Hans Journal of Medicinal Chemistry 药物化学, 2026, 14(4), 287-293
Published Online November 2026 in Hans. https://www.hanspub.org/journal/hjmce
https://doi.org/10.12677/hjmce.2026.144028

Hans汉斯

基于人工智能与计算化学的多维度靶向药物虚拟筛选

崔韵婕，陈　涵，刘婉鑫，权贞熙，符一诺，王爽悦，冬海洋*

郑州师范学院化学化工学院，河南　郑州

收稿日期：2026年10月3日；录用日期：2026年10月16日；发布日期：2026年11月10日

摘　要

现代药物的研发存在着周期长、成本大、传统的高通量筛选成功率低等诸多问题。为了快速找到高效低毒靶向药物(抗肿瘤微管抑制剂)的候选物，本文从人工智能、计算化学两方面来综合论述虚拟筛选的方法。该计算管线以生物活性数据库为基础，用机器学习分类预测模型建立基于配体的预测体系，并采用SHAP归因分析破解算法“黑箱”方式来解释结果，给药物微观构效关系赋予了直观的物理指引。其次使用核心药效团模型和SMARTS规则对千万级的商业化化合物库做骨架硬编码筛选，可以快速缩减化学空间。接着利用基于受体的半柔性三维分子对接技术来确定候选小分子和受体蛋白空腔的构象契合程度以及结合自由能。最终使用无监督结构聚类技术来规避结构风险，同时配合ADMET早期预测，从中筛选出有较高靶点亲和力以及较好的临床转化前景的先导化合物。这样一种漏斗式高效的集成计算管线，很好地克服了传统研发的盲目性，给精准药物设计提供了一个可靠的数字化闭环方案。

关键词

机器学习，分子对接，虚拟筛选，药效团过滤，ADMET预测，计算辅助药物设计

A Multi-Dimensional Virtual Screening for Targeted Drugs Based on Artificial Intelligence and Computational Chemistry

Yunjie Cui, Han Chen, Wanxin Liu, Zhenxi Quan, Yinuo Fu, Shuangyue Wang, Haiyang Dong*

College of Chemistry and Chemical Engineering, Zhengzhou Normal University, Zhengzhou Henan

Received: October 3, 2026; accepted: October 16, 2026; published: November 10, 2026

*通讯作者。

文章引用: 崔韵婕, 陈涵, 刘婉鑫, 权贞熙, 符一诺, 王爽悦, 冬海洋. 基于人工智能与计算化学的多维度靶向药物虚拟筛选[J]. 药物化学, 2026, 14(4): 287-293. DOI: 10.12677/hjmce.2026.144028

Abstract

Current problems in the development of new drugs include a long cycle, high cost and a low success rate for traditional high-throughput screening. To accelerate the discovery of candidates for high-efficiency and low-toxicity targeted drugs (anti-tumor microtubule inhibitors), this paper introduces a multi-dimensional virtual screening strategy based on artificial intelligence and computational chemistry. First, the computational pipeline builds ligand-based machine learning classification models using bioactivity databases, and then applies SHAP attribution analysis to interpret the "black box" of the algorithm and provide intuitive physical guidance for microscopic structure-activity relationships. Second, it uses core pharmacophore models and SMARTS rules to perform hard-coded scaffold filtering on commercial compound libraries containing tens of millions of compounds and achieve a rapid reduction of chemical space. Next, receptor-based semi-flexible three-dimensional molecular docking technology is employed to determine the degree of conformational fit between candidate small molecules and the receptor protein cavity and calculate the binding free energy. Finally, unsupervised structural clustering technology is employed to mitigate structural risks; in conjunction with early ADMET prediction, high-affinity target-binding lead compounds with good clinical translation potential are carefully selected. This high-efficiency funnel-shaped integrated computational pipeline can effectively solve the problem of blindness in traditional R&D and provide a reliable digital closed-loop solution for precision drug design.

Keywords

Machine Learning, Molecular Docking, Virtual Screening, Pharmacophore Filtering, ADMET Prediction, Computer-Aided Drug Design (CADD)

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

现代临床肿瘤学及重大疾病治疗领域，寻找高效低毒的靶向药物一直为医学界所追求[1]。抗肿瘤微管阻滞剂虽然在化疗中占主导地位，但是由于其引起的系统性毒性和易引发多药耐药性的特点，给临床治疗带来了很大的困扰[2]。为了克服以上缺点，研究人员把目光投向了秋水仙碱结合位点等一些蛋白靶点区域[3]，该类抑制剂因为具有独特的破坏微管聚合机制，可以避开耐药泵的识别并且减少不良生理反应的发生，因此是很有发展前景的新型药物研发方向[4]。

但是传统的药物研发方式很困难。从早期的靶点确定、高通量物理筛选到先导化合物的优化以及临床试验，一般新药诞生要耗时十年以上并用掉大量资金，成功率很低[5]。传统的高通量筛选依靠在巨大的化学库上做盲目的试验，研究人员在无尽的化学参数和结构空间里迷路了[6]。伴随着人工智能、计算化学的迅猛发展，药物发现范式也发生着改变[7]。目前，以配体为基础的机器学习活性预测[8]、根据特定药效团进行骨架硬编码过滤、基于受体的三维分子对接三者相串联，形成了一个多维度、逐级推进的虚拟筛选线路，已经成为加快新型靶向抑制剂发现、削减研发成本的主导前沿手段[9]。

2. 基于配体的药物设计：数据驱动与算法预测

根据配体来设计虚拟筛选(LBVS)，LBVS 并不需要考虑目标蛋白的三维结构，它依靠已经知道的活

性分子的构效关系(SAR)来训练算法，从而预测未知化合物的生物活性[10]。

2.1. 数据库挖掘与数据标准化

机器学习模型的上限是由数据质量来决定的。研究流程的起点一般是从 BindingDB [11]、ChEMBL [12]这些大的公共生物活性知识库里做大量的数据搜索。此阶段研究者会得到有关特定靶点具有明显抑制活性的大量小分子信息。数据清理不可小视，它至关重要[13]。原始数据里常常包含着信息缺失，测试类别不明或者活性标签不一致的噪声。经过严格的过滤后得到有效的数据，并对所有化合物执行统一的化学标识符(SMILES 字符串)去重处理，从而得到精炼的化合物集[14]。为了适应二分类机器学习模型的训练，连续的生物活性指标(IC50 或者 Ki 值)一般被转换成对数值(pIC_{50})，并设置一个严格的科学阈值，把化合物明确分为“高活性”和“非活性”，这样就把复杂的生化问题转化为清晰的数学分类问题[15]。

2.2. 分子表征与特征提取

计算机不能直接识别化学结构式，所以要把分子的三维或者二维拓扑结构转成机器可以识别的数学向量。扩展连通性指纹(ECFP)是目前应用最广泛的分子表征技术之一[16]。根据 Morgan 算法[17]，分子中每一个原子以及该原子所处的一定范围内的化学环境都会被映射成一个固定的长度二进制位向量。该种表征方式既保留了分子的核心子结构信息，又因为统一的维度尺度，不需要额外的数据标准化就可以直接输入到树状算法模型中[18]。

2.3. 集成学习与算法优化

在得到高质量分子指纹之后，一般会并行训练多种先进的机器学习算法进行交叉验证来评价预测的效果。常见的算法包括随机森林(RF) [19]、极端随机树(ET) [20]、XGB [21]以及 LGBM [22]等。以决策树为基础的集成算法对于高维稀疏特征(分子指纹)以及化学空间中的非线性关系有较大的优势。用贝叶斯优化等自动化调参的方法，在训练过程中使模型可以找到最佳的超参数组合，用多折交叉验证等手段防止模型对训练集出现过度拟合的情况，从而保证模型对于未知的化学空间具有很强的泛化能力以及预测精度[23]。

2.4. 破解黑盒：SHAP 归因与微观构效关系

机器学习经常受到指责是黑箱式模型，无法给出指导分子设计的直接信息。为此，现代计算流程引入了源自博弈论的 SHAP 归因分析[24]。SHAP 可以得到分子指纹里每一个具体的结构位点对于最终模型预测结果的边际影响[25]。将抽象的数学权重反向映射到具体的化学官能团或者分子骨架上，就可以清楚地发现哪些结构片段(例如特定的杂环、取代基等)是产生生物活性的主要驱动力，哪些片段会导致活性的丧失。强大的解释能力把冰冷的算法预测变成了具体的化学修饰蓝图，给合理的药物设计提供准确的物理导航[26]。

3. 领域知识融合：骨架硬编码与高通量空间缩减

现代商业化合物库(ZINC 等数据库)里分子数目是成百上千万的，有的甚至多达上亿[27]。如果直接让机器学习模型去处理整个庞大的数据库，那么计算资源就会变得非常昂贵。所以，在算法预测之前加入人类专家的先验知识做骨架硬编码过滤，是提高效率的关键。

3.1. 核心药效团模型的建立

就某一个蛋白靶点空腔而言，化学家一般已有某些结构偏好方面的经验。对于微管蛋白的某些结合

位点来说，含有一定药效团(比如磺酰胺结构)的骨架可以提高分子的代谢稳定性以及靶点亲和力，没有它就无法达到这种目的。一个理想的候选框架一般要符合一定的空间拓扑要求，例如必须要有疏水相互作用的芳基区域、有柔性的中心骨架和能和靶点深部的关键氨基酸残基形成紧密氢键网络的杂芳基环末端[28]。

3.2. SMARTS 匹配与化学维度骤降

根据以上建立的三维药效团模型，就会将其转换成严谨的 SMARTS 字符串(用以描述分子模式的一种语言)。利用刚性物理过滤规则把它们部署到海量化合物库中，在很短时间内就可以去掉数以百万计不符合主要结合特征的无效分子[29]。高通量过滤操作大幅度减小了化学空间的维度，得到一个非常浓缩、有极高成药潜力的靶向子集，并且大大减少了后面机器学习模型预测和物理分子对接所消耗的计算资源。

4. 基于受体的确证：三维分子对接与构象探索

机器学习、药效团过滤(基于配体的方法)主要从二维角度评价分子的潜力，但候选药物最后还要在三维空间里同真实的蛋白靶点发生物理嵌合。因此，基于受体的分子对接是不可或缺的验证环节[30]。

4.1. 靶点受体与配体的三维准备

分子对接第一步就是对获得的高分辨率靶点蛋白晶体结构(一般从 PDB 数据库中获取)做精细的预处理。包括除去结晶过程中带有的游离水分子、去掉原来共结晶的配体、加上必要的极性氢原子，准确地分布电荷的分布状况，进而得到一个接近生理状态的蛋白结合口袋。同时需要使用力场算法对筛选出的配体分子进行构象计算和能量最小化，得到最有利的三维结构，并给各个结构分配合适的电荷状态，以便进行对接模拟。

4.2. 半柔性对接与结合自由能评估

对接计算时一般使用半柔性对接方法，即保持受体蛋白大分子骨架不变，只让小分子配体单键自由旋转、折叠[31]。算法会先在事先确定好的靶点三维网格空间内对配体所有可能的空间姿态进行穷举，并且求出它们和受体之间的氢键、范德华力、静电引力、疏水作用等所产生的结合自由能。采用严格的结合能阈值来筛选出管线可以剔除那些虽然算法评分很高，但是在三维空间中存在空间位阻或者不能和重要的氨基酸残基形成有效互作的假阳性分子。只有在受体空腔里找到最合适的构象，结合能量释放得越多的候选分子才被选中。

5. 成药性精选与化学空间多样化

一个高亲和力的分子并不等于一个合格的药物。进入昂贵的湿实验阶段之前，必须从药代动力学、系统毒性、化学多样性等各方面对候选者做最后的严格审查。

5.1. 理化性质约束与类药性评估

候选分子需要具备基本的类药性原则，最著名的便是 Lipinski 的“类药五原则”[32]。筛选程序对分子的物理参数有严格要求，分子量要控制在容易被膜吸收的范围内，氢键供体和受体会限制分子的极性和溶解度，可旋转键会影响分子的柔性和口服生物利用度，脂水分配系数(LogP)决定分子在水相和脂相中分布。违反这些刚性界限的框架就会被无情地淘汰。

5.2. 无监督结构聚类与骨架跳跃

如果最终选出的分子都具有相同的骨架结构，在之后的体内动物实验中发现该骨架无法克服毒性或

者出现专利壁垒，那么这个研发项目就会陷入灭顶之灾。为了分散这种结构坍塌的风险，用分子指纹相似性做无监督聚类算法(Butina 聚类)代替传统聚类算法[33]。可以将大候选库分成很多个物理性质互相不交叉的独立集合。通过对每一个截然不同的类簇中选取出代表性的核心结构，研究人员人为地增大了化合物样本在化学空间中的分布范围，从而达到了骨架跳跃的效果，大大提高了研发管线的稳健性。

5.3. ADMET 属性早期预测

在计算流程的最后，先进的数字环境平台(ADMET、pkCSM 等)被用来对分子的吸收、分布、代谢、排泄以及毒性(ADMET)进行全方位的评价[34]。这些评价指标对于判断分子在临床中存活率具有非常大的意义。通过预测肠道吸收率(HIA)来评价口服可性；通过血脑屏障(BBB)穿透性预测来决定药物是否适合于中枢神经系统靶点或者防止出现中枢毒性；评价分子是否是 P-糖蛋白(P-gp)的外排底物，从而判断它是否会轻易产生耐药性。从毒理学角度来说，还要提前进行致突变危险，心血管毒性，以及急性组织毒性等的筛查。只有那些药代动力学特征清晰、毒性最小、合成可及性最好的少数几个“天选之子”才会进入实体实验室的合成和生物活性测试阶段。

6. 结论与展望

因此，现代药物研发已经完全摒弃了以前那种靠盲目试错来寻找药物的时代。巧妙地将大容量的结构骨架过滤、基于机器学习的准确活性预测和靶点受体三维对接及成药性评价三者结合在一起，就形成了一个漏斗式的、高度集成化的计算药物发现管线[35]。技术融合跨学科的计算方面做到高精确性、高速度的同时，也从源头上给发现新抑制剂提供了条件。

就目前而言，高度自动化的虚拟筛选体系会越来越完善。随着后面候选分子的体外细胞实验和体内动物模型的真实数据不断产生，这些湿实验的结果会反过来补充到机器学习模型中进行再训练。除此之外，借助更加深入的计算手段，比如分子动力学模拟、自由能微扰计算等，研究者可以对药物和耐药蛋白之间对抗的过程有更动态的认识。数字孪生同实体实验交替开展起来的“设计、制造、测试、分析”闭环，会重新绘制出抗癌靶向药物以及其它复杂的疾病治疗药物的设计图谱。

基金项目

本研究得到郑州师范学院大学生创新创业训练计划(编号：DCZ2026025、DCZ2026012)的支持。

参考文献

[1] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., *et al*. (2019) Applications of Machine Learning in Drug Discovery and Development. *Nature Reviews Drug Discovery*, **18**, 463-477. https://doi.org/10.1038/s41573-019-0024-5

[2] McLoughlin, E.C. and O'Boyle, N.M. (2020) Colchicine Binding Site Inhibitors from Chemistry to Clinic: A Review. *Pharmaceuticals*, **13**, Article No. 8. https://doi.org/10.3390/ph13010008

[3] Sargsyan, A., Sahakyan, H. and Nazaryan, K. (2023) Effect of Colchicine Binding Site Inhibitors on the Tubulin Intersubunit Interaction. *ACS Omega*, **8**, 29448-29454. https://doi.org/10.1021/acsomega.3c02979

[4] Ekins, S., Puhl, A.C., Zorn, K.M., Lane, T.R., Russo, D.P., Klein, J.J., *et al*. (2019) Exploiting Machine Learning for End-to-End Drug Discovery and Development. *Nature Materials*, **18**, 435-441. https://doi.org/10.1038/s41563-019-0338-z

[5] Macalino, S.J.Y., Gosu, V., Hong, S. and Choi, S. (2015) Role of Computer-Aided Drug Design in Modern Drug Discovery. *Archives of Pharmacal Research*, **38**, 1686-1701. https://doi.org/10.1007/s12272-015-0640-5

[6] Shoichet, B.K. (2004) Virtual Screening of Chemical Libraries. *Nature*, **432**, 862-865. https://doi.org/10.1038/nature03197

[7] Sliwoski, G., Kothiwale, S., Meiler, J. and Lowe, E.W. (2014) Computational Methods in Drug Discovery. *Pharmacological*

Reviews, **66**, 334-395. https://doi.org/10.1124/pr.112.007336

[8] Patel, L., Shukla, T., Huang, X., Ussery, D.W. and Wang, S. (2020) Machine Learning Methods in Drug Discovery. *Molecules*, **25**, Article 5277. https://doi.org/10.3390/molecules25225277

[9] Lavecchia, A. (2015) Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today*, **20**, 318-331. https://doi.org/10.1016/j.drudis.2014.10.012

[10] Carpenter, K.A. and Huang, X. (2018) Machine Learning-Based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *Current Pharmaceutical Design*, **24**, 3347-3358. https://doi.org/10.2174/1381612824666180607124038

[11] Liu, T., Lin, Y., Wen, X., Jorissen, R.N. and Gilson, M.K. (2007) Bindingdb: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Research*, **35**, D198-D201. https://doi.org/10.1093/nar/gkl999

[12] Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., *et al*. (2011) ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research*, **40**, D1100-D1107. https://doi.org/10.1093/nar/gkr777

[13] Yang, X., Wang, Y., Byrne, R., Schneider, G. and Yang, S. (2019) Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews*, **119**, 10520-10594. https://doi.org/10.1021/acs.chemrev.8b00728

[14] Lo, Y., Rensi, S.E., Torng, W. and Altman, R.B. (2018) Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today*, **23**, 1538-1546. https://doi.org/10.1016/j.drudis.2018.05.010

[15] Dara, S., Dhamercherla, S., Jadav, S.S., Babu, C.M. and Ahsan, M.J. (2022) Machine Learning in Drug Discovery: A Review. *Artificial Intelligence Review*, **55**, 1947-1999. https://doi.org/10.1007/s10462-021-10058-4

[16] Rogers, D. and Hahn, M. (2010) Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, **50**, 742-754. https://doi.org/10.1021/ci100050t

[17] Morgan, H.L. (1965) The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, **5**, 107-113. https://doi.org/10.1021/c160017a018

[18] Axen, S.D., Huang, X., Cáceres, E.L., Gendelev, L., Roth, B.L. and Keiser, M.J. (2017) A Simple Representation of Three-Dimensional Molecular Structure. *Journal of Medicinal Chemistry*, **60**, 7393-7409. https://doi.org/10.1021/acs.jmedchem.7b00696

[19] Biau, G. (2012) Analysis of a Random Forest Model. *Journal of Machine Learning Research*, **13**, 1063-1095.

[20] Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely Randomized Trees. *Machine Learning*, **63**, 3-42. https://doi.org/10.1007/s10994-006-6226-1

[21] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[22] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3149-3157.

[23] Wu, J., Kong, L., Yi, M., Chen, Q., Cheng, Z., Zuo, H., *et al*. (2022) Prediction and Screening Model for Products Based on Fusion Regression and XGBoost Classification. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 4987639. https://doi.org/10.1155/2022/4987639

[24] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[25] Wojtuch, A., Jankowski, R. and Podlewska, S. (2021) How Can SHAP Values Help to Shape Metabolic Stability of Chemical Compounds? *Journal of Cheminformatics*, **13**, Article No. 74. https://doi.org/10.1186/s13321-021-00542-y

[26] Bernal, L., *et al*. (2023) Improving Machine Learning Classification Predictions through Interpretability. *Molecules*, **28**, 1-15.

[27] Irwin, J.J. and Shoichet, B.K. (2005) ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling*, **45**, 177-182. https://doi.org/10.1021/ci049714+

[28] Wang, J., Miller, D.D. and Li, W. (2022) Molecular Interactions at the Colchicine Binding Site in Tubulin: An X-Ray Crystallography Perspective. *Drug Discovery Today*, **27**, 759-776. https://doi.org/10.1016/j.drudis.2021.12.001

[29] Wójcikowski, M., Zielenkiewicz, P. and Siedlecki, P. (2015) Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *Journal of Cheminformatics*, **7**, Article No. 26. https://doi.org/10.1186/s13321-015-0078-2

[30] Muegge, I. and Oloff, S. (2006) Advances in Virtual Screening. *Drug Discovery Today*: *Technologies*, **3**, 405-411. https://doi.org/10.1016/j.ddtec.2006.12.002

[31] Trott, O. and Olson, A.J. (2009) AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry*, **31**, 455-461. https://doi.org/10.1002/jcc.21334

[32] Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (2001) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews*, **46**, 3-26. https://doi.org/10.1016/s0169-409x(00)00129-0

[33] Butina, D. (1999) Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences*, **39**, 747-750. https://doi.org/10.1021/ci9803381

[34] Daina, A., Michielin, O. and Zoete, V. (2017) SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Scientific Reports*, **7**, Article No. 42717. https://doi.org/10.1038/srep42717

[35] Eberhardt, J., Santos-Martins, D., Tillack, A.F. and Forli, S. (2021) Autodock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling*, **61**, 3891-3898. https://doi.org/10.1021/acs.jcim.1c00203