药物分子多晶型预测技术发展与展望

郭 妹, 丁文星, 彭 勃, 刘劲风, 苏沂菲, 朱 彬*, 任国宾*

(华东理工大学, 生物反应器工程国家重点实验室, 上海市新药设计重点实验室, 上海市细胞代谢光遗传学技术前沿科学研究基地, 制药工程与过程化学教育部工程研究中心, 药物晶体工程技术研究实验室, 上海 200237)

摘要: 大部分的化学药物都存在多晶型现象。药物多晶型的理化性质差异直接影响固态药物制剂产品的稳定性、有效性和安全性, 因此药物多晶型的研究是药物化学、制造和控制的重要组成部分, 也是影响高端原料药及制剂质量的关键因素。多晶型预测技术可以高效指导试错性实验的筛选, 降低传统筛选实验遗漏稳定晶型带来的风险。药物分子多晶型预测技术正在不断发展进步, 最初是基于量子力学和计算化学等理论计算, 后有应用人工智能的机器学习关键技术, 以及联合理论计算和机器学习两者优势共同预测晶体结构。目前, 准确预测药物分子晶型依旧具有挑战性, 但有望借鉴并综合现有技术, 开发更加精确且高效的预测晶型技术。

关键词: 药物多晶型; 晶体结构预测; 机器学习; 计算化学

中图分类号: R917 **文献标识码**: A **文章编号**: 0513-4870(2024)01-0076-08

Development and prospects of predicting drug polymorphs technology

GUO Mei, DING Wen-xing, PENG Bo, LIU Jin-feng, SU Yi-fei, ZHU Bin*, REN Guo-bin*

(*State Key Laboratory of Bioreactor Engineering, Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, Engineering Research Centre of Pharmaceutical Process Chemistry, Laboratory of Pharmaceutical Crystal Engineering & Technology, East China University of Science and Technology, Shanghai 200237, China*)

Abstract: Most chemical medicines have polymorphs. The difference of medicine polymorphs in physico-chemical properties directly affects the stability, efficacy, and safety of solid medicine products. Polymorphs is incomparably important to pharmaceutical chemistry, manufacturing, and control. Meantime polymorphs is a key factor for the quality of high-end drug and formulations. Polymorph prediction technology can effectively guide screening of trial experiments, and reduce the risk of missing stable crystal form in the traditional experiment. Polymorph prediction technology was firstly based on theoretical calculations such as quantum mechanics and computational chemistry, and then was developed by the key technology of machine learning using the artificial intelligence. Nowadays, the popular trend is to combine the advantages of theoretical calculation and machine learning to jointly predict crystal structure. Recently, predicting medicine polymorphs has still been a challenging problem. It is expected to learn from and integrate existing technologies to predict medicine polymorphs more accurately and efficiently.

Key words: medicine polymorphs; crystal structure prediction; machine learning; computational chemistry

收稿日期: 2023-04-11; 修回日期: 2023-06-17.

基金项目: 国家自然科学基金资助项目 (22078094, 21776073, 21908055); 中国博士后科学基金资助项目 (2021M701188, 2019M661410); 苏州市社会发展科技创新资助项目 (2022SS27).

*通讯作者 Tel / Fax: 86-21-64252579, E-mail: zhubin@ecust.edu.cn;

Tel / Fax: 86-21-64253406, E-mail: rgb@ecust.edu.cn

DOI: 10.16438/j.0513-4870.2023-0450

多晶型行为是一种化合物能以两种或两种以上不同的晶体结构存在的现象。大部分的化学药物都存在多晶型行为[1]。在微观层面, 多晶型化合物的内部晶格中分子、原子的排列或构象不同; 从宏观层面, 多晶型化合物的理化性质存在差异, 如: 熔点、溶解度、溶出速率、渗透性、生物利用度等[2-4]。多晶型行为导致的理化性质差异, 直接影响固态药物制剂产品的稳定性和安全性[5,6], 以及在临床治疗中的药效、不良反应[7-10]。治疗艾滋病重磅药物利托那韦, 在上市后意外发现市售晶型是亚稳晶型, 会慢慢转变为稳定晶型, 导致药物溶解度和溶出速率降低, 影响药物的吸收和生物利用度, 这一意料之外的晶型转变造成了上亿美元的经济损失[11]。

目前, 国内外研究药物多晶型行为主要还停留在高通量筛选实验层面[12]。药物科学家普遍采用试错方法, 通过设计不同的实验方法和变量因素[13-15]进行高通量筛选。但高通量筛选存在以下的不足: 第一, 需要投入大量人力、物力、财力和时间; 第二, 通过筛选实验获取多晶型效率低, 并不具有绿色经济性; 第三, 获得的晶型不一定是热力学上的稳定晶型, 给上市药物带来潜在风险。

目前, 有三种主流的理论计算策略研究药物多晶型和预测晶体结构。第一种, 是基于量子力学和计算化学等理论计算预测晶体结构, 已经成功预测分子晶体结构。在色散校正密度泛函理论 (dispersion-corrected density functional theory, DFT-D) 支撑下, 英国伦敦大学学院的Hulme等[16]发展的基于分布式多极的DMACRYS晶体结构优化器, 以及Neumann等[17]发展的GRACE (generation ranking and characterisation engine) 已经能够对部分药物分子进行成功的晶体预测, 但它有自身缺陷, 如计算量大, 成本高, 产生无效结构较多, 效率低等[18-20]。第二种, 是基于机器学习预测晶体结构。如Ryan等[21]选用晶体分子指纹描述符构建训练集, 训练深度神经网络模型能识别数据集中结构相似原子位点, 分析已有晶体结构作为模板, 预测新化合物晶体结构。但基于机器学习的方法目前主要用于预测原子晶体, 很难准确预测具有复杂分子结构和多样晶体堆积方式的药物分子的晶体结构。第三种, 联合理论计算和机器学习方法预测晶体结构, 采用可承受成本的理论计算生成数据集, 在高质量数据集中训练生成高性价比的预测晶体结构机器学习模型。Wengert等[22]联合DFT-D和机器学习, 构建精确预测有机分子晶体结构模型。

1　基于理论计算预测晶体结构

基于理论计算预测晶体结构大都需要在计算之前生成候选晶体结构, 方法包括密度泛函理论, 模拟退火和遗传算法, 这些方法使用不同的势能搜寻最低能量结构, 即基能态结构[23]。基能态结构是热力学上最稳定的晶体结构[19]。因此, 理论计算的重心是能量最小化, 只能计算基态结构, 不能用于确定亚稳态或需要外部温度或压力来保持稳定的结构[19]。

1.1　密度泛函理论　密度泛函理论 (density functional theory, DFT) 是材料科学家认为目前最著名的预测算法[19]。在第一性原理方法中, DFT是研究分子晶体多态性最常用的方法。DFT研究多元系统基能态的电子结构, 计算原子的核子势和电子密度近似值, 依赖空间的电子密度进行量子力学建模, 评估候选晶体结构的电子密度确定每个结构或构型的系统能量, 最小化系统能量以确定最稳定的晶体结构[24,25]。DFT的缺陷是计算成本高, 需要大量的试验来确定适当的能量泛函, 且对非键相互作用存在缺陷[26]。DFT-D能对这一缺陷进行修正, 在占用较少计算资源下, 提高计算系统能量精度, 使准确能量排位成为可能。

Neumann[27]在晶体结构预测 (crystal structure prediction, CSP) 中采用一种方法, 对不同构象有机分子的晶体结构进行DFT计算和经验范德华校正, 生成近似晶格能函数, 成功预测1,4-环己二酮的最稳定晶体结构, 并推广应用于有机分子CSP第四次盲测[28]的四种化合物XII、XIII、XIV和XV, 化学结构式如图1, 该方法计算的晶格能与文章提到的分子力场计算的晶格能之间的均方根误差仅为每原子0.034 kcal·mol^{-1}。

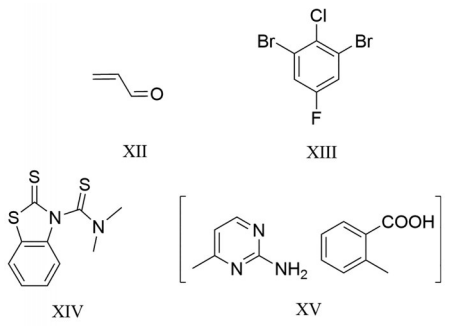

Figure 1　Chemical structures of compound XII, XIII, XIV and XV

1.2　分子力场　分子力场是原子尺度上的一种势能场, 用于描述体系中原子间相互作用的参数化的经验势能函数。在分子以及凝聚相内部, 化学键通常都会有“松弛”的键长值和键角值。当满足这些条件时, 体系的能量及原子间的相互作用会使体系自身调整它的几何形状 (构象), 使其键长值和键角值尽可能接近松弛值, 也使非键相互作用能较小, 从而使得体系的总能

量位于一个较低的状态。

在复杂系统, 经验力场模型以独特的方式模拟原子间的相互作用, 计算原子构型能量, 规避从头计算高昂成本[29]。同时, 经验力场模型被认为是预测纳米粒子结构的关键 (纳米粒子可以是规则的晶体结构, 或无定形, 或以任何晶体学空间群无法描述的方式堆积), 极大程度减少计算时间[29]。力场模型计算的准确性在很大程度上依赖于模型。Swamy等[30]使用了两个独立的力场模型来预测所有已知的TiO_2的晶型, 其成功预测取决于特定的晶型, 即一种力场模型能准确预测低压态的晶型, 但难以预测高压态的晶型; 另一种模型能准确预测高压态的晶型, 但难以预测低压态的晶型。

Zhang等[31]开发的CSP平台根据有机分子的二维结构信息预测最低晶格能的三维晶体结构, CSP平台在经典分子力场的基础上补充附加参数, 用于模拟复杂构象和晶体结构。Zhang等[31]应用CSP平台预测三种柔性药物分子的晶体结构, 如图2的化合物a、b和c, 分别含有2、3和8个可旋转键, 它们实验晶体结构的晶格能均在CSP最低能量结构的4.0 kJ·mol^{-1}窗口以内。

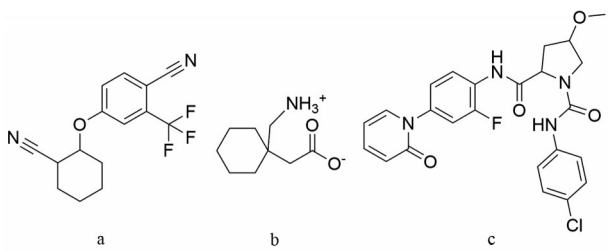

Figure 2 Chemical structures of drug molecules a, b, and c

1.3 全局能量优化算法 全局能量优化算法具体有模拟退火, 遗传算法和粒子群优化算法等, 先随机生成一代候选晶体结构种群, 通过各种结构生成操作迭代更新, 采用量子力学方法评估晶体结构, 直到满足某收敛准则, 找到全局能量最小值[32,33]。

模拟退火利用蒙特卡罗统计和分子动力学等技术模拟原子的运动, 搜寻最低能量晶体结构[34,35]。模拟退火设置的初始温度 (是能量单位, 与实际温度无关), 使高动能的系统原子能克服局部能量势垒, 从而找到全局最小值[36]。模拟退火已经成功地预测了无机、有机分子结构, 可以预测部分无序的材料。模拟退火搜寻整个能量全景图以找到能量最小值, 需要较多的计算时间和资源[19]。

Yang等[37]应用蒙特卡罗阈值算法分析有机分子晶体结构的能量全景图, 评估已知多晶型分子晶体结构之间的能量势垒, 提供晶体结构能量最小值等信息。Yang等[37]计算了刚性分子晶体结构的能量全景图, 有单组分的靛蓝、2-丁炔酸和triptycene trisbenzimidazolone (TTBI), 也有多组分的烟酰胺与苯甲酸的1∶1共晶 (GAZCES), 其化学结构式如图3。

遗传算法是模拟达尔文进化论自然选择的计算模型。遗传算法随机生成初始候选晶体结构为第一代种群, 运行“交叉”“变异”和“复制”等N次迭代操作, 每次迭代选择性保留在一定程度符合适应度函数的结构, 直到寻找到全局能量最小值的晶体结构[38,39], 如全局空间群优化 (GSGO)[40]。但遗传算法同样面临计算时间长问题[19]。

粒子群优化算法 (particle swarm algorithm, PSO), 是Kennedy和Eberhart[41]受到鸟群在空中集体编舞行为启发而提出, 是基于种群的随机全局优化方法, 是一种通过多维搜索的分布式行为算法。鸟群个体飞行受到局部最优或全局最优的影响, 个体学习过去经验来调整飞行速度和方向, 以帮助其在密集快速的群体中飞行。在求解空间中, 根据目标函数值调节粒子参数, 搜索得个体最优解和群体最优解。虽然PSO能高效收敛全局最优解, 但由于晶体的晶格能表面存在大量能量极小值, 导致陷入局部最优解[42]。

Sun等[43]开发的基于云计算和具有海量计算能力的CSP平台, 集成了多种方法和创新技术, 结合使用蒙特卡罗和粒子群全局能量优化算法生成晶体结构和加速能量的收敛。CSP平台辅助预测阿斯利康用于治疗心血管疾病药物AZD1305的多晶型, AZD1305含有

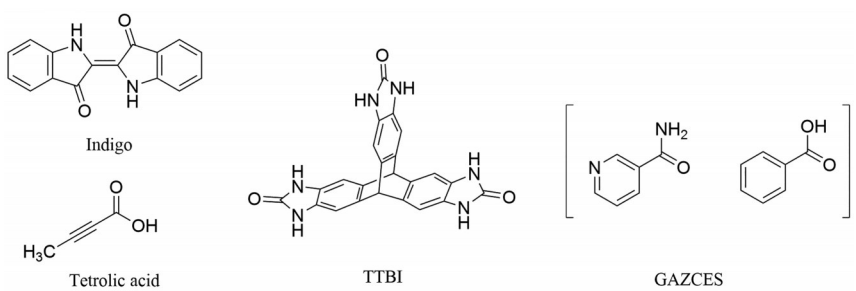

Figure 3 Chemical structures of indigo, tetrolic acid, TTBI and GAZCES. TTBI: Triptycene trisbenzimidazolone, GAZCES: 1:1 co-crystal of nicotinamide and benzoic acid

10个可旋转键 (图4), 成功预测晶型A和晶型B, 根据能量排名确定晶型B为热力学最稳定的晶体结构。

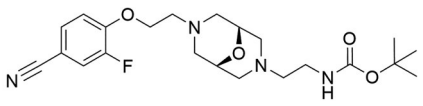

Figure 4　Molecular structure of AZD1305

2　基于机器学习预测晶体结构

机器学习分为监督学习、无监督学习和强化学习。监督学习是目前材料科学中使用最广泛的机器学习形式, 是以大量统计数据集为基础, 提取现有数据信息之间相互关联的描述符, 再采用这些描述符训练机器学习模型预测全新样本[44]。机器学习算法可以被认为是一个使用大量数据的复杂启发式模型的拟合过程[45]。机器学习算法采用包含大量晶体结构信息的数据库, 通过机器学习算法表达晶体数据之间的内禀关联, 对于给定的新化合物预测其晶体结构。

预测晶体结构的机器学习算法, 有基于线性的回归/分类方法, 如支持向量机; 有基于非线性的回归方法, 如高斯过程回归; 基于决策树的方法, 如随机森林; 基于神经网络的非线性模型, 如多层感知器等。

2.1　支持向量机　支持向量机 (support vector machines, SVM) 是基于分割数据的机器学习算法, 分为回归和分类模型。SVM通过将数据绘制在N维空间, 尝试找出分类数据的超平面。SVM通过核函数和调整维度优化算法预测晶体结构[46]。这一方法可以有效避免依靠DFT带来的高昂计算成本。Honrao等[47]用偏径向分布函数表示Li-Ge晶体结构信息, 采用通用的基于核学习算法, 如核岭回归 (kernel ridge regression, KRR) 和支持向量回归 (support vector regression, SVR), 优化训练集上KRR和SVR两个机器学习模型, 以预测测试集上晶体结构的最小能量构型的生成能。通过对比预测能量和DFT评估的晶体结构能量, 机器学习预测能量的精度达到1 kcal·mol^{-1} [47]。

2.2　随机森林　随机森林 (random forests, RFs) 从数据集中训练众多独立的决策树, 建立一个具有预测能力的模型。RFs利用数据中的特征子集训练不同决策树, 分类或回归输入数据。决策树分为回归和分类两大类。分类决策树预测离散值, 回归决策树预测连续值。在RFs算法中, “森林”中所有的决策树具有不同的组成结构。Graser等[19]应用RFs对Pearson晶体数据库中24 215种化合物晶体结构分类, 研究不同的数据集预处理方法对模型分类性能的影响。将数据集中实例数少于某个截止数的原型归为“其他”类。根据截止数的选择, “其他”类占数据集的92.51%至64.1%, 大大降低了数据集的复杂度, 研究了模型的预测能力随截止数的变化情况 (图5)。

2.3　神经网络　神经网络 (neural networks, NNs) 能发现大量数据间复杂内禀关系。NNs由输入层、隐藏层和输出层组成, 由神经元的相互连接, 通过调整连接权和阈值来训练模型, 直到输出值接近训练集中的原数值[48]。Ryan等[21]选用晶体原子指纹描述符构建训练集, 训练深度神经网络模型能识别数据集中结构相似原子位点, 以已有晶体结构为模板, 预测新化合物晶体结构。在Mn-Ge二元体系或Li-Mn-Ge三元体系中 (表1), 化合物的晶体结构参照ICSD数据集中已有化合物的结构模板, 计算两者之间相似度。

Kilgour等[49]应用图神经网络 (graph neural

	$Ca(Ca_{0.5}Nd_{0.5})_2NbO_6$	$Ca_2Nb_2O_7$	$CaTiO_3$	$GeAl_2Ga_2$	Cu	CuZrSiAs	FeAs	$GdFeO_3$	K_2NiF_4	$LaAlO_3$	$MgAl_2O_4$	$MgCu_2$	NaCl	$NaFeO_2$	TiNiSi	Other	Recall
$Ca(Ca_{0.5}Nd_{0.5})_2NbO_6$	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	0.686
$Ca_2Nb_2O_7$	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0.655
$CaTiO_3$	0	0	133	0	0	0	0	12	0	5	0	0	0	0	1	105	0.522
$GeAl_2Ga_2$	0	0	0	161	0	0	0	0	0	0	0	0	0	0	0	28	0.847
Cu	0	0	0	0	56	0	0	0	0	0	0	0	0	0	0	70	0.444
CuZrSiAs	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	21	0.816
FeAs	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	15	0.854
$GdFeO_3$	0	0	9	0	0	0	0	454	0	19	0	0	1	0	0	120	0.753
K_2NiF_4	0	0	0	0	0	0	0	3	81	2	0	0	0	0	0	56	0.570
$LaAlO_3$	0	0	2	0	0	0	0	33	1	92	0	0	0	0	0	27	0.594
$MgAl_2O_4$	0	0	0	0	0	0	0	0	0	0	315	0	0	0	0	69	0.820
$MgCu_2$	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	53	0.523
NaCl	0	0	0	0	0	0	0	1	0	0	1	0	140	1	0	81	0.625
$NaFeO_2$	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0	34	0.755
TiNiSi	0	0	0	1	0	0	3	0	0	0	0	0	0	0	45	65	0.395
Other	0	5	29	5	37	2	18	59	21	16	31	14	21	12	8	20984	0.986
Total	105	100	173	167	93	95	109	562	103	134	103	72	162	118	54	21821	0.950
Precision	0.895	0.950	0.769	0.964	0.602	0.979	0.807	0.808	0.786	0.687	0.908	0.806	0.864	0.890	0.833	0.962	

Figure 5　Confusion matrix of algorithm with a cutoff size of 100

Table 1 Chemical compositions predicted for the Mn-Ge and Li-Mn-Ge systems

Mn-Ge system			Li-Mn-Ge system		
Composition	Template structure (ICSD code)	Likelihood	Composition	Template structure (ICSD code)	Likelihood
1. $MnGe_2$	NiS_2 (68169)	0.890	1. $LiMn_2Ge_2$	Li_2CuO_2 (108666)	0.819
2. $MnGe_3$	$CoSb_3$ (62111)	0.886	2. $LiMn_2Ge_4$	$MgCr_2O_4$ (75623)	0.813
3. $MnGe_4$	CrP_4 (2790)	0.859	3. Li_6MnGe_6	Li_6UO_6 (48209)	0.772
4. MnGe	NiS (151599)	0.831	4. Li_2MnGe	PrCuSO (96345)	0.759
5. Mn_2Ge_3	Fe_2O_3 (164008)	0.785	5. $Li_3Mn_2Ge_3$	$Ge_3Rh_2Se_3$ (261240)	0.755
6. Mn_3Ge_4	Mn_2ZnO_4 (166522)	0.780	6. $LiMnGe_2$	$ScAgSe_2$ (155115)	0.754
7. Mn_7Ge_6	Li_6UO_6 (48209)	0.762	7. $LiMnGe_3$	$MnCoO_3$ (31854)	0.753
8. Mn_3Ge_5	Sm_3Ge_5 (416581)	0.739	8. $LiMn_6Ge_6$	Li_6UO_6 (48209)	0.746
9. $MnGe_7$	Li_6TeO_6 (40247)	0.729	9. LiMnGe	CoAsS (69129)	0.738
10. Mn_6Ge	B_6As (68151)	0.727	10. Li_3MnGe_2	$CoCu_2O_3$ (33996)	0.723
12. Mn_2Ge	Cs_2Se (41687)	0.701			
18. Mn_3Ge	$LiNbO_2$ (75880)	0.608			
26. Mn_5Ge_3	Fe_5Si_3 (99973)	0.521			
27. Mn_5Ge_2	V_4SiSb_2 (82564)	0.520			
239. $Mn_{11}Ge_8$	K_8TI_{11} (370009)	0.008 52			
332. $Mn_{23}Ge_9$	$Li_6Ca_{17}Hg_9$ (420846)	0.001 04			

networks, GNN) 建模预测有机分子晶体结构的稳定性和密度, MolXtalNet-S 模型输入不同晶胞结构信息预测晶体结构的稳定性, MolXtalNet-D 模型输入分子构象信息预测晶体结构的密度, 两个模型训练的数据为分子特征, 如分子片段和表面积, 而不是分子几何结构中原子坐标, 但这些结果显示了 GNN 作为 CSP 管道一部分的巨大潜力。将几何深度学习方法应用于有机分子晶体的研究, 这些方法结合了速度、质量和广泛的适用性, 使其成为加速分子晶体结构预测的强大工具。下一步的关键是开发更先进的分子晶体结构生成模型, 通过极快速地采样高质量初始候选结构来加速 CSP。

Liang 等[20]利用多层感知层开发神经网络模型 (CRYSPNet), 在无机晶体结构的 10 万多条目中进行了训练和验证机器学习模型, 输入无机材料的化学组成预测晶体结构的布拉维晶格、空间群和晶格参数, 模型具有良好的预测性能。表 2 为训练预测 14 种布拉维晶格空间群的交叉验证精度和前 3 位精度, 预测模型的性能明显高于随机选择。

Table 2 The cross-validated accuracy and top-3 accuracy of models predicting the space group

Bravais lattice	Accuracy model/%	Accuracy random/%	Top-3 accuracy model/%
Cubic (F)	90.7 ± 0.6	36.5	98.8 ± 0.2
Cubic (I)	87.1 ± 1.7	17.6	95.1 ± 1.2
Cubic (P)	87.3 ± 1.6	30.4	95.8 ± 0.1
Hex. (P)	74.8 ± 0.9	13.3	87.9 ± 0.6
Rhom. (P)	81.7 ± 1.4	26.6	94.5 ± 1.0
Tetra. (I)	81.8 ± 1.1	29.1	92.7 ± 1.0
Tetra. (P)	78.2 ± 1.4	13.6	86.9 ± 1.0
Ortho. (F)	72.7 ± 4.7	24.8	93.0 ± 0.3
Ortho. (I)	78.0 ± 3.1	40.6	91.9 ± 2.6
Ortho. (C)	72.2 ± 1.9	33.0	90.8 ± 0.8
Ortho. (P)	63.8 ± 1.4	25.6	81.0 ± 1.0
Mono. (C)	74.0 ± 1.1	40.1	94.0 ± 0.7
Mono. (P)	79.4 ± 1.1	58.0	81.0 ± 1.0
Triclinic (P)	94.7 ± 0.9	87.1	100

3 联合理论计算和机器学习方法预测晶体结构

化合物分子通过不同排列和堆积方式形成不同的晶体结构称为多晶型行为。预测多晶型主要的挑战在于大量可能共存的晶型之间的能量差异小[50, 51]。除此之外还需应用高层次的理论, 充分描述不同分子间相互作用如氢键、静电力、色散效应之间的细微相互作用[22]。因此预测多晶型行为, 需要在分子构象空间内高精度且高效计算能量, 预测同一化合物的不同的晶体结构。近些年, 涌现一种精度和速度并存的高性价比的预测晶体结构方法——联合理论计算和机器学习的策略。该方法采用可承受成本的能量计算方法生成数据集, 如 DFTB (density-functional tight-binding)[52,53], 在高质量数据集中训练生成高性价比的预测晶体结构的机器学习模型。

势能面 (potential energy surface, PES) 表示原子间相互作用势, 能识别具有最低能量的晶体结构, 从而取代计算昂贵的 DFT[54]。机器学习的训练集来源于量子力学计算, 机器学习算法能用复杂的函数表示的 PES, 称为机器学习势 (machine learning potential, MLP)[55]。MLP 具有独立计算原子能量的优点, 使其有可能应用于不同原子数的系统。开发 MLP 的流程有两个步骤[56]。第一步, 设置描述符用于量化原子邻域构型的变换坐标, 类似于结构预测方法中表示结构的方法。第二步, 输入描述符, 用量子力学的参考数据建立拟合 PES 的回归模型, 预测结构的能量, 识别具有最低能量的晶体结构。采用 MLP 用于结构预测是一个很有前途的方法, 但前提是量子力学能够准确计算参考数据[56]。

Podryabinkin等[57]利用进化算法USPEX[33]和MLP, 从头开始构建原子间相互作用模型, 取代计算昂贵的DFT, 加速晶体结构预测, 预测了元素的同素异形体, 正确生成碳元素的石墨, 钻石和方石结构; 正确生成所有已知的硼的同素异形体, 包括晶胞内含有106个原子的无序的β-硼, 其结构与文献[58]中提供的结构具有相同的能量 (DFT能量相差小于1 meV per atom)。

Wang等[59]基于NNs模型拟合PES预测合金的晶体结构, 采用DP+CALYPSO策略预测二元合金体系的晶体结构, 预测潜在的稳定金属间化合物的Al-Mg二元体系, 发现了6个亚稳态晶体结构具有热力学稳定、动力稳定和机械稳定。

Hong等[56]利用DFT能量计算构建数据集, 训练神经网络势 (neural network potential, NNP) 模型, 通过仅给定的化学组成信息能预测无机化合物晶体结构, 比较了Ba_2AgSi_3、Mg_2SiO_4、$LiAlCl_4$、$InTe_2O_5F$在实验阶段以及理论上生成的低能晶体结构中的NNP和DFT的能量, 证明了训练后的NNP在晶体结构预测中的适用性, 以及利用NNP的进化搜索方法可以更有效地识别亚稳态。

Wengert等[22]联合机器学习和DFT-D等理论能量计算方法, 构建了预测有机分子晶体结构的精确Δ-ML模型。采用四个代表性的分子验证Δ-ML模型精度, 分别为水 (H_2O), 吡嗪 (C_4N_2)、草酸 ($C_2O_4H_2$) 和丁炔酸 ($C_4O_2H_4$)。选用有机分子CSP第六次盲测目标XXII[60] (the tricyano-1,4-dithiino[c]-isothiazole, C_8N_4S), 化学结构式如图6, 验证了Δ-ML模型依靠理论计算提高了能量排序的相关性, 并能正确识别实验筛选确定的最稳定晶体结构。

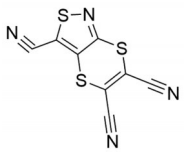

Figure 6　Chemical structure of compound XXII

Kapil等[61]提出了一个高效且稳健的“端到端”框架, 该框架结合了最先进的电子结构、MLP和自由能的计算方法, 以前所未有的准确性从头计算有机分子的吉布斯自由能。分别计算苯、甘氨酸和琥珀酸多晶型化合物, 其计算和预测的热力学稳定性结果排名均与实验保持高度一致, 因此他们认为工业相关有机分子的热力学稳定性预测研究不再是艰巨的任务。

4　总结与展望

药物分子的多晶型行为对药物稳定性、有效性和安全性以及药物制剂开发和药品运输储存具有深远的影响。不同晶型的药物分子的理化性质具有明显差异, 影响其生物利用度和治疗效果, 并且药物晶型专利是具有极强排他性的专利, 可以延长创新药生命周期。高通量筛选晶型实验需要耗费大量人力物力, 不能保证筛选到热力学上稳定晶型; 在预测晶型技术的研究发展进程中, 量子力学和计算化学等理论计算方法, 虽已经成功预测药物分子晶体结构, 但存在的计算成本高等缺陷; 机器学习方法主要预测原子晶体结构, 很难准确预测结构和晶体堆积复杂的药物分子; 联合机器学习和量子力学和计算化学等理论计算方法, 是一种有前景的预测药物分子晶体结构策略。

由于多晶型行为成因复杂, 由分子内部因素决定, 并有外部环境限制[62]。分子特征识别后形成单分子体或多分子体, 再以某些排列方式堆砌成热力学相对稳定的多晶型。而分子识别或堆砌又受外部因素诱导, 在不同的实验环境区间形成特定的晶体排列方式。因此, 借鉴机器学习在预测原子晶体结构的经验, 结合可接受成本的量子力学和计算化学等理论计算, 综合实验参数和结晶工艺产生大量准确数据, 有望能开发更精确且高效的预测晶型技术。

作者贡献: 郭妹负责文献检索、查找分析资料并撰写论文初稿、检查修改论文; 丁文星、彭勃、刘劲风和苏沂菲参与查找资料; 朱彬和任国宾负责提出研究内容和方向, 论文修改并定稿。

利益冲突: 所有作者均声明不存在利益冲突。

References

[1] Vippagunta SR, Brittain HG, Grant DJW. Crystalline solids [J]. Adv Drug Deliv Rev, 2001, 48: 3-26.

[2] Byrn SR. Solid State Chemistry of Drugs [M]. Pittsburgh: Academic Press, 1982: 346-350.

[3] Hilfiker R. Polymorphism in the Pharmaceutical Industry [M]. Weinheim: Wiley-VCH Verlag GmbH, 2006: 211-233.

[4] Fang ZY, Xing C, Xing WH, et al. Preparation, characterization, transformation and solubility of imatinib-oxalate salt polymorphs [J]. Acta Pharm Sin (药学学报), 2021, 56: 3153-3158.

[5] Matsuda Y, Akazawa R, Teraoka R, et al. Pharmaceutical evaluation of carbamazepine modifications: comparative study for photostability of carbamazepine polymorphs by using Fourier-transformed reflection-absorption infrared spectroscopy and colorimetric measurement [J]. J Pharm Pharmacol, 1994, 46: 162-167.

[6] Raijada DK, Prasad B, Paudel A, et al. Characterization of degradation products of amorphous and polymorphic forms of clopidogrel bisulphate under solid state stress conditions [J]. J

Pharm Biomed Anal, 2010, 52: 332-344.

[7] Singhal D, Curatolo W. Drug polymorphism and dosage form design: a practical perspective [J]. Adv Drug Deliv Rev, 2004, 56: 335-347.

[8] Aguiar AJ, Krc JJ, Kinkel AW, et al. Effect of polymorphism on the absorption of chloramphenicol from chloramphenicol palmitate [J]. J Pharm Sci, 1967, 56: 847-853.

[9] Aguiar AJ, Zelmer JE. Dissolution behavior of polymorphs of chloramphenicol palmitate and mefenamic acid [J]. J Pharm Sci, 1969, 58: 983-987.

[10] Raw AS, Furness MS, Gill DS, et al. Regulatory considerations of pharmaceutical solid polymorphism in abbreviated new drug applications (ANDAs) [J]. Adv Drug Deliv Rev, 2004, 56: 397-414.

[11] Bauer J, Spanton S, Henry R, et al. Ritonavir: an extraordinary example of conformational polymorphism [J]. Pharm Res, 2001, 18: 859-866.

[12] Feng ZQ, Deng W, Guo ZR. High-throughput crystallization in pharmaceutical research and development [J]. Acta Pharm Sin (药学学报), 2005, 40: 481-485.

[13] Ding XH. Industrial Crystallization (工业结晶) [M]. Beijing: Chemical Industry Press, 1985: 26-73.

[14] Threlfall T. Crystallisation of polymorphs: thermodynamic insight into the role of solvent [J]. Org Process Res Dev, 2000, 4: 384-390.

[15] Kitamura M. Strategy for control of crystallization of polymorphs [J]. CrystEngComm, 2009, 11: 949-964.

[16] Hulme AT, Price SL, Tocher DA. A new polymorph of 5-fluorouracil found following computational crystal structure predictions [J]. J Am Chem Soc, 2005, 127: 1116-1117.

[17] Neumann MA, Perrin MA. Can crystal structure prediction guide experimentalists to a new polymorph of paracetamol? [J]. CrystEngComm, 2009, 11: 2475-2479.

[18] Tong Q, Gao P, Liu H, et al. Combining machine learning potential and structure prediction for accelerated materials design and discovery [J]. J Phys Chem Lett, 2020, 11: 8710-8720.

[19] Graser J, Kauwe SK, Sparks TD. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons [J]. Chem Mater, 2018, 30: 3601-3612.

[20] Liang H, Stanev V, Kusne AG, et al. CRYSPNet: crystal structure predictions *via* neural networks [J]. Phys Rev Mater, 2020, 4: 123802.

[21] Ryan K, Lengyel J, Shatruk M. Crystal structure prediction *via* deep learning [J]. J Am Chem Soc, 2018, 140: 10158-10168.

[22] Wengert S, Csányi G, Reuter K, et al. Data-efficient machine learning for molecular crystal structure prediction [J]. Chem Sci, 2021, 12: 4536-4546.

[23] Woodley SM, Catlow R. Crystal structure prediction from first principles [J]. Nat Mater, 2008, 7: 937-946.

[24] Beran GJO. Modeling polymorphic molecular crystals with electronic structure theory [J]. Chem Rev, 2016, 116: 5567-5613.

[25] Aubrey-Medendorp C, Swadley MJ, Li T. The polymorphism of indomethacin: an analysis by density functional theory calculations [J]. Pharm Res, 2008, 25: 953-959.

[26] Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies *via Voronoi tessellations* [J]. Phys Rev B, 2017, 96: 1-12.

[27] Neumann MA. Tailor-made force fields for crystal-structure prediction [J]. J Phys Chem B, 2008, 112: 9810-9829.

[28] Neumann MA, Leusen FJJ, Kendrick J. A major advance in crystal structure prediction [J]. Angew Chem Int Ed, 2008, 47: 2427-2430.

[29] Shao GF, Tu NN, Liu TD, et al. Structural studies of Au-Pd bimetallic nanoparticles by a genetic algorithm method [J]. Phys E (Amsterdam, Neth), 2015, 70: 11-20.

[30] Swamy V, Gale JD, Dubrovinsky LS. Atomistic simulation of the crystal structures and bulk moduli of TiO_2 polymorphs [J]. J Phys Chem Solids, 2001, 62: 887-895.

[31] Zhang P, Wood GPF, Ma J, et al. Harnessing cloud architecture for crystal structure prediction calculations [J]. Cryst Growth Des, 2018, 18: 6891-6900.

[32] Hofmann DWM, Apostolakis J. Crystal structure prediction by data mining [J]. J Mol Struct, 2003, 647: 17-39.

[33] Glass CW, Oganov AR, Hansen N. USPEX—evolutionary crystal structure prediction [J]. Comput Phys Commun, 2006, 175: 713-720.

[34] Doll K, Schon JC, Jansen M. Structure prediction based on ab initio simulated annealing [J]. J Phys Conf Ser, 2008, 117: 012014.

[35] Huq A, Stephens PW. Subtleties in crystal structure solution from powder diffraction data using simulated annealing: ranitidine hydrochloride [J]. J Pharm Sci, 2003, 92: 244-249.

[36] Harris KJ, Foster JM, Tessaro MZ, et al. Structure solution of metal-oxide Li battery cathodes from simulated annealing and lithium NMR spectroscopy [J]. Chem Mater, 2017, 29: 5550-5557.

[37] Yang S, Day GM. Global analysis of the energy landscapes of molecular crystal structures by applying the threshold algorithm [J]. Commun Chem, 2022, 5: 86.

[38] Lloyd LD, Johnston RL, Salhi S. Strategies for increasing the efficiency of a genetic algorithm for the structural optimization of nanoalloy clusters [J]. J Comput Chem, 2005, 26: 1069-1078.

[39] Kim S, Orendt AM, Ferraro MB, et al. Crystal structure prediction of flexible molecules using parallel genetic algorithms with a standard force field [J]. J Comput Chem, 2009, 30: 1973-1985.

[40] Trimarchi G, Freeman AJ, Zunger A. Predicting stable stoichiometries of compounds *via* evolutionary global space-group optimization [J]. Phys Rev B, 2009, 80: 092101.

[41] Kennedy J, Eberhart RC. Proceedings of the Sixth International Symposium on Micro Machine and Human Science [C].

Nagoya: IEEE Service Center, 1995: 39-43.

[42] Wang Y, Lv J, Zhu L, et al. CALYPSO: a method for crystal structure prediction [J]. Comput Phys Commun, 2012, 183: 2063-2070.

[43] Sun G, Liu X, Abramov YA, et al. Current state-of-the-art in-house and cloud-based applications of virtual polymorph screening of pharmaceutical compounds: a challenging case of AZD1305 [J]. Cryst Growth Des, 2021, 21: 1972-1983.

[44] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor [J]. Phys Rev Lett, 2015, 114: 105503.

[45] Musil F, De S, Yang J, et al. Machine learning for the structure-energy-property landscapes of molecular crystals [J]. Chem Sci, 2018, 9: 1289-1300.

[46] Balachandran PV, Theiler J, Rondinelli JM, et al. Materials prediction *via* classification learning [J]. Sci Rep, 2015, 5: 13285.

[47] Honrao S, Anthonio BE, Ramanathan R, et al. Machine learning of ab-initio energy landscapes for crystal structure predictions [J]. Comput Mater Sci, 2019, 158: 414-419.

[48] Zhou ZH. Machine Learning (机器学习) [M]. Beijing: Tsinghua University Press, 2016: 97-115.

[49] Kilgour M, Rogal J, Tuckerman M. Geometric deep learning for molecular crystal structure prediction [J]. J Chem Theory Comput, 2023. DOI: 10.1021/acs.jctc.3c00031.

[50] Cruz-Cabeza AJ, Reutzel-Edens SM, Bernstein J. Facts and fictions about polymorphism [J]. Chem Soc Rev, 2015, 44: 8619-8635.

[51] Nyman J, Day GM. Static and lattice vibrational energy differences between polymorphs [J]. Crystengcomm, 2015, 17: 5154-5165.

[52] Brandenburg JG, Grimme S. Accurate modeling of organic molecular crystals by dispersion-corrected density functional tight binding (DFTB) [J]. J Phys Chem Lett, 2014, 5: 1785-1789.

[53] Mortazavi M, Brandenburg JG, Maurer RJ, et al. Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding [J]. J Phys Chem Lett, 2018, 9: 399-405.

[54] Wang YC, Ma YM. Perspective: crystal structure prediction at high pressures [J]. J Chem Phys, 2014, 140: 040901.

[55] Behler J. Perspective: machine learning potentials for atomistic simulations [J]. J Chem Phys, 2016, 145: 170901.

[56] Hong C, Choi JM, Jeong W, et al. Training machine-learning potentials for crystal structure prediction using disordered structures [J]. Phys Rev B, 2020, 102: 224104.

[57] Podryabinkin EV, Tikhonov EV, Shapeev AV, et al. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning [J]. Phys Rev B, 2019, 99: 064114.

[58] van Setten MJ, Uijttewaal MA, De Wijs GA, et al. Thermodynamic stability of boron: the role of defects and zero point motion [J]. J Am Chem Soc, 2007, 129: 2458-2465.

[59] Wang H, Zhang Y, Zhang L, et al. Crystal structure prediction of binary alloys *via* deep potential [J]. Front Chem, 2020, 8: 589975.

[60] Reilly AM, Cooper RI, Adjiman CS, et al. Report on the sixth blind test of organic crystal structure prediction methods [J]. Acta Crystallogr B Struct Sci Cryst Eng Mater, 2016, 72: 439-459.

[61] Kapil V, Engel Edgar A. A complete description of thermodynamic stabilities of molecular crystals [J]. Proc Natl Acad Sci U S A, 2022, 119: e2111769119.

[62] Gong JB, Sun J, Wang JK. Research progress of industrial crystallization towards intelligent manufacturing [J]. J Chem Ind Eng (化工学报), 2018, 69: 4505-4517.