Journal of proteome research

Article

pubs.acs.org/jpr

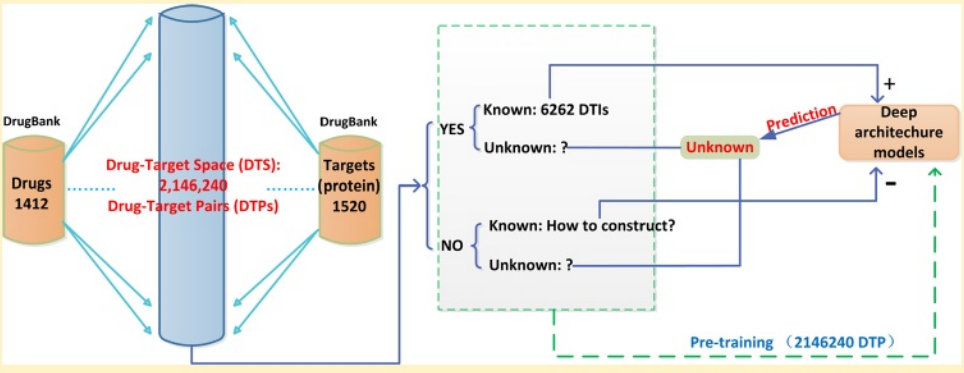
基于深度学习的药物-靶点相互作用预测

Ming Wen, † Zhimin Zhang, † Shaoyu Niu, † Haozhi Sha, † Ruihan Yang, † Yonghuan Yun, ‡ and Hongmei Lu*, †

† 中南大学化学化工学院，中国长沙 410083

‡ 中国热带农业科学院环境与植物保护研究所，中国海口 571101

* 支持信息



摘要：在药物再定位中，识别已知药物与靶点之间的相互作用是一项重大挑战。通过计算机模拟预测药物 - 靶点相互作用（DTI）能够提供最有效的 DTI，从而加快昂贵且耗时的实验工作。计算机模拟预测 DTI 还能为潜在的药物 - 药物相互作用提供见解，并促进对药物副作用的探索。传统上，DTI 预测的性能在很大程度上取决于用于表示药物和靶蛋白的描述符。在本文中，为了准确预测已批准药物与靶点之间新的 DTI，且无需将靶点分类，我们开发了一种基于深度学习的算法框架，名为 DeepDTIs。它首先使用无监督预训练从原始输入描述符中提取表示，然后应用已知的相互作用标签对来构建分类模型。与其他方法相比，发现 DeepDTIs 达到或超过了其他最先进的方法。DeepDTIs 还可用于预测新药是否针对某些现有靶点，或者新靶点是否与某些现有药物相互作用。

关键词：深度学习、深度信念网络、特征提取、药物-靶点相互作用预测、半监督学习

引言

药物与靶点相互作用的识别是药物发现和药物再定位的关键领域。[1,2] 由于已获批药物具有明确的可获得性和已知的安全性特征，将药物重新用于新的适应症不仅能够降低药物开发成本，还能降低药物安全性风险。[3] 尽管有各种生物学检测技术可用，但大规模药物 - 靶点相互作用实验仍存在局限性。此外，实验成本极高且公开的药物再定位检测数据很少，因此有必要开发能够精确检测药物与靶点相互作用的适当计算工具。

如今，已提出了许多计算机模拟方法来识别新的药物 - 靶点相互作用（DTI）。基于配体和基于结构的方法是两种最常用的方法。[4] 最为人熟知的基于配体的方法是应用定量构效关系（QSAR）来预测分子对靶点的生物活性。QSAR 基于这样一个假设，即结构相似的分子具有相似的生物活性。[5] 给定一个确定一定数量的目标，每个目标利用其已知的活性分子构建预测模型。然后利用这些构建好的模型对所有药物进行筛选，以预测药物与目标之间的相互作用。不幸的是，如果某个目标已知的活性分子数量不足，所构建的定量构效关系（QSAR）模型的性能就会很差，并且大多数 QSAR 模型缺乏特异性，或者仅能预测对单一目标的活性。为解决这一问题，已有一些方法被应用，例如构建多靶点 QSAR（mt-QSAR）分类模型。与基于配体的方法不同，基于结构的方法（即分子对接）利用目标的晶体结构来筛选小分子。当目标的三维（3D）结构信息可用时，分子对接方法是精确的。然而，对于大多数目标，尤其是膜蛋白（如 G 蛋白偶联受体），其 3D 结构信息目前仍不可用。近来，提出了多种基于网络的方法来推断药物 - 目标相互作用。在

Received: July 3, 2016
Published: March 6, 2017

DOI: 10.1021/acs.jproteome.6b00618
J. Proteome Res. 2017, 16, 1401−1409

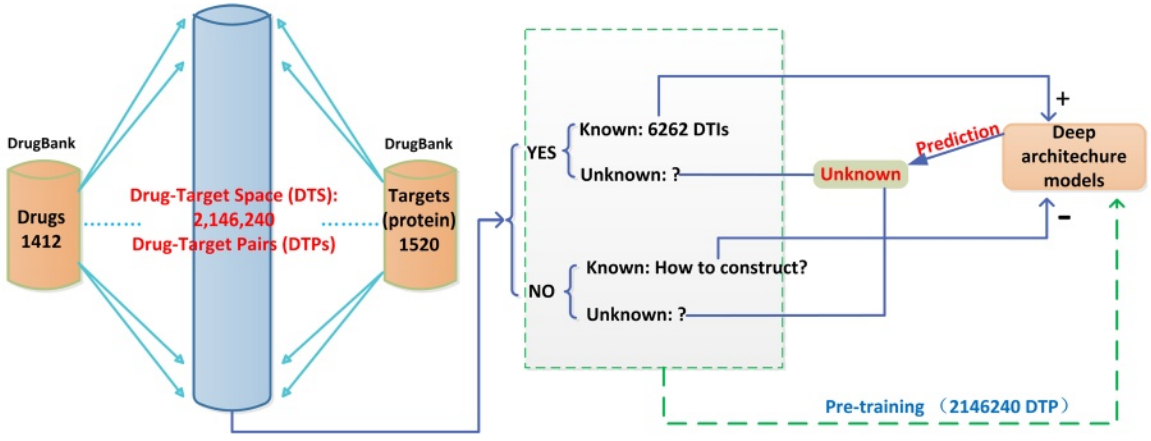


图 1.DeepDTIs 流程图

药物 - 靶点相互作用网络中，药物和靶点由节点表示；已知的药物与靶点相互作用则对应连接节点的连线。新的药物 - 靶点相互作用（DTIs）是从已知网络中推断出来的。例如，Cheng 等人 [12] 开发了一种基于网络的推断模型（NBI）来推断新的 DTIs。NBI 仅基于药物 - 靶点二分网络拓扑相似性。使用了一个评分函数来评估药物与靶点之间的关联。NBI 的一个缺点是，对于训练集中没有任何已知靶点信息的新药，它无法应用。还可以整合其他一些网络来提高网络方法的性能。例如，Chen 等人 [13] 使用了一个异质网络，该网络整合了药物 - 药物相似性网络、蛋白质 - 蛋白质相似性网络和药物 - 靶点相互作用网络，开发了一个名为 NRWRH 的有效模型。与仅使用一个网络相比，其性能有了显著提升。此外，诸如药物副作用相似性网络 [14] 和疾病相关分子网络 [15] 等网络也可用于推断 DTIs。近年来，随着实验数据的增多，众多机器学习方法被用于预测药物 - 靶点相互作用（DTIs）。常用的机器学习方法是构建分类模型，以药物 - 靶点对（DTPs）作为输入，输出为该药物 - 靶点对（DTP）之间是否存在相互作用。应用最广泛的机器学习模型是二分类器，如随机森林（RF）、支持向量机（SVM）和人工神经网络（ANN）。深度学习方法是一种具有多个隐藏层和更复杂参数训练过程的人工神经网络。由于其相对较好的性能以及能够学习具有多层抽象的数据表示，深度学习方法受到了广泛关注。深度学习已在生物学和化学的许多领域得到应用。例如，Frey 等人将一种名为 DeepBind 的深度学习方法应用于预测 DNA 和 RNA 结合蛋白的序列特异性任务。结果表明，即使在使用体外数据训练和体内数据测试的情况下，深度学习也优于其他最先进的方法。程等人 [21] 开发了一种深度学习网络方法（DN-Fold），极大地提高了蛋白质折叠识别的性能（即预测给定的查询模板蛋白质对是否属于相同的结构折叠）。

近来，随着诸如高通量实验和下一代测序等众多实验仪器和技术的发展，出现了整合多种资源以获取更多药物靶点相互作用（DTIs）信息的趋势。例如，Nascimento 等人 1 将多种异质信息源整合起来以识别新的药物靶点相互作用。Yamanishi 等人 22 利用核方法预测药物靶点相互作用，该方法整合了多种信息源，并测量了药物和靶点的相似性。整合不同资源的优点在于，它能够检测到计算特征未体现的信息，或者检测到其他资源未保存的信息。然而，如果这些资源并非权威来源，这可能会带来错误。此外，从这些资源中提取特征需要更多的领域知识。例如，手头有几十种化学结构和蛋白质序列描述符时，很难决定在特定任务中哪种描述符是有用的。23从原始数据中提取和组织信息的能力以及学习具有多层抽象的数据表示的能力，使得深度学习方法在药物 - 靶点相互作用（DTIs）预测方面特别有效。此前，大多数方法都使用了由 Yamanishi 等人首次提出的“黄金标准”数据集来评估其方法。在该数据集中，目标被分为四类：酶、离子通道、G 蛋白偶联受体（GPCR）和核受体。对应的蛋白质数量分别为 664、204、95 和 26 种。相应的相互作用数量分别为 1515、776、314 和 44 种。这些数据集在许多研究中被广泛使用，并达到了相对较高的准确率。然而，基于不同类别目标构建的模型，无法预测药物的原始目标和新目标属于不同类别时的药物 - 目标相互作用（DTI）。例如，血清素和血清素能药物既能与 5-HT（一种 G 蛋白偶联受体蛋白）又能与 59HT3A（一种离子通道蛋白）相互作用。5-HT 和 59HT3A 属于不同的类别。如果 G 蛋白偶联受体类别包含血清素而离子通道类别不包含，那么使用 G 蛋白偶联受体类别模型就无法预测血清素与 59HT3A 的相互作用。因此，所构建的模型应用范围相对有限，应当重新考虑使用一个全局数据集。

在本文中，一种有效的深度学习方法——深度信念网络（DBN）被用于准确预测已获美国食品药品监督管理局（FDA）批准的药物与靶点之间新的药物-靶点相互作用（DTIs），且无需将靶点划分为不同的类别。所开发的方法被命名为 DeepDTIs。药物和靶点的特征是从简单的化学亚结构和序列顺序信息中自动提取的。据我们所知，这是首次将深度学习方法用于预测药物-靶点相互作用。我们使用独立测试集对我们的方法进行了测试，并与一些可接受的算法（如随机森林（RF）、伯努利朴素贝叶斯）进行了比较。

DOI: 10.1021/acs.jproteome.6b00618
J. Proteome Res. 2017, 16, 1401−1409

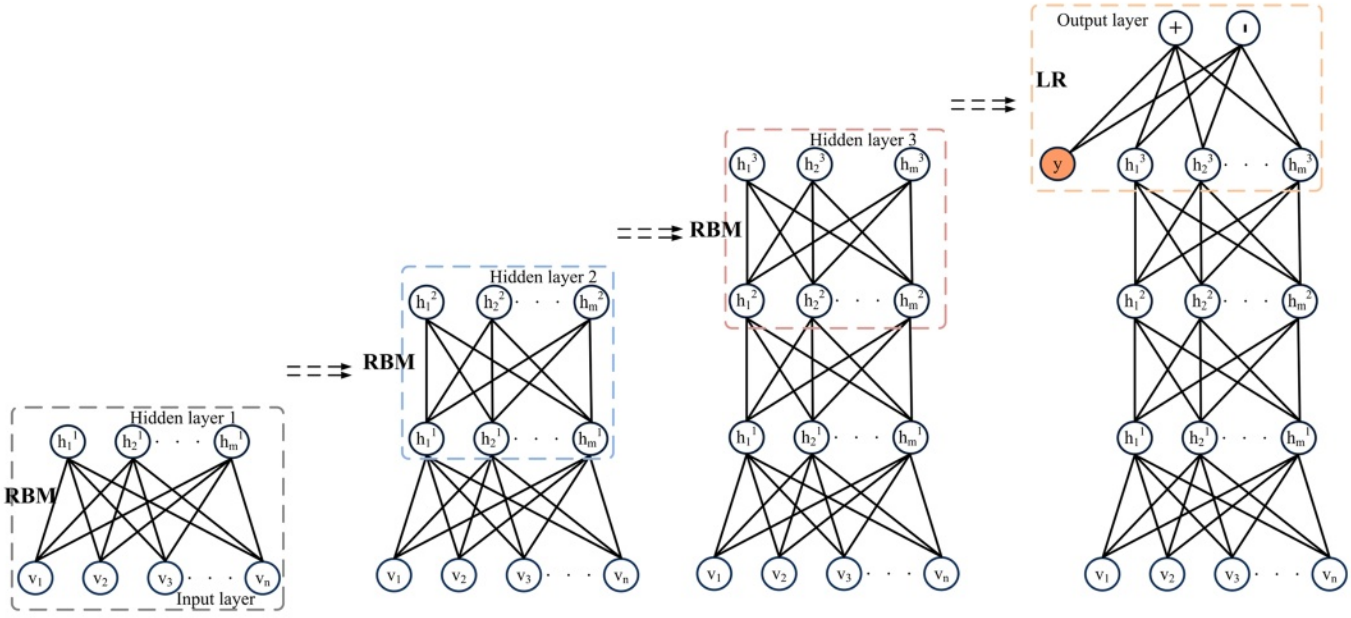


图 2.深度信念网络（DBN）架构。DBN 是通过堆叠多个受限玻尔兹曼机（RBM）构建而成的。

此外，我们还采用了 BNB（朴素贝叶斯）和 DT（决策树）算法。[26] 此外，我们的算法还在从 DrugBank 数据库中提取的外部 EDTIs 数据库（实验性药物 - 靶点相互作用数据库）上进行了测试。最后，通过 DeepDTIs 成功预测了药物 - 靶点空间（DTS）中所有可能的相互作用，并且在文献中已知实验的基础上对预测出的 10 种最有可能的药物 - 靶点相互作用进行了部分验证。

■ 方法

数据集与药物靶点空间（DTS）

药物和靶点数据是从 DrugBank 数据库（http://www.drugbank.ca/）中提取的。[27] DrugBank 数据库是一个独特的生物信息学和化学信息学资源，它将详细的药物数据与全面的药物靶点信息相结合。本研究中使用的数据于 2016 年 2 月 15 日发布。药物与靶点的相互作用是从 DrugBank 下载网站（https://www.drugbank.ca/releases/latest#protein-identifiers）的“药物靶点标识符”类别下的“蛋白质标识符”中下载的。已批准的药物结构和已批准的靶点序列分别从 https://www.drugbank.ca/releases/latest#structures 和 https://www.drugbank.ca/releases/latest#target-sequences 下载。我们手动剔除了无机化合物（如锂（DB01356））或非常小的分子化合物（如二氧化碳（DB09157））类药物。计算了所有已批准药物结构和靶点序列的描述符。如果下载的药物 - 靶点相互作用对具有计算出的蛋白质和药物描述符，则用于构建模型。这些数据集也可在 https://github.com/Bjoux2/DeepDTIs 上获取。药物 - 靶点空间（DTS）被定义为所有可能的药物 - 靶点对（DTPs）。如图 1 所示，DTS 包含 1412 种药物和 1520 个靶点。DTS 共有 2146240（即 1412×1520）个 DTPs。其中，如图 1 所示，有些对是已知的药物 - 靶点相互作用（标记为“是”），有些对则不是（标记为“否”）。目前，已知相互作用的 DTPs 有 6262 对，其余的则未知。已知的 6262 对 DTPs 被设为正数据集。由于无相互作用的对数远多于有相互作用的对数，因此可以从 DTS 中随机选取负数据集。在本研究中，我们从 DTS 中随机选取了 6262 对 DTPs 作为负数据集。因此，整个数据集共有 12524 个样本。

实验性药物数据集

我们使用了一个外部实验性药物 - 靶点对（EDTPs）数据集来测试我们的模型。该数据集源自 DrugBank。其下载过程和网站与上述训练数据集类似。实验性药物 - 靶点对由实验性药物和实验性靶点组成，或者由实验性药物和已批准的靶点组成。实验性药物是指已被实验表明能与特定蛋白质结合的药物（http://www.drugbank.ca/documentation）。尽管 EDTPs 还不是真正的药物 - 靶点相互作用（DTIs），但它们成为 DTIs 的可能性很高。对这些 EDTPs 的预测可以在一定程度上评估我们模型的性能。EDTPs 数据集包含 7352 个实验性药物 - 靶点相互作用对。该数据集包含 2528 个靶点和 4383 种实验性药物。在这 7352 对中，2444 对被设为 EDTPs1，包含 504 个靶点（这些靶点包含在训练数据集中）和 2003 种药物；4908 对被设为 EDTPs2，包含 2818 种药物和 2024 个靶点（这些靶点未包含在训练数据集中）。在 EDTPs2 中，这些药物和靶点与训练集中的不同，被用于评估模型的适用性。

化学结构与蛋白质序列表示

化学化合物的表示方法可分为两类：分子描述符（MDs）和分子指纹（MFs）。分子描述符是通过实验定义或理论推导得出的分子属性。分子指纹是分子的属性概要。分子指纹通常由位向量或计数向量表示，向量元素分别表示某些属性的存在与否或出现频率。[28] 在本文的其余部分，为避免混淆，我们将用“特征”一词来同时表示描述符和指纹。

在本研究中，我们选择了最常见且简单的特征：药物的扩展连接指纹（ECFP）和蛋白质序列组成描述符（PSC）。

DOI: 10.1021/acs.jproteome.6b00618 J. Proteome Res. 2017, 16, 1401−1409

目标表示。ECFP 是一类拓扑指纹，用于表示特定子结构的存在。[29] ECFP 描述了由每个原子及其直径范围内的圆形邻域组成的子结构的特征。我们结合了 ECFP2、ECFP4 和 ECFP6（分别对应直径为 2、4、6），以确保不丢失结构信息。每个化合物有 6144 个指纹。PSC 包含氨基酸组成（AAC）、二肽组成（DC）和三肽组成（TC）。AAC 是每个氨基酸的统计频率。DC 是每两个氨基酸组合的统计频率。TC 是每三个氨基酸组合的统计频率。每个蛋白质序列有 8420 个描述符。与上述分子和蛋白质的其他特征不同，ECFP 和 PSC 是两个简单的特征，仅包含分子子结构和蛋白质子序列。它们包含了分子结构和序列顺序的信息。最终，每个 DTP 有 14564 个特征。分子指纹和蛋白质描述符分别使用开源化学信息学软件 rdkit 和蛋白质描述符生成器 propy 进行计算。

受限玻尔兹曼机

受限玻尔兹曼机（RBM）是一种图形模型，能够从输入数据中学习概率分布。如图 2 所示，RBM 由两层组成：一层可见层和一层隐藏层。每个可见单元都与整个隐藏层单元相连。同一层内不存在可见单元与可见单元之间以及隐藏单元与隐藏单元之间的连接。对于给定的 RBM 网络，RBM 的能量 E（v,h） 定义为

$$E(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta}) = -\mathbf{b}'\mathbf{v} - \mathbf{c}'\mathbf{h} - \mathbf{h}'\mathbf{W}\mathbf{v} \quad (1)$$

其中 θ = {W,b,c}。W 表示连接隐藏层和可见层单元的权重。b 和 c 分别是可见层和隐藏层的偏置。当参数 θ 给定后，基于 E（v,h），（v,h） 的概率分布为

$$P(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta}) = \frac{1}{z(\theta)} e^{-E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})} \quad (2)$$

$$z(\boldsymbol{\theta}) = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})} \quad (3)$$

其中，\（z(θ)\） 是一个归一化因子，由于与物理系统中的情况类似，所以被称为配分函数。网络赋予可见层 \（v\） 的概率是通过对所有可能的隐藏向量求和来给出的。

$$P(\mathbf{v}|\boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{\theta})} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})} \quad (4)$$

这个函数也被称为似然函数。

受限玻尔兹曼机（RBM）模型可以通过对训练数据的经验负对数似然进行随机梯度下降（SGD）来学习。损失函数被定义为负对数似然函数。

$$\mathbf{L}(\theta, T) = -\frac{1}{N} \sum_{\mathbf{v}\in T} \log P(\mathbf{v}|\boldsymbol{\theta}) \quad (5)$$

其中 T 是随机梯度下降（SGD）中使用的样本集。然后我们通过以下方式更新 θ：

$$\mathbf{W} \leftarrow \mathbf{W} - \frac{\partial \mathbf{L}(\theta, T)}{\partial \mathbf{W}} \quad (6)$$

$$\mathbf{h} \leftarrow \mathbf{h} - \frac{\partial \mathbf{L}(\theta, T)}{\partial \mathbf{h}} \quad (8)$$

深度信念网络

深度信念网络（DBN）是一种神经网络，由多层受限玻尔兹曼机（RBM）堆叠而成，并以贪婪方式训练。图 2 展示了 DBN 模型的架构。它由 5 层组成。第一层是输入层，包含计算得出的特征。第二、第三和第四层是隐藏层。最后一层是输出层。除最后一对层外，每相邻两层构成一个 RBM。DBN 是一种图形模型，用于学习提取训练数据的深度分层特征。它将训练样本向量 x 与 l 个隐藏层之间的联合分布建模如下：

$$P(\mathbf{x}, \mathbf{h}^1, ..., \mathbf{h}^l) = (\prod_{k=0}^{l-2} P(\mathbf{h}^k|\mathbf{h}^{k+1}))P(\mathbf{h}^{l-1}, \mathbf{h}^l)$$

其中 x = h⓪，$P(h^{k-1}|h^k)$ 是条件分布，表示在第 k 层受限玻尔兹曼机（RBM）中，可见单元在给定隐藏单元条件下的分布，而$P(h^{l-1},h^l)$ 是顶层 RBM 中可见单元与隐藏单元的联合分布。

深度置信网络（DBN）的训练过程可以分为两个连续的步骤：贪婪逐层无监督训练过程和有监督微调过程。贪婪逐层无监督训练过程如下：

1. 通过随机生成器初始化参数 W、b 和 c。
2. 将第一层和第二层作为受限玻尔兹曼机（RBM）进行训练。使用原始输入向量 x 作为其可见层。
3. 将第二层和第三层作为受限玻尔兹曼机（RBM）进行训练，将第二层作为可见层，并获取第三层的表示。对于所需的层数重复此操作。

监督微调过程如下：

1. 将深度置信网络（DBN）的最后一个隐藏层的输出作为逻辑回归分类器（LR）的输入。
2. 通过监督随机梯度下降法对深度信念网络对数似然代价进行微调，以优化所有受限玻尔兹曼机（RBM）和逻辑回归（LR）的参数。

预测质量的度量

本研究中使用了四个常用的评估指标——受试者工作特征曲线下面积（AUC）、准确率（ACC）、真正例率（TPR，灵敏度/召回率）和假正例率（FPR，特异度）来评估模型的性能。ACC、TPR 和 FPR 的计算公式如下：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

其中 TP、FP、TN 和 FN 分别代表真阳性、假阳性、真阴性和假阴性。在二分类预测问题中，结果被标记为正类（p）或负类（n）。如果预测值和实际值均为正类（p），则称为真阳性（TP）；如果预测值为正类（p）而实际值为负类（n），则称为假阳性（FP）。相反，如果预测值和实际值均为负类（n），则称为真阴性（TN）；如果预测值为负类（n）而实际值为正类（p），则称为假阴性（FN）。

DOI: 10.1021/acs.jproteome.6b00618 J. Proteome Res. 2017, 16, 1401−1409

对于所有预测值为 n 而实际值为 n 的情况，称为真阴性（TN）；如果预测值为 n 而实际值为 p，则称为假阴性（FN）。

实施

DBN 算法是在 Python（版本 2.7）中实现的，使用了著名的 DeepLearningTutorials 包（https://github.com/lisa-lab/DeepLearningTutorials）。该算法基于 Theano 编写。[34] 该算法在 GPU（GEFORCE GTX-TITAN-X 6GD5）上使用 CUDA 进行加速。操作系统为 Ubuntu Kylin 15.04，配备 4.4 GHz Intel 酷睿 i7 处理器和 32G 内存。代码可在 https://github.com/Bjoux2/DeepDTIs 网站免费获取。

■ 结果

确定深度药物相互作用模型的架构

DeepDTIs 有 3 个超参数：（i）小批量大小（mbs）；（ii）预训练学习率（plr）；（iii）微调学习率

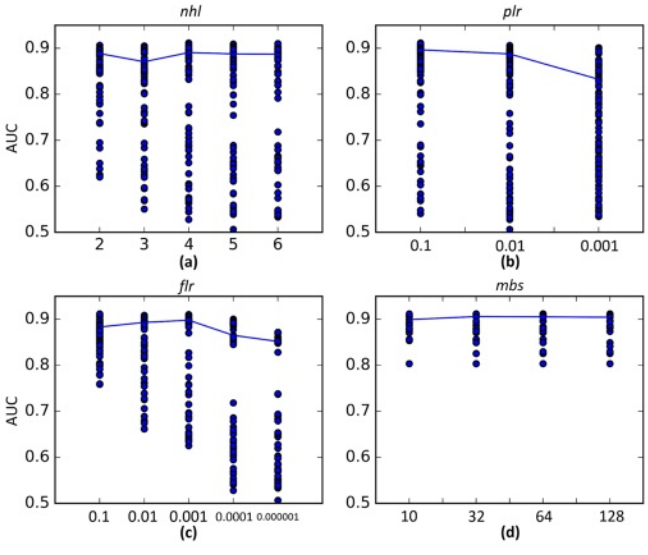


图 3.参数值与模型性能的关系图。

调整学习率（flr）。DeepDTIs 的架构主要由两个因素决定：（i）隐藏层的数量（nhl）和（ii）每层的节点数量（nnl）。这些超参数和因素是通过在验证误差上进行网格搜索来确定的（在本文的其余部分，为避免歧义，我们使用参数来表示上述提到的超参数和因素）。在网格搜索中，mbs 的取值范围为 [10, 32, 64, 128]；plr 的取值范围为 [10−1, 10−2, 10−3]；flr 的取值范围为 [10−1, 10−2, 10−3, 10−5]；nhl 的取值范围为 [2, 3, 4, 5, 6]。为了减少网格中的点数，我们任意将 nnl 设为 2000。因此，总共构建了 300（4×3×5×5×1）个模型来优化这些参数。为了确定每个参数对 DeepDTIs 性能的影响是否呈线性关系，绘制了参数与模型性能的关系图。图 3a 是 nhl 与 300 个模型性能的关系图，其中每个点代表在特定 nhl 下模型的性能。图中的线代表不同 nhl 下性能排名前 25%（从大到小排序）的模型。从图 3a 可以看出，模型性能与 nhl 之间不存在线性关系。与 nhl 类似，plr、flr 和 mbs（图 3b - d）与模型性能之间也没有明显的关联。每个参数值都能实现较高的性能。因此，仅凭经验难以确定模型架构和训练参数，建议采用网格搜索策略来优化模型架构和训练参数。网格搜索结果见补充表 S1。优化后的 plr、flr、nhl 和 mbs 分别为 0.1、0.1、4 和 128。

整体性能

整个数据集（12524 个样本）被分为三个子集，分别是训练集、验证集和测试集，其比例分别为 0.6（7514 个样本）、0.2（2505 个样本）和 0.2（2505 个样本）。训练集用于预训练和微调模型。验证集用于优化参数。测试集用于评估模型。此外，为了评估每种方法的稳健性并避免从 DTS 随机生成的数据中获得偶然结果，我们随机生成了 10 个负样本，并给出了平均结果。DBN 模型的性能列于表 1 中。测试集的 AUC、准确率、灵敏度和特异度分别为 0.9158、0.8588、0.8227 和 0.8953。此外，为了将我们的模型与其他机器学习方法进行比较，还使用了伯努利朴素贝叶斯（BNB）、决策树（DT）和随机森林（RF）来构建分类模型。这些方法的性能也列于表 1 中。从表 1 可以看出，DBN 在 AUC、ACC、TPR 和 TNR 方面均优于 BNB 和 DT（p 值 < 0.05）。由于在药物靶点相互作用数据集（DTS）中，正向药物靶点相互作用的数量远少于负向的，而模型的目的是预测真正的正向药物靶点相互作用，因此在四个评估指标中，真阳性率（TPR）是一个更重要的评估指标。与随机森林（RF）相比，尽管深度置信网络（DBN）的曲线下面积（AUC）仅比 RF 高 0.58%（p 值 < 0.05，DeepDTI 对比 RF），但其真阳性率（TPR）却比 RF 高 1.71%（p 值 < 0.05，DeepDTI 对比 RF）。图 4 展示了四种方法的真阳性率（TPR）、真阴性率（TNR）、准确率（ACC）和曲线下面积（AUC）的曲线图。从图 4 可以看出，10 次运行中的性能波动范围很小，这表明随机数据分割过程是稳定的。总体而言，DBN 在真阳性率（TPR）、真阴性率（TNR）、准确率（ACC）和曲线下面积（AUC）方面均表现最佳。这表明所构建的 DBN 模型是可靠的，并且可以进一步应用于新的药物靶点相互作用预测。

表 1.BNB、DT、RF 和 DeepDTIs 的总体性能

	TPR	TNR	ACC	AUC
BNBa	0.5913±0.0123	0.8654 ± 0.012	0.7272 ± 0.0620	0.7544 ± 0.007
DTa	0.7949 ± 0.0123	0.7418 ± 0.0335	0.7684 ± 0.0147	0.7683 ± 0.0153
RFb	0.8056 ± 0.0105	0.8636 ± 0.0108	0.8342 ± 0.0068	0.9100±0.0053
深度信念网络	0.8227 ± 0.0065	0.8953 ± 0.0130	0.8588 ± 0.0049	0.9158 ± 0.0059

[a]The 超参数设置为默认值。[b]The 在对数 2（q）、平方根（q）、q/3、q/2 和 q 中优化超参数最大特征值，其中 q 为变量数量。其他超参数设置为默认值。给定一个列表 X，其中包含使用 10 个负数据集得出的 10 个预测结果，最终结果表示为 mean(X) ± max(max(X) - mean(X), mean（X) - min(X))。

DOI: 10.1021/acs.jproteome.6b00618 J. Proteome Res. 2017, 16, 1401−1409

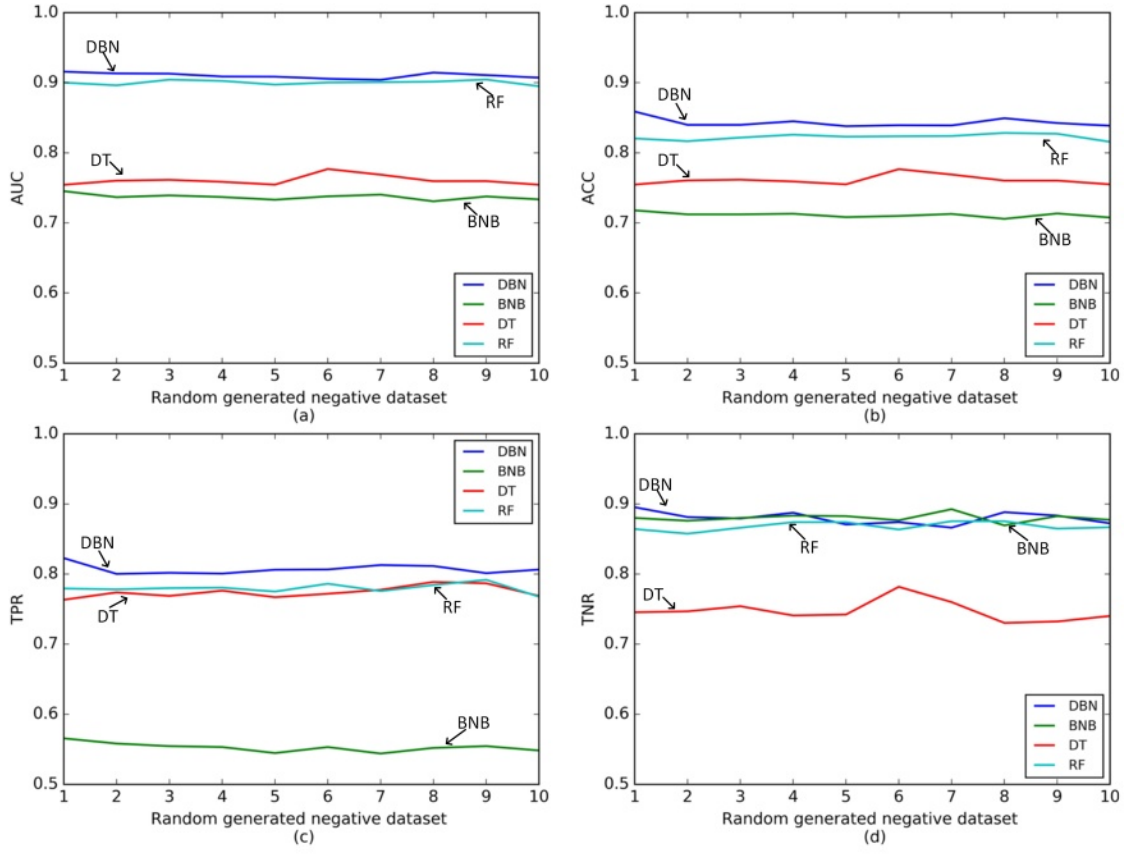


图 4.使用随机数据生成程序进行 10 次实验，4 种方法的真阳性率（TPR）、真阴性率（TNR）、准确率（ACC）和曲线下面积（AUC）。

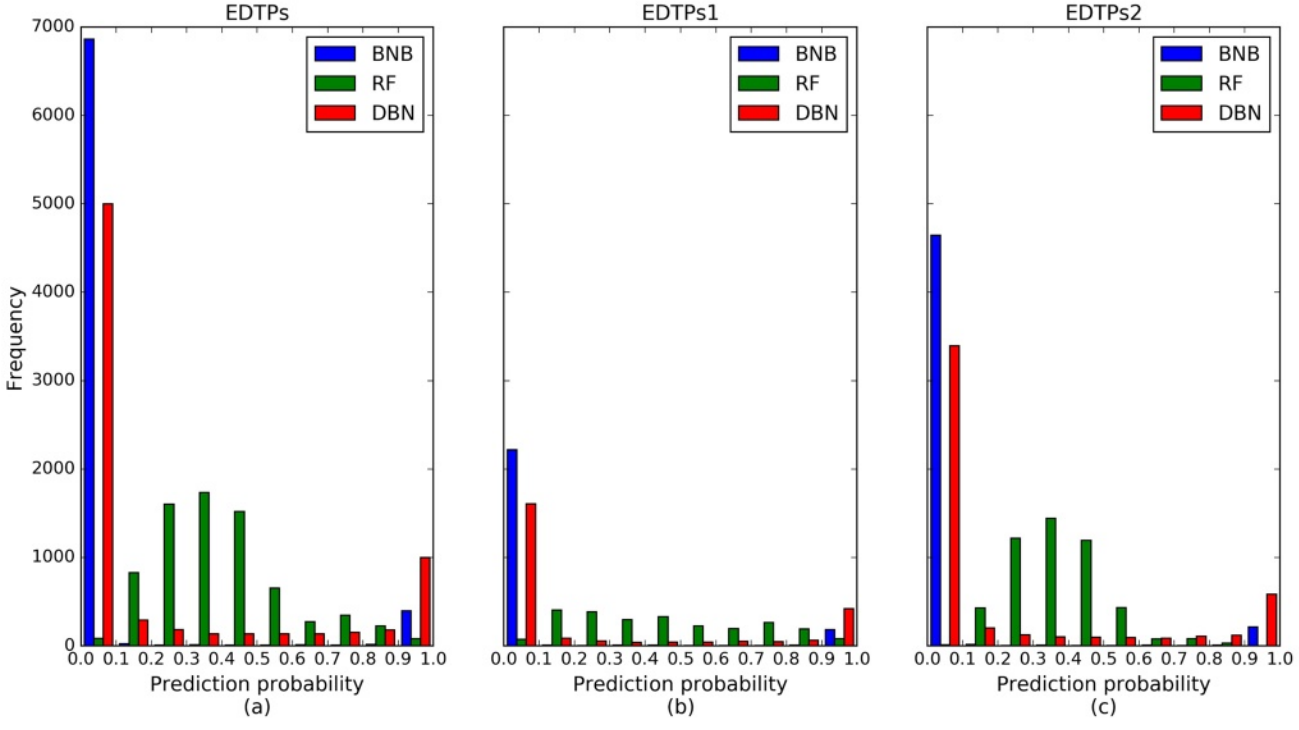


图 5.在 EDTPs、EDTPs1 和 EDTPs2 数据集中 BNB、RF 和 DBN 的预测概率分布。

EDTP 的性能及适用性

图 5 展示了 BNB、RF 和 DBN 在 EDTPs 数据集上的预测概率分布。预测概率越大，表明药物 - 靶点对具有正向相互作用的可信度越高。由于 DT 是单树模型，无法计算预测概率，其结果未在图 5 中展示。如图 5 所示，BNB 和 DBN 的预测概率倾向于接近 0 或 1，而 RF 则倾向于接近

DOI: 10.1021/acs.jproteome.6b00618
J. Proteome Res. 2017, 16, 1401−1409

表 2.BNB、RF 和 DBN 在 EDTPs1、EDTPs2 和 EDTPs 数据集上的预测结果

方法	数据集	NOSa	recall[b] (p > 0.5)c	召回（p > 0.9）d
币安币	EDTPs1	2444	202 (8.3%)	184 (7.5%)
	EDTPs[2]	4908	214 (4.4%)	236 (4.8%)
	EDTPs	7352	438 (5.9%)	398 (5.4%)
射频	EDTPs1	2444	956 (39.1%)	80 (3.3%)
	EDTPs[2]	4908	622 (12.7%)	0 (0%)
	EDTPs	7352	1578 (21.4%)	80 (1.1%)
深度信念网络	EDTPs1	2444	618 (25.3%)	419 (17.2%)
	EDTPs[2]	4908	988 (20.1%)	582 (11.9%)
	EDTPs	7352	1606 (21.8%)	1001 (13.6%)

[a]Number 样本的 [b]The 召回率定义为预测为正样本的对数/非药物靶点对数。[c]Drug−target 若预测概率大于 0.5，则该对被视为药物靶点相互作用（DTIs）。[d]Drug−target 若预测概率大于 0.9，则该对被视为药物靶点相互作用（DTIs）。

阈值设为 0.3。BNB、RF 和 DBN 的预测结果列于表 2 中。若将实验药物 - 靶点对的预测概率大于 0.5 设为正向药物 - 靶点相互作用（DTI），则 DBN 和 RF 在 EDTPs 中的召回率相似（分别为 21.8% 和 21.4%）。在 EDTPs1 中，RF 的召回率高于 DBN（表 2）。然而，从图 5 可以看出，RF 预测的大多数正向 DTI 并不十分确定，因为其预测概率接近 0.5。若将实验药物 - 靶点对的预测概率大于 0.9 设为正向 DTI，DBN 在所有 EDTPs1、EDTPs2 和 EDTPs 数据集中的召回率最高（远高于 RF 和 BNB）。此外，EDTPs2 中的药物和靶点在训练数据集中是新的，这表明 DeepDTIs 可用于预测新药是否作用于某些已知靶点，或者新靶点是否与某些已知药物相互作用。

新预测的相互作用

在确认了所构建模型的可靠性之后，我们对 DTS 中剩余的两百万对未知配对进行了预测，并根据其概率进行了排序。所有预测概率大于 0.9 的药物 - 靶点相互作用（DTIs）均列于补充表 S2 中。预测概率的分布情况如图 6 所示。如图 6 所示，大多数配对被预测为无相互作用。然而，88.15% 的配对具有预测概率。

表 3.我们模型预测的前 10 个概率评分最高的药物靶点相互作用（DTIs）

等级	对；双；成对的	描述	证据
1	DB00246，P08909	齐拉西酮，5-羟色胺受体 1C（2C）（HTR2C）	针
2	DB09016，P21728	布替林，D⊠A 多巴胺受体（DRD1）	-
3	DB00894，P04150	睾酮内酯，糖皮质激素受体（NR3C1）	-
4	DB01395，P04150	屈螺酮，糖皮质激素受体（NR3C1）	无互动
5	DB00404，P28476	阿普唑仑，γ-氨基丁酸受体亚基 ρ-2（GABRR2）	针
6	DB00363，P08909	氯氮平，5-羟色胺受体 1C（2C）（HTR2C）	针
7	DB00246，P41595	齐拉西酮，5-羟色胺受体 2B（HTR2B）	-
8	DB00831，P21728	三氟拉嗪，D⊠A 多巴胺受体（DRD1）	-
9	DB00990，P04150	依西美坦，糖皮质激素受体（NR3C1）	-
10	DB04839，P04150	醋酸环丙孕酮，糖皮质激素受体（NR3C1）	

概率小于 0.5。如果阈值设为 0.9，则仅有 5.53% 的配对被预测为正向药物 - 靶点相互作用（DTI）。这与事实相符，即无相互作用的配对数量远多于有相互作用的配对数量。表 3 展示了 DBN 预测的前 10 个概率最高的 DTI 列表。在前 10 个预测的 DTI 中，有 4 个在 STITCH 数据库或文献中被发现。文献中发现一种药物——屈螺酮（DB01395）与糖皮质激素受体（P04150）的相对结合亲和力较低。[35] 但在其余预测的 5 个 DTI 中，我们未从数据库和文献中找到任何实验证据，它们仍有可能是真正的正向 DTI。例如，齐拉西酮（DB00246）被发现对 5-羟色胺受体 1A、2A、1B、2C 和 1D 有作用，这些受体与 5-羟色胺受体 1C 具有很高的序列同源性。这些新预测的结果表明，DBN 模型在预测新的 DTI 方面具有实际应用价值，并且在药物再定位方面具有潜在的应用前景。

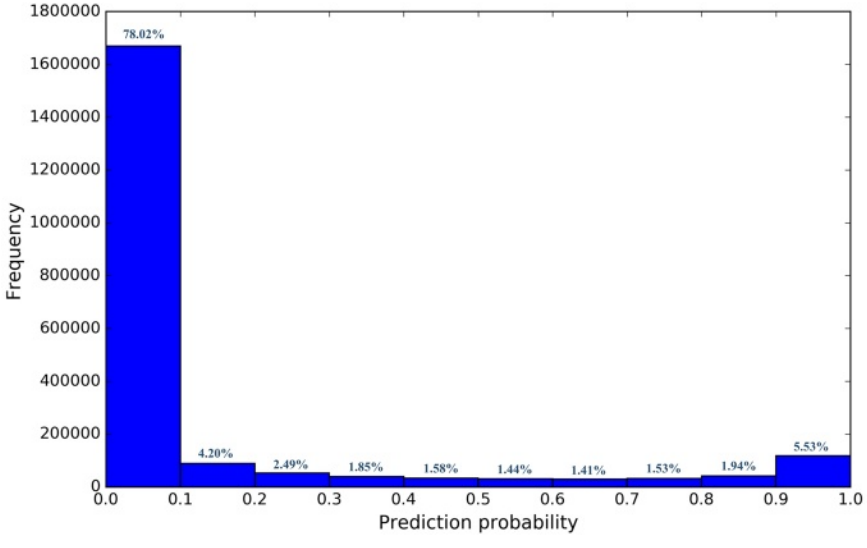


图 6. 通过深度信念网络（DBN）预测的 DTS 概率分布。

DOI: 10.1021/acs.jproteome.6b00618 J. Proteome Res. 2017, 16, 1401−1409

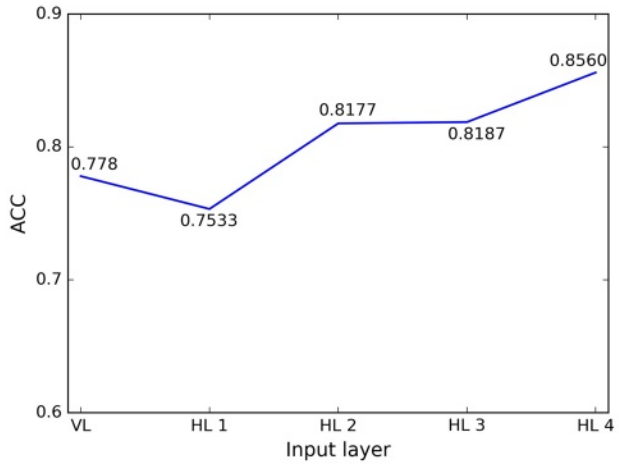


图 7. 五层逻辑回归（LR）模型准确率的曲线图。VL 表示可见层（原始输入数据集）。HL 表示隐藏层。每层的转换数据集均用于训练逻辑回归模型，并使用测试数据集的准确率来评估每个模型的性能。

无监督预训练的影响

机器学习算法的成功通常取决于用于表示和描述数据的描述符。这是因为不同的描述符会或多或少地纠缠和隐藏数据背后不同的解释性变化因素。这促使专家设计更强大的描述符，以帮助算法在特定任务上表现得更好。描述符的重要性无疑凸显了传统学习算法的弱点：它们无法从数据中提取和组织有区分度的信息。专家设计更强大的描述符是利用人类的聪明才智和先验知识来弥补算法弱点的一种方式。此外，无监督学习与有监督学习相结合对于药物靶点相互作用（DTIs）的预测特别有益。在药物靶点空间（DTS）中，已知的药物靶点对（DTPs）只占很小的一部分（不到 0.3%），其余的都是未知的；此外，无相互作用的配对数量远远多于有相互作用的配对数量。因此，仅用 0.3% 的样本很难代表整个样本空间，模型的适用性可能会出现偏差。在 DeepDTIs 中，深度信念网络（DBN）将数据空间中的所有样本用于学习数据空间的分布。DBN 能够捕捉到观测输入的解释性因素的后验分布。为了生成更抽象和有用的表示，DBN 的隐藏层由数据的多个非线性变换组合而成。图 7 展示了我们 DBN 模型中 5 层逻辑回归（LR）模型的准确率曲线。我们利用每一层的变换数据作为输入数据，并应用 LR 训练分类模型。然后使用测试集评估 LR 模型的性能。如图 7 所示，除了隐藏层 1 之外，隐藏层 2、3 和 4 的准确率都明显高于原始输入数据。随着隐藏层深度的增加，LR 的准确率也随之提高。这表明 DBN 的预训练过程具有从原始输入数据集中抽象信息的强大能力。

■ 结论

药物 - 靶点相互作用（DTIs）对于当前的药物发现过程至关重要。已知药物与靶点的相互作用有助于推断药物适应症、药物不良反应、药物 - 药物相互作用以及药物作用模式。在本研究中，我们提出了一种名为 DeepDTIssta 的深度学习方法来预测药物 - 靶点相互作用。我们的方法使用深度信念网络（DBN）模型有效地提取原始输入向量，并准确预测药物 - 靶点相互作用。在一组测试数据集和一组外部 EDTPs 数据集上的结果表明，我们的算法能够实现相对较高的预测性能。对新预测的药物 - 靶点相互作用的进一步分析表明，我们的方法能够推断出一系列新的药物 - 靶点相互作用，这对药物再定位具有实际应用价值。

■ 相关内容

* 支持信息

支持信息可在 ACS 出版网站免费获取，网址为 DOI: 10.1021/acs.jproteome.6b00618 。

补充表 S1：DeepDTIs 中参数的网格搜索结果（XLSX）

补充表 S2：预测概率大于 0.9 的药物 - 靶点相互作用（XLSX）

■ 作者信息

通讯作者

电子邮件：hongmeilu@csu.edu.cn。电话：0731 - 8830830。

ORCID

陆红梅：0000-0002-4686-4491

笔记

作者声明无竞争性经济利益。

■ 致谢

作者感谢曾华亮先生对论文的修改帮助。作者衷心感谢国家自然科学基金对本研究项目的资助（项目编号：81402853、21175157、21375151、21305163 和 21675174），同时也感谢中南大学中央高校基本科研业务费专项资金（项目编号：2015zzts163）的支持。本研究已获得学校审查委员会的批准。

■ 参考文献

(1) Nascimento, A. C.; Prude☒ ncio, R. B.; Costa, I. G. A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinf. 2016, 17, 1.

(2) Chen, X.; Yan, C. C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug−target interaction prediction: databases, web servers and computational models. Briefings Bioinf. 2016, 17 (4), 696−712.

(3) Novac, N. Challenges and opportunities of drug repositioning. Trends Pharmacol. Sci. 2013, 34 (5), 267−272.

(4) Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 2010, 26 (12), i246− i254.

(5) Zanni, R.; Galvez-Llompart, M.; Galvez, J.; Garcia-Domenech, R. QSAR multi-target in drug discovery: a review. Curr. Comput.-Aided Drug Des. 2014, 10 (2), 129−136.

(6) Gonza lez-Díaz, H.; Prado-Prado, F.; García-Mera, X.; Alonso, N.; Abeijo n, P.; Caamano, O.; Ya n☒ez, M.; Munteanu, C. R.; Pazos, A.; Dea-Ayuela, M. A.; et al. MIND-BEST: Web Server for Drugs and Target Discovery; Design, Synthesis, and Assay of MAO-B Inhibitors and Theoretical− Experimental Study of G3PDH Protein from Trichomo-nas gallinae. J. Proteome Res. 2011, 10 (4), 1698−1718.

(7) Vina, D.; Uriarte, E.; Orallo, F.; Gonza lez-Díaz, H. Alignment-Free Prediction of a Drug− Target Complex Network Based on Parameters

of Drug Connectivity and Protein Sequence of Receptors. Mol. Pharmaceutics 2009, 6 (3), 825−835.

(8) Dura n, F. J. R.; Alonso, N.; Caaman□ o, O.; García-Mera, X.; Yan□ ez, M.; Prado-Prado, F. J.; Gonza lez-Díaz, H. Prediction of multi-target networks of neuroprotective compounds with entropy indices and synthesis, assay, and theoretical study of new asymmetric 1, 2-rasagiline carbamates. Int. J. Mol. Sci. 2014, 15 (9), 17035−17064.

(9) Chen, Y.; Zhi, D. Ligand−protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins: Struct., Funct., Genet. 2001, 43 (2), 217−226.

(10) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discovery 2004, 3 (11), 935−949.

(11) Periole, X.; Knepp, A. M.; Sakmar, T. P.; Marrink, S. J.; Huber, T. Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers. J. Am. Chem. Soc. 2012, 134 (26), 10959−10965.

(12) Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. 2012, 8 (5), e1002503.

(13) Chen, X.; Liu, M.-X.; Yan, G.-Y. Drug−target interaction prediction by random walk on the heterogeneous network. Mol. BioSyst. 2012, 8 (7), 1970−1978.

(14) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. Science 2008, 321 (5886), 263−266.

(15) Yang, K.; Bai, H.; Ouyang, Q.; Lai, L.; Tang, C. Finding multiple target optimal intervention in disease-related molecular network. Mol. Syst. Biol. 2008, 4 (1), 228.

(16) Cao, D. S.; Zhang, L. X.; Tan, G. S.; Xiang, Z.; Zeng, W. B.; Xu, Q. S.; Chen, A. F. Computational Prediction of Drug□ Target Interactions Using Chemical, Biological, and Network Features. Mol. Inf. 2014, 33 (10), 669−681.

(17) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J. Chem. Inf. Comput. Sci. 2003, 43 (6), 1882−1889.

(18) Romero-Dura n, F. J.; Alonso, N.; Yan□ ez, M.; Caaman□ o, O.; García-Mera, X.; Gonza lez-Díaz, H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. Neuropharmacology 2016, 103, 270−278.

(19) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521 (7553), 436−444.

(20) Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. 2015, 33 (8), 831−838.

(21) Jo, T.; Hou, J.; Eickholt, J.; Cheng, J. Improving protein fold recognition by deep learning networks. Sci. Rep. 2015, 5, Article No. 17573.

(22) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug−target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 2008, 24 (13), i232−i240.

(23) Sawada, R.; Kotera, M.; Yamanishi, Y. Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. Mol. Inf. 2014, 33 (11 12), 719− 731.

(24) Cao, D.-S.; Liu, S.; Xu, Q.-S.; Lu, H.-M.; Huang, J.-H.; Hu, Q.-N.; Liang, Y.-Z. Large-scale prediction of drug−target interactions using protein sequences and drug topological structures. Anal. Chim. Acta 2012, 752, 1−10.

(25) Bleakley, K.; Yamanishi, Y. Supervised prediction of drug−target interactions using bipartite local models. Bioinformatics 2009, 25 (18), 2397−2403.

(26) Quinlan, J. R. Induction of decision trees. Machine Learning 1986, 1 (1), 81−106.

(27) 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.

(28) Dong, J.; Cao, D.-S.; Miao, H.-Y.; Liu, S.; Deng, B.-C.; Yun, Y.-H.; Wang, N.-N.; Lu, A.-P.; Zeng, W.-B.; Chen, A. F. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J. Cheminf. 2015, 7 (1), 1−10.

(29) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 2010, 50 (5), 742−754.

(30) Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z. propy: a tool to generate various modes of Chou’s PseAAC. Bioinformatics 2013, 29 (7), 960−962.

(31) Hinton, G. E. A practical guide to training restricted boltzmann machines. In Neural Networks: Tricks of the Trade; Montavon, G., Orr, G. B., Mu□ ller, K.-R., Eds.; Springer: Berlin, 2012; pp 599−619.

(32) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. Science 2006, 313 (5786), 504−507.

(33) Hinton, G. E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. Neural computation 2006, 18 (7), 1527−1554.

(34) Bastien, F.; Lamblin, P.; Pascanu, R.; Bergstra, J.; Goodfellow, I.; Bergeron, A.; Bouchard, N.; Warde-Farley, D.; Bengio, Y. Theano: new features and speed improvements. 2012. arXiv preprint arXiv:1211.5590. arXiv e-print archive. https://arxiv.org/pdf/1009. 3589.pdf

(35) Krattenmacher, R. Drospirenone: pharmacology and pharmaco-kinetics of a unique progestogen. Contraception 2000, 62 (1), 29−38.

(36) Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. Pattern Analysis and Machine Intelligence, IEEE Transactions on 2013, 35 (8), 1798−1828.

DOI: 10.1021/acs.jproteome.6b00618
J. Proteome Res. 2017, 16, 1401−1409