JOURNAL OF CHEMICAL INFORMATION AND MODELING

文章

pubs.acs.org/jcim

靶向位置过滤集合体的研制及其对虚拟筛选配体富集的促进作用

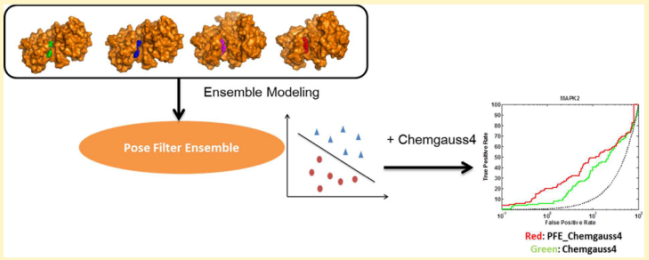
±胡、宋武*、†和王*

中国医学科学院中国北京协和医学院研究院新药研发部天然药物生物活性物质和功能国家重点实验室北京+100050 中国 ±凯利政府解决方案，北卡罗来纳州三角研究园，27709，美国

哥伦比亚特区艾滋病研究中心分子建模和药物发现核心实验室，药物科学系，霍华德大学，华盛顿特区，20059，美国

*s 支援资料

摘要:在药物发现的早期阶段，基于结构的虚拟筛选识别技术已经成为一种不可或缺的技术。然而，目前评分功能的准确性不够高，不足以确保每个目标的成功，因此仍有待提高。在此之前，我们已经开发了二元位姿滤波器(pf)的知识来自于蛋白质-配体界面的单一 x 射线结构的特定目标。这种新方法已被证明是改善配体富集的有效途径。接着，在目前的工作中，我们试图将知识结合起来从不同的蛋白质-配体界面收集相同目标的多个晶体结构，构建 PF 系综(PFEs)。为此，我们首先构建了一个完整的数据集，以满足集成建模和验证的要求。这套设备包括 10 个不同的目标，118 个精心准备的蛋白质配体复合物的 x 射线结构，以及大型的基准活性/诱饵设备。值得注意的是，我们基于集成学习分类的概念设计了一个独特的两层分类器的工作流程，并将其应用于所有目标的 PFEs 的构造。通过广泛的基准研究，我们证明



Pfe 与 Chemgauss4 偶联显著改善了 Chemgauss4 自身的早期富集，(2)PFE 在促进早期富集方面表现出较大的一致性和较大的总富集量。此外，我们还分析了用于构建 PFEs 的同源配体之间的成对拓扑相似性，发现同源配体的化学多样性提高了 PFEs 的性能。综上所述，迄今为止的研究结果证明，通过集成模型将不同蛋白质-配体界面的知识结合起来，能够提高 SBVS 评分功能的筛选能力。

引言

虚拟筛选是一种强大的早期药物发现技术，因为它能够快速而廉价地从大规模的化学文库中发现特定靶点的新型支架活性化合物。1 基于结构的虚拟筛选需要靶点的三维结构，并利用分子对接/打分来识别符合结合位点的命中化合物。2,3 由于近年来高质量的大分子结构数量的快速增长，4vs 在现代药物发现中发挥着关键作用，并已导致许多命中识别的成功案例

在 VS 过程中，一个对接算法首先在目标的结合部位采集每个化合物的可能姿态，然后一个评分函数预测每个配体姿态的结合自由能 8,10-12 错误的评分函数错误地评分和排序构成诱饵/天然的姿势和结合诱饵/真正的结合物，13 这导致有限的配体浓缩(即筛选能力)。14 为了提高现实世界药物筛选的命中率，有必要开发新的方法来提高现有评分函数的准确性，特别是在排序结果表的顶部，丰富不同支架的配体的能力。

分子对接和评分后的视觉检测是提高命中率的一种经验而有效的方法，迄今为止已有许多成功案例报道这种方法。15-21 这主要是指根据每种化合物的结合模式对其类天然姿势进行人工识别，然后根据分数从中选择姿势。这个过程非常耗时，因此

收到:2016 年 12 月 12 日
发表日期:2017 年 5 月 16 日

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

表一。目前研究的主要目标、构建位置滤波系列的蛋白质-配体复合物的精细 x 射线结构、DUD-E 中的代表性结构及其基准参数(即 actives/decoys)概述

target class (no.)	target name	abbr.	structures/PFE no.	structures/PFE name	structure/DUD-E	no. of actives/decoys[a]
other enzymes (33)	adenosine deaminase	ADA	19	1a4l, 1a4m, 1add, 1krm, 1ndv, 1ndy, 1ndz, 1o5r, 1qxl, 1uio, 1uip, 1uml, 1v79, 1wxy, 1wxz, 2e1w, 2z7g, 3iar, 3km8	2e1w	84/5443
	HMG-CoA reductase	HMDH	18	1dq8, 1hw8, 1hw9, 1hwi, 1hwj, 1hwl, 2q1l, 2q6b, 2q6c, 2r4f, 3bgl, 3cct, 3ccw, 3ccz, 3cd0, 3cd5, 3cda, 3cdb	3ccw	166/8735
kinases (25)	MAP kinase-activated protein kinase 2	MAPK2	10	1ny3, 3a2c, 3ka0, 3kc3, 3kga, 3m2w, 3m42, 3r2b, 3r2y, 3wi6	3m2w	99/6130
	insulin-like growth factor I receptor	IGF1R	10	2oj9, 3d94, 3f5p, 3i81, 3lvp, 3nw5, 3nw6, 3nw7, 3o23, 3qqu	2oj9	144/9272
proteases (13)	leukotriene a4 hydrolase	LKHA4	17	2r59, 3b7r, 3b7u, 3cho, 3chp, 3chq, 3chs, 3fh5, 3fh7, 3fh8, 3fuf, 3fui, 3fuk, 3ful, 3fum, 3fun, 3u9w	3chp	166/9426
nuclear receptors (11)	progesterone receptor	PRGR	13	1a28, 1e3k, 1sqn, 1sr7, 1zuc, 2w8y, 3d90, 3g8o, 3hq5, 3kba, 3zra, 3zrb, 4a2j	3kba	274/15431
cytochromes 450 (1)	cytochrome P450 3A4	CP3A4	6	3nxu, 3ua1, 4i4h, 4k9t, 4k9v, 4k9w	3nxu	170/11781
ion channels (2)	glutamate receptor ionotropic kainate 1	GRIK1	13	1vso, 2f35, 2ojt, 2pbw, 2qs1, 2qs2, 2qs3, 2qs4, 2znt, 2znu, 3fvo, 3gbb, 4dld	1vso	96/6537
GPCRs (3)	β2 adrenergic receptor	ADRB2	6	2rh1, 3d4s, 3ny8, 3ny9, 4lde, 4ldl	3ny8	229/14986
miscellaneous	fatty acid binding	FABP4	6	1tou, 1tow, 2nnq, 3fr2, 3fr4, 3fr5	2nnq	42/2711

蛋白脂肪细胞

用于基准测试计算的校准。

人类专家能够实际检查的姿势/化合物的数量是有限的。此外，通过目视检查做出的决定通常是主观的，而且结果随着人类专家的经验水平而变化。Sbvs 社区通过开发自动的姿态过滤器(pf)/分类器来帮助合理选择潜在的命中来解决这个问题。早期的 PFs 是通过直接定义必要的蛋白质-配体相互作用作为约束(例如，氢键约束)或应用来自蛋白质-配体相互作用的类药效团-特征约束来构建的。这些 PFs 中最流行的一类是相互作用指纹，它们的发展通常伴随着蛋白质-配体相互作用编码成一维二元结构。位串的成对相似度被用来定量度量姿态相似度。在分子操作环境中实现的高引用指纹相互作用的 PFs 包括 SIFt25 和 w-SIFt，26 个 APIF，27 个 TIFP，28 个 PLIF，29 个和 SPLIF。和他的同事设计了几何化学描述符，即，基于 Delaunay 分割的 Pauling 电负性，然后更新它们以获得基于 Delaunay 分割的蛋白质-配体双原子最大电荷转移势。在此基础上，我们应用支持向量机(SVM)算法来构建特定于目标的可以区分自然姿态和姿态诱饵的光谱支持向量机。这类以知识为基础的植物生长因子也在第二次世界大战期间进行了评估 Csar2013/2014 基准测试，并成功地识别了原生姿势。

将多个蛋白质-配体界面结合到分子对接中也是 SBVS 成功的关键。最广泛使用的策略，即集成对接多个受体结构/构象，处理多个亲配体界面在计算廉价的方式。大量的文献介绍了集合对接的可行方法，并展示了它们在改善配体富集方面的潜力

报告。鉴于环境保护署

我们提出了一个假设，即利用多个 x 射线结构构建 PFs 可以提出正确的姿态分类并最终改善配体的富集。在这项研究中，我们从 DUD-E 目标列表中选择了 10 个不同的目标作为实例，并编制了 10 个包含蛋白质配体复合物的多重 x 射线结构和每个目标的基准活性/诱饵的现成数据集。利用这些数据集，我们建立了一个新颖而全面的工作流程，将从不同的蛋白质-配体界面收集到的知识结合起来，也就是说，将多个蛋白质-配体界面结合到构建 PFs 中。然后，我们将一种新型的 PF 系综(PFE)应用于经验/回归的评分函数 45，即 Chemgauss4，它不同于我们以前使用的 MedusaScore，一种基于力场的评分函数。本研究是我们以前关于 PF 模型的工作的延续，旨在回答三个尚未探索的问题:(1)如何将不同蛋白质-配体界面的知识整合到 PFs 的构建中，以提高配体的富集度;(2)PFs 除了基于力场的评分函数之外，是否还具有经验评分函数(如 Chemgauss4);(3)PFE 是否比单独的 PF 表现更好。

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

方法

目标选择。这项研究包括两个连续的部分，即目标特异性聚合物的构建和评估其在 SBVS 配体富集中的表现。前者需要蛋白质配体复合物的多重 x 射线结构，而后者依赖于基准数据集。由于 sc-PDB 数据库收集了准备使用的高质量的辅晶结构的药物结合位点(2015 年 6 月访问)，46 它成为我们的蛋白质配体复合物结构的来源。此外，DUD-E 还包括适合分子对接的黄金标准基准数据集 47 和 48，因此，它被用于 PFEs 的性能评估。为了满足上述数据要求，即蛋白质-配体复合物和基准数据集，只选择 DUD-E 和 sc-PDB 之间的重叠目标进行进一步分析。由于 sc-PDB(cf)不包括 102 个靶点中的 3 个，即组蛋白脱乙酰酶 2(HDAC2)、胰蛋白酶 i(TRY1)和类胰蛋白酶 1(TRYB1)，它们被排除在我们的靶点列表之外。

选择目标的其他标准如下。(1)蛋白质-配体复合物必须具有至少两种可用的 x 射线结构才能构建聚四氟乙烯。为此，DUD-E 配体涵盖的所有 UniProt 加入码(即 UniProt.txt)都是从 DUD-E(2015 年 6 月访问)获得的。然后从 UniProt(2015 年 6 月访问)中检索他们的 UniProt 名称。以这些 Uniprot 名称为输入，计算了 sc-PDB 中每个靶蛋白配体复合物的 x 射线结构。只有一种 x 射线结构的 6 种靶点，即单胺氧化酶 B(AOFB)、CYP2C9(CP2C9)、CXC 趋化因子受体 4 型(CXCR4)、多巴胺 D3 受体(DRD3)、-葡萄糖脑苷酶(GLCM)和蛋白激酶 C(KPCB)被排除在外。(2)其余 93 个目标包括 8 个类别，即其他酶、激酶、蛋白酶、核受体、杂质、GPCRs、离子通道和细胞色素 P450。为了在不同的目标上构建和测试 PFEs，从每个类中选择一个或两个有代表性的目标作为例子。具体来说，对于包含 20 多个目标的类别，即其他酶和激酶，选择了两个目标。对于其他类别，即蛋白酶、核受体、细胞色素 P450、离子通道、GPCRs 和其他，只选择了一个目标。表 1 列出了 10 个选定的目标，包括其他酶的腺苷脱氨酶(ADA)和羟甲基戊二酸单酰辅酶 A 还原酶(HMDH)，激酶的 MAP 激活蛋白激酶 2(MAPK2)和胰岛素样生长因子 1 受体(IGF1R)，蛋白酶的白三烯 A4 水解酶(LKHA4)，核受体的 PRGR(PRGR),细胞色素 P450 的 CP3A4，离子通道的 GRIK1，离子通道的 ADRB2，脂肪酸结合蛋白脂肪细胞的 FABP4，CYP3A4 的 CP3A4，谷氨酸受体离子通道的 GRIK1，脂肪酸结合蛋白脂肪细胞的 ADRB2。

蛋白质-配体复合物的制备与筛选。对于每个选定的目标，所有可用的蛋白质配体复合物的 x 射线结构对应的 UniProt 名称(参考)是从 sc-PDB。然后，蛋白质及其同源配体的结构被下载。虽然 sc-PDB 中的蛋白质结构已经得到整理，因此很容易与分子对接相容，但蛋白质的制备和选择还需要 46 个步骤

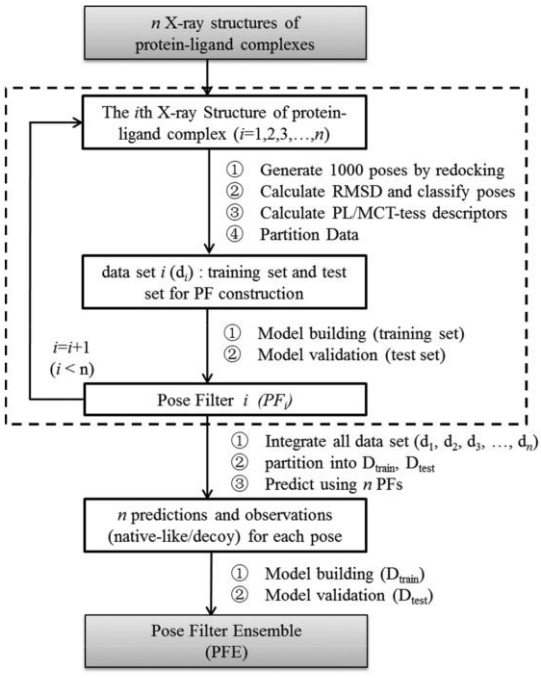
(1)不涉及蛋白质-配体相互作用的链，如 T4 溶菌酶融合蛋白(ADRB2)被去除。(2)所有的水分子都被排除在蛋白质结构之外，但是在结合位点与同源配体相互作用的必要辅助因子(如 ADA 和 LKHA4 中的锌离子)被保留了下来。(3)将各种蛋白质结构及其同源配体组装成蛋白质-配体复合物。以 DUD-E 基准设定的蛋白质结构为参照，对蛋白质-配体复合物的所有 x 射线结构进行叠加，以检测所有同源配体是否结合在同一位点。同源配体与参考位点(如变构位点)结合的配合物被排除在外(如 IGF1R，PDB 条目 3lw0;cf)。(4)为了进一步检测蛋白质结构，我们使用 OMEGA(2.5.1.4 版本，OpenEye 科学软件)49 和 FRED(3.0.1 版本，OpenEye 科学软件)50-52 进行了每个同源配体与其结合位点的重新对接。对于那些 FRED 由于与氢原子的问题而无法产生对接姿态的结构，氢原子通过 DiscoveryStudio(2.5 版本;AccelrysSoftware)重新进行了对接，之后又重复进行了对接。本文通过重新测量氢原子的数量和重复的对接过程，研究了 1ndv、1ndy、1ndz、1qxl 和 2e1w(ADA)的晶体结构，为分子对接提供了可能。通过上述步骤，制备了蛋白质和配体的结构，为模型的建立奠定了基础。

目标特异性聚合物聚合物的构建。一般工作流程。工作流的输入是 n 个蛋白质-配体复合物的 x 射线结构，输出是一个构造的可以对姿态进行分类的聚四氟乙烯。在此，n 表示每个 PFE 的结构数(参见表 1 和)。产生的 PFE 是一个两层链式分类器的方案，其中由 n 个第一层模型(即 n 个 PFs)预测的姿态类构成第二层模型的输入。因此，工作流包括两个基本步骤:(1)建立第一层模型，即 n 个基于 svm 的 PFs，其中每个模型仅利用来自一个 x 射线结构的知识;(2)建立第二层模型，使用使用所有第一层模型预测的姿态类作为输入(参考方案 1)

基于每个 x 射线结构的第一层模型/pf 的建立。在这项研究中构建 PFs 的协议是我们之前构建特定目标 PFs 的方法的修改版本。(1)应用 OMEGA 和 FRED 生成 1000 个几何造型。具体来说，OMEGA 被用来为每个同源配体建立一个多构象数据库，其最大构象数被设置为默认值 200。Fred 用于将同源配体从其多通道数据库中停靠到由同源配体自身定义的结合位点。因此，得分最高的 1000 个姿势被保留下来作进一步分析。(2)「RMSD 计算器」(7.5 版本;accelryssoftware)，用以计算同源配体在 1000 个姿态(即 x 射线结构中的姿态)与其本身姿态之间的重原子均方根差(RMSD)值。根据 RMSD 值，将 1000 个姿态分为两类，即本地姿态(RMSD≤4)和姿态诱饵(RMSD4)。X 射线结构不超过 10 个自然姿态产生的 x 射线结构不包括在结构/pfe 列表中(如 1v7a(ADA)、1dqa)

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

计划 1。构建姿态滤波器集成(PFE)的工作流程:从单个 x 射线结构构建 PF 的主要步骤显示在虚线框中



(HMDH)、1w0 及 2v0m(CP3A4);。(3)利用我们的内部程序 ENTESS(2016 年 10 月访问)计算 PL/MCT-tess 描述子，以表征生成的同源配体的 1000 个蛋白质-配体界面。该程序使用 Delaunay 分割将每个蛋白质-配体界面分割成 Delaunay 四面体，然后根据最大电荷转移(MCT)将每种类型的 Delaunay 四面体进行成对原子势拥有属性化

$$\text{PL/MCT- tess}^{m} = \sum_{k=1}^{n} \sum_{pl}^{1-3} \sum_{l}^{1-3} \text{MCT}^{p} \times \text{MCT}/d^{l}_{pl}{}_{k} \quad (1)$$

在这个方程中，m 是一个描述类型的指数，即，四面体类型(m1,2,3，...，554)，pl/mct-tessm 代表 mth 四面体类型的潜力，n 是蛋白质-配体络合物中 mth 四面体类型的出现次数，k 是 mth 四面体类型(k1,2，...，n)的四面体指数，p 是界面四面体中蛋白质表示的原子的指数，l 是 delaay 四面体的指数，dpl 是 pth 蛋白质与 lth 的对应距离。由于 pl/mct-tess 的计算是基于每个蛋白质-配体界面四面体，所以蛋白质原子或配体原子的最大数目为 3 个(即 p1,2,3 和 l1,2,3)。

根据自然姿态与假目标姿态的比值，将数据集定义为平衡(比值≤2)或非平衡(比值 2)。每个平衡数据集被随机分割成一个用于建模的训练集(80%的

在 MATLAB(version7.6.0.324)中，使用“cvpartition”(即在一组指定大小的数据上创建随机分区的函数)进行模型验证的测试集(20%的姿势)。对于每个由一个主类和一个次类组成的不平衡数据集，在数据划分之前都采用了下抽样策略。设 p 和 q 分别为主要类和次要类中摆姿势的总数。首先，根据 pl/mct-tess 描述符，主要类别中的每个姿势的拥有属性是它的欧几里得度量到次要类别中最近的邻居的距离。然后用欧几里德距离对主要类中的所有 p 构象进行排序。第四个姿势的距离 Circa 被设置为一个阈值，那些在主类中距离不超过阈值的姿势被选择为小类的新对应物。用这种方法构造了一个新的平衡数据集。随后，根据随机数据划分生成训练集和测试集。(5)使用开源程序 LibSVM(2015 年 6 月访问)建立和验证基本的二进制分类器，即 PF.55 在训练集的基础上，使用五径向基核函数交叉验证(CV)对参数值和径向基函数(RBF)值进行网格搜索(grid.py)。网格搜索的输出，即导致最高 CV 精度的参数值，被选中并用于构建 PF。对于模型验证，然后应用 PF 预测测试集中的姿态类。

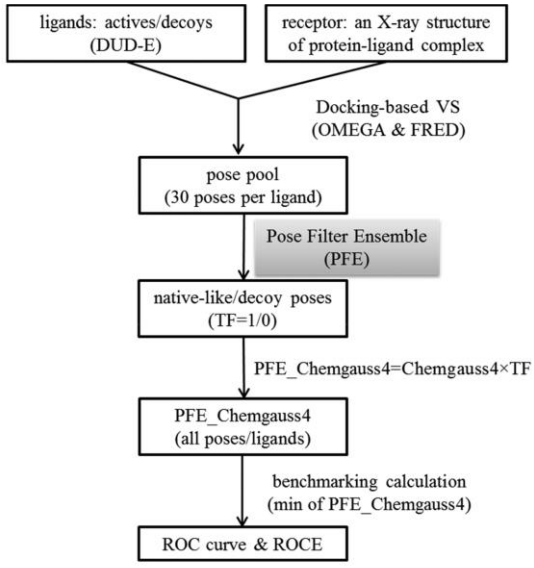
利用 PFs 预测作为输入建立第二层模型/pfe。通过集成以前用于 PF 构造(di)的所有数据集，构建了一个大型数据集(d)。是的
也随机分成一个训练集(Dtrain，80%)和一个测试集(Dtest，20%)。以 pl/mct-tess 描述子为基础，预测 dtrain/dtest 中每个姿态的潜在类别(即类天然姿态和诱饵姿态)，然后构造一个 dtrain/dtest 阵列
从多个 PFs 中预测的姿态类成为 PFE 建模的二进制描述符。在 Dtrain 上使用 LibSVM 执行相同的网格搜索(grid.py)，以搜索参数的最优值
建立了二层模型，并在 Dtest 上进行了验证。这个模型是一个集合，使用 n 个第一层模型(即 PFs)作为基本分类器，因此被称为 PFE。

Pfes 在 SBVS 中的应用及标杆管理研究。一般工作流程。为了检验 PFE 对 SBVS 的潜在影响，将 PFE 与基于结构的经验评分函数整合到对接程序 FRED 中，即 Chemgauss4。然后通过基准研究评价了使用 PFE 后 Chemgauss4 的配体富集情况。在这里，“配体富集”是指通过区分活性物质和诱饵/随机分子，将活性物质富集到筛选列表的顶级的能力。48 具体的工作流程见方案 2。首先，由选定目标的真实活动和绑定诱饵组成的基准测试集从(2015 年 6 月访问)下载。为了避免标杆管理研究中的潜在偏差，将与培训 PFE 的同源配体重叠的真实活动排除在标杆管理集之外。将清理后的基准测试集作为一个小规模的化学库，使用 OMEGA 和 FRED 根据该目标的参考结构(参考结

构/表 1 中的 dud-e)执行基于对接的 VS。Omega 的参数设置为默认值。在对接模拟中，每个配体产生最多 30 个对接姿态，并利用 Chemgauss4 的评分函数对所有姿态进行评分

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

计划 2。位姿滤波器集成(PFE)与 Chemgauss4 的耦合作为当前的基准研究



造成了大量的姿势。其次，PFE 预测每个姿态的潜在等级(TF)，即天然类(TF1)或姿态诱饵(TF0)。通过简单地将 Chemgauss4 与 TF 相乘，得到了一个新的“PFE-耦合 Chemgauss4”(PFEChemgauss4)评分。最后，每个配体保留 PFE 趋化因子 4 最小值的姿势。根据 PFEchemgauss4 的真活性和结合诱饵的排序表，计算了 PFEchemgauss4 的配体富集。

基于本文研究的具体目标，我们用相应的方法代替了原始工作流的第二步和最后一步。为了评估 PFE 的影响，先前的评分函数，即，Chemgauss4 没有使用 PFE，也是基准。在基准研究中，每个配体保留最小 Chemgauss4 评分而不是 PFEChemgauss4 评分的姿势。为了探索集成建模的潜在优势，将基于对接 VS 蛋白质结构的单个 PF 作为姿态分类器，与 Chemgauss4 结合，得到一个“PF-coupchemgauss4”评分(PFchem-gauss4)。然后进行了 PFchemgauss4 配体的富集对比分析。

基准测试计算度量。来自 ROC 曲线(ROC)曲线的参数，即 ROC 曲线下面积(AUC)和 ROC 浓度(ROCE)为 1%/0.5%，是基准测试计算的主要指标。Auc 值是配体富集总体性能的指标，ROCE1%/ROCE0.5%的值用于测定早期富集。Roce 的定义是在已知诱饵的一定百分比下，真正阳性率与假阳性率的比值(例如，ROCE1%为 1%)。56,57 除主要指标外，还计算了回收活性百分比(AR%)，以显示识别活性的性能。

为了探索每一种方法在新型支架鉴定中的潜力，我们利用“生成碎片(MurckoAssem-)”技术生成了 58 个早期回收的活性物质(即 1%的结合诱饵)在 PipelinePilot 中，只有唯一的部分被计算在内。根据活性物质和独特的 Murcko 支架的数量，计算了 PFEChemgauss4 与 PFChemgauss4 和 Chemgauss4 的重叠。为了对 PFEchemgauss4 和 PFchemgauss4 进行比较分析，设计并应用了两个指标，即未识别的%和唯一的%。前者代表 Chemgauss4 恢复的活性物/支架物的百分比，后者代表 pfe/pf 未鉴定的活性物/支架物的百分比。后者表示使用 pfe/pf 后附加活性物/支架的百分比。根据定义，未识别的%的较小值和唯一的%的较大值表示方法的理想性能。

结果及讨论

用于集成建模和基准测试的综合数据集。如表 1 所示，我们编译并用于集成建模和基准测试的数据集包括涵盖 8 个类别的 10 个代表性目标。具体来说，ADA 和 HMDH 代表其他类别的酶，而 MAPK2 和 IGF1R 是这类激酶的代表性靶标。Lkha4、PRGR、CP3A4、GRIK1、ADRB2 和 FABP4 分别是蛋白酶类、核受体类、细胞色素 P450 类、离子通道类、GPCRs 类和其他类的例子。每个目标的子集包括两个不可缺少的组成部分:(1)蛋白质-配体复合物的多重 x 射线结构用于构建 PFE(即结构/PFE)和(2)活性物质/诱饵用于基准研究。用于构建 PFEs 的蛋白质配体复合物的数量从 6 个到 19 个，跨越了 10 个不同的靶点。对于 10 个目标中的 7 个，这个数字不少于 10 个。对于每一个靶标，配合物中所有的同源配体都结合在同一位点，而该结合位点的蛋白质原子坐标随同源配体化学结构的不同而不同。此外，所有包含的 x 射线结构对于 FRED 分子对接是可行的，并且能够产生类自然姿态。另一个组成部分，即该数据集的活性/诱饵，是从 DUD-E 中的现成基准测试集中检索出来的，并经过处理，以确保活性与同源配体没有重叠。

虽然所有的相关数据最初都是从公开的数据库 sc-PDB 和 DUD-E 中获得的，但是我们将它们重新编译成一个可以随时使用的数据集，其中包括蛋白质-配体复合物的结构及其相应的基准集。在本研究中，我们使用它来建立 PFE 并且评估它对 SBVS 的影响。我们预计，社区对这些数据集的公开访问将促进与集成建模和基准测试相关的方法开发。

类自然姿态分类器与诱饵姿态分类器:PFs 和 PFEs。方案框架的特性。使用 LibSVM 程序，我们根据每个目标蛋白质配体复合物的 x 射线结构构建了多个 PFs。模型建立和验证的所有细节，包括训练集和测试集，模型建立的参数，模型在姿态分类中的性能。以下几点值得注意:(1)训练集和测试集中本机类姿态和姿态诱饵的分布是均衡的，

说明我们的下采样策略能有效地将非均衡分布转化为均衡分布。(2)参数 c 和的取值表明，只保留一个径向基函数模型作为姿态

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

表二。姿态滤波器集成建模:每个姿态滤波器的参数值、CV 精度和预测精度

	D_{train}		RBF model			D_{test}		
targets	no. of native-like poses	no. of pose decoys	C	γ (10^{-4})	CV accuracy (%)	no. of native-like poses	no. of pose decoys	prediction accuracy (%)
ADA	2682	2977	32	1250.0	96.3	648	766	96.1
HMDH	4158	4436	2	5000.0	88.6	977	1171	88.5
MAPK2	2991	2977	8	1250.0	84.6	775	717	84.4
IGF1R	1969	1947	512	312.5	86.1	489	490	87.3
LKHA4	3677	3145	128	1250.0	86.7	911	794	85.7
PRGR	4155	3571	2048	312.5	88.1	1060	871	88.0
CP3A4	961	1014	8	78.1	93.1	244	249	93.1
GRIK1	3275	2729	512	78.1	85.8	840	660	86.9
ADRB2	2084	2256	2048	312.5	85.1	496	589	87.2
FABP4	1525	1496	2	78.1	88.3	361	394	88.7

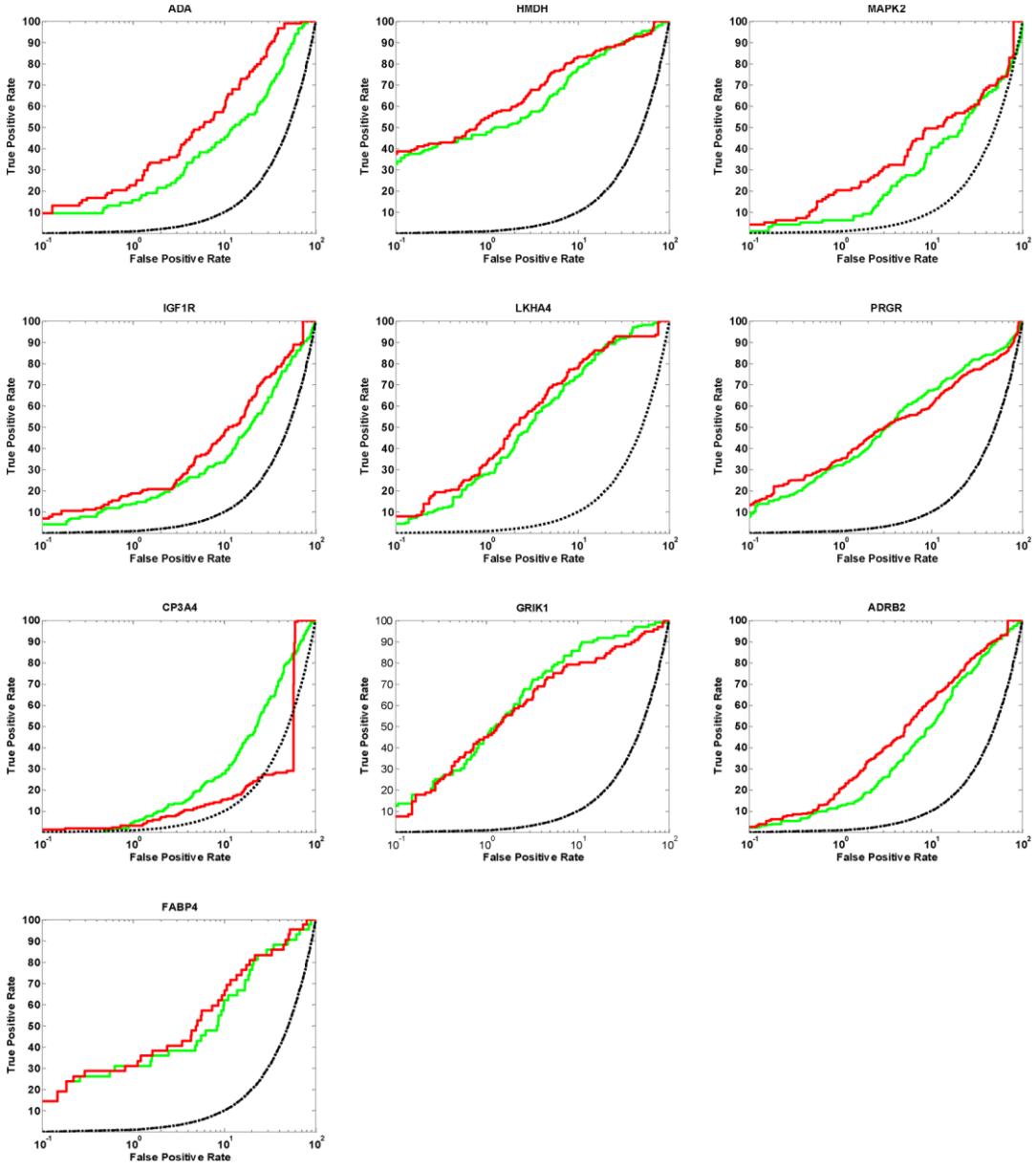


图 1。配体富集用 ROC 曲线表示。颜色代码:红色，PFEChemgauss4;绿色，Chemgauss4;黑色虚线，随机分布。

每个 x 射线结构的分类器。这不同于我们早期的策略，即保持所有模型的合格简历准确性。32

为了节省计算时间，我们进行了网格搜索，寻找 c 的最佳值，然后使用它们

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

表三。配体富集 Chemgauss4，PFEChemgauss4，和 PFChemgauss4 作为衡量 ROCE1%，ROCE0.5%，AUC 和活性回收百分比(AR%)

	ROCE$_{1\%}$			ROCE$_{0.5\%}$			AUC			AR%		
target	C4[a]	PFE_C4[a]	PF_C4[a]	C4	PFE_C4	PF_C4	C4	PFE_C4	PF_C4	C4	PFE_C4	PF_C4
ADA	15.48	22.62	39.29	23.81	35.72	57.16	0.78	0.88	0.95	100.00	98.81	96.43
HMDH	47.59	54.22	42.17	89.17	90.37	78.32	0.90	0.92	0.90	100.00	93.37	96.99
MAPK2	6.06	20.20	31.31	10.10	22.23	60.62	0.65	0.71	0.67	100.00	82.83	52.53
IGF1R	13.89	18.75	20.14	23.61	26.39	29.17	0.72	0.79	0.83	100.00	88.89	87.50
LKHA4	27.71	33.13	36.75	33.74	44.58	44.58	0.91	0.90	0.89	100.00	92.77	88.55
PRGR	32.12	34.67	27.37	48.18	55.48	39.42	0.83	0.80	0.74	100.00	93.43	71.90
CP3A4	4.71	2.94	4.71	1.18	3.53	3.53	0.71	0.55	0.37	100.00	28.24	7.65
GRIK1	44.79	44.79	34.38	58.35	68.76	45.84	0.94	0.90	0.80	100.00	98.96	68.75
ADRB2	12.23	20.52	15.72	17.47	19.21	19.21	0.81	0.85	0.84	100.00	93.01	91.70
FABP4	30.96	30.96	33.34	52.45	57.22	57.22	0.84	0.87	0.71	100.00	100.00	78.57
average	23.55	28.28	28.52	35.81	42.35	43.51	0.81	0.82	0.77	100.00	87.03	74.06

4; PFE c4，PFE Chemgauss4; PF c4，PF Chemgauss4.

表四。用 Chemgauss4、PFEChemgauss4 和 PFChemgauss4 回收 1%结合诱饵活性物质的统计分析

	actives			overlap		unidentified%		unique%	
target	C4[a]	PFE_C4[a]	PF_C4[a]	C4 and PFE_C4	C4 and PF_C4	PFE_C4	PF_C4	PFE_C4	PF_C4
ADA	13	19	33	12	13	5.0	0.0	35.0	60.6
HMDH	79	90	70	77	67	2.2	14.6	14.1	3.7
MAPK2	6	20	31	6	5	0.0	3.1	70.0	81.3
IGF1R	20	27	29	17	15	10.0	14.7	33.3	41.2
LKHA4	46	54	60	38	36	12.9	14.3	25.8	34.3
PRGR	88	94	74	78	59	9.6	28.2	15.4	14.6
CP3A4	8	5	8	0	0	61.5	50.0	38.5	50.0
GRIK1	43	43	31	36	21	14.0	41.5	14.0	18.9
ADRB2	28	47	36	22	24	11.3	10.0	47.2	30.0
FABP4	13	13	14	12	12	7.1	6.7	7.1	13.3
average	34	41	39	30	25	8.9	18.8	24.4	29.2

4; PFE c4，PFE Chemgauss4; PF c4，PF Chemgauss4.

建立一个能够提高简历准确性的模型。(3)在 118 个二元位姿分类器(即 PFs)中，有 106 个是最优的位姿预测模型，其 CV 准确率和 PA 准确率均大于 90%。在其他 12 个 PFs 中，7 个 x 射线结构(ADA 的 1ndy、1uip 和 2e1w，IGF1R 的 3nw5，GRIK1 的 2f35、2qs1 和 2qs3)的模型预测结果较好，因为这些 PFs 的 CV 准确度或 PA 都大于 90%。5 个模型(HMDH 为 1dq8 和 3cdb，CP3A4 为 4k9t 和 4k9v，FABP4 为 1to4)的变异系数和 PA 均达到可接受的范围，均在 70%和 90%之间。每个目标的所有模型都用于集成建模。

聚四氟乙烯的特性。表 2 显示了 10 个选定目标的 PFEs，它们的参数值，以及它们在姿态预测中的模型性能。如表所示，在所有的数据集中，为了构建 PFEs，天然类姿态和姿态诱饵的分布是平衡的。同样，每个目标只保留最高 CV 准确性的模型。各模型的变异系数准确率均在 80%以上，ADA 最高为 96.3%，ADA 最低为 84.6%
Mapk2.各模型的 PA 值(Dtest)与 CV 值(Dtrain)接近。Mapk2 的最小值为 84.4%，ADA 的最大值为 96.1%。

配体富集:PFEchemgauss4 与 Chem-gauss4 的比较。基准测试计算。对于每一个目标，我们基于一个基准设置进行回顾性 SBVS，并评估 Chemgauss4 在使用 PFE 前后的配体丰度。值得注意的是，一些配体(例如，活性配体或

因为它们不确定的立体化学性质，或者因为它们的大小不适合场地而不能停靠，因此，它们不包括在基准计算中。如表 1 和所示，用于基准测试计算的 actives/decoys 数量少于 DUD-E(2015 年 6 月访问)中列出的初始数量。为了获得指标，我们绘制了 ROC 曲线(参见图 1)并计算了它们的参数(参见表 3 和)来评估不同方法的配体浓度。在实践中，只有一小部分化合物而不是来自 VS 的所有化合物被提交用于生物测定，因此，早期富集被认为是评估配体富集的一个重要指标

早期充实。如表 3 和所示，PFEChemgauss4 的 ROCE1%值比 Chemgauss4 的 7 个目标(即 ADA、HMDH、MAPK2、IGF1R、LKHA4、PRGR 和 ADRB2)的 ROCE1%值大。就这些目标而言，Pfechemgauss4 的 ROCE0.5%值也较大。对于目标 GRIK1 和 FABP4，尽管 ROCE1%值为 PfeChemgauss4 与 Chemgauss4 相当，ROCE0.5%值大于 Chemgauss4。对于 CP3A4，虽然 PFEchemgauss4 的 ROCE1%值是与 Chemgauss4 相比，ROCE0.5%的值较大。在 10 个指标中，PFEChemgauss4 的 ROCE1% 和 ROCE0.5% 均高于 Chemgauss4(ROCE1%，28.28vs23.55;ROCE0.5%，42.35vs35.81)。ROC 曲线的假阳性率在 0.1%到 1%之间的部分显示了类似的趋势(参见图 1)。所有

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

表 5。Chemgauss4、PFEChemgauss4、PFChemgauss4 对 1%粘结剂回收的 Murcko 支架的统计学研究

	scaffolds			Overlap		unidentified%		unique%	
target	C4[a]	PFE_C4[a]	PF_C4[a]	C4 and PFE_C4	C4 and PF_C4	PFE_C4	PF_C4	PFE_C4	PF_C4
ADA	11	13	19	10	11	7.1	0.0	21.4	42.1
HMDH	77	88	68	75	65	2.2	15.0	14.4	3.8
MAPK2	6	19	26	6	5	0.0	3.7	68.4	77.8
IGF1R	20	27	29	17	15	10.0	14.7	33.3	41.2
LKHA4	46	54	59	38	36	12.9	14.5	25.8	33.3
PRGR	79	80	67	69	54	11.1	27.2	12.2	14.1
CP3A4	8	5	8	0	0	61.5	50.0	38.5	50.0
GRIK1	24	23	17	19	14	17.9	37.0	14.3	11.1
ADRB2	27	46	35	21	23	11.5	10.3	48.1	30.8
FABP4	4	5	6	4	4	0.0	0.0	20.0	33.3
average	30	36	33	26	23	10.0	17.5	25.0	25.0

4; PFE c4，PFE Chemgauss4; PF c4，PF Chemgauss4.

上述数据，即 ROC 曲线和相应的参数表明，聚四氟乙烯的使用改善了 Chemgauss4 的早期富集。

我们进一步分析了聚四氟乙烯(PFE)在 Chemgauss4 中应用前后早期(即约束诱饵的 1%)恢复的活性。此外，我们生成了 Murcko-Bemis 框架来描述这些活动并探索它们的脚手架多样性。根据表 4 和表 5，对于除 CP3A4 以外的所有目标，PFEChemgauss4 比 Chemgauss4 恢复了更多或相当数量的活性/支架。这表明 PFEchemgauss4 能够识别出最初没有被 Chem-gauss4 复原的其他活性物/支架物。以 CP3A4 为靶点，PFEChemgauss4 比 Chemgauss4 识别出较少的活性物质/支架。这很好地解释了 CP3A4 的情况，即 PFEchem-gauss4 的 ROCE1%值小于 Chemgauss4。然而，值得注意的是，Chemgauss4 和 PFEChemgauss4 在这个目标上没有重叠。在整体性能方面，PFEChemgauss4 和 Chemgauss4 鉴定的活性/支架平均数分别为 41/36 和 34/30，由于 PFE 的使用，表现出更大的性能。在上述数据的基础上，可以得出结论:PFE 的使用通过识别附加活性物质和 Murcko 支架改善了早期富集。

全面提升。除早期铀浓缩外，代表整体铀浓缩的 ROCAUC 值亦载于表 3。正如 PFEChemgauss4 和 chemgauss410 个目标的 ROCAUC 值的平均值(即 0.82vs0.81)所示，PFE 的使用对整体浓缩没有显著影响。对于 9 个单独的目标，PFEChemgauss4 和 Chemgauss4 的 AUCs 差别很小。Cp3a4 有显著变化，AUC 值从 0.71(Chemgauss4)下降到 0.55(PFEChemgauss4)。

为了阐明 AUC 变化的潜在机制，我们计算了活性物质回收率(AR%)在整个筛选中的百分比。10 个指标的 AR%平均值为 87.03%。对于七个目标，AR%值大于 90%，对于两个目标，AR%值在 80%到 90%之间。Cp3a4 的 AR%值非常低(28.24%;参见表 3 和)。根据 PFEchem-gauss4 的评分方法，AR%与 PFE 的预测(即 TF1/0)相关。如果 PFE 预测一个有源作为姿态诱饵的所有姿态，那么这个有源作为姿态诱饵将不会被恢复。所以

Pfe 无法识别 CP3A4 大部分活性物质的类天然构象，导致 CP3A4 的 AR%极低，总体富集。总之，PFE 的加入在大多数情况下保持了 Chemgauss4 的总体浓缩能力。

影响聚四氟乙烯效能的关键因素。如上所述，PFE 的疗效因不同目标而异。例如，PFE 对 MAPK2 的富集效果最好，使 Chemgauss4 的配体富集率提高了 233%。然而，CP3A4 表现最差，增长率为-38%(参考文献)。我们假设用于模型建立的同源配体的化学多样性可能影响早期富集。因此，我们使用最大直径 6 的函数类指纹图谱(fcfp6)来表征这些同源配体，并用 Tanimoto 系数(Tc)计算它们的两两拓扑相似性。所有 Tc 值的平均值，即平均值(Tc)，然后与 ROCE1%相关，以揭示化学多样性和早期富集之间的潜在关系。在此之前，我们使用库克的距离来识别数据中的异常值。59,60 案例顺序图清楚地显示 MAPK2 是一个异常值，因为它的库克距离远大于阈值，即 3 倍的平均库克距离(参见参考文献)。根据未释放数据，即 9 对平均值(Tc)/ROCE1%，早期富集增加率(ROCE1%)与平均值(Tc)之间的皮尔逊相关系数(r)为-0.68(参见图 2)。这种强相关性表明，总的来说，同源配体之间较低的成对相似性导致 Chemgauss4 的早期富集得到更大的改善。为了

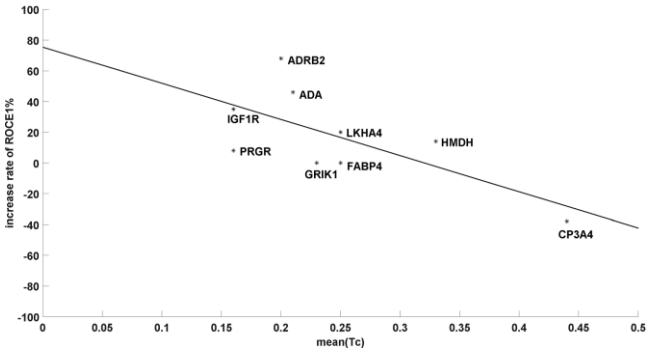


图 2。平均 Tc 值与 ROCE1%增长率的回归曲线，皮尔逊相关系数为-0.68。

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

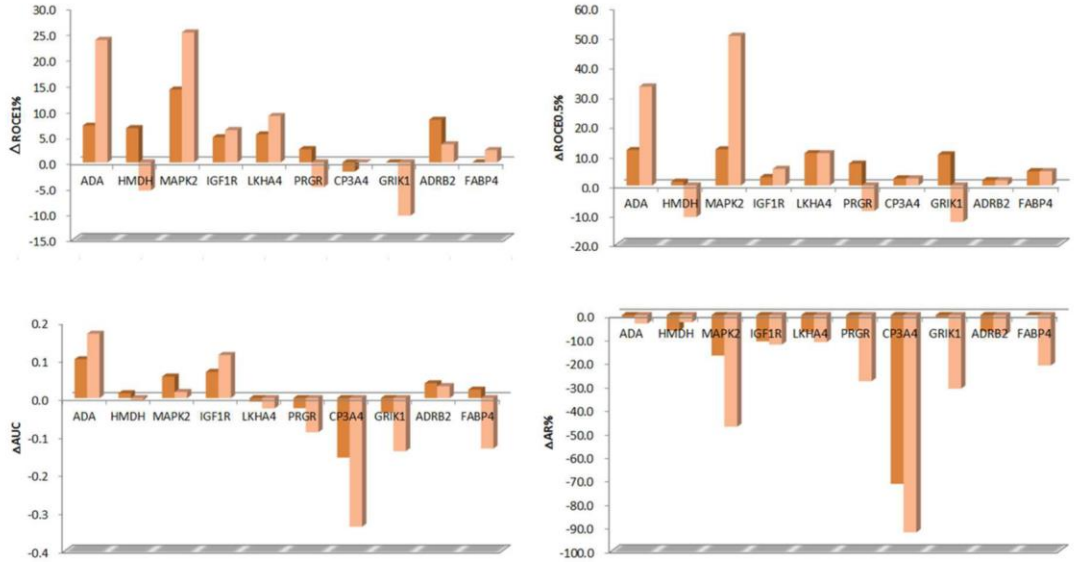


图 3。根据 ROCE1%，ROCE0.5%，AUC 和 AR%比较分析 PFEs 和 PFs 的潜在影响。颜色代码:深橙色，PFE;浅橙色，PF。

例如，ADRB2(0.20)的平均(Tc)值比 HMDH(0.33)的平均(Tc)值小得多，因此，它对 ROCE1%的改善相应更大(增长率为 68%比 14%)。以上数据表明，影响聚四氟乙烯功效的主要因素是用于系综建模的同源配体的化学多样性。

集成建模的优势:ppes 相对于 PFs。定标比较结果的内容。为了探索集成模拟的潜在优势，我们以 PFchemgauss4 为基准，并与 PFEchemgauss4 进行了广泛的比较研究。评估结果载于表 3、、图 3(ROCE0.5%、ROCE1%、AUC 和 AR%)、表 4(已恢复的活动)和表 5(Murcko)脚手架复原)。

共同特点:早期认识的全面提高。根据表 3，我们发现 PFE 与 PF 具有共同的特征，对早期富集有全面的影响。PfeChemgauss4 的 ROCE1%和 ROCE0.5%的平均值高于 Chemgauss4(即分别为 28.28vs23.55 和 42.35vs35.81)，表明总体上 PFE 的加入改善了 Chemgauss4 的早期富集(参见表 3)。如表 4 和表 5 所示，PFEChemgauss4 回收的 1%结合诱饵中活性物和 Murcko 支架的平均数分别为 41 和 36，高于 Chemgauss4 回收的活性物和 Murcko 支架的平均数 34 和 30。因此，PFEchemgauss4 鉴定的活性物/支架的数量较多，以及 PFEchemgauss4 鉴定的附加活性物/支架的数量大于未鉴定的活性物/支架的数量，这些都表明了早期富集效果的改善。如表所示，所鉴定的活性物/支架的对数为 41/36，与 PFChemgauss4(39/33)相近，但与 Chemgauss4(34/30)相差较远。以上数据表明，整体 PFEchemgauss4 具有 PFchemgauss4 的基本能力，即通过比 Chem-gauss4 丰富更多的活性物/支架来提高早期识别能力。

聚四氟乙烯对早期富集有较一致的改善作用。用平均法测定早期富集

Pfe-Chemgauss4 和 PFChemgauss4 之间的差异很小(即 28.28
Roce1%为 28.52，ROCE0.5%为 42.35，ROCE0.5%为 43.51。基于这些一般数据，本文提出了一种基于数据挖掘技术的数据挖掘方法
首先，PFE 与 PF 相比没有显示出优势。为了发现 PFE 的优点，我们进一步比较了使用 PFE 和 PF 对每个目标的结果，通过计算它们各自在 ROCE1%、ROCE0.5%、AUC 和 AR%方面的改进程度。值得注意的是，我们观察到早期富集的增加
使用 PFE 后的所有目标。与聚四氟乙烯不同，只有七个目标使用 PF 改善早期富集。对于其他三个目标，它显著削弱了原始记分功能，即 Chemgauss4 的早期识别能力。具体来说，HMDH 的 ROCE1%/ROCE0.5%下降了 5.42/10.84，PRGR 下降了 4.74/8.76，GRIK1 下降了 10.42/12.50(参见图 3)。相应地，它们的相关参数也比原始计分函数的相关参数差得多。例如，PFChemgauss4 对这些目标识别的活性物/支架明显少于 Chemgauss4。这可能是导致 PFchemgauss4 与 PFEchemgauss4 相比平均减少两个活性物和三个支架的主要因素(参见表 4 和表 5)。在这一点上，PFE 是一种更加健壮的方法，因为它以更加一致的方式改进了早期富集。

Pfes 显示 Chemgauss4 活性/支架回收能力的改进维护。我们还比较了 PFEchemgauss4 和 PFchemgauss4 在未确定的%和唯一的%(参见表 4 和表 5)。除 GRIK1 和 PRGR 外，所有指标的活体及其 Murcko 支架早期恢复的结果基本一致。根据 10 个指标的平均得分，PFEchemgauss4 的未识别率远低于 PFchemgauss4，尤其是活性物(8.9vs18.8)。同时，PFEchemgauss4 在活性物质和支架材料中的平均唯一度与 PFchemgauss4 接近。理想情况下，一种方法应该能够(1)恢复所有原先由 Chemgauss4 确定的活性物质/支架(即，未确定的%0)和(2)确定更多的附加物

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

动物/支架尽可能(即，更大的独特%)。目前的数据确实表明，与使用聚四氟乙烯相比，使用聚四氟乙烯能够更有效地维持 Chemgauss4 的初始富集，并具有同等的能力发现更多的活性/支架。

聚四氟乙烯能提高总体浓度。根据先前的分析，在大多数情况下，颗粒型聚合物保持了 Chemgauss4 的整体富集。图 3 中 AUC 和 AR%的图表进一步说明了这一点。为了区分 PFE 和 PFs 对总体富集的影响，我们首先比较 PFE 和 PF 对哪些目标表现出积极的影响(参见图 3)。一方面，PFE 改善了个目标的 AUC，而 PF 只改善了个目标的 AUC。另一方面，PFE 能够识别 9 个目标超过 80%的活动，而 PF 只能识别 5 个目标。接下来，我们比较了 PFEs 和 pf 引起的变化的重要性(参见表 3)。平均而言，PFs 使 AUCs 保持在一个恒定的水平，而 PFs 使 AUCs 从 0.81 下降到 0.77。同时，使用 PFs 后的 AR%值由 100%降至 87.03%，而使用 PFs 后的 AR%值则降至 74.06%。在上述两点的基础上，可以得出结论，在维护方面，PFEs 比 pff 更有优势全面充实的高水平。

结论

在目前的研究中，我们汇编了一个涵盖 10 个不同目标的综合数据集。其相应目标的每个子集包含精心策划的蛋白质-配体复合物的 x 射线结构和基准活性/诱饵。这个即时可用的数据集不仅用于整体建模和基准测试(例如，本研究中所做的 PFEs 的构建和验证)，而且也适用于需要蛋白质-配体相互作用和基准测试集知识的其他类型的研究。最重要的是，我们设计了一种基于多个 x 射线结构的集成建模方法来构造位姿滤波器集成。Pfes 可以与基于结构的经验评分函数耦合，即 Chemgauss4。作为一种验证手段，我们将这种方法应用于数据集中的所有不同目标，并通过对活性物/诱饵进行基准测试，广泛评估了每种特定目标 PFE 对 Chemgauss4 配体富集的影响。

对比分析了使用 PFEs 前后 Chem-gauss4 对配体的富集情况，结果表明:使用 PFEs 前后配体的富集程度不同

随着 ROCE1%/ROCE0.5%和等效 ROC 曲线的增加，PFEs 的使用可以改善 Chemgauss4 的早期富集，同时保持其整体富集，(2)根据早期活性恢复的 Murcko 支架的数量，PFEs 的使用提高了 Chemgauss4 鉴定新型支架活性的机会。此外，我们还观察到 PFE 的疗效因靶点不同而不同。进一步的拓扑相似性分析，即基于 fcfp6 指纹的同源配体内的 Tc 分布，表明在 PFE 构建中使用更多的同源配体可以使 Chemgauss4 的配体富集得到更大的改善。

通过对 PFEchemgauss4 和 PFchemgauss4 配体富集的比较，评价了系综模拟的效果。评价结果表明，使用 PFEs 能更好地提高早期识别率

同时保持 Chemgauss4 的活性/支架恢复能力，并显示出比使用 PFs 更大的总体富集能力。这些特征验证了我们的假设，即从蛋白质-配体界面引入额外的知识来构建姿态分类器/滤波器，将大大改进现有的 SBVS 评分函数。

本研究是我们利用 pl/mct-tess 描述子改进配体富集的目标特异性 PF 构建工作的延续。据我们所知，基于集成建模技术并与 Chemgauss4 相结合构造 PFEs 的工作流程至今尚未见报道。有趣的是，PFEs 的使用也有利于经验评分函数。由于它工作于多种不同的目标，我们期待它在未来的 SBVS 活动中得到广泛的应用和成功。我们最直接的目标是将这种新颖的方法应用于发现用于治疗糖尿病的法尼酯 x 受体(FXR)激动剂，因为目前已有多种 x 射线晶体结构可用于这一重要靶点。

相关内容

*s 支援资料

支持信息可以在 DOI:上免费获得。

Dud-e 的目标列表和 sc-PDB 所覆盖的蛋白质-配体复合物的 x 射线结构数目(表 S1);排除 x 射线结构的 PDBid(表 S2);真活性和同源配体之间的重叠(表 S3);每个支持向量机基于 PF 的参数(表 S4);未用于基准测试计算的配体数量(表 S5);同源配体中 ROCE1%的增长率和平均(Tc)值(表 S6);Pfe/pf 改善了配体富集(表 S7);蛋白质-配体复合物的叠加 x 射线结构(图 S1);Chemgauss4/PFEChemgauss4 配体富集(图 S2);Cook 距离的案例顺序图(图 S3)

作者信息

通讯作者

*x.s.w.: 电邮:。地址:霍华德大学药学院，FourthStreetNW2300 号。

*s.w.:电邮:。地址:中国医学科学院北京协和医学院药物研究所新药研发部天然药物活性物质与功能国家重点实验室，北京 100050。

Orcid

解霞:

王:

注释

两位作者声称没有相互竞争的经济利益。

用于集成建模和基准研究的综合数据集可在。可根据要求提供 PFE 建设的代码和手册。

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

鸣谢

这项工作得到了中央大学基础研究基金(2016ZX350036)和国家自然科学基金委员会81603027的部分支持。我们非常感谢OpenEye科学软件公司为他们的软件包提供学术许可。我们也感谢张博士(北京大学药学院)免费使用他的团队的计算机设备。我们还要感谢哥伦比亚特区艾滋病研究发展中心(P30AI087714)、中美生物医学合作研究行政补充国立卫生研究院(5P30A10877714-02)以及获得G12MD007597奖项的美国国家少数民族健康与国立卫生研究院差异研究所提供的支持。内容完全是作者的责任，不一定代表国立卫生研究院的官方观点。

缩写

基于Delaunay镶嵌的蛋白质配体最大电荷转移电位;基于结构的VS;基于结构的VS;基于虚拟筛选的PF;基于PFE，PF集成;基于Delaunay镶嵌的ENTess，电负性;基于Delaunay镶嵌的pl/mct-tess，蛋白质配体双原子最大电荷转移电位;基于QSBAR，定量结构-结合亲和关系;基于支持向量机的SVM;基于组蛋白脱乙酰酶的HDAC2;1，胰蛋白酶i;TRYB1，胰蛋白酶1;AOFB，单胺氧化酶B;CP2C9，CYP2C9;CXCR4，CXC趋化因子受体亚型4;DRD3，多巴胺D3受体;GLCM,-葡萄糖脑苷酶;KPCB，蛋白激酶C;ADA，腺苷脱氨酶;HMDH，羟甲基戊二酸单酰辅酶A还原酶;MAPK2，MAP激活蛋白激酶2;1r，胰岛素样生长因子1受体;LKHA4，白三烯A4水解酶;PRGR，孕酮受体;CP3A4，CYP3A4;GRIK1，谷氨酸受体，离子型海因酸1;ADRB2,2;肾上腺素能受体Fabp4，脂肪酸结合蛋白脂肪细胞;MCT，最大电荷转移;RBF，径向基核函数;PFEChemgauss4，PFE耦合Chemgauss4;PFChemgauss4，PF-耦合Chemgauss4;ROC，ROC曲线;AUC，曲线下面积;ROCE，ROC富集;AR，活性恢复;PA，预测精度;FCFP6，最大直径指纹功能级指纹图;Tc，Tanimoto系数

参考文献

虚拟筛选和快速自动对接方法。今日药物发现2002,7,64-70。

蛋白质-配体对接和基于结构的虚拟筛选研究的突出挑战。威利安迪斯普。牧师:康普特。分子。科学。2011,1,229-259.

基于结构的药物发现虚拟筛选:一个以问题为中心的评论。2012,14,133-141.

郑，h.;Handing，k.b.;Zimmerman，M.d.;Shabalin，I.g.;Almo，S.c.;Minor，w.X光散射技术在过去十年的新药发现--我们下一步要去哪里？专家Opin。药物发现2015,10,975-989。

基于结构的虚拟配体筛选:最近的成功故事。梳子。化学。高通量筛选2009,12,1000-1016。

基于结构的药物发现虚拟筛选:原理、应用和最新进展。柯尔。页首。地中海。化学。2014,14,1923-1938.

基于对接的虚拟筛选:最新进展。梳子。化学。高通量筛选 2009,12,303-314。
基于结构的虚拟筛选蛋白质-配体相互作用指纹图谱:方法和基准研究。北京杰凯化工技术有限公司。Inf.模型。二○一四年，54,2555-2561。
不连续的分子动力学状结合体与困难目标中的诱饵的区别。Biophys.J.2012,102,144-151.
马丁内斯-马约尔加，肯塔基州;兰格，肯塔基州;坎纳洛-孔特雷拉斯，肯塔基州;阿格拉菲奥蒂斯，肯塔基州认识虚拟筛选的陷阱:一个批判性的评论。北京杰凯化工技术有限公司。Inf.模型。2012,52,867-881.
几种分子对接程序的比较:姿态预测和虚拟筛选精确度。北京杰凯化工技术有限公司。Inf.模型。2009,49,1455-1474.
沃伦，g.l.;安德鲁斯，c.w.;卡佩利，A.M.;克拉克，b.;拉隆德，j.;兰伯特，m.h.;林德瓦尔，m.;内文斯，n.;塞穆斯，s.f.;辛格，s.;特德斯科，g.;沃尔，I.d.;Woolven，j.m.;佩绍夫，c.e.;海德，m.s.f.。对接程序和评分函数的评价。J.Med.化学。2006,49,5912-5931.
格拉夫斯，a.p.;布伦克，r.;肖切特，B.k.诱饵对接。J.Med.化学。2005,48,3714-3728.
李勇勇;韩良良;刘;王，r。更新基准评分函数比较评估:2。评估方法及一般结果。北京杰凯化工技术有限公司。Inf.模型。二○一四年，54,1717-1736。
李，t;尹，n;刘，h.;裴，j.;赖，l.新型毒素 Hipa 抑制剂减少耐多药持留菌。美国化学会医学会。化学。莱特。二○一六年 7,449-453。
通过基于结构的酶抑制剂虚拟筛选研究，发现了一种新的活性化合物 Nedd8 与 Piperidin-4-Amine 脚手架相互作用。美国化学学会。Biol.2016,11,1901-1907.
鉴定 n-苯基-2-(n-苯基磺酰胺基)乙酰胺类新型 rgamma 反向激动剂:虚拟筛选、基于结构的优化和生物学评价。欧盟。J.Med.化学。2016,116,13-26.
针对趋化因子-蛋白偶联受体界面中多个位点的新型配体识别。J.Med.化学。二○一六年，59,4342-4351。
对接虚拟筛选中的蛋白质柔性:利用多晶体结构发现新型淋巴细胞特异性酪氨酸磷酸酶抑制剂。北京杰凯化工技术有限公司。Inf.模型。2015,55,1973-1983.
高通量虚拟筛选确定新型的 n′-(1-苯基乙烯基)-苯甲酰肼为有效的、特异的和可逆的 Lsd1 抑制剂。J.Med.化学。二○一三年，56,9496-9508。
利用基于结构的虚拟筛选发现 Brd4 的新型小分子抑制剂。J.Med.化学。二○一三年，56,8073-8088。
基于结构的虚拟筛选分析中的目标偏向评分方法和专家系统。柯尔。Opin.化学。Biol.2004,8,359-364.
最小化激酶虚拟筛选中的假阳性。
蛋白质:Struct.，Funct.，Genet。2006,64,422-435.
有没有可能通过药效团过滤来提高基于结构的虚拟筛选的命中率？后过滤技术的优缺点探讨。J.Mol.GraphicsModell.2008,26,1237-1251.
结构相互作用指纹(Sift):一种新的三维分析方法

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425

蛋白质-配体结合作用。J.Med.化学。2004,47,337-344.
基于加权蛋白质-配体相互作用指纹图谱的虚拟筛选位置特异相互作用评分技术。北京杰凯化工技术有限公司。Inf.模型。2009,49,1185-1192.
一个新的基于原子对的交互指纹及其在虚拟筛选中的应用。北京杰凯化工技术有限公司。Inf.模型。2009,49,1245-1260.
编码蛋白质-配体相互作用在指纹和图形中的模式。北京杰凯化工技术有限公司。Inf.模型。二○一三年，53,623-637。
分子操作环境(MOE)，2010.10 版;化学计算小组:蒙特利尔，QC，2010。
基于蛋白质-配体界面新几何化学描述的定量结构-结合亲和关系模型的建立。J.Med.化学。2006,49,2713-2724.
化学信息学和物理力场评分函数的联合应用改进了 Csar 数据集的结合亲和力预测。北京杰凯化工技术有限公司。Inf.模型。2011,51,2027-2035.
基于知识的姿势评分和基于物理力场的打击评分功能的联合应用提高了基于结构的分子力学虚拟筛选的准确性。北京杰凯化工技术有限公司。Inf.模型。2012,52,16-28.
针对特定目标的天然/诱骗式分类器提高配体在 Csar2013 基准中排名的准确性。北京杰凯化工技术有限公司。Inf.模型。2015,55,63-71.
在 Csar2014 基准练习中，特定目标姿势分类器的对接和评分成功地识别了天然姿势，但没有结合亲和力预测。北京杰凯化工技术有限公司。Inf.模型。二○一六年，56,1032-1041。
通过考虑蛋白质灵活性评估基于对接的蛋白激酶靶标整体虚拟筛选策略。北京杰凯化工技术有限公司。Inf.模型。二○一四年，54,2664-2679。
Bolia，a;Gerek，z.n.;Ozkan，s.b.B.-dock:基于自由结构探索蛋白质-配体相互作用的一个灵活的对接方案。北京杰凯化工技术有限公司。Inf.模型。二○一四年，54,913-925。
达，c.;Mooberry，S.l.;Gupton，j.t.;Kellogg，G.e。如何处理低分辨率的目标结构:使用 Sar，集合对接，水道测量分析，和 3d-Qsar 确定地图的 α-微管蛋白秋水仙素网站。J.Med.化学。二○一三年，56,7382-7395。
科斯康纳蒂，s.;马里内利，l.;DiLeva，f.s.;LaPietra，v.;DeSimone，a.;Mancini，f.;Andrisano，v.;Novellino，e.;Goodsell，d.;Olson，a.j.蛋白质柔性在虚拟筛选:TheBace-1 案例研究。北京杰凯化工技术有限公司。Inf.模型。2012,52,2697-2704.
科尔布;奥尔森;鲍登;霍尔，新泽西;韦尔东克，m.l.;利贝舒茨，j.w.;科尔，j.c.对接的潜力和局限性。北京杰凯化工技术有限公司。Inf.模型。二○一二年，52,1262-1274。
进化出一个互补的口袋构想团队而不是一个单一的领导者。
北京杰凯化工技术有限公司。Inf.模型。2012,52,2705-2714.
rettenmaier，t.j.;Fan，h.;Karpiak，j.;Doak，a.;Sali，a.;Shoichet，b.k.;Wells，j.a.来自基于结构的对接的蛋白质激酶 Pdk1 的小分子变构调节剂。J.Med.化学。2015,58,8285-8291.
化学图书馆与蛋白质构象系综对接，并应用于 aldehyde2 号。北京杰凯化工技术有限公司。Inf.模型。2014,54,2105-2116.
李勇勇;金大江;马文;卢贝特，r.a.;博德，A.M.;董，z.发现新的检查点激酶 1 抑制剂作者:Virtual

基于多晶体结构的筛选。
2011,51,2904-2914.
基于结构的虚拟筛选发现亚型选择性 Janus 激酶(Jak)抑制剂。北京杰凯化工技术有限公司。Inf.模型。二○一六年，56,234-247。
现行评分功能的分类。北京杰凯化工技术有限公司。Inf.模型。二○一五年，55,475-482。
龙 南 , d.;Kellenberger , e.sc-PDB:a3ddatabaseofLigandableBindingSites10YearsOn。核酸。2015,43，D399-D404.
Mysinger， m.;Carchia， m.;欧文， j.j.;Shoichet，b.k.目录有用诱饵，增强(Dud-E):更好的配体和诱饵更好的基准。J.Med.化学。2012,55,6582-6594.
中国虚拟筛选配体富集评价的基准方法和数据集。方法 2015,71,146-157。
从数据集中学习和失败的分析。北京杰凯化工技术有限公司。Inf.模型。2012,52,2919-2936.
麦克甘恩;阿蒙德;尼克尔斯;格兰特;布朗，高斯对接函数。生物聚合物 2003,68,76-90.
姿态预测和虚拟筛选精度。北京杰凯化工技术有限公司。Inf.模型。2011,51,578-596.
虚拟筛选的拓扑、形状和对接方法的比较。北京杰凯化工技术有限公司。Inf.模型。2007,47,1504-1519.
;Szentpaly，l.v.;Liu，s.亲电性指数。女名女子名。化学。Soc.1999,121,1922-1924.
刘议员，S.-B。概念密度泛函理论及其最新进展。美国物理学会。罪。2009,25,590-600.
支持向量机的一个库。AcmTrans.Intell.Syst.泰诺科技。2011,2,27.
尼科尔斯，a。我们知道什么，什么时候知道？计算机辅助模型。女名女子名。2008,22,239-255.
基于配体的虚拟筛选的最优分配方法。J.Cheminf.2009,1,14.
比米斯，G.w.;穆尔科，M.a。已知药物的性质。分子框架。J.Med.化学。1996,39,2887-2893.
线性回归的影响力观察。女名女子名。统计。1979,74,169-174.
线性回归影响观测的检测。技术测量 2000,42,65-68。

DOI: 10.1021/acs.jcim.6b00749
J. Chem. Inf. Model. 2017, 57, 1414−1425