,068 ambiguous pairs. Of the 198,712 chemicals, 722 were found to be FDA-approved drugs. Furthermore, drug-gene relationships were extracted from the DrugBank and integrated into the ZINC_ChEMBL dataset above. A total of 199,338 unique chemicals and 6,277 unique genes were obtained from the combination of ZINC, ChEMBL, and DrugBank with 233,378 unique chemical-gene active pairs. Drug-disease and gene-disease associations were directly obtained from the Comparative Toxicology Database (CTD) [12].

Chemical-chemical similarity scores are one of the required inputs of tREMAP. Although there are a number of metrics developed for chemical-chemical similarity, a recent study showed that Jaccard index-based similarity is highly efficient for fingerprint-based similarity measurement [16]. The fingerprint of choice in this study is the Extended Connectivity Fingerprint (ECFP), which has been successfully applied to chemical structure-based target prediction method, PRW [17]. Jaccard index is used to calculate a similarity score between two chemicals, c_1 and c_2.

Gene-gene similarity scores are also one of the required inputs for tREMAP. The similarity between two proteins encoded by genes was calculated based on their amino acid sequence similarity using NCBI BLAST [18] with an e-value threshold of 1×10^{-5} and its default options. A similarity score for query protein p_1 to target protein p_2, $d_{bit\ (p_1,\,p_2)}$, was calculated by the ratio of a bit score for the pair compared to the bit score of a self-query. To be specific, for the query protein p_1 to the target protein p_2, protein-protein similarity score was defined such that $T_{(p_1,p_2)} = d_{bit\ (p_1,\,p_2)}/d_{bit\ (p_1,\,p_1)}$.

Disease-disease similarity is required for tREMAP to infer chemical-disease associations and can be calculated using distributed word representations [19]. In this work, we do not infer the chemical-disease association directly using tREMAP, since only less than 0.4 percent of chemicals have observed associations with one or more diseases. Instead, we use ENTS and target binding profile of a chemical, which is derived from tREMAP, to infer the chemical-disease associations.

4.3 tREMAP Algorithm

Our prediction method tREMAP is based on a tri-factorization one-class collaborative filtering algorithm. In the case of chemical-gene association, it assumes that similar chemicals will interact with similar genes, and unobserved associations are not necessarily negative. Assuming that a fairly low number of factors (i.e., smaller number of features than the number of total chemicals or genes) may capture the characteristics determining the drug-gene associations, two low-rank matrices, F (drug side) and G (gene side), were approximated such that $\sum_i^n \sum_j^m \{R - (F \cdot S \cdot G')\}$ is minimized where R is the matrix for known drug-gene interactions and G' is the transposition of the gene side low-rank matrix G. The two low rank matrices, $F_{n \times r_1}$ with the rank of r_1 and $G_{m \times r_2}$ with the rank of r_2, and their connectivity matrix $S_{r_1 \times r_2}$ are obtained by iteratively minimizing the objective function.

$$\begin{aligned} \min_{F,S,G \geq 0} \sum_{(u,i)} W_{(u,i)} \left(R_{(u,i)} + P_{(u,i)} - (FSG')_{(u,i)} \right)^2 \\ + \lambda_r \left(\|F\|^2 + \|S\|^2 + \|G\|^2 \right) \\ + \lambda_F tr(F'(D_M - M)F) \\ + \lambda_G tr(G'(D_N - N)G) \end{aligned} \tag{1}$$

Here, $W_{(u,i)}$ is the penalty weight on the observed and unobserved associations which indicate the reliability of the

assigned probability of true association, $P_{(u,i)}$ is the imputed value (i.e., the probability of unobserved associations as real associations), M and N is the symmetric chemical-chemical similarity matrix and gene-gene similarity matrix, respectively. D_M and D_N are the degree matrix of M and N, respectively. λ_r is the regularization parameter to prevent overfitting, λ_F is the importance parameter for chemical-chemical similarity, λ_G is the importance parameter for gene-gene similarity, and $tr(\mathrm{A})$ is the trace of matrix A. The weight and imputation values can be determined by *a priori* knowledge or from the prediction of other machine learning algorithms. The first term in (1) forces the approximation FSG' to be close to the observation matrix R. The second term is regularization term preventing overfitting. The third and fourth terms force the low-rank feature vectors close to each other according to their chemical-chemical or protein-protein similarity score. Thus, the optimal low-rank matrix F was obtained after minimizing the sum of Euclidean distances for each row weighted by the chemical-chemical similarity score. The derivation of the formula can be found in [20].

Similar to the bi-factorization problem in [20], the optimization problem defined in (1) is non-convex. Thus, we seek to find a local optimum by the block coordinate descent method. In (1), D_M, M, D_N, and N are non-negative matrices. The derivative of (1) with regard to F, G, and S with the non-negativity constraint has a fixed-point solution. To scale up tREMAP in terms of both time and storage, we propose efficient multiplicative updating rules as follows:

$$\mathrm{F}_{(\mathrm{u,r})} \leftarrow \mathrm{F}_{(\mathrm{u,r})} \sqrt{\frac{[(1-wp)\mathrm{RGS}' + wp\mathbf{1}_{m\times n}\mathrm{GS}' + \lambda_F \mathrm{MF}]_{(\mathrm{u,r})}}{\left[(1-w)\widetilde{R_1}\mathrm{GS}' + w\mathrm{F}(\mathrm{SG}'\mathrm{GS}') + \lambda_r \mathrm{F} + \lambda_F \mathrm{D_M F}\right]_{(\mathrm{u,r})}}} \tag{2}$$

$$\mathrm{G}_{(\mathrm{i,s})} \leftarrow \mathrm{G}_{(\mathrm{i,s})} \sqrt{\frac{[(1-wp)\mathrm{R}'\mathrm{FS} + wp\mathbf{1}_{n\times m}(\mathrm{FS}) + \lambda_G \mathrm{NG}]_{(\mathrm{i,s})}}{\left[(1-w)\widetilde{R_1}\,(\mathrm{FS}) + w\mathrm{G}(\mathrm{S}'\mathrm{F}'\mathrm{FS}) + \lambda_r \mathrm{G} + \lambda_G \mathrm{D_N G}\right]_{(\mathrm{i,s})}}} \tag{3}$$

$$S_{(\mathrm{r,s})} \leftarrow \mathrm{S}_{(r,\mathrm{s})} \sqrt{\frac{[(1-wp)\mathrm{F}'\mathrm{RG} + wp(\mathrm{F}'(\mathbf{1}_{m\times n})G)]_{(\mathrm{r,s})}}{\left[(1-w)\mathrm{F}'\widetilde{R_1}\mathrm{G} + w\mathrm{F}'(\mathrm{FSG}')G\right]_{(r,\mathrm{s})} + \lambda_r \mathrm{S}}}, \tag{4}$$

Where w and p are weighted and imputed value, respectively. They are either set based on a priori knowledge (e.g., the false positive rate of high-throughput screening experiments) or can be tuned as hyper-parameters. $\widetilde{R_1}_{(u,i)}$ is the sparse matrix in which the value of elements is predicted by F and G on the observed cases Θ in R, i.e.,

$$\widetilde{R_1}_{(u,i)} = \begin{cases} \mathrm{FSG}'_{(\mathrm{u,i})} & \text{if } (u,i) \in \Theta \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We use a block-coordinate descent algorithm to iteratively update F, G, and S.

The raw predicted score for the ith chemical to bind the jth protein can be calculated by $P_{(i,j)} = F_{(i,:)} \cdot S \cdot G'_{(j,:)}$. Also, the matrix $F_{n\times r_1}$ is referred to as a low-rank drug profile since its ith row represents the ith drug's behavior in the drug-gene association network as well as drug-drug similarity spaces compressed to r_1 number of features.

4.4 ENTS Algorithm

The rationale of ENTS is that when clusters of instance share common features, a cluster ranked closely together is more likely similar to the new instance than a cluster ranked randomly or spread out across the ranking. In addition, network topological similarity provides more robust and accurate global ranking across an entire hypothesis space than pairwise similarity does. Unlike conventional local ranking (e.g., k-nearest neighbors), global instance ranking can support statistical enrichment analysis because it draws valuable information on the ranking for all instances in a cluster from lower, non-randomly ranked cases.

4.4.1 Classification or Clustering of Database Instances

To initialize ENTS, part or all of the instances in the database (training set) are classified based on target feature T. In ANTENNA, the T is the disease associated with a drug. If database instances are not pre-classified, clusters of training data are assembled using T features under unsupervised clustering techniques [21] such as k-means [22], mean-shift [23], affinity propagation [24], or p-median model [25] etc. After the classification or clustering, each instance cluster will be assigned with a unique label (i.e., a specific disease in ANTENNA). These instance clusters are applied to the next step. It is noted that the instance clusters are not necessarily disjointed. They can overlap.

4.4.2 A Weighted Graph Represents Training Instance Similarity by T-Features

After the initialization, ENTS builds a database instance graph; a weighted graph with one node for the T-feature of each training instance and an edge between two nodes only if their pairwise similarity exceeds a certain threshold. The threshold depends on the features and the pairwise similarity metric. Any similarity metric (e.g., Euclidean distance, Jaccard index, Hidden Markov Model, kernel-based similarity etc.) can be applied here. In ANTENNA, we use cosine similarity of low-rank profile of drugs to measure the distance between drugs.

4.4.3 Network Topological Similarity

Given a query with known K-feature and the goal to predict its unknown T-feature, ENTS first links the query to all nodes in the training instance graph, where new edges are not found in the training instance graph. The weights of these new edges are only based on K-feature similarity. Then Random Walk with Restart (RWR) is applied to perform a probabilistic traversal of the instance graph across all paths leading away from the query, where the probability of choosing an edge will be proportional to its weight. The algorithm will output a list of all instances in the graph, ranked by the probability that a path from the query will reach the node. In this way, RWR can capture global relationships that may be missed by pair-wise similarity [26].

We modified the RankProp algorithm [27], a variant of RWR. The graph is represented as an adjacency list to save memory and speed up the iterative algorithm. The current implementation is scalable to a graph with millions of nodes and hundreds of millions of edges.

4.4.4 Statistical Significance of Network Topological Similarity

A network topological search only ranks instances based on their similarity but gives no information on the reliability of the ranking. To assess the statistical significance of the ranking of an instance cluster C_i generated previously, ENTS compares the score distribution of the cluster C_i with that of a randomly drawn cluster of the same size. When the mean of global topological similarity scores $\bar{X}$ in a cluster is used as the statistic, an efficient random-set method is used for the parametric approximation of the null distribution [28]. The random-set method compares an enriched cluster of size m with all other distinct clusters of size m drawn randomly from a case graph on N nodes. The exact distribution of $\bar{X}$ is intractable, but can be approximated with the normal distribution with mean and variance as follows:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} p_j$$

$$\sigma^2 = \frac{1}{m}\left(\frac{N-m}{N-1}\right)\left[\left(\frac{1}{N}\sum_{j=1}^{N} {p_j}^2\right) - \left(\frac{1}{N}\sum_{j=1}^{N} p_j\right)^2\right],$$

Where p_j is the global topological similarity score of the structure j in the graph to the query. The enrichment score of the cluster C_i is then normalized with $Z = (\bar{X} - \mu)/\sigma$.

A p-value and Benjamini-Hochber adjusted false discovery rate (FDR) is then calculated for each Z-score.

4.5 Combining tREMAP and ENTS to Predict Drug-Disease Association

In ANTENNA, we firstly use tREMAP to generate chemical-side low rank matrix F and gene side low-rank matrix G. The ith row of F contains the gene association profile for the ith drug. Then, we calculated drug-drug cosine similarities based on the matrix F, and construct a drug-drug similarity graph. For each row of F for FDA approved drugs, the cosine similarity of drug c_1 and drug c_2 can be calculated by, $S_{cos,(c_1,c_2)} = \frac{U_{c_1} \cdot U_{c_2}}{\sqrt{|U_{c_1}|}\sqrt{|U_{c_2}|}}$. To search for possibly undiscovered uses of the drugs, we focus on drugs that are found to have high cosine similarity but low chemical structural similarity (< 0.5). Finally, we cluster drugs based on their directly or indirectly associated diseases annotated in CTD database [12], and use ENT to assess and rank the statistical significance of novel drug-disease associations. The final output of ANTENNA is the ranked list of predicted drug-disease association based on FDR.

4.6 Experimental Validation

4.6.1 Kinase Binding Assay

Kinase is an enzyme that catalyzes the transfer of a chemical group phosphate to another biomolecule. It functions as a molecular switch in many biological processes. The malfunction of kinases is responsible for many diseases such as cancer. There are more than 400 kinases in the human genome, which is termed as kinome. To rigorously validate the performance of ANTENNA, we employed a competition binding assay to detect the binding of selected drugs to a set of 438 kinases (human kinome). The proprietary KinomeScan assay was performed by DiscoverX (CA). The assay tested the capacity for a drug to disrupt the binding of each DNA-tagged kinase to a support which one was in turn bound to the kinase's known ligand. If binding between the kinase and its known ligand was disrupted in the presence of the drug, this indicated that the drug either competed directly with the known ligand or allosterically altered the kinase's ability to bind to that ligand. DMSO was used as a positive control and a pico-molar kinase inhibitor was used as a negative control. Binding levels were quantitated by performing real-time polymerase chain reaction (qPCR) on the DNA tag of the ligand-bound kinases. The qPCR is a molecular biology technique to amplify a single copy or a few copies of DNA segment in several orders of magnitudeand to measure the reaction in a real time. The tests were performed at 100 μM concentration of tested drug, and results were reported as percentControl, calculated as follows, where a lower percentControl score indicates a stronger interaction.

$$\frac{(test\ compound\ signal - positive\ control\ signal)}{(negative\ control\ signal - positive\ control\ signal)}\ X\ 100.$$

4.6.2 Cancer Cell Viability Assay

MCF-7 cells from ATCC and MDA-MB 468 cells (a gift of Dr. R Sullivan from Queens Community College, the City University of New York) were used for this study. MCF-7 is breast cancer cell line. MDA-MB 468 is triple negative breast cancer cell line which does not express estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor (Her2/neu). Cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Thermo Fisher Scientific) supplemented with 10 percent fetal bovine serum (Thermo Fisher Scientific) and $50\,\mu g/ml$ gentamicin (Thermo Fisher Scientific) at $37^{\circ}C$ 5 percent CO_2 incubator.

Cell viability was determined by neutral red assay which is based on the lysosome uptake of neutral red dye [29]. Briefly, cells (2×10^4 cells per well) were plated onto 96-well plate in a total volume of $200\,\mu l$ on the day before chemical treatments. Chemicals were dissolved in dimethyl sulfoxide (DMSO) to obtain 0.1 M stock solution 15 minutes before chemical treatments. Then, various concentrations $(0.1 - 150\,\mu M)$ of chemicals were prepared in fresh media. The final concentration of DMSO in each well was equal to or less than 0.15 percent which is considered non-toxic to cells [30].

After 24 hours of chemical treatments, $20\,\mu l$ of 0.33 percent Neutral Red Solution (Sigma Aldrich) was added onto wells. After 2 hours incubation at $37^{\circ}C$ 5 percent CO_2 incubator, dye solution was carefully removed and cells were rinsed with $200\,\mu l$ Neutral Red Assay Fixative (0.1 percent $CaCl_2$ in 0.5 percent formaldehyde) (Sigma Aldrich) twice. The absorbed dye was then solubilized in $200\,\mu l$ of Neutral Red Assay Solubilization Solution (1 percent acetic acid in 50 percent ethanol) (Sigma Aldrich) for 10 minutes at room temperature on a shaker. Absorbance at 540 nm and 690 nm (background) was measured by BioTek Synergy Mx microplate reader.

Each concentration in each experiment was done in at least triplicate. Multiple experiments were done to obtain IC_{50} values for each drug and each cell line. The viability was determined based on a comparison with untreated cells which were set as 100 percent cell viability. The IC_{50} values

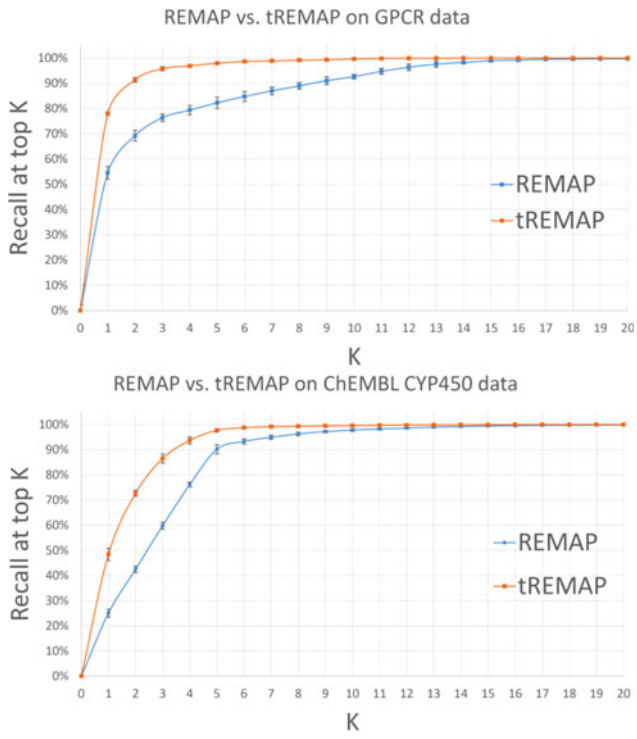


Fig. 3. Performance comparison of tREMAP with REMAP for GPCR (top) and CYP450 (bottom), respectively. Performance is measured by the recall at the top rank K.

which represent the chemical concentration needed to inhibit 50 percent cell proliferation were calculated from the dose-response curve.

5 Results and Discussions

5.1 Performance Evaluation of tREMAP

In our published study [8], single rank REMAP outperformed state-of-the-art methods: a chemical similarity-based method (PRW [17]), the best performed matrix factorization methods so far (NRLMF [31] and KBMF with twin kernels (KBMF2K) [32]), combination of WNN and GIP (WNNGIP [33]), and another type of collaborative filtering algorithm (Collaborative Matrix Factorization (CMF) [34]). Here we compare the performance of tREMAP with that of REMAP using two benchmarks. The first benchmark includes 3,494 chemicals, 25 G-protein coupled receptors (GPCRs), and 4,494 observed chemical-GPCR associations. The second benchmark includes 33,684 chemicals, 31 Cytochrome P450 enzymes (CYP450), and 51,699 observed chemical-CYP450 associations.

As shown in Fig. 3, tREMAP clearly outperforms REMAP when evaluated by both benchmarks. tREMAP identifies around 96 percent and 87 percent true associations ranked on the top 3 for GPCR and CYP450, respectively, while REMAP can only identify around 78 percent and 60 percent true hits ranked on top 3 respectively.

When evaluated by the application to sequence-structure similarity search, ENTS is superior to Hidden Markov Model and RWR [5].

5.2 Time Complexity of tREMAP

Empirically, the running time of tREMAP is linearly dependent on the number of chemicals and genes, as shown in Fig. 4. When evaluated in a machine with 2 cores of

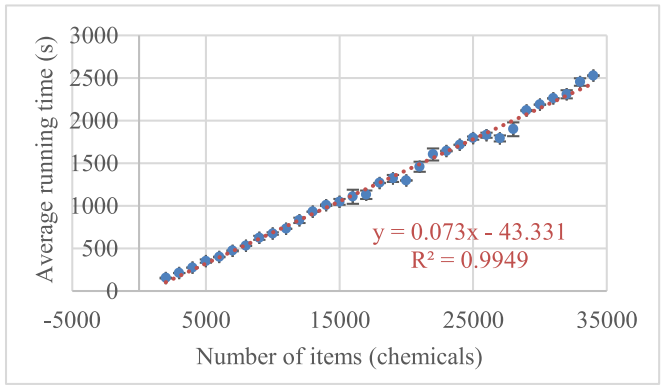


Fig. 4. Running time of tREMAP vs the number of items. The computational time was measured using two cores of 2.18 GHz CPU, for a matrix with 200 genes and varied number of chemicals. The ranks for chemical and gene are fixed as 1,000 and 200, respectively

2.18 GHz CPU. It takes around 1,000 seconds for a matrix with 15,000 chemicals, 200 genes, chemical-side rank of 1,000, and gene-side rank of 200 to converge.

5.3 ANTENNA Predictions

By combining tREMAP with ENTS, ANTENNA predicted that 21,921 novel drug-disease associations with Benjamini-Hochberg adjusted false discovery rate (FDR) less than 0.02. We selected a drug-disease pair for further experimental evaluation based on the following criteria. First, the drug was predicted to bind kinases, as the genome-wide binding assay for kinases is accessible. Second, the associated disease does not have effective therapy, so that the repurposed drug will have the biggest clinical impact. Third, the cell-based disease model is available, so that we can evaluate the efficacy of the drug.

Based on above criteria, diazoxide, a safe FDA-approved drug for hypertension, was selected. Diazoxide was predicted to interact with protein kinases. Furthermore, ANTENNA predicted that diazoxide was associated with Triple Negative Breast Cancer (TNBC) with Benjamini-Hochber adjusted false discovery rate (FDR) of 0.0108. Thus, diazoxide may be repurposed for the treatment of TBNC which is the most aggressive type of breast cancer and cannot be treated by any existing targeted therapy. It notes that the FDR of predicted diazoxide-TNBC association is not particular statistically significant. If this prediction is experimentally validated, we will have more confidence in predictions with lower FDRs.

5.4 Kinase Binding Assay

We validated the binding of diazoxide to kinases using KinomeScan assay. Fig. 5 displays the binding profile of

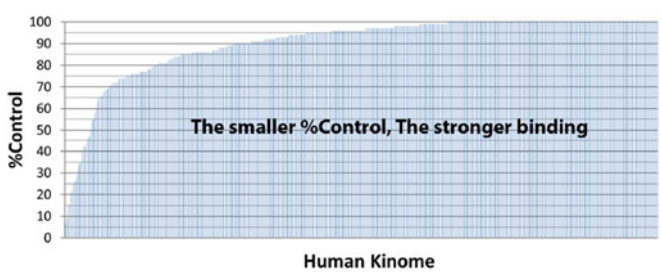


Fig. 5. Binding profile of FDA-approved drug diazoxide (100 μM) on 438 kinases determined by KinomeScan™ assay.

TABLE 1
Gene-Disease Associations of Three Kinases Having Highest Inhibition Percentage by Diazoxide

Kinase	KinomeScan %Control	Gene-Disease Association
DYRK1A	7.0	Multiple cancer drug-resistance, Alzheimer's disease
IRAK1	8.9	Breast cancer metastasis, herpesvirus lymphoma, Alzheimer's disease
TTK	15	TNBC, Hepatocellular Carcinoma

diazoxide across 438 kinases (kinome). Diazoxide has the highest percentage inhibition of kinases DRYK1A, IRAK1, and TTK with 7.0 percent, 8.9 percent, and 15.0 percent control. It is noted that the lower %Control, the higher inhibition of kinase activity.

As shown in Table 1, the malfunction of DYRK1A, IRAK1, and TTK is associated with multiple diseases, especially cancers and Alzheimer's disease. To verify our predictions, we tested the effect of diazoxide on breast cancer cells.

5.5 Cancer Cell Viability Assay

The cytotoxicity of diazoxide was determined by neutral red cell viability assay. The IC_{50} values obtained from Estrogen positive breast cancer MCF-7 cells and TNBC MDA-MB-468 cells treated with chemicals for 24 hours were shown in Table 2. Diazoxide was much more effective in inhibiting the cell proliferation of TNBC cancer MDA-MB 468 cells as compared to MCF-7 breast cancer cells with the values of $IC_{50} 0.87 \pm 0.39\,\mu M$ and $130.0 \pm 70.0\,\mu M$, respectively. The IC_{50} is the concentration of diazoxide that inhibits the cell proliferation of 50 percent cancer cells. The smaller the IC_{50} value is, the stronger anti-cancer activity diazoxide has. It is accepted that a chemical compound is active when the IC_{50} is less than $10\,\mu M$. Thus, diazoxide could be a highly effective targeted therapy for the treatment of TNBC at a low concentration.

6 Conclusions

In summary, we have developed a reliable and accurate multi-rank, multi-layered recommender system ANTENNA. Using ANTENNA, we predicted that FDA-approved safe medicine diazoxide could bind to kinases whose malfunction is associated with TNBC. KinomeScanTM assay confirmed the kinase binding of diazoxide. Cancer cell viability assay further validated that diazoxide is highly effective in inhibiting the proliferation of TNBC cancer cells. These findings suggest that diazoxide can be repurposed as an effective targeted therapy for the treatment of TNBC. Furthermore, diazoxide may be effective in the treatment of other diseases such as hepatocellular carcinoma and Alzheimer's disease. We are carrying out experiments to verify these predictions. This study demonstrates that big data analytics provides new opportunities for accelerating drug discovery and development, and realizing the full potential of precision medicines.

TABLE 2
IC_{50} Values of Diazoxide on Cancer Cells

Cell line	IC_{50} (Mean $\pm$ SEM)
MCF-7 (ER positive)	$130.0 \pm 70.0\,\mu M$
MDA-MB-468 (TNBC)	$0.87 \pm 0.39\,\mu M$

Acknowledgments

This work was partly supported by Grant Number R01LM011986 from the National Library of Medicine (NLM) of the National Institute of Health (NIH), Grant Number R01GM122845 from the National Institute of General Medical Sciences (NIGMS) of the National Institute of Health (NIH), Grant Number R21TR001722 from the National Center for Advancing Translational Sciences of NIH, and Grant Number MD007599 from the National Institute on Minority Health and Health Disparities (NIMHD) of NIH. A. Wang and H. Lim contributed equally to this work.

References

[1] G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason, and A. L. Hopkins, "Global mapping of pharmacological space," *Nat. Biotechnol.*, vol. 24, no. 7, pp. 805–815, Jul., 2006.

[2] A. L. Hopkins, "Network pharmacology," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1110–1111, 2007.

[3] A. L. Hopkins, "Network pharmacology: The next paradigm in drug discovery," *Nat. Chem. Biol.*, vol. 4, no. 11, pp. 682–690, Nov., 2008.

[4] C. Chen, H. Tong, L. Xie, L. Ying, and Q. He, "FASCINATE: Fast cross-layer dependency inference on multi-layered networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, California, USA, 2016, pp. 765–774.

[5] J. Lhota, R. Hauptman, T. Hart, C. Ng, and L. Xie, "A new method to improve network topological similarity search: Applied to fold recognition," *Bioinf.*, vol. 31, no. 13, pp. 2106–2114, Jul. 1, 2015.

[6] A. R. Maia, J. de Man, U. Boon, A. Janssen, J. Y. Song, M. Omerzu, J. G. Sterrenburg, M. B. Prinsen, N. Willemsen-Seegers, J. A. de Roos, A. M. van Doornmalen, J. C. Uitdehaag, G. J. Kops, J. Jonkers, R. C. Buijsman, G. J. Zaman, and R. H. Medema, "Inhibition of the spindle assembly checkpoint kinase TTK enhances the efficacy of docetaxel in a triple-negative breast cancer model," *Ann. Oncol.*, vol. 26, no. 10, pp. 2180–2192, Oct., 2015.

[7] V. Maire, C. Baldeyron, M. Richardson, B. Tesson, A. Vincent-Salomon, E. Gravier, B. Marty-Prouvost, L. De Koning, G. Rigaill, A. Dumont, D. Gentien, E. Barillot, S. Roman-Roman, S. Depil, F. Cruzalegui, A. Pierre, G. C. Tucker, and T. Dubois, "TTK/hMPS1 is an attractive therapeutic target for triple-negative breast cancer," *PLoS One*, vol. 8, no. 5, 2013, Art. no. e63712.

[8] H. Lim, A. Poleksic, Y. Yao, H. Tong, D. He, L. Zhuang, P. Meng, and L. Xie, "Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing," *PLoS Comput. Biol.*, vol. 12, no. 10, Oct., 2016, Art. no. e1005135.

[9] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: A knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D901–D906, Jan., 2008.

[10] J. J. Irwin and B. K. Shoichet, "ZINC–a free database of commercially available compounds for virtual screening," *J. Chem. Inf. Model*, vol. 45, no. 1, pp. 177–182, Jan.-Feb., 2005.

[11] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, "ChEMBL: A large-scale bioactivity database for drug discovery," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1100–D1107, Jan., 2012.

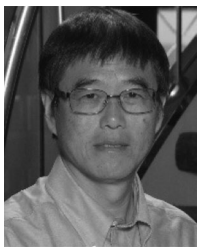
[12] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database's 10th year anniversary: update 2015," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D914–D920, Jan., 2015.

[13] A. Ma'ayan, A. D. Rouillard, N. R. Clark, Z. Wang, Q. Duan, and Y. Kou, "Lean big data integration in systems biology and systems pharmacology," *Trends Pharmacol. Sci.*, vol. 35, no. 9, pp. 450–460, Sep., 2014.
[14] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: A free tool to discover chemistry for biology," *J. Chemical Inform. Model.*, vol. 52, no. 7, pp. 1757–1768, 2012.
[15] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, and S. McGlinchey, "The ChEMBL bioactivity database: An update," *Nucleic Acids Res.*, 2013, Art. no. gkt1031.
[16] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *J. Cheminformatics*, vol. 7, no. 1, pp. 1–13, 2015.
[17] A. Koutsoukas, R. Lowe, Y. KalantarMotamedi, H. Y. Mussa, W. Klaffke, J. B. Mitchell, R. C. Glen, and A. Bender, "In silico target predictions: Defining a benchmarking data set and comparison of performance of the multiclass naïve bayes and parzen-rosenblatt window," *J. Chemical Inform. Model.*, vol. 53, no. 8, pp. 1957–1966, 2013.
[18] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: Architecture and applications," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. 1.
[19] S. Dider, J. Ji, Z. Zhao, and L. Xie, "Molecular mechanisms involved in the side effects of fatty acid amide hydrolase inhibitors: A structural phenomics approach to proteome-wide cellular off-target deconvolution and disease association," *NPJ Syst. Biol. Appl.*, vol. 2, 2016, Art. no. 16023.
[20] Y. Yao, H. Tong, G. Yan, F. Xu, X. Zhang, B. K. Szymanski, and J. Lu, "Dual-regularized one-class collaborative filtering," *Proc. 23rd ACM Int. Conf. Inform. Knowl. Manag*, 2014, pp. 759–768.
[21] V. Estivill-Castro, "Why so many clustering algorithms—A position paper," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 1, pp. 65–75, 2002.
[22] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. Royal Statistical Society Series C*, vol. 28, no. 1, pp. 100–108, 1979.
[23] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 24, no. 5, pp. 603–619, 2002.
[24] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Sci.*, vol. 315, no. 5814, pp. 972–976, Feb. 16, 2007.
[25] M. J. Brusco and H. F. Kohn, "Comment on "Clustering by passing messages between data points," *Sci.*, vol. 319, no. 5864, Feb. 8, 2008, Art. no. 726; author reply 726.
[26] H. Tong and C. Faloutsos, "Center-piece subgraphs: Problem definition and fast solutions," *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 404–413
[27] I. Melvin, J. Weston, C. Leslie, and W. S. Noble, "RANKPROP: A web server for protein remote homology detection," *Bioinform.*, vol. 25, no. 1, pp. 121–122, Jan. 1, 2009.
[28] M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, and P. Ahlquist, "Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis," *Ann. Appl. Stat.*, vol. 1, no. 1, pp. 85–106, 2007.
[29] E. Borenfreund and J. A. Puerner, "Toxicity determined in vitro by morphological alterations and neutral red absorption," *Toxicol. Lett.*, vol. 24, no. 2-3, pp. 119–124, Feb.-Mar., 1985.
[30] J. Galvao, B. Davis, M. Tilley, E. Normando, M. R. Duchen, and M. F. Cordeiro, "Unexpected low-dose toxicity of the universal solvent DMSO," *FASEB J.*, vol. 28, no. 3, pp. 1317–1330, Mar., 2014.
[31] Y. Liu, M. Wu, C. Miao, P. Zhao, and X.-L. Li, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS Comput. Biol.*, vol. 12, no. 2, 2016, Art. no. e1004760.
[32] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinf.*, vol. 28, no. 18, pp. 2304–2310, 2012.
[33] T. van Laarhoven and E. Marchiori, "Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile," *PloS One*, vol. 8, no. 6, 2013, Art. no. e66952.
[34] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1025–1033.

Annie Wang is currently a senior with the Bronx High School of Science, New York, NY. Her current research interests include drug development, computational biology, and machine learning.

Hansaim Lim received the BA degree in chemistry from the Hunter College of the City University of New York, NY, in 2014. Since 2015, he has been working toward the Ph.D. degree in biochemistry at the Graduate Center of the City University of New York, NY. He joined Dr. Lei Xie's Laboratory, Hunter College, for his thesis project. His research interest focuses on machine learning-based drug activity prediction.

Shu-Yuan Cheng received the MS and PhD degrees in toxicology from St. John's University, Jamaica, NY, in 1996 and 2003, respectively. She has worked in the John Jay College of Criminal Justice, the City University of New York, NY, since 2008, and is currently an associate professor in toxicology. She is a member of the Society of Toxicology and the Society for Neuroscience. She is also a member of the editorial board of the *Journal of Cell Science and Apoptosis*. She has received grants from NSF (RUI), NIH (SCORE), and DOJ to support her research. She has authored or coauthored 20 scientific papers in the field of toxicology, neuroscience, forensic toxicology, biochemistry, and cancer research. Her research interests include the pathogenesis study of neurodegeneration, the epidemiology study of abused drugs in wastewater in NYC, the pharmacological mechanism study of mitomycin C and its analog, and the cytotoxicity study of the potential anticancer drugs.



Lei Xie received the BS degree in polymer physics from the University of Science and Technology of China, China, in 1990, and the MSc degree in computer science and the PhD degree in chemistry from Rutgers University, in 2000. He was an associate scientist with Columbia University and the Howard Hughes Medical Institute. He has worked in pharmaceutical and biotechnology companies Roche and Eidogen for several years. He was a principal scientist with the San Diego Supercomputer Center from 2006 to 2011. He is currently an associate professor with the Department of Computer Science, Hunter College, and The Graduate Center, The City University of New York, NY. His research interests include data mining, machine learning, biophysics, systems biology, and drug discovery with more than 50 technical publications.