JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

GESSE: Predicting Drug Side Effects from Drug−Target Relationships

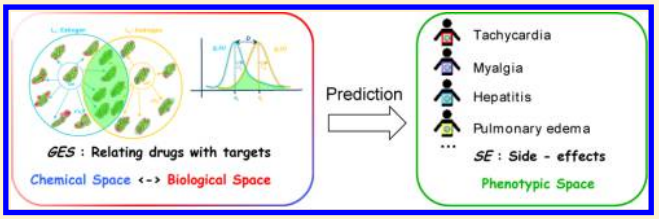
Violeta I. Pérez-Nueno,*[,†] Michel Souchet,[†] Arnaud S. Karaboga,[†] and David W. Ritchie[‡]

[†]Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique, 54600 Villers-les-Nancy, France

[‡]INRIA Nancy − Grand Est, Equipe Capsid, 615 rue du Jardin Botanique, 54600 Villers-les-Nancy, France

Supporting Information

ABSTRACT: The in silico prediction of unwanted side effects (SEs) caused by the promiscuous behavior of drugs and their targets is highly relevant to the pharmaceutical industry. Considerable effort is now being put into computational and experimental screening of several suspected off-target proteins in the hope that SEs might be identified early, before the cost associated with developing a drug candidate rises steeply. Following this need, we present a new method called GESSE to predict potential SEs of drugs from their physicochemical properties (three-dimensional shape plus chemistry) and to target protein data extracted from predicted drug−target relationships. The GESSE approach uses a canonical correlation analysis of the full drug−target and drug−SE matrices, and it then calculates a probability that each drug in the resulting drug−target matrix will have a given SE using a Bayesian discriminant analysis (DA) technique. The performance of GESSE is quantified using retrospective (external database) analysis and literature examples by means of area under the ROC curve analysis, "top hit rates", misclassification rates, and a χ^2 independence test. Overall, the robust and very promising retrospective statistics obtained and the many SE predictions that have experimental corroboration demonstrate that GESSE can successfully predict potential drug−SE profiles of candidate drug compounds from their predicted drug−target relationships.



INTRODUCTION

The in silico prediction of unwanted side effects (SEs) caused by the promiscuous behavior of drugs and their targets is highly relevant to the pharmaceutical industry. A particular challenge is to achieve an appropriate balance between drug efficacy and possible adverse effects as early as possible in order to reduce the possibility that safety issues might appear during clinical trials or even after a drug has reached the market.

Considerable effort is now being put into computational and experimental screening of suspected off-target proteins in the hope that SEs might be identified early, before the cost associated with developing a drug candidate rises steeply. In this regard, Bowes et al.[1] described rational strategies and methodologies for in vitro pharmacological profiling at four major pharmaceutical companies (AstraZeneca, GlaxoSmithKline, Novartis, and Pfizer). They shared their knowledge and experience of the use of existing screening technologies to detect off-target interactions of compounds as well as to define a minimum panel of targets that should be considered. These companies have been screening compounds for up to 10 years and have generated robust data to create this panel. Here, the term "robust" is understood to mean that all of the assays (binding or functional) included in the early profiling panel produce reliable and reproducible results and that they have predictive value for safety. According to Bowes et al., although in vitro pharmacological profiling can be used to predict SEs, detecting drug SEs experimentally remains challenging and costly.[1,2] Therefore, in silico prediction of SEs early in the drug discovery process promises to complement and speed up (or even perhaps avoid) the long and expensive process of in vitro safety profiling.

In regard to in silico methods, several computational approaches have been developed recently to identify possible SEs and to use SEs to predict drug−target relationships. They can be classified as pathway-based or chemical-based. Pathway-based approaches deal with molecular network information such as the proteins targeted by a given drug, gene−disease−drug connections, drug−drug interactions, and clinically known SEs combined with known drug−disease relationships. They build different pharmacological networks and then train models on them in order to predict adverse drug reactions for unknown drug−SE associations. For example, Campillos et al.[3] used phenotypic SE similarities to infer whether two drugs share a target, enabling new targets for known drugs to be found from drugs with similar SEs. Lee et al.[4] proposed a process−drug−SE network to automatically discover the relationship between biological processes and SEs using a co-occurrence-based multilevel network. Cheng et al.[5] developed a drug−SE similarity inferencing method to predict drug−target interactions (DTIs) from a known DTI network of approved drugs and target proteins. Yang and Agarwal[6] proposed a drug repositioning approach based on associations between diseases and SEs. They built naïve Bayes models using SEs as features in order to predict indications for diseases. Scheiber et al.,[7] Xie et al.,[8] Wallach et al.,[9] and Takarabe et al.[10,11] also proposed ways to link drug SEs and biological pathways. Scheiber et al. compared biological pathways affected by toxic compounds and those affected by nontoxic compounds. Xie et al. and Wallach et al. predicted potential SEs by

Received: March 5, 2015
Published: August 7, 2015

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

docking drugs into protein binding pockets similar to those of their primary targets and then mapping the proteins with the best docking scores to known biological pathways. Takarabe et al. defined pharmacological similarity for all possible drugs using the U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) and developed a method to predict unknown drug−target interactions on a large scale by relating the pharmacological similarities of drugs to the genomic sequence similarities of target proteins.

On the other hand, compound-based approaches relate the chemical structures of drug molecules with drug SEs. The basic idea is that similar ligands are likely to interact with similar proteins, allowing predictions to be made by comparing drug chemical structures, protein sequences, and known drug−protein interactions. For example, Scheiber et al.[12] developed a method that associates chemical substructures with SEs. Simon et al.[13] related drug−protein interaction profiles (binding free energies computed from docking) with SE profiles compiled from the literature using canonical correlation analysis (CCA) followed by linear discriminant analysis (LDA). Yamanishi et al.[14] proposed a method to predict the pharmacological effects of drugs using their chemical structures in order to interpret drug−target interactions as well as a method to predict potential SE profiles of drug candidate molecules by correlating sets of chemical substructures and SEs using Kernel regression models.[15,16] In a similar way, Pauwels et al.,[17] Mizutani et al.,[18] and Atias and Sharan[19] predicted potential SE profiles by correlating sets of chemical substructures and SEs using CCA.

Several databases are available for studying relationships between drugs, targets, biological processes, and SEs. For example, Lamb et al.[20] developed a connectivity map approach to create and analyze a drug−gene−disease network from large-scale experimental gene expression responses to drugs, data from the SIDER side-effects resource[21] (which associates drugs with their observed adverse drug reactions), PharmGKB[22] (which provides drug−disease associations), and the KEGG database[23] (which maps proteins to biological pathways). Concerning terminology, the majority of in silico approaches use gene ontology (GO) terms[24] to describe biological processes and medical subject headings (MeSHs) or Unified Medical Language System (UMLS) vocabularies[25] to describe diseases.

Unfortunately, biology and chemistry are often considered separately, and this can lead to incomplete models that do not provide a unified view of SEs. Most in silico methods for predicting SEs focus on the use of information about only protein targets or only drug chemical structures. However, for the prediction of drug SEs, it intuitively seems more desirable to consider target protein and chemical structure information simultaneously. In this regard, Duran-Frigola and Aloy[26] analyzed both approaches by investigating the molecular bases of over 1600 SEs. They gave mechanistic explanations for most of the SEs and emphasized the need to combine biology and chemistry to capture complex phenomena not covered in the molecular biology view. Yamanishi et al.[15] developed an integrated framework to predict potential SEs of drugs from their chemical structures and target protein information on a large scale. With a view to joining biology and chemistry, Mizutani et al.[27] used FAERS, which provides a valuable resource for pharmacoepidemiology (the study of the uses and effects of drugs in human populations). They used a biclustering approach to calculate the relationships between drugs and adverse reactions from a large FAERS data set and demonstrated a systematic way to find cases where different drug administration regimes resulted in similar adverse reactions and where the same drug could cause different reactions in different patients.

In the present work, we also aimed to integrate the chemical space of drug structures and the biological space of drug target proteins. We previously introduced the "Gaussian ensemble screening" (GES) approach for rapid and reliable quantitative prediction of polypharmacological relationships between drug classes[28] and the GES "computational polypharmacology fingerprint" (CPF) for encoding drug promiscuity information.[29] In this paper, we present a new method, called GESSE, to predict potential SEs of drugs from their physicochemical properties (i.e., three-dimensional (3D) shape plus chemistry) and target protein data extracted from GES-predicted drug−target relationships.

To our knowledge, no other computational method has been reported for predicting drug SEs by associating SEs with predicted drug−target relationships using each drug's physicochemical properties. The two most similar previous approaches are those of Pauwels et al.,[17] who predicted drug SEs by associating SEs with the presence of certain chemical substructures, and Simon et al.,[13] who predicted drug SEs by associating SEs with predicted drug−target interactions according to calculated binding free energies.

To demonstrate the usefulness of our approach, we predicted the SEs for a set of DrugBank[30] drugs for which drug−target relationships had been calculated for 777 targets using GES. The performance of the approach was quantified using retrospective analysis and literature examples. Overall, the robust and very promising retrospective statistics obtained and the many SEs predictions having experimental corroboration demonstrate that GESSE can successfully predict drug−SE profiles of candidate drug compounds from predicted drug−target relationships.

METHODS

SE Profile Matrix. To build a SE profile matrix, or "*SE matrix*", we used our *NetworkDB* relational database[31] containing 554 DrugBank drugs linked to 1077 SEs. This database integrates data about drugs and their targets from several data sources (including DrugBank, UniProt, KEGG, and GO) with their related SEs compiled from SIDER version 2.[21] Terms describing individual side effects reported in SIDER are clustered into 112 term clusters (TCs) using an expert-validated[32] semantic similarity measure derived from MedDRA.[33]

In fact, two SE matrices were built, which we call "*NetworkDB_TC*" and "*NetworkDB_SE*" for brevity. In *NetworkDB_TC*, each drug is represented by a binary profile whose 112 elements encode the presence or absence of a TC, thus giving a SE profile matrix of 554 × 112 elements. Similarly, in *NetworkDB_SE* each drug is represented by a vector of 1077 elements encoding the presence or absence of an individual SE for each drug, giving a matrix of 554 × 1077 elements.

GES Drug−Target Relationship Matrix. In the GES approach, a "ligand set" is defined as a cluster of high-affinity ligands that bind to a specific target. The main novelty of GES is to represent such a cluster as a Gaussian distribution with respect to a selected center molecule (CM).[28] Through the use of spherical harmonic (SH) surface shapes, it is straightforward to calculate the CM of a ligand set. However, because Gaussian functions require a distance coordinate rather than a similarity score, we calculate the normalized SH distance ($0.0 \leq x \leq 1.0$) between the CM and each cluster member using the assumption that it is valid to let Distance = 1 − Similarity.[28]

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

We extracted from the DrugBank database 6353 drug entries that are in clinical trials or are on the market, and the SH shape-plus-chemistry representations of these drugs were calculated using the Harmonic Pharma chemistry coefficient (HPCC).[34] These drugs were grouped into 778 ligand sets according to the targets to which they bind. Ten conformations for each molecule were computed, and the HPCC representation was calculated for each conformation. We then used the CAST clustering algorithm[35] to cluster the members of each ligand set using a PARAFIT Tanimoto similarity score of 0.65 as described previously.[28] This gave 777 ligand-set shape clusters. Thus, all of the ligand sets remained unsplit except for one that had substantially different drug scaffolds. We then calculated the CM for each cluster to obtain our simplified CM representation of each ligand set.

A matrix of HPCC similarity scores between the 554 drugs of *NetworkDB* and the CMs of the 777 target ligand sets were calculated using the Gaussian overlap score as described previously.[28,29] The results from the all-versus-all comparisons were recorded as a matrix of GES p values, which we call the "GES ligand−target relationship matrix", or more simply the "*GES matrix*".

Canonical Correlation Analysis. In canonical correlation analysis (CCA), two sets of variables are studied, and a new set of "canonical variables" that are as far as possible correlated with each set and orthogonal to each other is extracted.[36] CCA has been used previously to predict drug−SE profiles using different input matrices, such as chemical substructure profiles[17,19] or binding free energies calculated from ligand−protein docking.[13] Our objective here is to extract a group of canonical variables that capture common features from our two sets of variables: one containing information about drug−target relationships (the *GES matrix*) and the other containing drug side effects (the *SE matrix*). In order to try to predict SEs as reliably as possible, we explore three variations of CCA, namely, ordinary canonical correlation analysis (OCCA), sparse canonical correlation analysis (SCCA), and regularized canonical correlation analysis (RCCA).

Ordinary Canonical Correlation Analysis. OCCA aims to find linear combinations of the variables in a vector of features $\boldsymbol{x}$ that correlate maximally with linear combinations of the variables in some other feature vector, $\boldsymbol{y}$. These linear combinations are the so-called *canonical variables* $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, and are given by $\boldsymbol{u} = \boldsymbol{\alpha x}$ and $\boldsymbol{v} = \boldsymbol{\beta y}$, in which $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)$ are the weight vectors. For our specific case with n drugs, we consider an $n \times p$ matrix $\boldsymbol{X}$ containing p target relationship variables and an $n \times q$ matrix $\boldsymbol{Y}$ containing q SE variables. Thus, each drug is represented by a drug−target relationship vector $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)^T$ and a side-effect feature vector $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_q)^T$. If we now consider linear combinations of drug−target relationships and of drug SEs for the kth drug, $\boldsymbol{u}_k = \boldsymbol{\alpha}^T\boldsymbol{x}_k$ and $\boldsymbol{v}_k = \boldsymbol{\beta}^T\boldsymbol{y}_k$, then the vectors $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the *canonical components* of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$, respectively. The goal of ordinary CCA is to find the optimal weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (or *canonical coefficients*) by maximizing the correlation between the canonical variable pairs $(\boldsymbol{u}_k, \boldsymbol{v}_k)$, which is also known as the canonical correlation:

$$\rho = \mathrm{corr}(u, v) = \frac{\sum_{i=1}^{n} \alpha^T x_i \cdot \beta^T y_i}{\sqrt{\sum_{i=1}^{n} (\alpha^T x_i)^2}\sqrt{\sum_{i=1}^{n} (\beta^T y_i)^2}} \quad (1)$$

where

$$\sum_{i=1}^{n} u_i = 0, \qquad \sum_{i=1}^{n} v_i = 0$$

The residuals of pairs of variables $(\boldsymbol{u}_k, \boldsymbol{v}_k)$ are analyzed progressively in order to find the weights that maximize the correlation. This process then continues until a "significance" cutoff is reached or the maximum number of pairs (which equals the smaller of q and p) has been found. As each pair of canonical variables is calculated from the residuals of the preceding pair(s), the resulting canonical variables are orthogonal. Because the change in the canonical correlation decreases with the number of variable pairs, it is important to choose a good dimension (number of variables) for good predictive performance. OCCA is performed in GESSE using the R package PMA.[37]

Sparse Canonical Correlation Analysis. One drawback of OCCA is that it can be difficult to interpret the results when there are many nonzero elements in the weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Consequently, SCCA aims to change small weights into zeros for easier interpretation. More specifically, given the above matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, SCCA maximizes the product $\boldsymbol{u}'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{v}$ subject to constraints on $\boldsymbol{u}$ and $\boldsymbol{v}$ (penalty_x and penalty_y). These penalty terms are parameters to control the sparsity and are restricted to the ranges $0 < \text{penalty\_}x \leq 1$ and $0 < \text{penalty\_}y \leq 1$. For simplicity, we use the same value, c, for the two terms penalty_x and penalty_y. In GESSE, SCCA is also calculated using the R package PMA.[37]

Regularized Canonical Correlation Analysis. RCCA is an improved version of CCA that aims to prevent overfitting when there are insufficient training data. RCCA seeks a correlation between two data matrices when the number of columns (variables) exceeds the number of rows (observations). Let p and q denote the number of features in X and Y, and let n be the sample size. When $n < p$ or $n < q$, the features in X and Y tend to be highly colinear. This leads to ill-conditioned covariance matrices C_{xx} and C_{yy} (which denote the covariance matrix of X with itself and that of Y with itself, respectively), such that their inverses are no longer reliable, thus resulting in an invalid computation of the CCA. In RCCA, the condition placed on the data to guarantee that C_{xx} and C_{yy} will be invertible is $n \geq p + q + 1$. However, this condition is usually not met in domains where the number of samples (n) is limited and the numbers of features (p and q) are large. Consequently, in RCCA, small positive quantities (λ_1 and λ_2) are added to the diagonals of C_{xx} and C_{yy} to guarantee their invertibility. RCCA is performed in GESSE using the R package mixOmics.[38]

Scoring Prediction for a New Drug from CCA. Following the above setting with n drugs, the $n \times p$ GES matrix $\boldsymbol{X}$ of p target relationship variables, and the $n \times q$ SE matrix $\boldsymbol{Y}$ of q SEs and given that we have used GES to predict the drug−target relationship profile $\boldsymbol{x}$ of a new drug candidate, we now want to predict its SE profile $\boldsymbol{y}$ from the extracted m canonical variables defined by the weight vectors $\{\alpha_k\}_{k=1}^{m}$ and $\{\beta_k\}_{k=1}^{m}$. Following Pauwels et al.,[17] we calculated three prediction scores, Score 1 (s_1) using singular values (eq 2), Score 2 (s_2) using correlation values (eq 3), and Score 3 (s_3) using the pseudoinverse (eq 4), and we analyzed their SE prediction performance.

As before, the $\boldsymbol{x}$ and $\boldsymbol{y}$ vectors may be expressed as linear combinations of the canonical variables using $\boldsymbol{u} = \boldsymbol{A}^T\boldsymbol{x}$ and $\boldsymbol{v} = \boldsymbol{B}^T\boldsymbol{y}$, respectively, where $\boldsymbol{A} = [\alpha_1, ..., \alpha_m]$ and $\boldsymbol{B} = [\beta_1, ..., \beta_m]$. Since $\boldsymbol{x}$ has been calculated by GES but $\boldsymbol{y}$ is unknown, we wish to estimate $\boldsymbol{y}$ such that $\boldsymbol{u}$ and $\boldsymbol{v}$ remain orthogonal. This can be done by minimizing $\|\boldsymbol{u} - \boldsymbol{v}\|_2^2 = \|\boldsymbol{A}^T\boldsymbol{x} - \boldsymbol{B}^T\boldsymbol{y}\|_2^2$, which leads to the following equations:[17]

$$s_1(x) = \sum_{k=1}^{m} \beta_k d_k \alpha_k^T x = \boldsymbol{B\Lambda A}^T\boldsymbol{x} \quad (2)$$

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

where $\mathbf{\Lambda}$ is the diagonal matrix whose elements are singular values d;

$$s_2(x) = \sum_{k=1}^{m} \beta_k \rho_k \alpha_k^{\mathrm{T}} x = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x} \tag{3}$$

where $\mathbf{\Lambda}$ is the diagonal matrix whose elements are canonical correlation coefficients ρ; and

$$s_3(x) = \boldsymbol{B}^{-\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x} \tag{4}$$

where $\boldsymbol{B}^{-\mathrm{T}}$ is the pseudoinverse matrix of $\boldsymbol{B}^{\mathrm{T}}$. It should be noted that $s_i(x)$ is the q-dimensional vector whose jth element represents a prediction score for the jth side effect. If the jth element in $s_i(x)$ has a high score, the new molecule x is predicted to have the jth side effect.

Canonical Correlation Analysis Followed by Discriminant Analysis. Another statistical method with good interpretability is discriminant analysis (DA) using Bayesian probabilities. In general, DA aims to predict the group membership for a number of subjects from a set of predictor variables. In our particular case, we have two groups (i.e., those that possess a given SE and those that do not) and p predictor variables (the canonical variables from the CCA) for each of a number of drugs. However, DA can be applied to only one response variable or CCA component at a time. Hence, we apply DA repeatedly for each SE using the CCA canonical variables.

Two models of DA are used according to whether or not the covariance matrices of X and Y are assumed to be equal. If equality is assumed, we use linear discriminant analysis (LDA). Otherwise, we use quadratic discriminant analysis (QDA). DA constructs one or more discriminant equations D_i (linear combinations of the predictor variables $\boldsymbol{x}_k$) such that the different groups differ as much as possible on D (eq 5).

$$D_i = b_0 + \sum_{k=1}^{p} b_k x_k \tag{5}$$

Letting SS denote the sum of squared differences from the mean, the weights of the discriminant function are calculated in such a way that the ratio (between-group SS)/(within-group SS) is as large as possible. The number of discriminant functions is equal to the smaller of p and $g - 1$, where g is the number of groups. Hence, the first discriminant function D_1 distinguishes the first group from groups 2, 3, ..., g, the second discriminant function D_2 distinguishes the second group from groups 3, 4, ..., g, and so on.

To calculate the optimal weights, a training set containing the correct classification for a group of subjects is used. To classify a new group of subjects for which we do not yet know the group members, we can use the previously calculated discriminant weights to obtain their discriminant scores. To do this, we first calculate the probability that a subject belongs to a certain group using the estimated discriminant model. This calculation assumes that a subject should be assigned to a group if it has the highest probability of belonging to that group. To classify future observations, a Bayesian approach is used, in which the posterior probability of group membership for each group k is computed as

$$p(\text{group} = k|x) \frac{\exp[-\frac{1}{2}D_k^{*2}(x)]}{\sum_{i=1}^{g} \exp[-\frac{1}{2}D_i^{*2}(x)]} \tag{6}$$

where g is the number of groups and D_i^{*2} is given by one of the following expressions:

for QDA with unequal priors:

$$D_i^{*2}(x) = (x - \bar{x}_i)' S_i^{-1} (x - \bar{x}_i) + \ln|S_i| - 2 \ln p_i$$

for QDA with equal priors:

$$D_i^{*2}(x) = (x - \bar{x}_i)' S_i^{-1} (x - \bar{x}_i) + \ln|S_i|$$

for LDA with unequal priors:

$$D_i^{*2}(x) = (x - \bar{x}_i)' S_i^{-1} (x - \bar{x}_i) - 2 \ln p_i$$

for LDA with equal priors:

$$D_i^{*2}(x) = (x - \bar{x}_i)' S_i^{-1} (x - \bar{x}_i)$$

in which p_i is the prior probability of being from population or group i and S_p is the pooled estimate of the common variance−covariance matrix in LDA, given by

$$S_p^{-1} = \frac{\sum_{i=1}^{g} (n_i - 1) S_i^{-1}}{\sum_{i=1}^{g} (n_i - 1)}$$

where n_i is the number of members of group i and S_i is the variance−covariance matrix for group i.

CCA Followed by Linear Discriminant Analysis. We perform the CCA between the *GES matrix* and each of the SEs. Because LDA deals with just one variable at a time, the CCA canonical variables are extracted in pairs, and the LDA is applied to each variable in turn. Then we perform the Kullback test for equality of covariance matrices[39] using the R package asbio.[40] In all cases, the Kullback test suggests that we can assume the equality of the covariance matrices, so we always apply LDA. LDA is performed in GESSE using the R package MASS.[41]

To predict the SEs of a new drug by LDA, we first predict the factor pairs by CCA for the *GES test matrix*, and we then predict the LDA probabilities using the discriminant function calculated for the training *GES matrix* and the factor pairs obtained for the *GES test matrix*.

CCA Followed by Partial Least-Squares Discriminant Analysis. Partial least-squares (PLS) regression is a statistical method that shares some similarity with principal component regression. Instead of finding hyperplanes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bilinear factor models. Partial least-squares discriminant analysis (PLS-DA) is a variant used when Y is categorical. PLS-DA is performed in GESSE using the R package DiscriMiner.[42]

On of the advantages of projection-based methods such as principal component analysis (PCA) and PLS-DA is that the decomposition of the data matrix into loading vectors and latent variables makes it easy to visualize the results. We perform PLS-DA using the factor pairs extracted from the *GES matrix* and each of the SEs in order to explore visually the possibility of cluster formations in the DA and to see how well clusters of SEs are distinguished. Similarly, we can visualize the PLS components obtained from PLS-DA to explore the difference between using term clusters from *NetworkDB_TC* and discrete independent SEs from *NetworkDB_SE*.

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

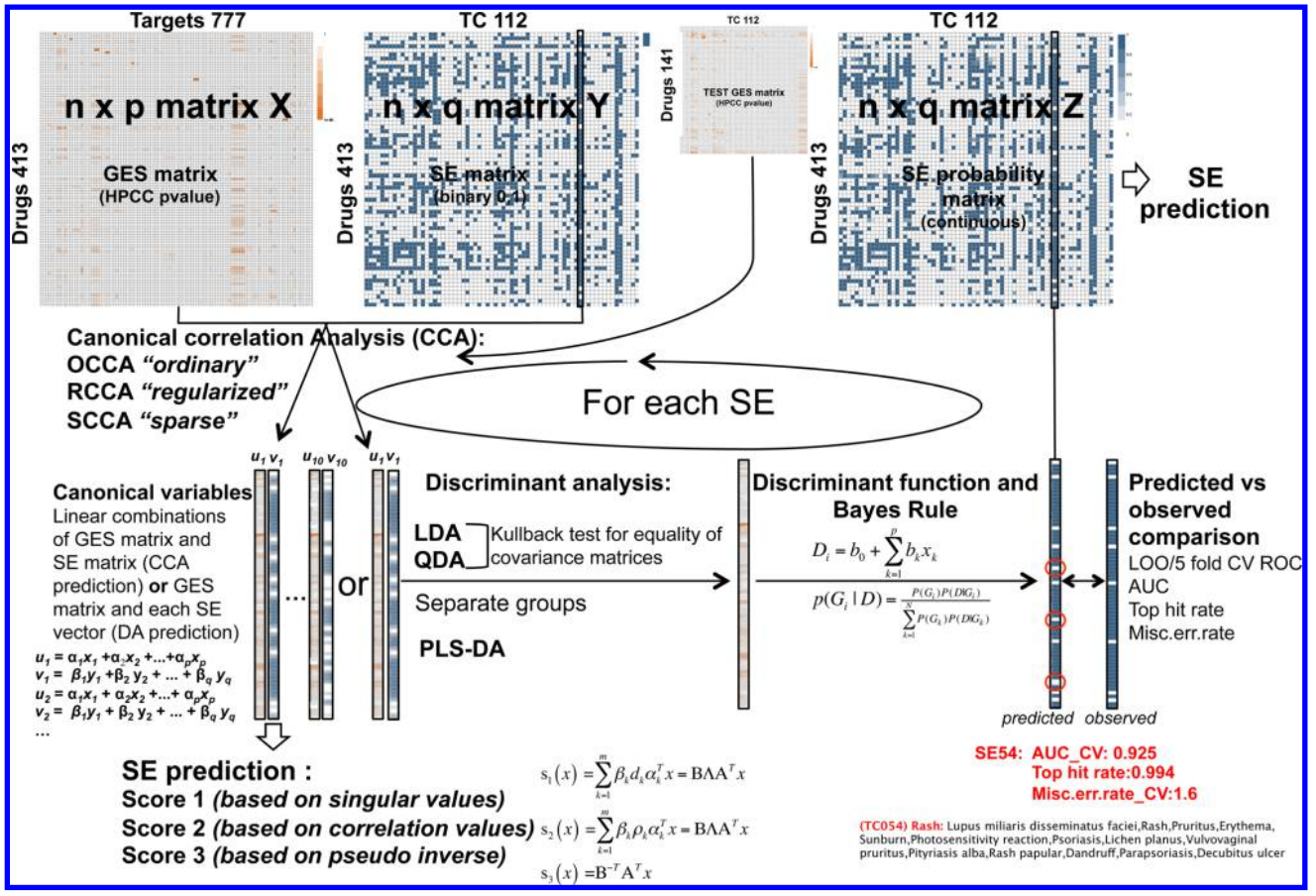


Figure 1. GESSE workflow.

Validation by Retrospective Analysis. The GESSE approach was validated using leave-one-out (LOO) cross-validation (CV) and 5-fold CV using a training set of 413 drugs and a test set of 141 drugs selected randomly from the 554 drugs stored in our *NetworkDB* database.

The utility of the classification functions was assessed by receiver operating characteristic (ROC) analysis, determining the true positive rate (TPR) and the false positive rate (FPR) of the predicted SEs and TCs. The misclassification rate and the χ^2 independence test were also calculated. Pearson's χ^2 test with Yates' continuity correction data was calculated using the R package MASS.[41] If the calculated p value is lower than the conventionally accepted significance level (i.e., $p < 0.05$), then we assume that there is no statistically significant difference between the observed and expected values. In the current setting, $p < 0.05$ indicates that a predicted SE is consistent with an SE known from the literature.

The top hit rate was also calculated. The entire set of 554 drugs was listed in descending order by the probability of possessing a given SE (or TC), and the top of the list was cut at the number of drugs known to possess the studied SE (or TC). This top list may contain known and unknown drugs that possess the given SE (or TC) because the unknown drugs can also be assigned a high probability and known drugs can have a low value. The classification accuracy can be characterized using the proportion of the known drugs that possess the SE (or TC) in the top list. Therefore, the following top hit rate was calculated for each of the 1077 SEs and 112 TCs:

$$\text{top hit rate} = \frac{N_{\text{top}}}{N_{\text{total}}} \quad (7)$$

where N_{top} is the number of known drugs that possess the SE (or TC) in the top list and N_{total} is the number of all known drugs that possess the SE (or TC).

Figure 1 shows a schematic representation of the full GESSE workflow.

Validation by Literature Examples: Prediction of SE Profiles for New Molecules. We compiled the SE data from the literature[15,43] for around 300 drugs of diverse molecular structure involved in 50 different SEs and grouped into 11 pathological manifestations, namely, "breathing", "dermal", "endocrine", "muscular−skeletal", "metabolic", "gastrointestinal", "cardiovascular", "hematological", "neurological", "systemic phenomena", and "psychiatric".

To assess the quality of our SE predictions, we made a prediction table. For the 554 drugs from *NetworkDB* and a further set of 235 DrugBank drugs not in this database, we compared the predicted SEs with known SEs from the literature. More specifically, we calculated the percentage well-predicted ("% WP") for both the SE and TC databases. In other words, for a given SE (or TC), the % WP was calculated as the percentage of the drugs correctly predicted to possess the given SE (or TC) with respect to the total number of known drugs that have that SE (or belong to the TC). We considered that a drug was predicted to have an SE if its CCA score was >0.4 and either its LDA probability was >0.8 or its PLS-DA probability was >0.8.

RESULTS AND DISCUSSION

Statistical Analyses. *Canonical Correlation Analysis.* We first performed CCA between the *GES matrix* and the *SE matrix* in order to correlate the two sets of variables and to extract from

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

them a set of canonical variables that are correlated as far as possible with both matrices and are also orthogonal to each other. In order to do this, we preprocessed the data by deleting any null columns or duplicates (this was necessary only for some SEs in *NetworkDB_SE*) and performed data normalization by centering and scaling the data to unit variance. Figure S1 in the Supporting Information shows the correlation within the *GES matrix* and the *SE matrix* and the cross-correlation between these two matrices before calculation of the CCA for both the 112 TC matrix (which we subsequently call the "TC dataset") and the 1077 independent SE matrix (which we subsequently call the "SE dataset"). It can be seen that the two matrices do not correlate before CCA analysis.

Because the canonical correlation decreases with the number of canonical variable pairs, it is important to choose a suitable number of canonical variables, or dimension (M), for reliable predictive performance. Figure 2 shows the canonical correlation as a function of M. We can see that after extraction of 10 canonical pairs, the correlation decreases significantly. Therefore, we extract from CCA only the 10 most representative canonical pairs. Figure S2 shows the canonical correlation between the first two canonical variables. It can be seen that the GES HPCC p value correlates well with the SEs for both the TC and SE data sets. Even with the SE data set, which has a considerably higher number of variables than the TC data set, the correlation is good except for some independent SEs that disperse (456, 342, 125, 221, 67). Figure S3 shows the canonical correlation between the first three canonical variables, which can be seen to be very similar to that in Figure S2.

The performances of OCCA, RCCA, and SCCA were analyzed according to the area under the ROC curve (AUC) obtained for 5-fold CV using the three aforementioned scores (s_1, s_2, s_3). It is important to optimize the λ_1 and λ_2 terms for a proper SE prediction using RCCA. Figure S4 shows a heat map with the optimized λ values by CV for both the TC data set (Figure S4A) and the SE data set (Figure S4B). The three different scores are calculated for the optimized λ values with $M = 10$. We can see that score S_2 using the canonical correlation coefficients gives the highest 5-fold CV AUC. Table S1 in the Supporting Information shows the influence of the dimension in RCCA using the optimized λ values. It can be seen that $M = 10$ gives the highest AUC values. In the same way, it is important to choose the penalty terms to control the sparsity for a good SE prediction in SCCA. Table S2 shows the influence of the sparsity parameter c in SCCA 5-fold CV for the three scores using $M = 10$. OCCA is the specific case when c = penalty_x= penalty_y = 1. Again, score S_2 gives the best ROC performance. Table S3 shows the influence of the dimension for various sparsity parameters for the same SCCA (OCCA when $c = 1$) calculation. It can be seen that $M = 10$ gives the highest 5-fold CV AUC. Figure 3 shows a summary of the optimized parameters that give the highest AUC for 5-fold CV for OCCA, RCCA, and SCCA for the TC and SE data sets. This figure also shows ROC plots for OCCA, RCCA, and SCCA for both *NetworkDB_TC* and *NetworkDB_SE*. It can be seen that RCCA is the best-performing method, followed by SCCA with very similar values. Therefore, the canonical variables calculated by RCCA were used as input for the subsequent DA analyses.

CCA results can vary significantly according to the number of factor pairs used, the SCCA sparsity parameter, and the scaling of the data sets. When the variables are clustered and scaled to unit variance, rare or less-frequent features are extracted. When no scaling is performed, more common features are extracted. This observation agrees with that of Pauwels et al.[17]

It is worth noting that a CCA model can be misinterpreted if the degree of sparsity is not tuned carefully in SCCA and the regularization parameters are not well optimized in RCCA. The optimal parameter values depend on the definition of the objective function to be investigated in the cross-validation. On this point, we can observe that the optimal sparsity parameter is different when *NetworkDB_TC* or *NetworkDB_SE* is used. When *NetworkDB_TC* is used (i.e., when the goal is global accuracy in the SCCA), the best AUC is obtained with a low sparsity parameter ($c = 0.03$ for $M = 10$). The canonical components are associated with very few GES ligand−target relationships and SEs. On the other hand, when *NetworkDB_SE* is used (i.e., when the SCCA goal is local accuracy (accuracy of individual SEs)), the best AUC is obtained with a larger sparsity parameter ($c = 0.06$ for $M = 10$). The canonical components are associated with a large number of GES ligand−target relationships and SEs.

On the other hand, the CCA results also depend on the definition of GES-related targets and the choice of SE keywords. Figure S7 shows the index plots of weight vectors for GES p values (left) and SEs (right) given by OCCA, RCC, and SCCA using both *NetworkDB_TC* (Figure S7A−C, respectively) and *NetworkDB_SE* (Figure S7D−F, respectively). It can be seen that in OCCA, almost all of the elements of the weight vectors are nonzero and highly variable. In RCCA, almost all of the elements of the weight vectors are around zero and have some variability. In SCCA, almost all of the elements in the weight vectors are zero in each component, selecting a small number of features as informative GES p values and SEs. This observation agrees with that of Pauwels et al.[17]

Finally, we also compared the performance of GESSE with that of the approach of Pauwels et al.,[17] the most similar existing technique that also predicts drug SE activity. In order to do the comparison, we produced the PubChem chemical substructure fingerprint for *NetworkDB* compounds in the same way as used by Pauwels et al. In other words, each drug was represented by a binary profile of 881 elements encoding the presence or absence of each PubChem[44] substructure by 1 or 0, respectively.

Table S4 and Figures S5 and S6 show the CCA-optimized parameters to achieve the maximum performance when PubChem substructure fingerprints of *NetworkDB* compounds are used as input. Comparison of the performances of the method of Pauwels et al. and GESSE (Figure 3 and Figure S6) shows that GESSE retrieves results very similar to those of Pauwels et al., although not superior. RCCA is the best-performing method in both cases, with AUCs of 0.82632 and 0.88543 for TCs and independent SEs, respectively, when using GES p values as input and 0.83076 and 0.88621, respectively, when using PubChem substructure fingerprints as input.

Canonical Correlation Analysis followed by Discriminant Analysis. Because DA can be applied to only one response variable at a time, here we applied it successively to each SE or TC in turn. The validation of CCA followed by DA (or "CCA/DA") was done by LOO CV and 5-fold CV using the same training and test sets for both *NetworkDB_TC* and *NetworkDB_SE* as described above. Figure 4 illustrates this process schematically for the SE TC054 (Rash) from *NetworkDB_TC*. It can be seen that the DA separates groups starting from RCCA canonical score X by means of the discriminant equation. The 554 drugs are discriminated according to whether or not they are associated with TC054. The folders LDA_Plots_TC and LDA_Plots_SE in the Supporting Information provide additional plots ("RCCA Canonical Score: Score X vs Score Y", "RCCA Score X vs Index", "LDA Histogram: group 0 and group 1 distributions", "LOO Cross−Validated LD

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

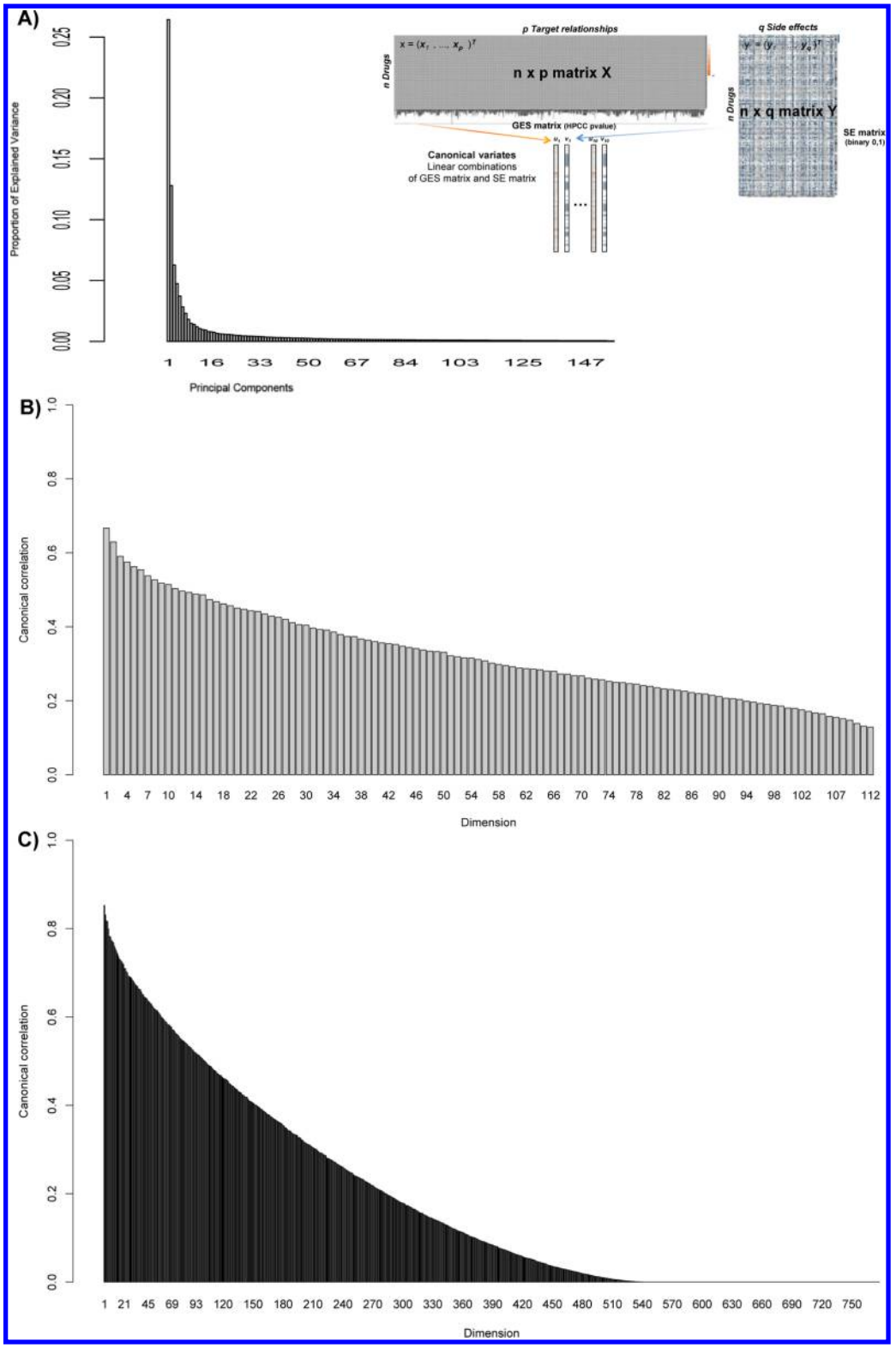


Figure 2. Choosing the CCA dimension. (A) Significance of the components showed in the PCA plot (proportion of explained variance vs principal components) for *NetworkDB*. Most of the data are explained with the first 10 principal components. The proportions of explained variance for the first 10 principal components (PC1−PC10) are 0.2644, 0.1279, 0.0627, 0.0475, 0.0372, 0.0284, 0.0231, 0.0180, 0.0149, and 0.0139, respectively. The cumulative proportions of explained variance for the first 10 principal components are 0.2644, 0.3923, 0.4551, 0.5026, 0.5398, 0.5683, 0.5914, 0.610, 0.6243, and 0.6382, respectively. (B, C) Plots of canonical correlation vs dimension for (B) *NetworkDB_TC* and (C) *NetworkDB_SE*. After component 10 the correlation decreases significantly. Hence, we extract from CCA 10 canonical pairs.

Classification: RCCA Score vs Index", and "ROC plot after LOO CV: sensitivity vs specificity") for all of the SEs in the *NetworkDB_TC* and *NetworkDB_SE* databases, respectively.

Table S5 shows a summary of the LOO CV statistics for CCA/LDA when applied to the *GES matrix* and the *NetworkDB_TC* database. The last column of this table ("Best Predicted 5fold")

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

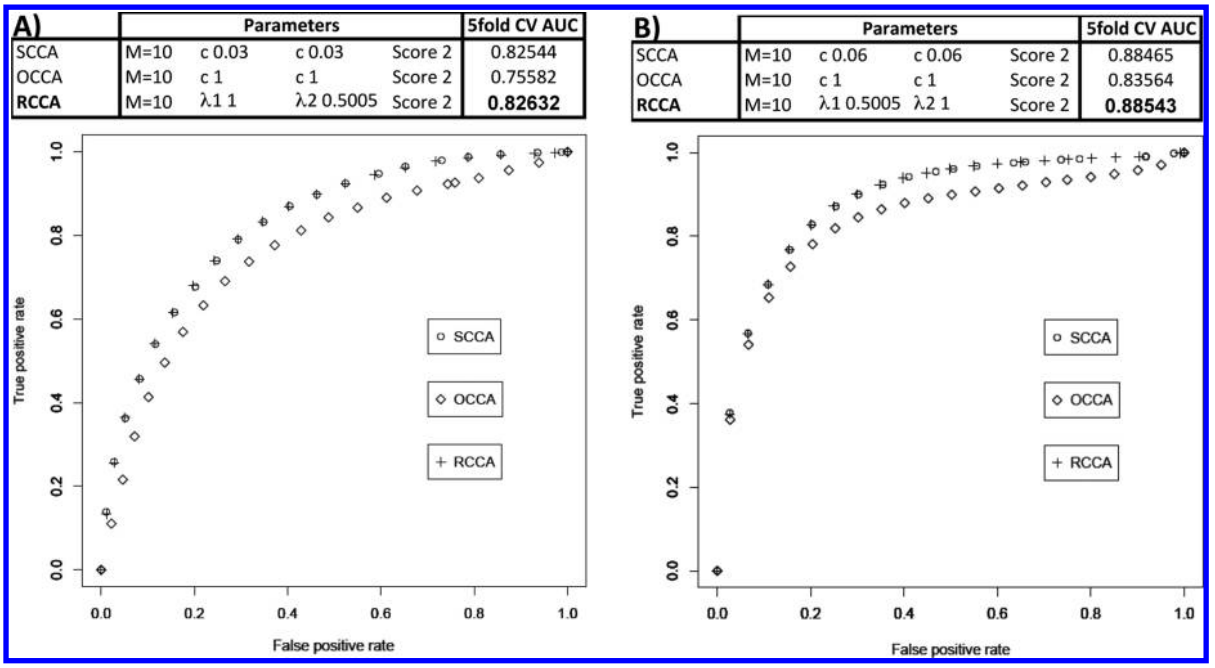


Figure 3. Summary of the optimized parameters for OCCA, RCCA, and SCCA using (A) 112 TCs or (B) 1077 independent SEs. The figure also shows ROC plots for OCCA, RCCA, and SCCA using (A) *NetworkDB_TC* and (B) *NetworkDB_SE*. RCCA is the best-performing method, followed by SCCA with very similar values.

shows the best-predicted TC from 5-fold CV. It can be seen in Table S5 that the maximum and minimum misclassification rates for LOO CV are 20.30% and 0.20%, respectively, while for 5-fold CV the corresponding rates are 26.40% and 0.00%, respectively. The column labeled "Best predicted SE" shows the TCs having the lowest misclassification rates and highest AUCs in the LOO CV. χ^2 values less than 0.05 in LOO CV indicate that the observed and predicted values agree with statistical significance. When the total number of compounds sharing a TC is low, it is more difficult to train a model in this way, and the χ^2 values are higher.

We also compared the "Best predicted SE" using the 554 drugs in *NetworkDB_TC* with an LOO CV prediction using a smaller database of 187 DrugBank drugs ("*DrugBank_187*"), representing CMs of target ligand sets in the *GES matrix*. The majority of the predictions for this smaller database (blue crosses) agree with the 5-fold CV prediction for *NetworkDB_TC* (black circles). From these coincident predictions, the 19 best-predicted TCs (out of 112) had a mean misclassification rate of less than 3.9% and a mean AUC of over 0.8 for the 5-fold CV.

Table S6 shows a summary of the LOO CV statistics for CCA/LDA of the *GES matrix* and *NetworkDB_SE* database. Like the TC results, the last column of this table ("Best Predicted 5fold") shows the best-predicted SEs according to 5-fold CV. It can be seen that the LOO CV predictions agree with the 5-fold CV predictions. The table shows the 324 best-predicted SEs (out of 1077 independent SEs), with a mean misclassification rate of less than 3.9% and a mean AUC of over 0.8 for the 5-fold CV. Red crosses and circles indicate "very good results" with a χ^2 test value of less than 10^{-90} for the LOO CV (crosses) and a mean misclassification rate of less than 0.7% and a mean AUC of over 0.95 for the 5-fold CV (circles). Black crosses and circles correspond to "good results" with low LOO CV misclassification rates and high LOO CV AUCs but lower χ^2 values (crosses) and mean misclassification rates of less than 3.9% and a mean AUC of over 0.8 for the 5-fold CV (circles). CCA/LDA of the *GES matrix* and *NetworkDB_SE* database gives maximum and minimum LOO CV misclassification rates of 16.90% and 0.00%, respectively, while the corresponding 5-fold CV rates are 26.40% and 0.00%, respectively. For the selected 324 best-predicted SE subset, the LOO CV misclassification rates are 3.80% and 0.00%, respectively.

Figure 5 shows plots of the distribution of the 554 *NetworkDB* drugs in the 1077 SEs and the top hit rate against the number of SEs for *NetworkDB_SE* LOO CV. The high average top hit rate obtained is worth noting.

Retrospective Analyses. Tables S5 and S6 show summaries of the retrospective analyses. It can be seen that GESSE can predict 24 TCs very well (coincidences between "Best predicted SE" and "Best Predicted 5fold" in Table S5), which have a LOO CV misclassification error rate of less than 3.7% and a LOO CV AUC of higher than 0.8. From the 1077 independent SEs we can predict 324 SEs very well (with a mean misclassification rate of less than 3.9% and a mean AUC of over 0.8 for the 5-fold CV). These 324 SEs cover 88 of the 112 TCs and belong to 16 of the 24 well-predicted TCs. Table 1 shows a summary of the best-predicted SE TCs according to the retrospective analysis statistics. It highlights the TCs to which the 324 best-predicted independent SEs belong. Only eight TCs (16, 25, 30, 32, 42, 61, 94, and 112) are calculated as "best-predicted" in *NetworkDB_TC* but are not related to the 324 best-predicted independent SEs from *NetworkDB_SE*. Blue crosses show the coincidences with the best-predicted SEs from the *DrugBank_187* database. It is worth noting that we predict well the SE TCs that have "NA" in the "Best Predicted 5fold" column because there are very few true positives, so it is easier to have better statistics. For this reason, these cases also have lower top hit rate values.

Experiment To Check the Robustness of GESSE. As described above, CCA and CCA/DA calculations were repeated for the *DrugBank_187* data set, which is distinct from the test set containing 235 DrugBank drugs (and which is again represented

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

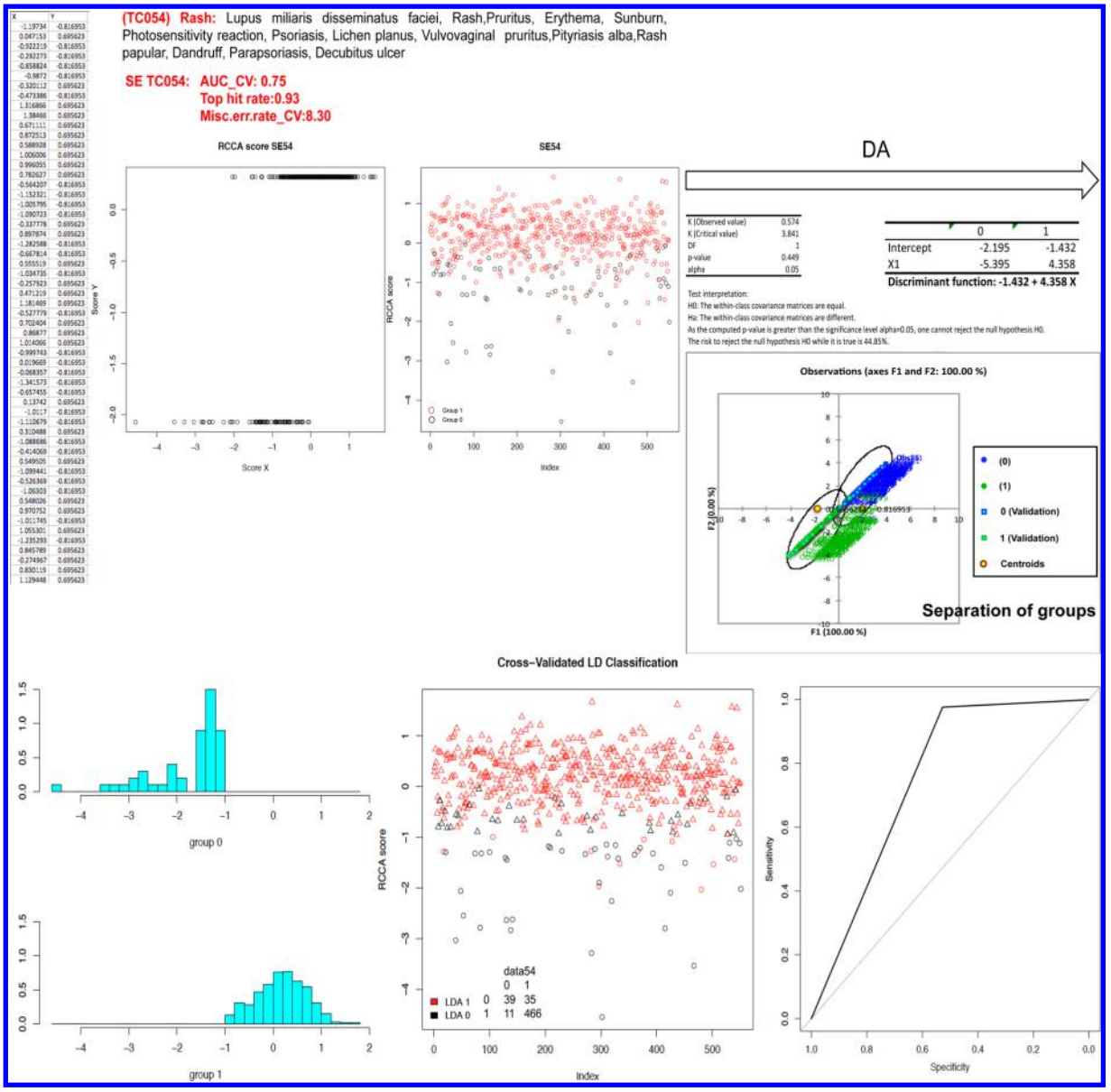


Figure 4. Scheme of the CCA/DA process for SE TC054 (Rash). The RCCA canonical pairs (Score X, Score Y) are presented in the table at the top left and the plot at the top left (Score X vs Score Y). The RCCA Score X is presented versus the molecule index (554 molecules in the *NetworkDB* database) in the plot at the top center. Before applying DA, we observe the groups to be mixed (Group 1 possesses the SE, Group 0 does not possess the SE). With DA (LDA, according to Kullback test), we perform separation of the groups by means of the discriminant equation. The plot at the bottom left shows histograms of the observations in each group on the first linear discriminant dimension after application of LDA. Rectangles are used to indicate how much data are in each of several "bins". The result is a picture that shows a rough "shape" of the distribution. We can see from the histograms that groups 0 and 1 are well-separated by the first discriminant function, since the values for group 0 are between −5 and −1 while the values for group 1 are between −1 and 2, so there is no overlap in the values. The plot at the bottom center shows the canonical score X (coming from RCCA) after application of LDA, with coloring determined by true classification and symbols determined by the resulting classification of leave-one-out LDA. This means, for example, in the case of SE TC054 that there were 11 molecules predicted like group 0 that were group 1 and 35 predicted like group 1 that were group 0. The ROC plot (sensitivity vs specificity) is shown at the bottom right. An AUC of 0.75 was retrieved using the R package pROC (does not perform smooth ROC curves).

using CMs). Figure 6 shows the CCA parameters calculated for this database using TCs. It can be seen that the best CCA parameters are the same as for *NetworkDB_TC* except for λ_1 and λ_2 in the RCCA. These parameters give 5-fold CV AUCs very similar to those calculated for the *NetworkDB_TC* database (5-fold AUC CVs for SCCA, OCCA, and RCCA: 0.825, 0.756, and 0.826, respectively, for *NetworkDB_TC* and 0.819, 0.713, and 0.820, respectively, for *DrugBank_187*).

Table S7 shows the LOO CV statistics for CCA/LDA applied to the *GES matrix* and the *DrugBank_187* TC database. The majority of the LOO CV predictions agree with the 5-fold CV predictions. CCA/LDA applied to the *GES matrix* and *DrugBank_187* gives maximum and minimum LOO CV misclassification rates of 5.90% and 0.00%, respectively, while the 5-fold CV gives values of 13.50% and 0.00%, respectively.

If we compare the results obtained with *Drugbank_187* with those from *NetworkDB*, we can first observe that in *NetworkDB* there are 554 drugs instead of 187, so we generally see higher misclassification rates, lower AUCs, and lower top hit rates, but we still get good results compared with using much more data.

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

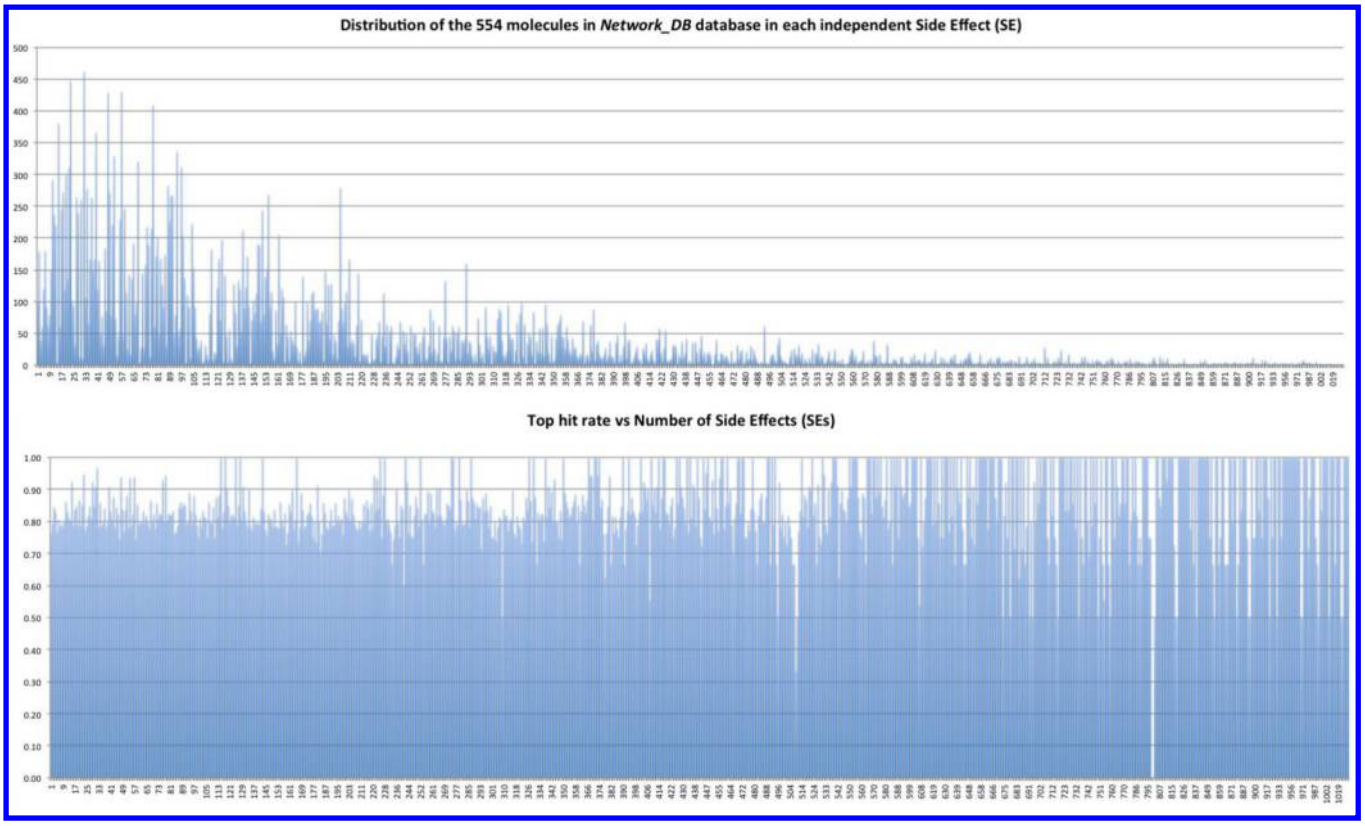


Figure 5. (top) Distribution of the 554 drugs in *NetworkDB* in each independent SE. (bottom) Top hit rate vs number of independent SEs for the LOO CV.

Comparing the results from the two databases, we can see that the tendencies are the same, i.e., the best predictions are obtained for the same TCs, and the worst predictions are also obtained for the same TCs. However, we naturally get lower AUCs and higher misclassification rates using *NetworkDB* because it contains much more data (LOO CV error rates of 20.30% and 0.20% for *NetworkDB* compared with error rates of 5.90% and 0.00% for *Drugbank_187*). Hence, we can claim that GESSE is robust because it can predict well the same TCs in both the small *Drugbank_187* and large *NetworkDB* databases.

The fact that we see the same tendency with both the *DrugBank_187* and *NetworkDB* databases can be explained because the 187 drugs are well chosen, i.e., they follow the same distribution of TCs and thus give similar results. Figure 7 compares the similar distributions of the 554 drugs in *NetworkDB* and the 187 drugs in *DrugBank_187* into the 112 TCs. This observation confirms the utility of using CMs to cluster and reduce data in a representative way, as described previously.[45,46,28] The high average top hit rate obtained, which is comparable to that using independent SEs, is also worth noting (Figure 5).

"Literature Examples" Analysis. Tables S8 and S9 show the TC and independent SE predictions, respectively, for the 413 drugs in the training set, the 141 drugs in the test set, and the test set of 235 DrugBank drugs not present in the *NetworkDB* database. The predicted SEs are compared to the SEs known from the literature for these molecules, covering 50 different SEs grouped into 11 pathological conditions. For each SE and TC, we calculated the percentage of drugs predicted to possess the SE (or TC) with respect to the total number of known drugs that have that SE (or belong to the TC), denoted as the percentage well-predicted (% WP). We considered that a drug was predicted to have an SE if it had a CCA score of >0.4 and either its LDA probability was >0.8 or its PLS-DA probability was >0.8. Comparing the results from the TC and SE data sets, Table S9 shows that using TCs gives better results (see the "% well predicted" and "% well predicted TC" columns). For example, with 112 TCs, 42 of the 50 SEs listed in the "literature examples" table are >50% WP, and 22 of the 50 SEs are 100% WP. With 1077 independent SEs, 25 of the 50 SEs listed in the "literature examples" table are >50% WP, and six of the 50 SEs are 100% WP. Some examples are "Photodermatitis" (20% compared with 100% WP with TCs), "Diverse erythema" (25% compared with 88% WP with TCs), and "Galactorrhea" (38% compared with 88% WP with TCs). However, one exception is "Hypopotassemia", which gives slightly better results for independent SEs than TCs, with SE 195 "Hypokalaemia" being 100% WP, compared with only 67% WP for TC034 "Blood sodium decreased" (which includes hypokalaemia).

Sometimes the name of an SE from the literature does not exactly match an SE name in *NetworkDB_SE,* and this can give an artificially low SE prediction rate. On the other hand, because TCs are often rather general, one TC can often cover several specific SE names, which would tend to give a higher SE probability for a group of molecules. Some examples of such cases are shown in the "Side-effect" column in Table S9. For example, the *NetworkDB_SE* database does not have a term for "Photodermatitis". Hence, molecules reported in the literature with this SE are classified using the next nearest term, which happens to be "Dermatitis.bullous". This gives 20% WP using the SE data set. However, in this case the TC data set gives 100% WP because the corresponding TC (TC065 "Dermatitis") covers every kind of dermatitis. Another example is "Arrhythmias". Again, the *NetworkDB_SE* database does not have this term, so molecules are classified using either "Bradyarrhythmia" or

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

Table 1. Summary of the Best-Predicted SE TCs According to Retrospective Statistics[a]

Independent SEs	SE TCs	misc.err.rate_CV	Chisqtest_CV	AUC_CV	Top hit rate	Best predicted SE	Best Predicted 5fold
	TC016: Hypothermia	2.50	4.1E-61	0.82	0.69	X	O
Wound dehiscence Pathological fracture Haemochromatosis Transfusion reaction	TC021: Pathological fracture	3.30	1.4E-56	0.85	0.71	X	O
Melanosis Leukoderma Vitiligo	TC022: Skin hyperpigmentation	3.80	7.6E-66	0.82	0.71	X	O
	TC025: Menorrhagia	3.40	7.7E-61	0.84	0.74	X	O
	TC030: Intra-ocular injection	0.50	4.0E-59	0.86	0.71	X	O
	TC032: Abortion spontaneous	1.10	1.4E-58	0.94	0.67	X	O
	TC042: Penile pain	0.20	1.3E-46	0.83	0.67	X	NA
Tenosynovitis Epicondylitis Trigger finger	TC045: Tendonitis	2.70	2.4E-75	0.88	0.80	X	O
Vitamin B12 deficiency	TC049: Night blindness	0.90	4.8E-85	0.94	0.89	X	O
Skin atrophy Skin striae	TC052: Hyperkeratosis	2.40	1.5E-62	0.79	0.79		O
Uterovaginal prolapse Uterine haemorrhage Endometriosis Endometrial cancer Endometrial hyperplasia	TC057: Uterine haemorrhage	3.80	6.1E-58	0.81	0.68	X	O
Sciatica	TC059: Sciatica	0.50	1.7E-84	0.88	0.77	X	O
	TC061: Malabsorption	0.50	7.4E-39	0.80	0.60	X	NA
Testicular atrophy Testicular pain	TC063: Testicular mass	2.40	6.4E-57	0.87	0.71	X	O
Wound Road traffic accident	TC067: Wound	1.10	1.5E-71	0.92	0.73	X	O
Methaemoglobinaemia Blood disorder	TC072: Methaemoglobinaemia	2.50	6.9E-67	0.87	0.77	X	O
Orchitis noninfective	TC073: Orchitis	0.90	1.1E-88	0.91	0.86	X	O
Synovitis	TC081: Bursitis	3.30	3.4E-64	0.84	0.79	X	O
Ovarian cyst Ovarian cancer Benign prostatic hyperplasia	TC082: Ovarian cyst	1.80	2.5E-63	0.89	0.79	X	O
Cryptosporidiosis infection	TC086: Tuberculosis	0.90	4.0E-36	0.83	0.50	X	NA
Porphyria	TC094: Porphyria	1.30	4.3E-61	0.78	0.75		O
Premature baby	TC097: Congenital anomaly	2.20	6.9E-52	0.84	0.68	X	O
Chondrodystrophy	TC101: Chondrodystrophy	0.50	7.4E-39	0.80	0.80	X	NA
Cervical dysplasia	TC112: Cervical dysplasia	1.10	5.5E-38	0.99	0.75	X	NA

[a]Abbreviations/headings: "SE", side effect; "TC", term cluster; "misc.err.rate_CV", misclassification rate for the LOO CV; "Chisqtest_CV", χ^2 independence test for the LOO CV; "AUC_CV", area under the ROC curve for the LOO CV; "Top hit rate", top hit rate values calculated according to eq 7; "Best predicted SE", best-predicted according to the lowest misclassification rate and highest AUC; "Best Predicted 5fold", best-predicted SE TC according to 5-fold CV statistics. "NA" appears in the "Best Predicted 5fold" column for the cases where there are very few true positives (drugs that possess the SE). The SE TCs selected have a "misc.err.rate_CV" lower than 3.9% (highlighted in green) and an "AUC_CV" higher than 0.8 (highlighted in red). The TCs to which the 324 best-predicted independent SEs belong are highlighted in yellow. In the "Chisqtest_CV" and "Top hit rate" columns, the 15% highest values and 15% lowest values are highlighted in red and green, respectively. Blue crosses show the coincidences with the best-predicted SEs for the drugs in the *DrugBank_187* database.

"Ventricular.tachyarrhythmia". This gives 14% WP using the SE data set. However, the TC data set gives 100% WP because the corresponding TC (TC074 "Syncope") includes general arrhythmia and all kinds of cardiac disorders. A further example is the TC "Tardive.dyskinesia", which is described as "Extrapyramidals effects" in the literature. Extrapyramidal symptoms (EPSs) are drug-induced movement disorders that include dystonia (continuous spasms and muscle contractions), akathisia (motor restlessness), parkinsonism (characteristic symptoms such as rigidity, bradykinesia, and tremor), and tardive dyskinesia (irregular, jerky movements). When the SE data set is used, only one of the above SEs may describe EPS, and we chose to use "Tardive.dyskinesia" because that gives 75% WP. However, when the TC data set is used, "Extrapyramidals effects" is described by TC001 "Tremor", which includes all of the EPS SEs, and this gives 100% WP.

A low % WP can also be due to poor predictions of the SEs that in the retrospective analysis have higher misclassification rates (see Tables S5–S7) and poor distributions of LDA discriminant scores (see folder LDA_Plots_TC). In other words, some overlap exists between the scores of group 1 (molecules possessing the SE) and group 0 (molecules not possessing the SE). One example of such a case is "Acute respiratory distress syndrome", represented by TC029 "Pulmonary oedema", for which the LOO misclassification rate is 11.40% (see Table S5) and there is some overlap between groups in the LDA histogram. Another example is "Polyps", represented by TC053 "Epistaxis", for which the LOO misclassification rate is 17.40% (see Table S5) and there is some overlap between groups in the LDA histogram. Finally, in the SE data set "Anaphylaxis" is represented by SE 80 "Anaphylactoid.reaction", for which the LOO misclassification rate is 9.8% and there is overlap with other groups in the LDA histogram.

Tables S8 and S9 show the targets known for each drug according to the literature, the highest related target for each drug predicted by GES, and the GES p value in the "literature examples" analysis. These tables show that the highest related target predicted by GES is in most cases one of its known related targets. Moreover, some of the highest related targets predicted by GES for drugs belonging to a specific pathological condition are present in the minimal pharmacological profiling panel of Bowes et al.[1] for this pathological condition. These targets are alpha-2A adrenergic receptor, beta-1 adrenergic receptor, and D(1A) dopamine receptor for the cardiovascular system;

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804–1823

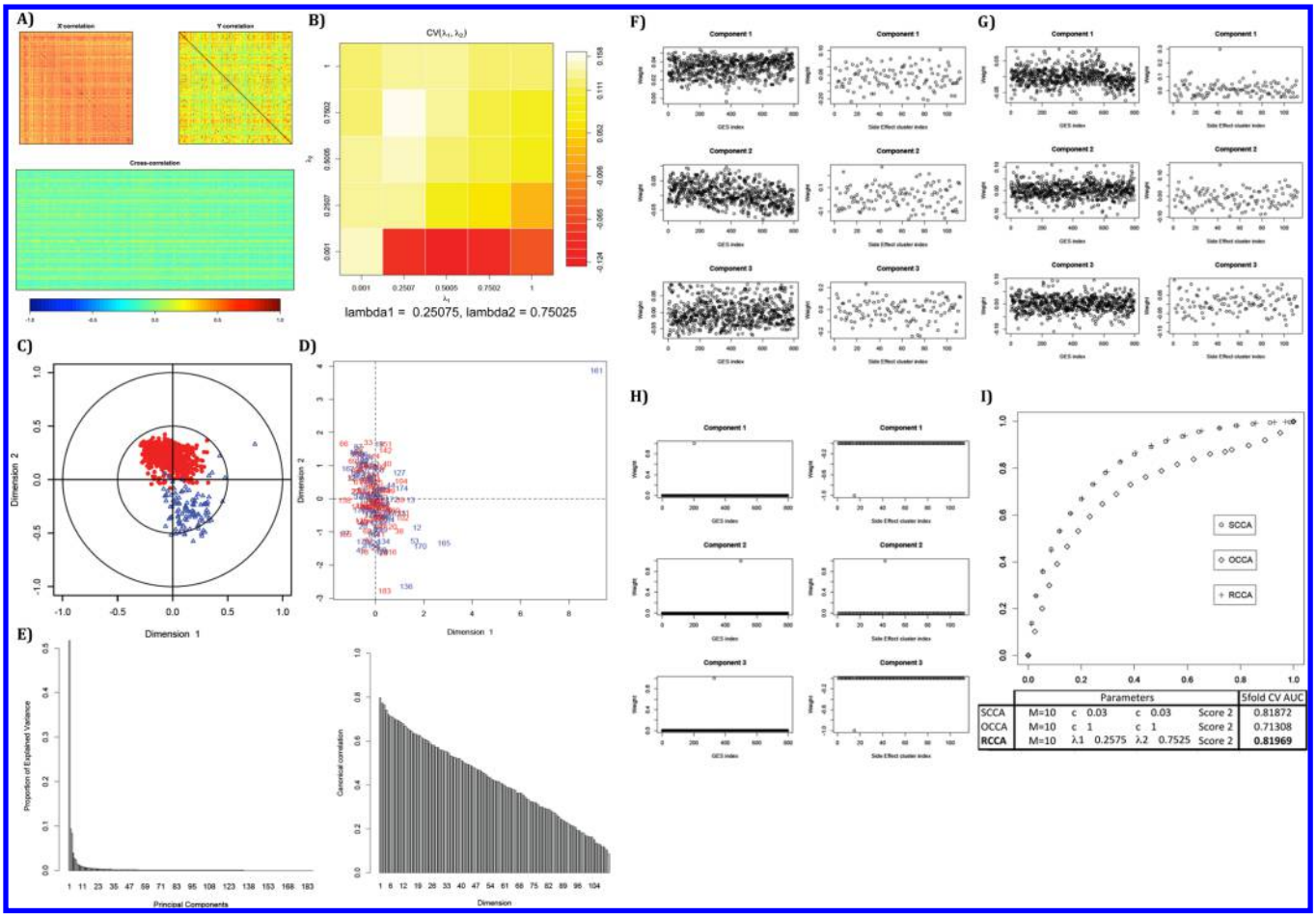


Figure 6. CCA and CCA/DA calculations with a smaller data set of 187 DrugBank drugs that are center molecules (CMs) of targets in the *GES matrix*. (A) *GES matrix* correlation (X), *SE matrix* correlation (Y), and cross-correlation between the *GES matrix* and the *SE matrix* before the CCA calculation (using 112 TCs). (B) Heat map with the CV-optimized λ values used in RCCA (using 112 TCs). (C) Canonical correlation between the first two dimensions of canonical variables (using 112 TCs). (D) Canonical correlation between the first three dimensions of canonical variables (using 112 TCs). (E) Choice of the CCA dimension. Most of the data are explained with the first 10 principal components. After component 10 the correlation decreases significantly. Hence, we extract from CCA 10 canonical pairs. (F–H) Index plots of weight vectors for GES *p* values (left) and side effects (right) extracted by OCCA, RCC, and SCCA, respectively (using 112 TCs). (I) Summary of the optimized parameters for OCCA, RCCA, and SCCA (using 112 TCs).

glucocorticoid receptor for the endocrine system; muscarinic acetylcholine receptor M3 for the gastrointestinal system; and D(1A) dopamine receptor, 5-hydroxytryptamine 1A receptor, and alpha-2A adrenergic receptor for the central nervous system. On the other hand, the highest related targets predicted by GES that are not associated with a drug and are not in the panel of Bowes et al. could be considered as potential new targets with unknown polypharmacology.

Figure 8 plots the first two PLS-DA components for the psychiatric, gastrointestinal, and cardiovascular manifestations for both the SE and TC data sets. The molecules known to have these pathological conditions according to the literature are shown in Tables S8 and S9. The ideal case would be to find clusters of these molecules divided according to the SEs to which they belong, that is to say, one cluster of true positive (TP) molecules for each SE (which corresponds to 100% WP in Tables S8 and S9). This is the case of the plot using 112 TCs for psychiatric manifestations (top right plot in Figure 8). All of the molecules belonging to this condition are depicted and clustered into just three groups (deliriums, somnolence, and dream dysfunctions), although there is some overlap between them because they are rather similar SEs belonging to the same manifestation.

For both the TC and independent SE predictions, Figure 8 shows that the SEs are grouped and differentiated but also separated by TP and true negative (TN) SE clusters, since we do not always have the ideal 100% WP (see Tables S8 and S9 for details). For example, in the plot using 1077 independent SEs for psychiatric manifestations (top left plot in Figure 8), all of the molecules are clustered into deliriums TP and TN (the TP cluster is much smaller than the TN cluster, with only 23% WP), somnolence TP and TN (the somnolence TP cluster is much bigger than the TN cluster, as it is 90% WP), and dream dysfunctions TN (both sleep.disorder and sleep.apnea.syndrome have 0% WP SE).

If we consider all of the TCs in *NetworkDB_TC*, we generally see clusters that are quite closely grouped and overlapping. However, the SE clusters in *NetworkDB_SE* are generally more distinct and less overlapping. These observations broadly agree with the nature of the two data sets, i.e., 1077 independent SEs (separated between them) and 112 TCs that associate several similar independent SEs (possibility of some overlap between TCs). Nonetheless, in both data sets we can visually differentiate well the pathological conditions and their SEs for psychiatric manifestations (dream dysfunctions, somnolence, and deliriums),

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804–1823

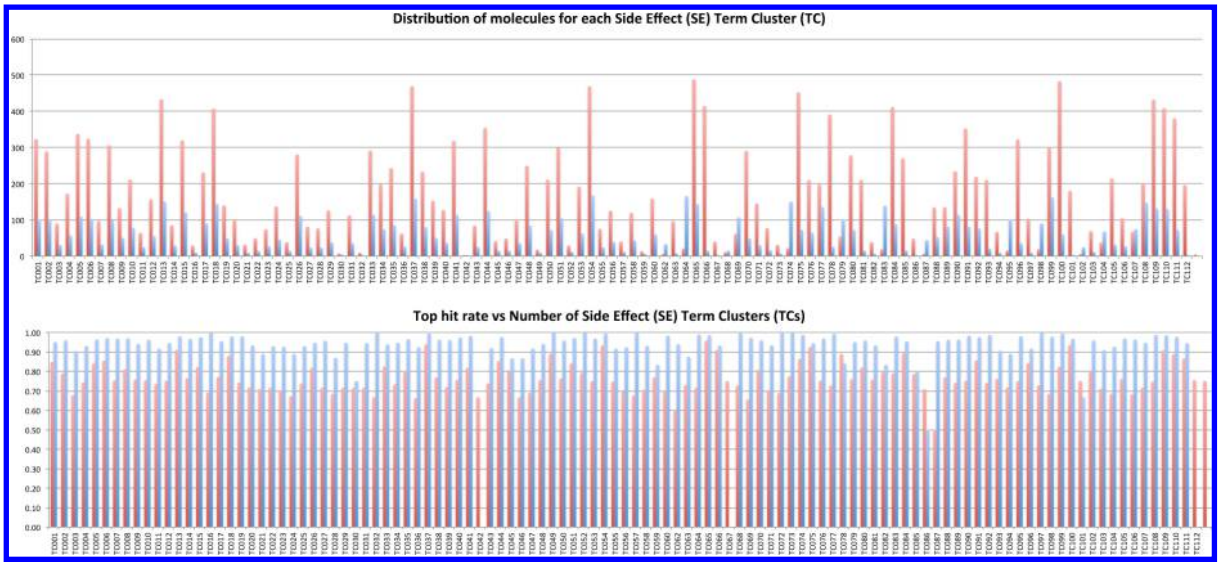


Figure 7. (top) Distributions of the 554 drugs in *NetworkDB* (red) and the 187 test molecules in *DrugBank_187* (blue) in the 112 SE TCs. (bottom) Top hit rate vs number of SE TCs for the LOO CV, for both the 554 drugs in *NetworkDB* (red) and the 187 test molecules in *DrugBank_187* (blue).

cardiovascular manifestations (exacerbations, arrhythmias, edema, hypertension, hypotension, and thromboembolism), and gastrointestinal manifestations (constipation, diarrhea, diffuse hepatocellular damage, mouth dryness, nausea, pancreatitis, and hemorrhagic ulceration).

Relating Targets to SEs. Joining the chemical space of drugs, the biological space of targets, and the phenotypic space of SEs allows targets and SEs to be related. Clearly, the ability to relate targets and SEs is of considerable interest to the pharmaceutical industry. In order to link targets and SEs quantitatively, we applied RCCA (i.e., the best-performing CCA method) to the GES drug−target relationships and *NetworkDB_TC* SE profiles using the optimized parameters found in the retrospective analysis. This provided us with 10 canonical components, ($\boldsymbol{u}_1$, $\boldsymbol{v}_1$) to ($\boldsymbol{u}_{10}$, $\boldsymbol{v}_{10}$). The correlated sets of targets and side effects were extracted from the weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (see Methods). A list of high-scoring drugs that contributed to the correlation for both targets and SEs was also obtained for each component. We refer to these correlated sets as canonical components (CCs). The contents of all 10 CCs are listed in the spreadsheet GESSE_Supporting_Information. For each CC are ranked the correlated sets of targets and side effects (highest positive weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively) and the high-scoring drugs for both targets and SEs ($\boldsymbol{u}$ and $\boldsymbol{v}$ canonical scores, respectively). To evaluate the biological relevance of the targeted proteins within the extracted 10 canonical components, we examined the top five targets and side effects with the highest positive weights extracted for each of the CCs with regard to biological pathways and molecular functions.

The GESSE_Supporting_Information spreadsheet also shows the target-related pathways extracted from KEGG[47] and REACTOME,[48] the disease-related genes extracted from OMIM,[49] and the related molecular functions extracted from Gene Ontology (GO)[24] for the top five high-ranked correlated targets (positive $\boldsymbol{\alpha}$ weights) and high-scoring drugs (for both targets and SEs) for the 10 CCs. The most frequently appearing KEGG pathways across the 10 CCs are "Caffeine metabolism" (hsa00232+7498), "Drug metabolism - other enzymes" (hsa00983+7498), "G0 and Early G1" (REACT_111214), "Glycosphingolipid metabolism" (REACT_116105), "Metabolic pathways" (hsa01100+6999 and hsa01100+7498), "Peroxisome" (hsa04146+7498), "Purine catabolism" (REACT_2086), "Purine metabolism" (hsa00230+7498), "Tryptophan catabolism" (REACT_916), and "Tryptophan metabolism" (hsa00380+6999).

Furthermore, visual inspection of the examples listed in GESSE_Supporting_Information suggests that the pathways related to the same CC seem to take part in the same global biological function. Similarly, we find that several of the high-scoring drugs for both targets and SEs for a given CC are related to the same target, which takes part in the associated pathways to the drugs. Moreover, we find that several of the high-scoring drugs for both targets and SEs for a given CC share the same molecular function. For example, in CC1 we find the targets "DNA topoisomerase 4 subunit B", "DNA topoisomerase 4 subunit A", and "3-phosphoinositide-dependent protein kinase 1", all of which are related to pathways that take part in the same molecular function "ATP binding" (GO:0005524). In regard to the high-scoring drugs in this CC, we find "Trimipramine" (DB00726), "Clomipramine" (DB01242), "Amoxapine" (DB00543), and "Desipramine" (DB01151), which are related to the same targets ("noradrenalin transporter inhibitor" and "serotonin transporter inhibitor"), which take part in the associated pathway to these drugs, "Serotonergic synapse" (hsa04726). Moreover, we find that these high-scoring drugs share the same molecular function ("Antidepressant"). Hence, it is coherent that these drugs are found to be highly correlated in GESSE with the SE TC "Anxiety" (see Figure 9C). In CC2, we find the glucorticoid drugs "Beclomethasone" (DB00394), "Betamethasone" (DB00443) and "Hydrocortisone" (DB00741), which are related to the same target ("glucocorticoid receptor agonist"), which participates in the associated "neuroactive ligand-receptor interaction" pathway. "Beclomethasone" and "Betamethasone" are predicted in GESSE to have the "Candidiasis" SE (see Figure 9D), which agrees with the fact that candidiasis is associated with the local deposition of inhaled glucocorticoids in the oropharynx and larynx. "Hydrocortisone" is linked to the "Furuncle" SE by GESSE (Figure 9D), which also agrees with several FDA report statistics showing that hydrocortisone causes furuncle.

In CC5, the GESSE groups "UMP-CMP kinase" and "Sphingomyelin phosphodiesterase" are related to "metabolic pathways" that take part in the "ATP binding" (GO:0005524)

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

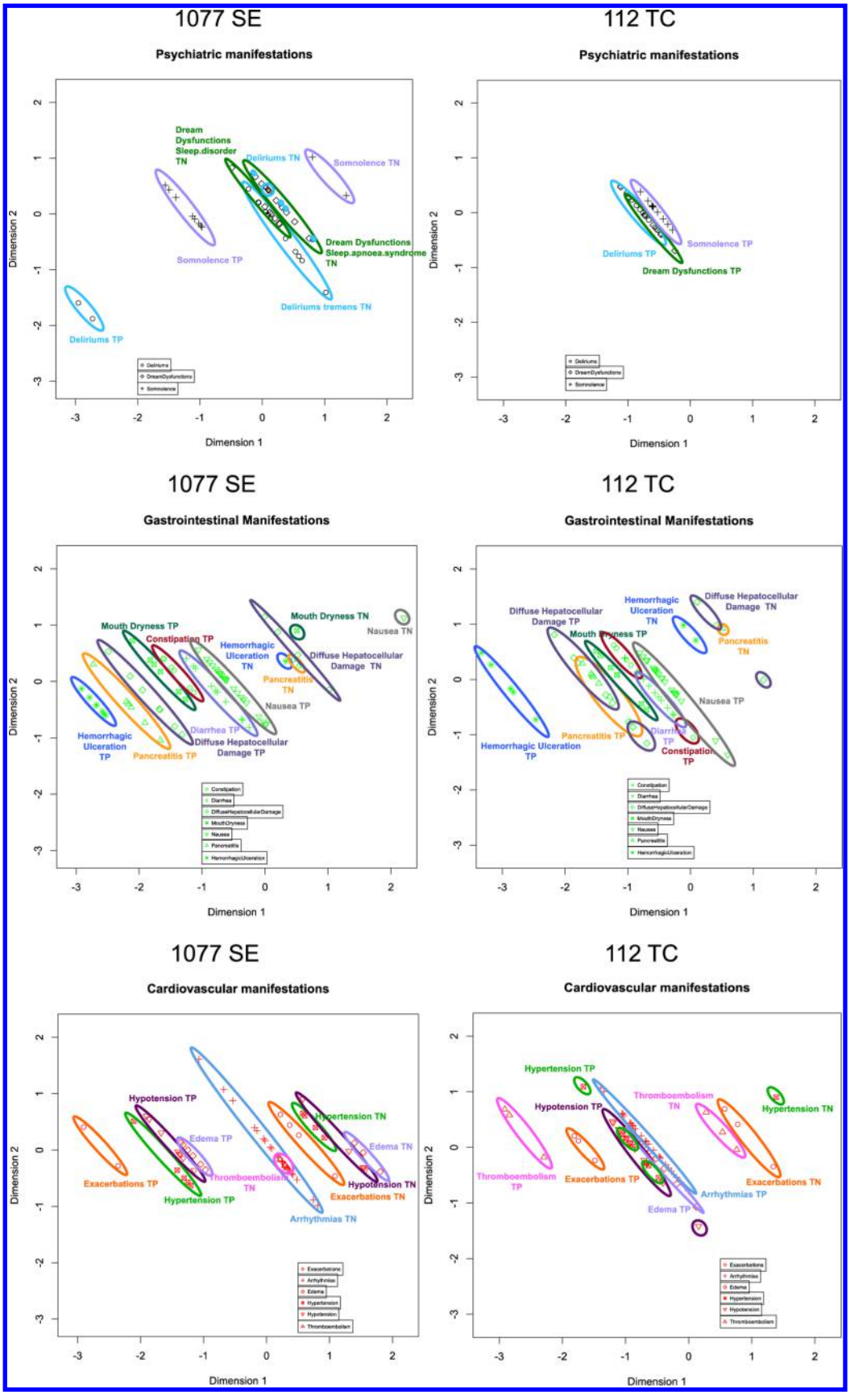


Figure 8. Visual representations of components 1 and 2 in PLS-DA analyses for psychiatric, gastrointestinal, and cardiovascular manifestations.

and "Protein binding" (GO:0005515) molecular functions, respectively. They are also grouped with "5′-AMP-activated protein kinase catalytic subunit alpha-1", which is related to the "Insulin signaling pathway" (hsa04910+5562) that takes part in

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

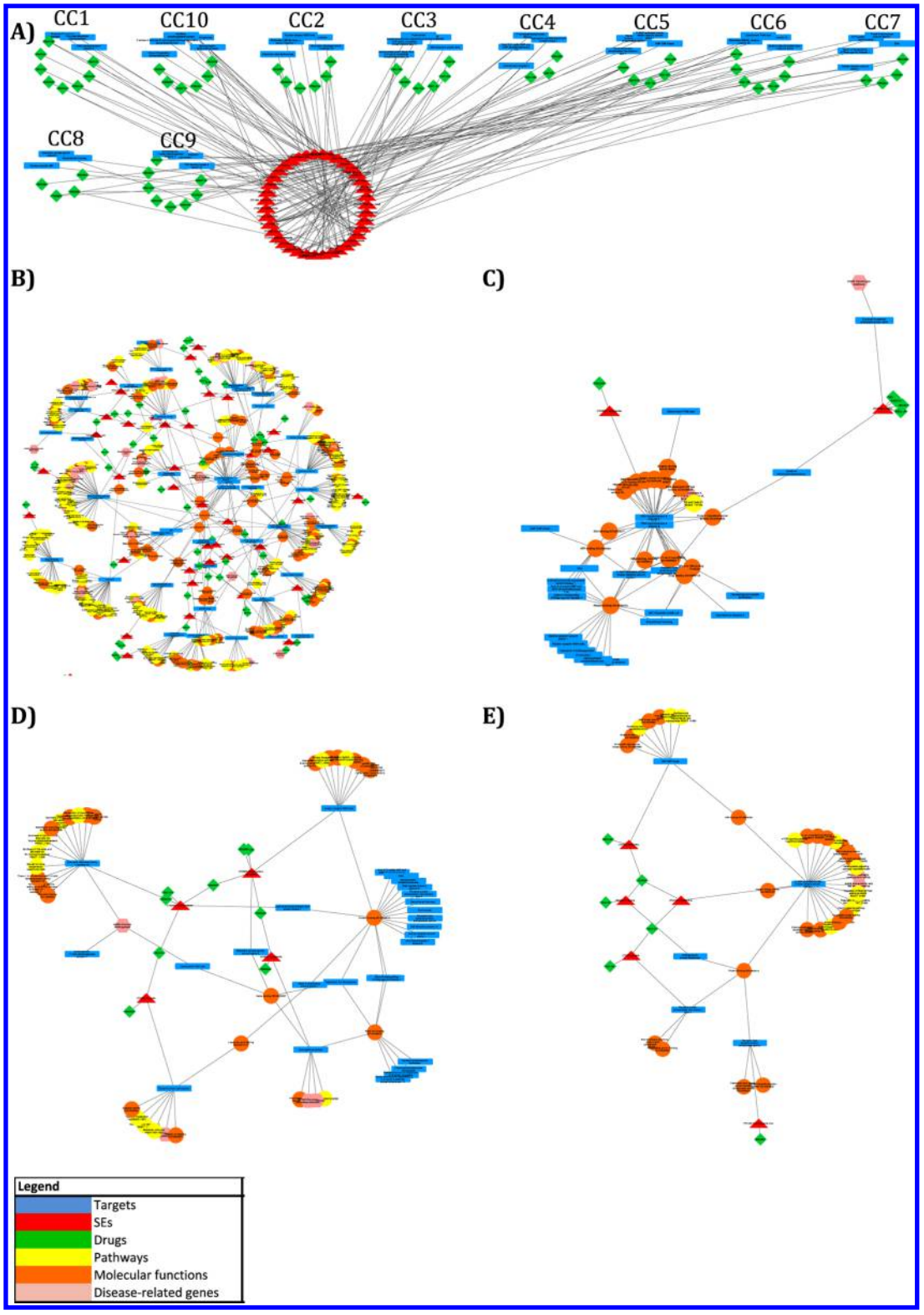


Figure 9. (A) Network of the top five highest-scored correlated targets−drugs−side effects in the extracted 10 CCs. The highest-positive-weighted proteins (blue rectangles), the high-scoring drugs for both targets and SEs (green diamonds), and the highest-positive-weighted side effects (red triangles) are connected if they appear in the same canonical component (CC). (B) Network of drug-targeted proteins and side effects in the extracted 10 CCs, with an edge-weighted spring-embedded layout in which the lengths of the edges in the network reflect the CCA $\boldsymbol{\alpha}$ weights when connecting targets and SEs and the CCA $\boldsymbol{u}$ CC scores when connecting drugs and SEs. The associated pathways to the drugs causing the side effects (yellow circles), the disease-related genes (pink hexagons), and the related molecular functions (orange circles) are also shown. We consider a high-scoring drug to perturb a pathway if the latter was related to at least one of the high-weighted target proteins in RCCA. (C−E) Zoomed subnetworks for easier visibility for (C) CC1, (D) CC2, and (E) CC5. High-resolution network graphs are available in the Supporting Information.

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

the shared molecular function "Protein binding" (GO:0005515) as well as "ATP binding" (GO:0005524). "Receptor-type tyrosine-protein phosphatase epsilon" and "Tyrosine-protein phosphatase non-receptor type 4" are also grouped in CC5, sharing "Protein binding" (GO:0005515) molecular function. In regard to the high-scoring drugs in this CC, we find "Gliclazide" (DB01120), "Tolazamide" (DB00839), and "Chlorpropamide" (DB00672), which are related to the same target ("sulfonylurea receptor 1 agonist"), which takes part in the pathways associated with these drugs ("ABC transporters", "Insulin secretion", and "Type II diabetes mellitus"). "Tolbutamide" (DB01124) also appears in this CC with a high score. This drug is related to "ATP-sensitive potassium channel" (SUR1/Kir6.2) blocker, which also takes part in the three aforementioned pathways. These four high-scoring drugs are related to "Type II diabetes mellitus" disease, and we find that they share the same molecular function ("Antidiabetic"). GESSE relates "Gliclazide"and "Tolazamide" to porphyria (see Figure 9E), which is consistent with the well-known and well-documented relationship between diabetes and porphyria.[50,51] "Tolbutamide" is predicted by GESSE to have the "Bursitis" SE, which also is in agreement with the fact that immune deficiencies, including diabetes, can also cause bursitis.[52]

As a last example, CC10 groups together the targets "Group IIE secretory phospholipase A2", "Sphingomyelin phosphodiesterase", "Tryptophan 2,3-dioxygenase", and "Xanthine dehydrogenase/oxidase". All of these are linked to metabolic pathways, and "Sphingomyelin phosphodiesterase" and "Tryptophan 2,3-dioxygenase" are involved in exactly the same metabolic pathway, "Metabolic pathways" (hsa01100+6999), and participate in the same molecular function, "Protein binding" (GO:0005515). Additionally, the high-scoring drugs "Norethindrone" (DB00717) and "Medroxyprogesterone Acetate" (DB00603) are related to the same target ("progesterone receptor agonist"), which takes part in the pathways associated with these drugs ("Oocyte meiosis" and "Progesterone-mediated oocyte maturation"). Moreover, these high-scoring drugs share the same molecular function ("Progestin").

Figure 9 shows the network of the top five targets and side effects with the highest positive weights extracted for each of the 10 CCs, where proteins (blue rectangles), side effects (red triangles), and high-scoring drugs for both targets and SEs (green diamonds) are connected if they appear in the same CC. The associated pathways to the drugs causing the side effects are also shown (yellow circles). We considered a high-scoring drug to perturb pathway if the latter was related to at least one of the high-weighted target proteins in RCCA. Similarly, we annotated drugs with molecular functions (orange circles).

We performed the same analysis for the CCA/DA method, in which DA is performed after CCA in order to classify in a more accurate manner the probability that a drug has a given SE. By CCA, *NetworkDB* drugs are ranked according to their RCCA canonical scores, whereas by CCA/DA, *NetworkDB* drugs are ranked according to their Bayes probability values. Because DA can be applied to only one response variable or CC at a time, we apply DA repeatedly for each SE using the CCA canonical variables. Hence, for each CC we obtain the ranked lists of highest positive $\boldsymbol{\alpha}$ weights on target proteins, highest positive $\boldsymbol{\beta}$ weights on SEs, and highest-scoring drugs for both targets and SEs ($\boldsymbol{u}$ and $\boldsymbol{v}$ canonical scores, respectively). For each of these four ranked lists, we then select the CCs (which correspond to SE TCs) that have the highest positive weights on target proteins and SEs and highest-scoring drugs. Any CCs common to the four lists are then selected as highly correlated SEs. The 20 highest-correlated targets and drugs for the highest-correlated CCs (corresponding to SEs) are listed in GESSE_Supporting_Information. This spreadsheet also shows the target-related pathways and disease-related genes extracted from REACTOME, KEGG, and OMIM and the related molecular functions extracted from GO.

Figure 10A illustrates the network of 20 extracted targets and drugs (with highest positive weights and highest canonical scores, respectively) for the highest-correlated CCs. The lengths of the edges in the network reflect the CCA $\boldsymbol{\alpha}$ weights when connecting targets and SEs and the CCA $\boldsymbol{u}$ CC scores when connecting drugs and SEs. Regarding the high-scoring drugs (CCA $\boldsymbol{u}$ scores), it can be seen that, for example, the high-scoring drug "Cyclophosphamide" (DB00531) is predicted by GESSE to have the SE TC100 "Erythema multiforme", which agrees with the fact that this drug is related to the disease "Systemic lupus erythematosus" and the "DNA" target with an "antineoplastic-alkylating" molecular function.

Regarding the zoomed subnetworks for some of the top proteins with highest positive weights (Figure 10C), "nuclear receptor 0B1" can be seen to be related to the diseases "46,XY disorders of sex development" (disorders of gonadal development, H00607) and "glycerol kinase deficiency (GKD)" (H00552) and the "nuclear receptor transcription" pathway [*Homo sapiens*] (REACT_15525.4). The CCA network built for "nuclear receptor 0B1", comprises the SE TC082 "Ovarian cyst", which agrees with the disorders of gonadal development associated with this target.

We can observe in the subnetwork for "Ornithine aminotransferase" that this target is related to the pathways "amino acid synthesis and interconversion (transamination)" (REACT_238), "arginine and proline metabolism" (hsa00330), "metabolic pathways" (hsa01100), and "biosynthesis of antibiotics" (hsa01130). It is also related to the molecular function "pyridoxal phosphate binding" (GO:0030170) and the diseases "ornithinaemia" (H00189) and "Secondary hyperammonemia" (H01400). Between the SEs that GESSE shows to be highly related to this target in the network we find TC028 "Furuncle". Since a furuncle is caused by infection by the bacterium *Staphylococcus aureus*, it is logical that this SE is related to this target, given that one of the related pathways is "biosynthesis of antibiotics".

From the subnetwork for "Group IIE phospholipase A2", we see that this target is related to the molecular functions "calcium ion binding" (GO:0005509) and "phospholipase A2 activity" (GO:0004623) and the pathways "alpha-linolenic acid metabolism" (hsa00592), "glycerophospholipid metabolism" (hsa00564), "ether lipid metabolism" (hsa00565), "arachidonic acid metabolism" (hsa00590), "linoleic acid metabolism" (hsa00590), "metabolic pathways" (hsa01100), "Ras signaling pathway" (hsa04014), "vascular smooth muscle contraction" (hsa04270), "pancreatic secretion" (hsa04972), and "fat digestion and absorption" (hsa04975). The SEs highly related to this target in the network as determined by GESSE are TC052 "Hyperkeratosis", TC061 "Malabsortion", and TC106 "Gastroenteritis viral", which are in agreement with the "fat digestion and absorption" pathway.

When we consider the subnetwork for "Fibroblast growth factor receptor 2", we find that according to GESSE this target is highly related to the SEs TC027 "Herpex simplex", TC101 "Chondrodystrophy", and TC112 "Cervical dysplasia". This target has been found in the literature to be related to these SEs. For example, "Fibroblast growth factor-activated receptor 3" has been related to skeletal dysplasia syndromes and cervical

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

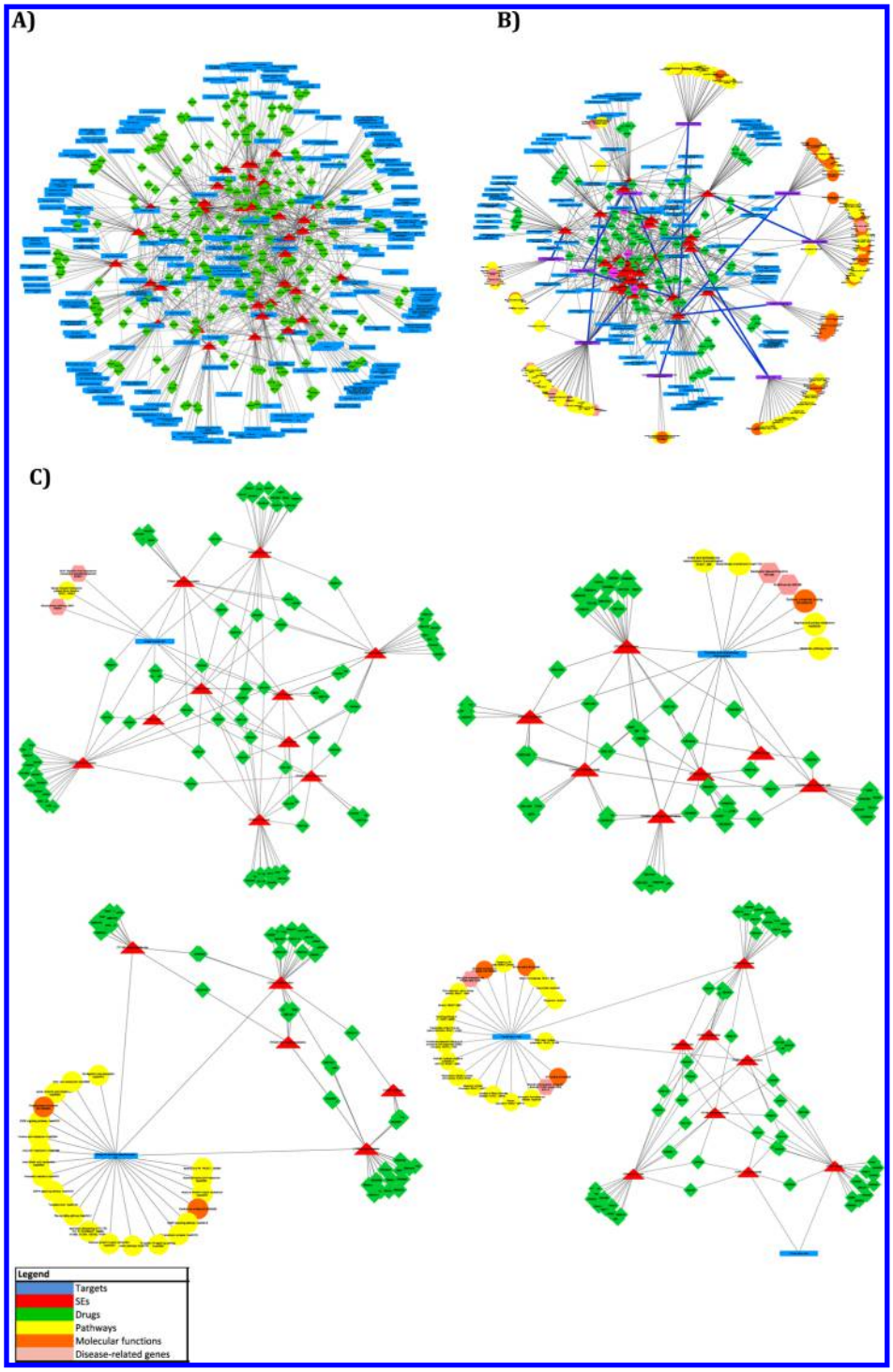


Figure 10. (A) Network of the extracted 20 targets and drugs (with highest positive weights and highest canonical scores, respectively) for the selected highest-correlated CCs (which correspond to SEs). The highest-positive-weight targets and highest-scoring drugs for targets are connected to the SE corresponding to the high-correlated CC. Target proteins are represented as blue rectangles, SEs as red triangles, and drugs as green diamonds. (B) Network for the top 10 proteins with highest positive weights, shown as purple rectangles with dark-blue-highlighted edges. The top high-scoring drugs are shown as fuchsia diamonds, the target-related pathways as yellow circles, the disease-related genes as pink hexagons, and the related molecular functions as orange circles. The lengths of the edges in the network reflect the CCA $\boldsymbol{\alpha}$ weights when connecting targets and SEs and the CCA $\boldsymbol{u}$ CC scores when connecting drugs and SEs. (C) Zoomed subnetworks for easier visibility for some examples of the highest-positive-weighted targets: (top left) "Nuclear receptor 0B1"; (top right) "Ornithine aminotransferase"; (bottom left) "Group IIE phospholipase A2"; (bottom right) "Tubulin beta-1 chain". High-resolution network graphs are available in the Supporting Information.

neoplasms.[53] "Fibroblast growth factor-activated receptor 2" has also been related to chondrodystrophy[54] and herpes simplex.[55]

As a final example, we consider the subnetwork for "Tubulin beta-1 chain". This target is related to the molecular functions "GTP binding" (GO:0005525), "GTPase activity" (GO:000392), and "structural constituent of cytoskeleton" (GO:0005200), among others. It is also related to the diseases "pathogenic escherichia coli infection" (KEGG Brite 05130) and "macrothrombocytopenia,

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

autosomal dominant, TUBB1-related" (OMIM 613112). This is in accordance with the highly related SE TC051 "Thrombocytopenia" that GESSE associates with this target's network.

CONCLUSION

We have presented a new approache called GESSE for predicting drug SEs. GESSE joins the physicochemical (3D shape + chemistry) space of drug molecules with the polypharmacologically relevant biological subspace of drug target proteins. Extending our earlier GES approach in this way has allowed the predictive triangle of *drugs−targets−SEs* to be closed, thus allowing targets and SEs to be related by the drugs that they share. We expect that this approach will be of interest to the pharmaceutical industry not only to decipher mechanisms of action related to SEs but also to investigate combination therapies in which the overall SEs experienced by patients might be lower than when the drugs are selected individually.

To demonstrate the utility of this approach, we have predicted the SEs for a set of DrugBank drugs whose polypharmacology was calculated for 777 targets using our previously described GES approach. The quality of the SE predictions was calculated using both retrospective analysis and literature examples. Keeping in mind that we used *predicted* drug−target interaction data, the GESSE approach has been shown to be rather robust.

Overall, GESSE gives good retrospective statistics and identifies high percentages of previously known SEs from the literature. Our results show that the use of TCs to describe clusters of SEs gives better performance than using unclustered SE relationships. For a data set of 112 TCs, a total of 42 out of 50 SEs listed in the literature examples are over 50% WP, and 22 out of the 50 are 100% WP. In comparison, using a data set of 1077 independent SEs, only 25 out of 50 SEs are over 50% WP and only six out of 50 are 100% WP. Moreover, in a majority of cases the highest-related target found for the literature examples is found to match one of the known related targets according to the literature. This demonstrates the usefulness of clustering SEs into TCs to create SE profiles. Hence, GESSE is a valuable aid for the prediction of SE profiles of drugs. We believe that the GESSE approach could assist the optimization of lead molecules and the drug development process and that the ability to relate drug targets to drug side effects could be of considerable interest to the pharmaceutical industry.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00120.

Figures S1−S7 and Tables S1−S9 (PDF)

Zip file LDA_Plots_TC containing the graphs "RCCA Canonical Score: Score X vs Score Y", "RCCA Score X vs Index", "LDA Histogram: Group 0 and Group 1 Distributions", "LOO Cross-Validated LD Classification: RCCA Score vs Index", and "ROC Plot after LOO CV: Sensitivity vs Specificity" for all of the SEs in the *NetworkDB_TC* database (ZIP)

Zip file LDA_Plots_SE containing the graphs "RCCA Canonical Score: Score X vs Score Y", "RCCA Score X vs Index", "LDA Histogram: Group 0 and Group 1 Distributions", "LOO Cross-Validated LD Classification: RCCA Score vs Index", and "ROC Plot after LOO CV: Sensitivity vs Specificity" for all of the SEs in the *NetworkDB_SE* database (ZIP)

Zip file High_Resolution_Network_Graphs containing high-resolution versions of the network graphs shown in Figures 9 and 10 (ZIP)

Excel spreadsheet GESSE_Supporting_Information providing a detailed analysis of the highest correlated drugs−targets−SEs (XLSX)

AUTHOR INFORMATION

Corresponding Author

*Tel: +33-354 958 604. Fax: +33-383 593 046. E-mail: pereznueno@harmonicpharma.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Cepos Insilico Ltd. for providing an Academic License for PARASURF and ChemAxon for a license for the Marvin and JChem toolkits.

REFERENCES

(1) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing Safety-Related Drug Attrition: The Use of in Vitro Pharmacological Profiling. *Nat. Rev. Drug Discovery* **2012**, *11*, 909−922.

(2) Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development. *Drug Discovery Today* **2005**, *10*, 1421−1433.

(3) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263−266.

(4) Lee, S.; Lee, K. H.; Song, M.; Lee, D. Building the Process-Drug-Side Effect Network to Discover the Relationship between Biological Processes and Side Effects. *BMC Bioinf.* **2011**, *12*, S2.

(5) Cheng, F.; Li, W.; Wu, Z.; Wang, X.; Zhang, C.; Li, J.; Liu, G.; Tang, Y. Prediction of Polypharmacological Profiles of Drugs by the Integration of Chemical, Side Effect, and Therapeutic Space. *J. Chem. Inf. Model.* **2013**, *53*, 753−762.

(6) Yang, L.; Agarwal, P. Systematic Drug Repositioning Based on Clinical Side-Effects. *PLoS One* **2011**, *6*, e28025.

(7) Scheiber, J.; Chen, B.; Milik, M.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L. Gaining Insight into off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 308−317.

(8) Xie, L.; Li, J.; Xie, L.; Bourne, P. E. Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network to Explain the Side Effects of CETP Inhibitors. *PLoS Comput. Biol.* **2009**, *5*, e1000387.

(9) Wallach, I.; Jaitly, N.; Lilien, R. A Structure-Based Approach for Mapping Adverse Drug Reactions to the Perturbation of Underlying Biological Pathways. *PLoS One* **2010**, *5*, e12063.

(10) Takarabe, M.; Shigemizu, D.; Kotera, M.; Goto, S.; Kanehisa, M. Network-Based Analysis and Characterization of Adverse Drug-Drug Interactions. *J. Chem. Inf. Model.* **2011**, *51*, 2977−2985.

(11) Takarabe, M.; Kotera, M.; Nishimura, Y.; Goto, S.; Yamanishi, Y. Drug Target Prediction Using Adverse Event Report Systems: A Pharmacogenomic Approach. *Bioinformatics* **2012**, *28*, i611−i618.

(12) Scheiber, J.; Jenkins, J. L.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Milik, M.; Azzaoui, K.; Whitebread, S.; Hamon, J.; Urban, L.; Glick, M.; Davies, J. W. Mapping Adverse Drug Reactions in Chemical Space. *J. Med. Chem.* **2009**, *52*, 3103−3107.

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

(13) Simon, Z.; Peragovics, A.; Vigh-Smeller, M.; Csukly, G.; Tombor, L.; Yang, Z.; Zahoránszky-Kohalmi, G.; Végner, L.; Jelinek, B.; Hári, P.; Hetényi, C.; Bitter, I.; Czobor, P.; Málnási-Csizmadia, A. Drug Effect Prediction by Polypharmacology-Based Interaction Profiling. *J. Chem. Inf. Model.* **2012**, *52*, 134−145.
(14) 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*, i246−i254.
(15) Yamanishi, Y.; Pauwels, E.; Kotera, M. Drug Side-Effect Prediction Based on the Integration of Chemical and Biological Spaces. *J. Chem. Inf. Model.* **2012**, *52*, 3284−3292.
(16) Yamanishi, Y.; Kotera, M.; Moriya, Y.; Sawada, R.; Kanehisa, M.; Goto, S. DINIES: Drug-Target Interaction Network Inference Engine Based on Supervised Analysis. *Nucleic Acids Res.* **2014**, *42*, W39−W45.
(17) Pauwels, E.; Stoven, V.; Yamanishi, Y. Predicting Drug Side-Effect Profiles: A Chemical Fragment-Based Approach. *BMC Bioinf.* **2011**, *12*, 169−182.
(18) Mizutani, S.; Pauwels, E.; Stoven, V.; Goto, S.; Yamanishi, Y. Relating Drug-Protein Interaction Network with Drug Side Effects. *Bioinformatics* **2012**, *28*, i522−i528.
(19) Atias, N.; Sharan, R. An Algorithmic Framework for Predicting Side Effects of Drugs - Journal of Computational Biology. *J. Comput. Biol.* **2011**, *18*, 207−218.
(20) 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. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313*, 1929−1935.
(21) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A Side Effect Resource to Capture Phenotypic Effects of Drugs. *Mol. Syst. Biol.* **2010**, *6*, 343−348.
(22) Altman, R. B. PharmGKB: A Logical Home for Knowledge Relating Genotype to Drug Response Phenotype. *Nat. Genet.* **2007**, *39*, 426−428.
(23) Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for Representation and Analysis of Molecular Networks Involving Diseases and Drugs. *Nucleic Acids Res.* **2010**, *38*, D355−D360.
(24) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25−29.
(25) Medical Subject Headings Home Page. http://www.nlm.nih.gov/mesh/meshhome.html (accessed Dec 5, 2014).
(26) Duran-Frigola, M.; Aloy, P. Analysis of Chemical and Biological Features Yields Mechanistic Insights into Drug Side Effects. *Chem. Biol.* **2013**, *20*, 594−603.
(27) Mizutani, S.; Noro, Y.; Kotera, M.; Goto, S. Pharmacoepidemiological Characterization of Drug-Induced Adverse Reaction Clusters towards Understanding of Their Mechanisms. *Comput. Biol. Chem.* **2014**, *50*, 50−59.
(28) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. W. Detecting Drug Promiscuity Using Gaussian Ensemble Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1948−1961.
(29) Pérez-Nueno, V. I.; Karaboga, A. S.; Souchet, M.; Ritchie, D. W. GES Polypharmacology Fingerprints: A Novel Approach for Drug Repositioning. *J. Chem. Inf. Model.* **2014**, *54*, 720−734.
(30) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901−D906.
(31) Bresso, E.; Grisoni, R.; Marchetti, G.; Karaboga, A. S.; Souchet, M.; Devignes, M.-D.; Smaïl-Tabbone, M. Integrative Relational Machine-Learning for Understanding Drug Side-Effect Profiles. *BMC Bioinf.* **2013**, *14*, 207−218.
(32) Bresso, E.; Benabderrahmane, S.; Smail-Tabbone, M.; Marchetti, G.; Karaboga, A. S.; Souchet, M.; Napoli, A.; Devignes, M.-D. Use of Domain Knowledge for Dimension Reduction—Application to Mining of Drug Side Effects. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*; SCITEPRESS Digital Library, 2011; pp 271−276.
(33) Benabderrahmane, S.; Smail-Tabbone, M.; Poch, O.; Napoli, A.; Devignes, M.-D. IntelliGO: A New Vector-Based Semantic Similarity Measure Including Annotation Origin. *BMC Bioinf.* **2010**, *11*, 588−604.
(34) Karaboga, A. S.; Petronin, F.; Marchetti, G.; Souchet, M.; Maigret, B. Benchmarking of HPCC: A Novel 3D Molecular Representation Combining Shape and Pharmacophoric Descriptors for Efficient Molecular Similarity Assessments. *J. Mol. Graphics Modell.* **2013**, *41*, 20−30.
(35) Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering Gene Expression Patterns. *J. Comput. Biol.* **1999**, *6*, 281−297.
(36) Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **1936**, *28*, 321−377.
(37) Witten, D. M.; Tibshirani, R.; Hastie, T. A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis. *Biostatistics* **2009**, *10*, 515−534.
(38) Lê Cao, K.-A.; González, I.; Déjean, S. integrOmics: An R Package to Unravel Relationships between Two Omics Datasets. *Bioinformatics* **2009**, *25*, 2855−2856.
(39) Kullback, S. *Information Theory and Statistics*; John Wiley and Sons: New York, 1959.
(40) Aho, K. Asbio: A Collection of Statistical Tools for Biologists. R Package, version 0.3-39. http://cran.r-project.org/web/packages/asbio/asbio.pdf (accessed January 2015).
(41) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*, 4th ed.; Statistics and Computing Series; Springer: New York, 2002.
(42) Sanchez, G. DiscriMiner: Tools of the Trade for Discriminant Analysis. R Package, version 0.1-29. http://cran.r-project.org/web/packages/DiscriMiner/DiscriMiner.pdf (accessed January 2015).
(43) Garcia, A. G.; Horga de la Parte, J. F. Reacciones Adversas a los Farmacos. In *Indice de especialidades S. A.*; INTERCON, Ed.; Sociedad Española de Farmacología, Fundación Teófilo Hernando: Madrid, Spain, 1994.
(44) Chen, B.; Wild, D.; Guha, R. PubChem as a Source of Polypharmacology. *J. Chem. Inf. Model.* **2009**, *49*, 2044−2055.
(45) Pérez-Nueno, V. I.; Ritchie, D. W.; Borrell, J. I.; Teixidó, J. Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146−2165.
(46) Pérez-Nueno, V. I.; Ritchie, D. W. Using Consensus-Shape Clustering to Identify Promiscuous Ligands and Protein Targets and to Choose the Right Query for Shape-Based Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 1233−1248.
(47) Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, Information, Knowledge and Principle: Back to Metabolism in KEGG. *Nucleic Acids Res.* **2014**, *42*, D199−D205.
(48) Milacic, M.; Haw, R.; Rothfels, K.; Wu, G.; Croft, D.; Hermjakob, H.; D'Eustachio, P.; Stein, L. Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome. *Cancers* **2012**, *4*, 1180−1211.
(49) Online Mendelian Inheritance in Man (OMIM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD. Updated 18 May 2015. http://omim.org/ (accessed January 2015).
(50) Burnham, T. K. Porphyria, Diabetes, and Their Relationship. *Arch. Dermatol.* **1961**, *83*, 717−722.
(51) Chakrabarty, A.; Norman, R. A.; Phillips, T. J. Cutaneous Manifestations of Diabetes. *Wounds* **2002**, *14*, 267−274.
(52) Vigorita, V. J.; Ghelman, B.; Mintz, D. *Orthopaedic Pathology*, 2nd ed.; Lippincott Williams & Wilkins: Philadelphia, 2008.
(53) Logié, A.; Dunois-Lardé, C.; Rosty, C.; Levrel, O.; Blanche, M.; Ribeiro, A.; Gasc, J.-M.; Jorcano, J.; Werner, S.; Sastre-Garau, X.; Thiery, J. P.; Radvanyi, F. Activating Mutations of the Tyrosine Kinase Receptor FGFR3 Are Associated with Benign Skin Tumors in Mice and Humans. *Hum. Mol. Genet.* **2005**, *14*, 1153−1160.

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823

(54) Van Wyk, J. J.; Smith, E. P. Insulin-Like Growth Factors and Skeletal Growth: Possibilities for Therapeutic Interventions. *J. Clin. Endocrinol. Metab.* **1999**, *84*, 4349−4354.
(55) Rancourt, C.; Rogers, B. E.; Sosnowski, B. A; Wang, M.; Piché, A.; Pierce, G. F.; Alvarez, R. D.; Siegal, G. P.; Douglas, J. T.; Curiel, D. T. Basic Fibroblast Growth Factor Enhancement of Adenovirus-Mediated Delivery of the Herpes Simplex Virus Thymidine Kinase Gene Results in Augmented Therapeutic Benefit in a Murine Model of Ovarian Cancer. *Clin. Cancer Res.* **1998**, *4*, 2455−2461.

DOI: 10.1021/acs.jcim.5b00120
J. Chem. Inf. Model. 2015, 55, 1804−1823