特约评述

DOI: 10.12211/2096-8280.2022-066

数据驱动的酶反应预测与设计

曾涛，巫瑞波
（中山大学药学院，广东 广州 510006）

摘要：酶催化已经在日用化学品、药物和功能材料等生产中得到越来越广泛的应用。酶，作为生物制造业的核心“芯片”，其催化反应的预测与设计是推动传统生物制造走向生物智造发展的核心驱动力之一。然而目前我们对大自然酶催化的了解仍然非常有限，这严重阻碍了我们对酶催化空间的探索和利用。随着大数据时代的到来，数据驱动的计算模拟已经成为酶催化新空间的挖掘及其功能优化设计的重要手段。各种计算工具和平台的开发正极大地加速并赋能于酶学相关领域的各类实验研究。本文针对酶催化过程中底物、产物和酶的预测及设计方法进行了综述，概述了近年来酶反应相关的数据库，汇总比较了数据驱动的酶反应设计工具，着重介绍了深度学习在该领域的应用，并从数据、模型、算法、平台等多方面展望和探讨了数据驱动型计算方法在酶反应预测与设计领域的发展前景。

关键词：大数据；机器学习；酶催化；酶设计；生物合成

中图分类号：Q814.9　　**文献标志码：**A

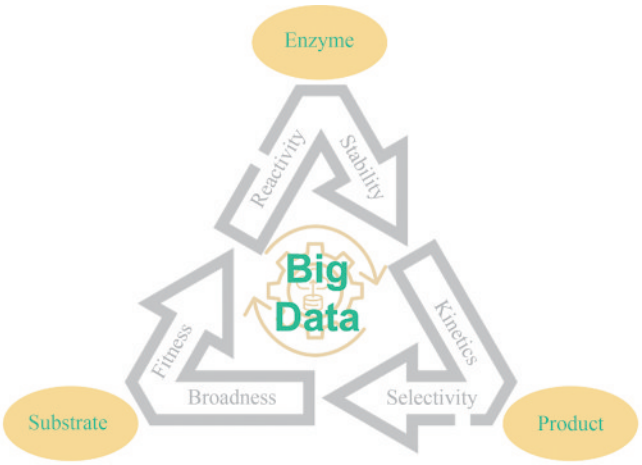
Data-driven prediction and design for enzymatic reactions

ZENG Tao，WU Ruibo
（*School of Pharmaceutical Science*，*Sun Yat-Sen University*，*Guangzhou 510006*，*Guangdong*，*China*）

Abstract: Enzymes are efficient catalysts with substrate specificity and stereo- and regioselectivity, which are widely used in producing chemicals, drugs and materials. Enzymes are cores for biocatalysis, and thus prediction on their functions and design of enzymatic reactions are driving forces for intelligent biomanufacturing through biocatalysis. So far limited understanding on enzymatic catalysis hinders the exploration of enzymatic reactions for industrial applications. For example, it is difficult to predict enzymatic activities on unreported substrates, to elucidate synthetic routes for newly found structures of enzymes, and to redesign enzymes for specific scenarios. In the era of big data, data-driven approaches have exhibited powerful capabilities for exploring enzymatic reactions, by filling gap between the large corpora of enzymatic data and limited understanding on functions of the enzymes. Recently, computational tools and platforms have greatly accelerated experimental research, and improved the design-build-test-learn cycle. Herein we review progress in computational tools for enzymatic reaction prediction and design, focusing on the

收稿日期：2022-11-23 修回日期：2022-12-27
基金项目：广东省重点研发计划（2022B1111080005）
引用本文：曾涛，巫瑞波. 数据驱动的酶反应预测与设计[J]. 合成生物学，2023，4(3)：535-550
Citation：ZENG Tao，WU Ruibo. Data-driven prediction and design for enzymatic reactions[J]. Synthetic Biology Journal，2023，4(3)：535-550

application of deep learning methods in this field. Referring to key elements (substrate, product and enzyme) for enzymatic reactions, related databases are summarized. Then, the data-driven approaches for forward and backward prediction of enzymatic reaction routes and functions of enzymes, their design and theoretical calculation for enzymatic catalysis are addressed. Finally, the status and prospective of data-driven approaches for enzymatic catalysis prediction and design, including the data, model, algorithm and platform, are discussed.



Keywords: big data; machine learning; enzymatic catalysis; enzyme design; biosynthesis

酶是自然界中的能工巧匠，其以高效、精准的手段催化生物体内大量化学反应[1]。酶催化的应用具有悠久的历史，最早可以追溯到古人的酿酒技术[2]。随着科学的进步，我们对酶催化的过程有了更深入的理解，同时在“碳中和”的大背景下，酶催化也因其高效环保、条件温和以及高立体选择性等优点被广泛应用于医药、化工等各领域[3-5]。此外，基于生物底盘的异源生物合成也非常依赖于由一系列酶催化反应组成的生物合成路线的优化与设计[6]。因此，酶被视为生物制造领域的核心“芯片”，而酶反应的机制解析与优化设计是“芯片”升级换代的重要驱动力。

在酶反应机制解析方面，虽然当前通过实验和计算（如多尺度模拟方法[7]等）结合来解析酶的三维结构、功能及其催化反应机制越来越流行[8-10]，但因为直接验证反应机理的实验手段有限，而QM/MM等多尺度模拟的计算代价仍然较为昂贵，当前人们所探索的酶促过程只是酶反应空间中的冰山一角。而随着测序技术的发展，有大量酶序列的功能有待阐明[11]，现有天然产物数据库也是日益丰富，但其中大量结构的生物合成反应路线仍有待解析[12]，这些都严重制约了新酶的发现与天然产物的生物制造。在酶反应优化设计方面，尽管AlphaFold2[13]等蛋白结构预测工具为从一维序列到三维蛋白结构的理论预测提供了利器，但是基于序列的酶功能预测以及以功能为导向的蛋白序列设计相关算法进展则相对更滞后[14-15]。此外，在工业酶领域，如何拓宽酶的底物谱、改善酶反应选择性、提升酶催化效率或稳定性是重要的研究方向[16]，但目前这些研究在很大程度上仍然依赖于研究人员的知识和经验。而当前广泛采用的多轮次“设计-构建-测试-学习（DBTL）”循环策略，往往要消耗大量的时间和资源。

随着大数据时代的到来，利用计算机从已知的各类数据中挖掘背后隐藏的序列与酶反应相关性成为可能。例如，合成路线与酶功能的计算预测[17-18]可助力于生物合成途径的设计与优化，而基于代谢组和基因组数据的代谢网络模型[19]以及全细胞模型[20]则可以对物种或细胞的代谢生长过

程进行模拟，进而对上述设计路线进行计算测试。总之，近年来这些数据驱动的模型正在逐渐深入参与到传统DBTL的各个环节中，从而加速DBTL循环而缩短时间周期，抑或代替实验环节来缩减实验成本[21-22]。

基于上述现状，本文首先整理了常用的酶反应数据库，然后以反应底物、产物和酶为三个抓手对近年来酶反应预测和设计的计算工具进行了梳理，最后对数据驱动的酶反应预测与设计研究进行了展望。

1 酶反应数据库

在数字信息的时代，数据就是生产力，因此生物信息研究领域出现了许多高质量的数据库，不仅为传统的实验人员提供了信息服务，更是在数据驱动的计算工具开发中发挥了关键作用。表1汇总了常用的几个酶反应相关的数据库，这些数据库都有相应的Web服务器，可以直接在线访问和检索，并且除了Reaxys[32]外，其他数据库都可以免费下载使用。

在天然产物代谢领域常用的数据库为KEGG[23]和MetaCyc[24]，两个数据库中均搜集了大量的酶反应，并且以生物合成途径对反应进行了不同层级结构的注释，如MetaCyc中针对次级代谢产物生物合成中划分有萜类生物合成途径、聚酮生物合成途径等，而萜类合成途径中又有单萜生物合成途径、萜类生物碱合成途径等等。Rhea[25]是由瑞士生物信息学研究所建立并维护的专门针对酶反应的数据库，其共同参与维护的还有蛋白序列数据库Uniprot[33]，因此Rhea中的反应具有全面的酶信息注释，且与Uniprot高度关联。BRENDA[26]和SABIO-RK[27]则是致力于搜集酶反应动力学信息的数据库，包括米氏常数（K_m）、催化常数（k_{cat}）以及酶反应条件如温度、酸碱度（pH）等，而且BRENDA还提供了酶的详细分类（EC number等）和命名信息。Reactome[28]、PathBank[29]、HMDB[30]是具有不同侧重点的生物通路数据库，它们搜集了包括各种代谢反应、信号转导在内的各种信号通路数据。基于以上众多数据库各有侧重，但同时又有大量重复数据的情况，Pagni等[31]对KEGG、MetaCyc、HMDB等12个数据库的反应和酶进行汇总去重，构建了MetaNetX数据库，可用于基因组尺度的代谢网络模型的构建和分析。除了上述开源数据库以外，也有一些商业数据库可提供信息的检索和下载服务，如Elsevier旗下的Reaxys[32]数据库，包含了从各种专利和文献中提取的有机反应和酶反应数据。

2 酶反应预测与设计

反应底物、产物和酶是认知酶反应的三个核心要素，因此大部分酶反应的计算预测和设计方法都围绕这三点展开，且计算模型通常是通过其中之一（或之二）对剩余要素进行预测（图1）：围绕底物、产物的正向或逆向预测探索反应和代

表1 酶反应数据库

Table 1 Databases of enzymatic reactions

数据库	特点	网址
KEGG[23]	具有物种、基因组、酶等多水平注释的合成（代谢）反应数据库	https://www.kegg.jp/kegg
MetaCyc[24]	以全面的初级/次级代谢产物合成途径对反应进行注释	https://metacyc.org
Rhea[25]	全面的生物酶反应数据库，与Uniprot高度关联	https://www.rhea-db.org
BRENDA[26]	对酶的各项信息（如分类、反应、参数等）进行详细注释	https://www.brenda-enzymes.org
SABIO-RK[27]	包含酶反应的动力学参数、反应条件等信息	https://sabiork.h-its.org
Reactome[28]	综合的生物通路数据库，包括代谢、信号调控等通路数据	https://reactome.org
PathBank[29]	以常见模式物种为基础的代谢、调控通路数据库	http://www.pathbank.org
HMDB[30]	人体小分子代谢数据库，包含反应、MS、NMR谱图等信息	https://hmdb.ca
MetaNetX[31]	整合了多个来源的生化反应数据库用于代谢网络模型构建	https://www.metanetx.org
Reaxys[32]	从专利和文献搜集和整理的大量有机反应和酶反应路线(商业非开源)	https://www.reaxys.com

谢物空间，同时还能用于合成路线的预测；根据给定反应预测所需的酶，或者反过来对未知反应功能的酶进行酶功能分类或反应活性强度预测；根据反应和酶的信息对催化反应重要性质（如反应动力学参数）进行预测等。因此，接下来论文将以酶、底物和产物为酶反应的三个抓手，从酶反应的数据表征、酶反应路线的正逆向预测、未知酶功能的预测与设计、已知功能的酶反应性质预测等方面来分别介绍。

2.1 数据的表征

在构建计算模型之前，我们需要对数据（即小分子和蛋白质的结构与性质）进行表征，使其转化成计算机能够理解的语言。无论是小分子还是蛋白质，都有不同维度的表征方式，如对于小分子来说，有基于二维结构的SMILES表达式、分子图（graph）和分子指纹等，还有基于三维结构的像素表征等[34-35]，此外也能通过分子的一维理化性质如分子量、疏水性、电荷等进行表征[36]。对于蛋白结构来说，最常用的是一维的氨基酸序列表征，以氨基酸序列为基础的多序列比对（MSA）结果同样也可以作为表征。近年来多种蛋白质结构预测模型都表明MSA中序列共进化信息对于模型的预测精度有显著提升[37]。除此以外还能用二维的位置权重矩阵（PSSM）、接触图（contact map）、三维的像素点等对蛋白进行表征[17]。而对于化学反应，在深度学习模型发展起来之前，研究人员主要通过经验和知识对反应规则进行总结，并主要通过SMIRKS表达式（SMILES的一种拓展）来表示，其中包含了特定的反应位点信息和化学键的形成和断裂模式，一些常用的化学信息学工具如RDKit[38]等可以直接读取SMIRKS并将其应用于给定底物，从而判断其是否符合该反应规则并生成特定的产物。对于酶来说，其功能可直接由其催化的反应来表征，但除此以外，酶的分类学标签和基因本体论（gene ontology，GO）[39]

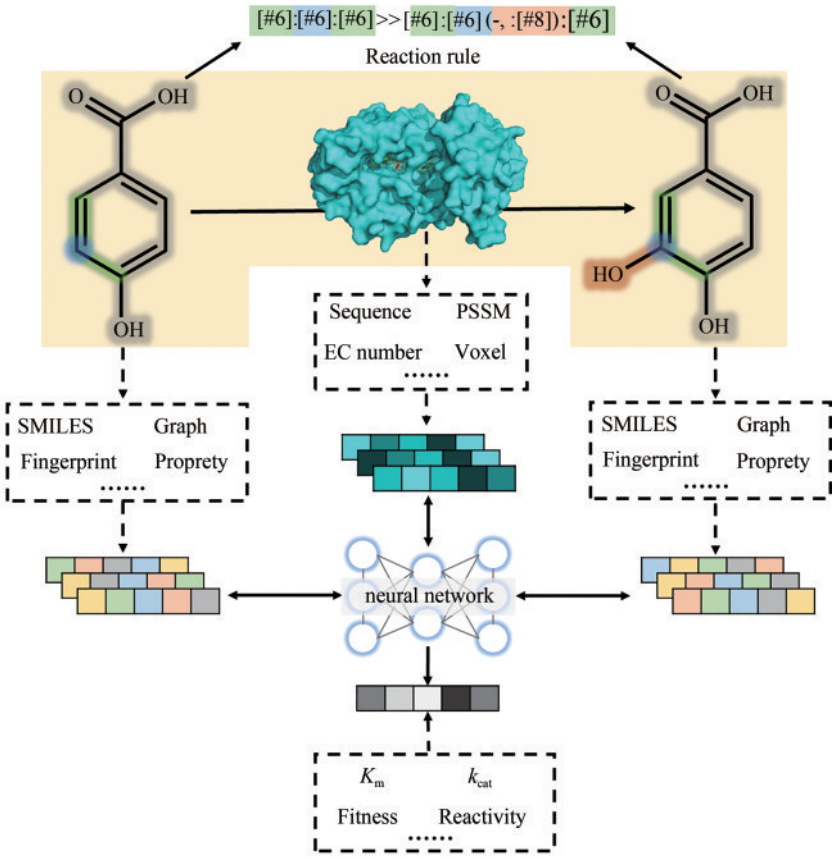


图1 酶反应的三个核心要素（底物、酶和产物）及其信息表征方式

Fig. 1 Key elements (substrate, enzyme and product) of enzymatic reactions and their information representations

注释也常用于描述酶的功能。酶的分类学标签通常指酶学委员会（Enzyme Commission）为酶所制作的一套编号分类法，该分类以化学反应的类型为基础。每个酶的 EC number 都由字母“EC”和四个数字组成，其中四个数字用点分隔，第一个数字使用数字 1 到 7 分别代表目前划分的七大类酶（氧化还原酶、转移酶、水解酶、裂解酶、异构酶、连接酶和转位酶）。后面三位数字将酶的分类逐级细分，由于不同大类下的子类数目不一，因此后三位数字的取值范围并不固定。而 GO 注释则是现代生物学从三个方面（分子功能、细胞组分、生物过程）对基因（及其表达的蛋白或 RNA）所进行的描述。和 EC number 类似，每个方面之中又有各种细分的描述，一般称为 GO term，如“GO：0005737”是细胞组分中的细胞质，表示某基因的产物是细胞质的组成成分。在机器学习模型中，数字表征（如分子量、电荷等）可以直接作为输入，而分子图、接触图等可转换为邻接矩阵进行输入，对于文本表征（如 SMILES、氨基酸序列等）则有多种输入方式，如独热编码（one-hot 编码）、词嵌入（word embedding）等。上述表征方式所提取出的特征各有侧重，因此在实际应用中通常需要根据任务的性质采用不同的表征方式进行模型训练。

2.2　基于反应物（以及酶）的产物预测

目前在自然界中仍然存在着大量未知的代谢过程，被称为“代谢暗物质”，阐明这些未知的代谢物和代谢反应能为新药发现和构建细胞工厂提供丰富的资源[40]。因此有许多工作聚焦于拓展现有分子的反应空间，即基于已知分子预测其潜在的各种代谢产物［图 2（a）］。以 Hatzimanikatis 课题组[40]的工作为例，他们将前期总结的约 500 条反应规则[41]应用于 150 万个生物来源小分子及活性小分子，构建了 ATLASx 数据库。该数据库中一共包含了 520 万条和现有的 8000 万小分子有关的反应，且其中有 148 万小分子此前并没有包含在任意反应中，即为“孤儿”分子。ATLASx 数据库极大

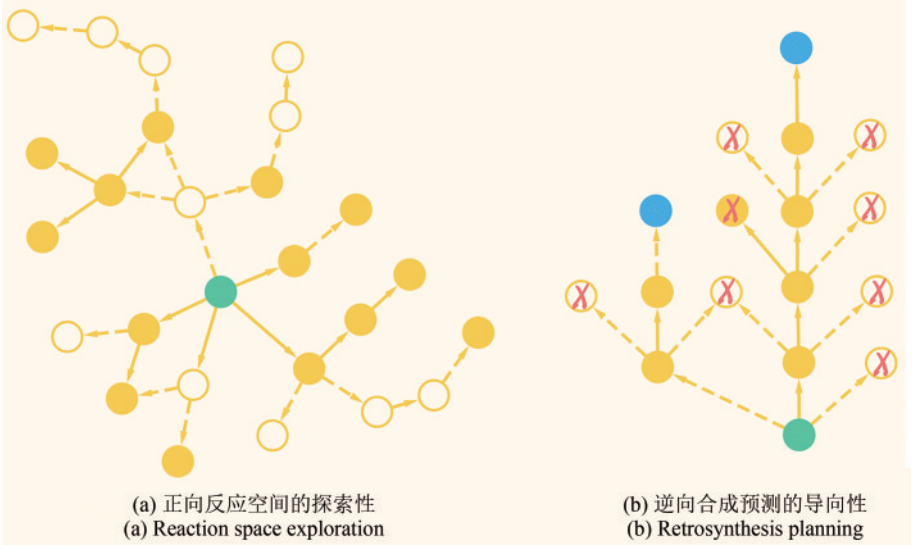


图 2　正向和逆向反应预测

［正向和逆向反应预测都是从一个分子（绿色）出发预测其潜在底物或产物（黄色），箭头表示两者之间能够通过反应进行转化，在（a）中箭头从反应物指向产物，（b）中则相反。经过多次迭代能够获得一个反应网络，网络中既能采样到已知的分子（实心）又能获得全新的结构（空心）。但不同的是正向反应预测每一次迭代方向都是随机的，而逆合成预测通常有一个终点条件（蓝色，如特定的原料分子），且会采取算法使得迭代过程朝着终点的方向进行］

Fig. 2　Prediction of forward and backward enzymatic reactions

[Prediction starts with an enzyme molecule (green node) to deduce its substrate or product (yellow nodes), the lines represent transformation reactions between two molecules, with arrow from substrate (enzyme) to product (a) and the reverse (b). A reaction network is developed after the iterative prediction in which both known (solid nodes) and unknown (hollow nodes) molecules are included. The forward prediction is generally random while a target (blue node, such as a building block) is specified in the backward prediction, and the exploration will lead to the target with the help of iterative algorithms.]

地丰富了代谢反应空间，同时也为许多未知合成途径的化学分子指明了潜在的生物合成途径。作者利用该方法对上市药物诺斯卡品（noscapine）的生物合成途径进行了拓展，发现了另一天然来源的上市药物分子延胡索乙素（tetrahydropalmatine）的潜在的生物合成途径并在酵母细胞中构建该途径并验证了其正确性[42]。Hu课题组[43]从文献中搜集了28万条反应数据并提取出其反应中心及其相邻原子的变化作为反应规则，并利用反应指纹（即底物的分子指纹减去产物的分子指纹）对上述反应规则进行去重，基于此反应数据库开发了BCSExplorer工具用于探索给定分子的合成或代谢空间。另外还有基于传统分子相似性的方法从数据库中查找已有的反应对目标分子反应空间进行探索（表2）。

除了上述传统的方法，Reymond课题组[45]利用SMILES和文本编码分别表征小分子和酶，将其用于深度学习模型Transformer[66]的输入，从而对产物进行预测。为了克服酶反应数据量不足的问题，作者采取了迁移学习的策略，先利用大量有机合成反应对模型进行预训练，再利用酶反应继续训练。研究结果表明有机反应的预训练确实对模型最终的预测能力有提升，并且和只使用反应物信息相比，酶信息的加入也有助于模型做出更加可信的预测。利用该模型不仅能对酶催化的产物做出预测，还能对酶的底物谱进行筛选，进一步阐明酶的催化功能。Kavraki及其合作者[46]则是利用深度学习构建了一个专门预测药物在人体内潜在代谢产物的预测模型，该模型同样是以底物SMILES作为输入，但不同的是该模型并没有包含酶的信息，因为对于药物代谢来说，所有可能的产物比特定酶催化得到的产物更加有指导意义，且作者测试发现在酶的信息数量有限的情况下，包含酶的信息对于模型的提升并不大。

2.3 基于产物的逆合成预测

和上述正向预测相比，逆合成预测[67-68]具有

表2 酶反应预测与设计工具汇总

Table 2 Tools for the prediction and design of enzymatic reactions

反应预测与酶设计工具			
	基于相似性	基于反应规则	基于机器学习
正向反应预测	BioSynther[44]（http://www.rxnfinder.org/）	ATLASx[40]（https://lcsb-databases.epfl.ch/Atlas2） BCSExplorer[43]（http://www.rxnfinder.org/）	Reymond等[45]（https://github.com/reymondgroup/OpenNMT-py） Kavraki等[46]（https:// github.com/KavrakiLab/MetaTrans）
逆合成预测	PrecursorFinder[47]（http://www.rxnfinder.org/）	RetroPath[48]（https://github.com/brsynth/RetroPathRL） RetroBioCat[49]（https://retrobiocat.com）	BioNavi-NP[50]（http://biopathnavi.qmclab.com/） Probst等[51]（https://github.com/rxn4chemistry/biocatalysis-model）
酶搜索和设计	EC-BLAST[52]（https://www.ebi.ac.uk/）	Selenzyme[53]（http://selenzyme.synbiochem.co.uk/） BridgIT[54]（https://lcsb-databases.epfl.ch/Atlas2） E-zyme2[55]（https://www.genome.jp/tools/e-zyme2/）	Faulon等[56]（tool not available） Ranganathan等[57]（https://github.com/ranganathanlab/bmDCA）
酶功能与性质预测工具			
酶功能预测 功能分类	DeepEC[58]（https://bitbucket.org/kaistsystemsbiology/deepec） Araki等[59]（https://github.com/nwatanbe/SVM_E_model） MTDNN[60]（http://bioinf.cs.ucl.ac.uk/downloads/mtdnn）		
功能优化	ECNet[61]（https://github.com/luoyunan/ECNet ） Gitter等[62]（https://github.com/gitter-lab/nn4dms）		
酶反应性质预测	Lercher等[63]（https://github.com/AlexanderKroll/KM_prediction） Palsson等[64]（tool not available） DLKcat[65]（https://github.com/SysBioChalmers/DLKcat）		

更强的目的性，它是对特定化合物的合成前体进行预测并将该过程循环迭代直到到达终止条件（如路线找到了一些常见的合成前体或容易获得的化学原料等）。由于逆合成预测的任务通常是找到目标分子和特定合成前体之间的合成路线，因此在每一步预测时并不会像正向预测那样任意拓展，而是需要进行评估和筛选以节约计算资源［图2(b)］。尽管如此，为了避免错过“正确”的合成前体，每一步逆合成预测依然会输出不止一个可能的结果，最终的路线组合数量会随着迭代步数增加呈指数级增长，因此在逆合成预测过程中仍然需要配合高效的搜索算法对路线分支进行“修剪”。Faulon团队[69]从MeteNetX反应数据库中自动提取出了超过上万条反应规则，并将其应用于生物逆合成路线的预测。由于许多分子可以同时应用多条反应规则，因此每一步逆合成预测都会产生大量的候选前体分子，为了从巨大的组合空间中高效搜索潜在的合成路线，作者首先采用了结构相似性和可用的蛋白序列数量对每一步结果进行打分，并结合蒙特卡洛树搜索[70]的策略优先选择更加可靠的前体分子进入后续的迭代预测[48]。为了使每一步的预测更加可靠，Turner及其合作者[49]则开发了RetroBioCat工具，通过人工总结常用的生物催化反应并编码反应规则，将其应用于生物催化级联合成路线的预测和设计，该工具很好地重现了文献报道的五十余条生物催化合成路线。

此外，深度学习模型凭借其无需构建反应规则就能捕捉反应信息的优势，也逐渐被应用于逆合成的预测[71]。Wu及其合作者[50]搜集了天然产物合成相关的3万余条反应并利用SMILES进行编码，用于天然产物生物逆合成模型BioNavi-NP的训练。作者还从有机反应数据中提取出了6万余条和天然产物结构类似的反应用于数据集的扩充并进行迁移学习，测试结果发现基于Transformer结构的模型表现要好于普通的神经网络和基于反应规则的模型。同时作者采用了基于与或树的Retro*搜索算法[72]用于多步反应路径的搜索，经过测试表明，该方法速度和精度远好于蒙特卡洛树搜索。该工具不仅可用于天然产物生物合成路线的预测，还能对已有的生物合成路线进行重构，有助于寻找更加高效的异源合成途径。Probst团队[51]也采取深度学习的方法，利用现有的酶催化反应分别将底物和产物作为模型输入构建了正向和逆向合成两个模型，在正向预测模型中酶的EC number也同时作为输入，而在逆合成预测中EC number则是作为输出，因此在逆合成模型中不仅能输出目标化合物的前体，还能对所需酶的类别进行预测。

上述模型能分别对代谢物和前体空间进行探索，但由于训练数据侧重的差异，不同模型有各自的应用范围，如药物代谢模型[46]仅能用于特定细胞色素P450酶的产物预测，RetroBioCat[49]和BioNavi-NP[50]则由于其训练数据的各自选择偏好与不足限制了其特定的一些适用范围。此外，和基于反应规则的方法相比，基于深度学习的方法虽然无需构建反应规则，但反应规则中的酶信息也同时被忽略了。尽管Probst团队[51]在模型中加入了EC number信息，但最多只能给出其前三位分类，对酶的预测仍然有限。

2.4 基于特定功能的酶搜索和设计

尽管正向和逆向反应模型在针对特定体系的预测中都有不错的精度，但在许多模型中酶的信息并没有被充分考虑，限制了其应用范围。在实际应用尤其是逆合成预测中获得潜在的合成途径之后，催化相应反应的酶对异源表达至关重要。尽管目前已有众多方法和平台可以用于酶蛋白元件的挖掘[73]，但获取催化特定反应的酶仍然是一个艰巨的任务。目前解决此类任务的思路主要是相似的反应往往可以通过同一个酶来催化完成，因此很多工具都是从已有数据库中搜索和目标反应相似的反应［图3(a)］，以催化该相似反应的酶作为候选序列进行后续的实验和改造。如EC-BLAST[52]可以通过键的变化、反应中心以及结构的相似性从KEGG数据库中寻找相似反应，并给出相应酶的EC number。Selenzyme[53]是一个基于结构相似性的酶搜索工具，用户输入反应规则或一条具体的反应后，Selenzyme会在MetaNetX反应数据库中搜索相似的反应，并以催化该相似反应的酶作为结果输出。用户还可以对所需酶的物种做出限制，该工具将会根据候选酶所在的物种

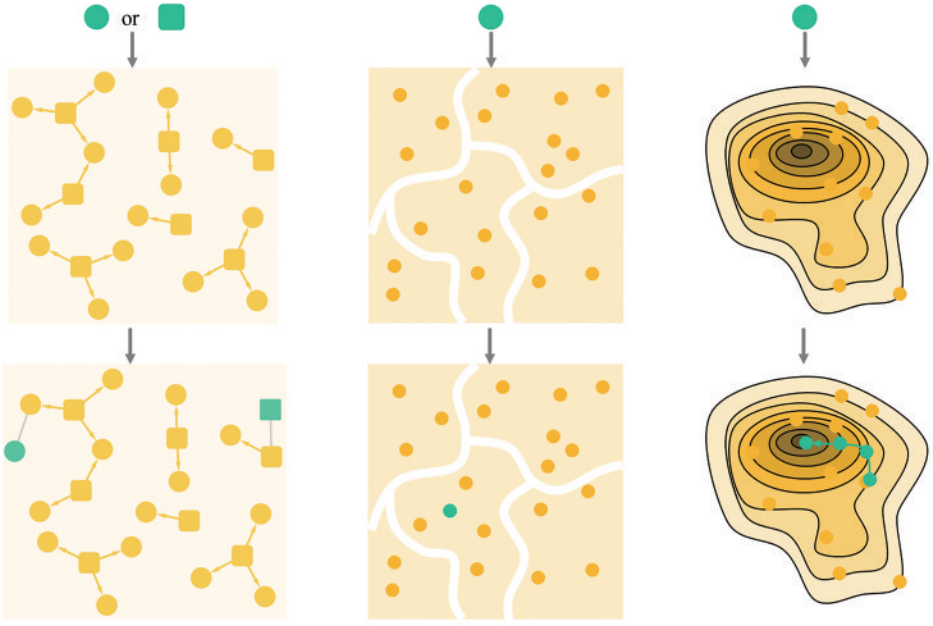


图3 不同类型的酶搜索和功能预测模型。

[圆形代表酶，方形代表反应，黄色节点代表已知数据，绿色代表待预测对象。基于相似性的预测方法（a）是从已知的酶-反应数据对中（图中相连的两个节点）寻找与目标对象相似的样本，从而对其反应（或酶）进行预测。功能分类模型（b）是将已知功能（通常是离散变量）的酶序列作为训练集，寻找其潜在分类规律（白色分界线），从而对目标序列进行预测。回归模型（c）则是对活性或稳定性强弱等连续变量进行建模预测，绘制适应度景观图从而对目标序列的功能进行预测，并用于酶设计]

Fig. 3 Models for searching and predicting enzymes

[Circular and square nodes represent sequences and reactions, respectively, and yellow filling indicates known data while green filling mean objects to be predicted. Similarity search (a) is to find a similar object in known enzyme-reaction pairs (connected nodes) to predict reactions (or enzymes) for target object. Classification model (b) is trained by enzymes with known function (usually discrete), in which the classification rule (white boundary) is clarified, and then the model can be used to classify an enzyme with unknown function. Regression model (c) is adapted to draw fitness landscape to predict continues variables such as the activity or stability of enzymes, which can then be used for enzyme design.]

与用户指定物种之间的系统发育距离对结果重新排序。其他工具如BridgIT[54]、E-zyme2[55]则分别通过反应指纹和反应模式对KEGG数据库进行搜索，从而对目标反应进行酶的预测。

除上述基于反应相似性的方法以外，近年来基于机器学习的工具也展现出潜在的应用价值。当前由于非冗余酶反应数据的不足，单纯依靠已知反应来寻找所需的酶难度极大。最近Faulon团队[56]采用了同时给定反应和酶，进而判断该酶是否能催化该反应的思路。作者利用反应中心原子环境的变化和氨基酸序列对反应和酶分别进行表征，将两者共同输入高斯过程分类模型进行训练。由于现有数据库中只有某种酶能够催化某些反应的数据（即正样本），而缺少负样本，因此作者有选择地从不同酶-反应数据对中分别挑选酶和反应组成新的样本，并作为无标签样本进行半监督学习。作者利用该模型挑选出了大肠杆菌中和*N*-乙酰-L-亮氨酸合成相关的酶并经实验成功验证。

虽然酶是高效、精准的催化剂，但大部分天然的酶却很难直接应用于工业生产[15]，传统的突变扫描实验尽管可以获得潜在的功能序列，但通常需要多轮次的实验筛选，不仅耗时长且成功率低。而数据驱动的酶计算改造和重设计可以绕过实验的突变筛选，对潜在的序列和活性空间进行高效搜索。如Ranganathan及其合作者[57]将分支酸变位酶（chorismate mutase，参与芳香族氨基酸生物合成的关键酶）的MSA信息用于玻尔兹曼机器（Boltzmann machine）学习的直接耦合分析（direct coupling analysis，DCA）[74]，得到的模型可捕捉序

列的保守位点和位点之间的相关性，因此可用于突变体的采样。近年来，蛋白质的从头设计（*de novo* design）[75]在深度学习的辅助下也取得了一些进展，诸如Baker课题组[76]和Liu课题组[77]都分别基于深度学习框架提出了逆向折叠（即利用蛋白结构生成序列）算法，有望针对催化特定反应的结构或活性位点设计出全新的蛋白序列[78]。

2.5 酶功能预测

随着测序技术的进步，数据库中有大量的序列已知但功能未知的待定蛋白元件，因此亟需可靠准确的方法对这些蛋白序列进行功能预测。由于只有序列信息已知，因此传统的方法便是通过序列相似性从数据库中搜索已知功能的蛋白（如BLAST[79]），从而对目标蛋白的功能做出预测[图3(a)]。而机器学习方法由于能够从更多特征中学习隐藏规律，更有可能取得精确的预测结果。目前针对酶功能最常用的标签是EC number，有许多机器学习方法通过序列进行EC number的分类预测［图3(b)］。如Lee课题组[58]开发了深度学习模型DeepEC，利用卷积神经网络对氨基酸序列进行编码并预测其EC number。经测试DeepEC比其他5种预测工具具有更高的准确率和更快的运行速度。Araki团队[59]则通过PROFEAT工具[80]从酶的氨基酸序列中提取性质特征（如氨基酸组成等），从而对其进行编码，然后输入机器学习分类模型中预测其EC number。测试结果显示支持向量机（SVM）模型的表现要优于随机森林（RF）、*k*-邻近算法（kNN）和多层感知机（MLP）模型。作者利用该模型对罂粟中苄基异喹啉生物碱合成路线中缺失的酶进行预测，成功解析了该物种中酪氨酸到下游生物碱的分支合成途径[81]。除了EC number以外，GO条目也常被用于蛋白功能预测模型中。Jones课题组[60]构建了一个多任务深度神经网络（MTDNN）用于预测给定序列的GO注释，其中多任务表示除了整个模型共有的隐藏层以外，对于每一个GO条目都有独立隐藏层网络来负责输出最后的预测结果。MTDNN采用了多种功能和结构描述符来编码蛋白序列如二级结构、跨膜组分等。MTDNN的预测精度不仅高于传统的BLAST方法，也比单纯的多标签分类神经网络要好。

和上述普适的功能分类预测不同的是，在蛋白质工程中，往往需要针对特定功能的酶进行活性强度的预测，即回归模型［图3(c)］，并利用该模型对突变体进行预测从而在多轮的突变实验中以筛选出最优序列。ECNet[61]是一个基于序列共进化信息进行活性预测的神经网络模型，蛋白序列表征由基于大数据库（如Uniprot[33]或Pfam[82]）预训练得到的全局特征和基于目标序列MSA得到的局部特征组合而成。将已有的深度突变扫描实验数据对模型进行训练，可以对未知的突变体活性进行准确预测，作者以β-内酰胺酶为例成功从模拟突变中筛选出活性强于野生型的新颖突变体。Gitter团队[62]则利用接触图作为蛋白表征，同时采用图卷积神经网络对蛋白活性进行预测。作者将随机重启爬山算法和该模型结合从链球菌蛋白G的GB1结构域（能与免疫球蛋白G结合，可用于抗体纯化）序列空间搜索得到了5个高亲和力的突变体，经实验测试，其中一个突变体确实表现出比野生型更强的结合亲和力。该类模型能够代替传统实验中的构建与测试环节，减少的DBTL循环迭代次数，从而针对特定的酶快速绘制出清晰的序列-功能图谱，将其应用于定向进化等蛋白质工程任务中可以极大地节约筛选时间和实验成本[83-84]。

2.6 酶反应性质预测

酶反应动力学研究对于理解酶反应机制，以及设计合理的反应条件具有重要作用[85-86]。其中K_m和k_{cat}是衡量酶催化效率最重要的两个参数，而实验是获得这些参数的主要手段。近年来，有不少工作利用深度学习模型来对酶反应动力学参数进行预测。如Lercher及其合作者[63]利用分子指纹、脂水分配系数以及分子量表征反应底物，用一个经过大量已知蛋白序列预训练过的特征提取模型UniRep[87]对酶进行表征，并使用梯度提升构建回归模型用于预测反应的K_m值。在独立的测试集中该模型也表现出良好的性能。此外，也有模型专门预测同一物种中同一个底物与不同酶作用

时的K_m值[88]，以及纤维二糖在不同β-葡萄糖苷酶催化下的K_m值[89]。Palsson及其合作者[64]利用代谢网络、蛋白结构、物质浓度等多种信息输入机器学习模型（包括随机森林、深度神经网络等）进行大肠杆菌代谢网络中酶反应k_{cat}的预测。尽管通过该模型对k_{cat}进行预测能够提高代谢网络模型的精度，从而对细胞生长状态进行更准确的预测，但其输入数据处理太复杂，很难应用于其他物种。因此Nielsen课题组[65]从BRENDA和SABIO-RK数据库中搜集了所有带有k_{cat}注释的酶反应，并用底物的分子图进行编码，以及用氨基酸序列对酶进行编码，两者分别用于图神经网络和循环神经网络的输入，并在拼接之后用于k_{cat}的预测。该模型除了在测试集中表现优越之外，对于酶序列具有细微变化的突变体的催化能力预测也有不错的准确率，且神经网络中注意力机制的加入能够让模型检测到和酶催化效率相关的关键残基。为了数据共享并服务于更多研究者，作者基于该预测模型的结果构建了一个在线的酶反应参数数据库，在目前的版本中可对计算机预测的反应k_{cat}值进行查询[90]。无论是K_m还是k_{cat}或是其他酶反应动力学参数，通过实验上的测量都有一定的难度，且受实验条件影响较大，因此计算模型的出现提供了巨大的便利。并且，这些参数对于目前各种类型基因组尺度的代谢网络模型[19]以及全细胞模型[91-92]至关重要，在未来，有望利用这些模型对物种或细胞整个生命周期进行模拟，不仅能让我们更好地理解生物体内的生长状态和生理过程，更能帮助我们设计和优化人工底盘细胞用于生物合成。

3 展 望

尽管随着各种实验技术的进步和计算模拟技术的引入，大自然酶催化这层神秘面纱正一点一点被揭开，但是在此面纱之下还有巨大的未知空间有待挖掘。以酶为核心的生物制造技术是极具潜力的发展方向。挖掘更多的酶催化元件、优化酶催化的功能与效率，提升工业酶的性能，是生物经济产业发展的重大需求。尽管蛋白结构的理论预测获得了重大突破，但酶催化机制的高效解析、酶功能的理性设计等仍然是领域内的难点。从蛋白的静态结构到酶反应的动态调控机理，从海量而冗余的酶反应数据到大数据驱动的酶分子设计，这中间还有不少鸿沟需要新理论、新技术的突破。本文旨在从大数据驱动的酶反应预测与设计的视角，汇总当前常用的酶反应数据库，并对近年来基于大数据和机器学习开发酶反应预测与设计工具进行了总结。上述一些成功的案例已经预示着未来，计算模型与算法的发展将是走向生物智造不可或缺的核心推力之一。然而我们也必须清醒意识到，现有计算工具在精度上还有很大的提升空间，要最大化地发挥计算的赋能作用还需要不断地探索（图4）。

首先，尽管当前已经有了诸多类型的数据库，但这些数据来源于世界各地的实验室，其实验条件、实验试剂等都会对实验数据造成影响，例如大多数酶的反应动力学参数对于实验温度、pH等都是敏感的[93]。幸运的是现有数据库中已经开始对这些实验条件进行了记录，而如何将这些条件纳入计算模型的构建中将是我们下一步要思考的问题[94]。此外，不管数据库的标注和更新是由人工还是计算机完成，都不可避免会有错误记录的产生，有文献就指出BRENDA数据库中有相当一部分酶的EC number注释是错误的[95]。因此尽可能避免错误标注，提高数据的质量是提升计算模型精度很重要的手段。此外，对于很多机器学习模型而言，训练集中负样本对于模型优化是极为重要的，但当前数据库收录的都是文献中报道的“阳性”结果（即能反应或有活性），而经实验验证的阴性结果同样包含重要信息，因此方便用户上传实验结果（无论是阴性还是阳性）的数据共享平台的建设也值得我们关注，当然这也需要领域内实验人员共同的努力。另外，智能化机器人也能在很大程度上解放实验人员，同时可以完成实验、记录、保存等一系列工作，加快正、负数据的积累从而加速实验与计算交互反馈的进程。

其次，无论是针对小分子还是蛋白的机器学习特征提取，我们都无法用数字解释其包含的所有信息，尽管我们也并不一定需要这样做，但是如何尽可能提取出我们所需要的输入信息并进行编码，也有待于进一步研究。近几年来自然语言

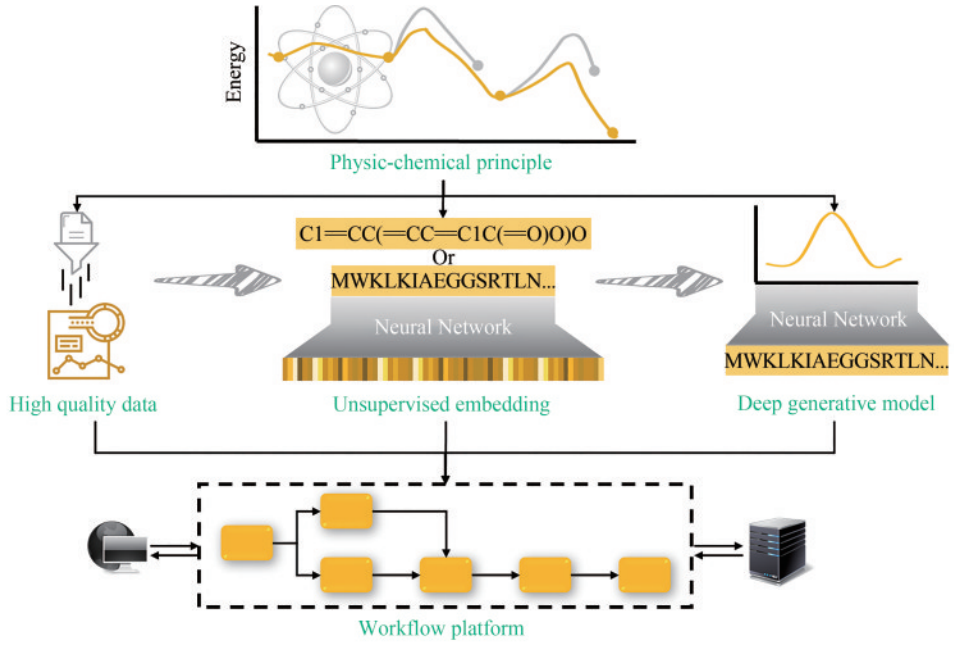


图 4　数据驱动的酶反应设计与预测在未来的展望

Fig. 4　Perspective of data-driven prediction and design of enzymatic reactions

处理模型的进步使我们能够利用非监督学习的方式从大量语料库（比如分子库或蛋白序列库）中训练出用于输出小分子和蛋白的嵌入表征的预训练模型[87, 96-98]。经分析这些模型能够很好地捕捉不同样本的潜在特征并用于输入，下游针对各种任务的机器学习模型性能有望从中得到提升[99]。

最后，在计算模型与算法层面，同样也是得益于自然语言处理模型的发展，越来越多生成式的端到端模型应用于反应预测和设计领域[71, 100]。和普通的回归或分类模型以标签为输出不同的是，端到端模型可以直接通过训练过的概率分布生成序列（在本文中即小分子或蛋白）。如前文提到的酶催化反应的正向和逆向预测的模型中，有不少就已经应用了该种端到端模型[45, 50]。而随着各种蛋白序列生成模型的提出[101]，将其应用于酶的预测将会大大扩大搜索空间，因为目前基于已知反应的酶预测还局限于从蛋白序列库中寻找潜在的酶或其突变体，生成模型则能在特定限制条件下产生具有潜在功能的新颖序列，此外还能生成更多的非天然序列用于蛋白优化与设计[102-105]。

以上是围绕数据驱动的策略在将来需要重点关注的三个方面，而对于酶反应来说，其背后的物理化学原理对于酶反应设计是至关重要的，对于酶反应物理化学规律的认知信息在上述数据驱动策略中往往被忽略或者未考虑周全。在过去的十几年中，基于物理模型的方法在酶反应预测和设计方面已经取得了许多成果。如基于量子力学和分子力学组合方法（QM/MM）的动力学模拟不仅用于探索分子的反应性，扩展反应空间[106-107]，还用于解析酶反应的机理，并为酶改造和重设计提供了极为重要的酶反应热动力学属性和酶催化调控的理论依据[8, 108]。以 Rosetta Design[109-110] 为代表的基于物理模型的计算方法则为酶设计提供了新范式，并成功应用于许多案例[111-113]。而数据驱动的方法尤其是深度学习模型近几年虽快速发展，但作为一种被称为“黑箱”模型的工具，目前仍无法参与到如微观机制解析这种复杂且动态的任务中。我们既需要依靠物理模型解决最底层的问题并积累更多的数据，用于数据模型的构建和训练，反过来数据模型因其高效预测能力使其能够参与到物理模型的框架中，实现优势互补。因此，物理模型与数据模型的结合将是酶反应预测和设计的新趋势，如结合动力学模拟和深度学习的反应空间探索[114]、借助深度学习势能的分子动力学模拟方法[115]、基于神经网络能量函数的氨基酸序列设计[77]等等。

最后，为上述模型与算法搭建高度集成的工程化平台也是非常有意义的。目前只有少数计算

模型和工具发布了在线服务器版本，其他大部分都是发布于各托管平台的源代码，并且目前很多深度学习模型体量庞大，需要一定的硬件支持才能运行，使用这些工具对于普通的实验人员来说有一定的壁垒。同时本文提到的反应预测、酶预测、酶反应性质预测在实际的实验中通常是链条式的流程，因此将这些工具整合在一个便捷友好的平台中将会给实验人员带来极大的便利。Hu课题组建立的RxnFinder（http：//www.rxnfinder.org/）商业化网站平台[116]开发并整合了反应探索、前体预测、逆合成分析、菌株设计等一系列计算工具，能为实验研究人员提供便利。Hatzimanikatis课题组和曼彻斯特大学精细化学品合成生物学研究中心分别搭建的LCSB（https://lcsb-databases.epfl.ch/Home）和SYNBIOCHEM（https://synbiochem.co.uk/）数据平台也包含了各自课题组开发的逆合成预测、酶选择等工具供学术界免费试用。但这些平台只是将工具汇总在一起，用户需要单独访问或下载使用某一模块，并且针对相似任务的不同算法部署在不同的平台，不利于用户进行直接比较研究。未来在以酶催化为基础的生物制造工业化应用中，我们可能更需要全链条式的设计平台，即所有工具以工作流的形式集成在平台中，用户可一键访问并自由组合使用。

参 考 文 献

[1] BENKOVIC S J, HAMMES-SCHIFFER S. A perspective on enzyme catalysis[J]. Science, 2003, 301(5637): 1196-1202.

[2] BORNSCHEUER U T, BUCHHOLZ K. Highlights in biocatalysis-historical landmarks and current trends[J]. Engineering in Life Sciences, 2005, 5(4): 309-323.

[3] SHELDON R A, WOODLEY J M. Role of biocatalysis in sustainable chemistry[J]. Chemical Reviews, 2018, 118(2): 801-838.

[4] ZIMMERMAN J B, ANASTAS P T, ERYTHROPEL H C, et al. Designing for a green chemistry future[J]. Science, 2020, 367(6476): 397-400.

[5] WINKLER C K, SCHRITTWIESER J H, KROUTIL W. Power of biocatalysis for organic synthesis[J]. ACS Central Science, 2021, 7(1): 55-71.

[6] LIN G M, WARDEN-ROTHMAN R, VOIGT C A. Retrosynthetic design of metabolic pathways to chemicals not found in nature[J]. Current Opinion in Systems Biology, 2019, 14: 82-107.

[7] SENN H M, THIEL W. QM/MM methods for biomolecular systems[J]. Angewandte Chemie International Edition, 2009, 48(7): 1198-1229.

[8] ZHA W L, ZHANG F, SHAO J Q, et al. Rationally engineering santalene synthase to readjust the component ratio of sandalwood oil[J]. Nature Communications, 2022, 13: 2508.

[9] WANG Y, MALACO MOROTTI A L, XIAO Y R, et al. Decoding the cytochrome P450 catalytic activity in divergence of benzophenone and xanthone biosynthetic pathways[J]. ACS Catalysis, 2022, 12(21): 13630-13637.

[10] LIANG M M, ZHANG F, XU J X, et al. A conserved mechanism affecting hydride shifting and deprotonation in the synthesis of hopane triterpenes as compositions of wax in oat[J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119(12): e2118709119.

[11] VAN DIJK E L, AUGER H, JASZCZYSZYN Y, et al. Ten years of next-generation sequencing technology[J]. Trends in Genetics, 2014, 30(9): 418-426.

[12] WANG H Y, GUO H, WANG N, et al. Toward the heterologous biosynthesis of plant natural products: gene discovery and characterization[J]. ACS Synthetic Biology, 2021, 10(11): 2784-2795.

[13] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873): 583-589.

[14] PEARCE R, ZHANG Y. Deep learning techniques have significantly impacted protein structure prediction and protein design[J]. Current Opinion in Structural Biology, 2021, 68: 194-207.

[15] CUI Y L, SUN J Y, WU B. Computational enzyme redesign: large jumps in function[J]. Trends in Chemistry, 2022, 4(5): 409-419.

[16] SHELDON R A, PEREIRA P C. Biocatalysis engineering: the big picture[J]. Chemical Society Reviews, 2017, 46(10): 2678-2691.

[17] JANG W D, KIM G B, KIM Y J, et al. Applications of artificial intelligence to enzyme and pathway design for metabolic engineering[J]. Current Opinion in Biotechnology, 2022, 73: 101-107.

[18] HADADI N, HATZIMANIKATIS V Design of computational retrobiosynthesis tools for the design of *de novo* synthetic pathways[J]. Current Opinion in Chemical Biology, 2015, 28: 99-104.

[19] 周静茹, 刘鹏, 夏建业, 等. 基于约束的基因组规模代谢网络模型构建方法研究进展[J]. 生物工程学报, 2021, 37(5): 1526-1540.
ZHOU J R, LIU P, XIA J Y, et al. Advances in the development of constraint-based genome-scale metabolic network models[J]. Chinese Journal of Biotechnology, 2021, 37(5): 1526-1540.

[20] MACKLIN D N, RUGGERO N A, COVERT M W. The future

of whole-cell modeling[J]. Current Opinion in Biotechnology, 2014, 28: 111-115.

[21] MAZURENKO S, PROKOP Z, DAMBORSKY J. Machine learning in enzyme engineering[J]. ACS Catalysis, 2020, 10 (2): 1210-1223.

[22] LIAO X P, MA H W, TANG Y J. Artificial intelligence: a solution to involution of design-build-test-learn cycle[J]. Current Opinion in Biotechnology, 2022, 75: 102712.

[23] KANEHISA M, GOTO S. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Research, 2000, 28(1): 27-30.

[24] CASPI R, BILLINGTON R, KESELER I M, et al. The MetaCyc database of metabolic pathways and enzymes-a 2019 update[J]. Nucleic Acids Research, 2020, 48(D1): D445-D453.

[25] BANSAL P, MORGAT A, AXELSEN K B, et al. Rhea, the reaction knowledgebase in 2022[J]. Nucleic Acids Research, 2022, 50(D1): D693-D700.

[26] CHANG A, JESKE L, ULBRICH S, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates[J]. Nucleic Acids Research, 2021, 49(D1): D498-D508.

[27] WITTIG U, REY M, WEIDEMANN A, et al. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics[J]. Nucleic Acids Research, 2018, 46(D1): D656-D660.

[28] GILLESPIE M, JASSAL B, STEPHAN R, et al. The reactome pathway knowledgebase 2022 [J]. Nucleic Acids Research, 2022, 50(D1): D687-D692.

[29] WISHART D S, LI C, MARCU A, et al. PathBank: a comprehensive pathway database for model organisms[J]. Nucleic Acids Research, 2020, 48(D1): D470-D478.

[30] WISHART D S, GUO A C, OLER E, et al. HMDB 5.0: the human metabolome database for 2022[J]. Nucleic Acids Research, 2022, 50(D1): D622-D631.

[31] MORETTI S, TRAN V D T, MEHL F, et al. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models[J]. Nucleic Acids Research, 2021, 49(D1): D570-D574.

[32] LAWSON A J, SWIENTY-BUSCH J, GÉOUI T, et al. The making of reaxys—towards unobstructed access to relevant chemistry information[M]//ACS Symposium Series: The Future of the History of Chemical Information. Washington, DC: American Chemical Society, 2014: 127-148.

[33] CONSORTIUM T U. UniProt: a worldwide hub of protein knowledge[J]. Nucleic Acids Research, 2019, 47(D1): D506-D515.

[34] XU Y J, LIN K J, WANG S W, et al. Deep learning for molecular generation[J]. Future Medicinal Chemistry, 2019, 11(6): 567-597.

[35] ELTON D C, BOUKOUVALAS Z, FUGE M D, et al. Deep learning for molecular design—a review of the state of the art[J]. Molecular Systems Design & Engineering, 2019, 4(4): 828-849.

[36] HAGHIGHATLARI M, LI J, HEIDAR-ZADEH F, et al. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods[J]. Chem, 2020, 6(7): 1527-1542.

[37] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.

[38] LANDRUM G. RDKit: Open-source cheminformatics software [EB/OL][2022-12-01]. https://rdkit.org/.

[39] The Gene Ontology Consortium. The gene ontology resource: enriching a GOld mine[J]. Nucleic Acids Research, 2021, 49 (D1): D325-D334.

[40] MOHAMMADIPEYHANI H, HAFNER J, SVESHNIKOVA A, et al. Expanding biochemical knowledge and illuminating metabolic dark matter with ATLASx[J]. Nature Communications, 2022, 13: 1560.

[41] HATZIMANIKATIS V, LI C H, IONITA J A, et al. Exploring the diversity of complex metabolic networks[J]. Bioinformatics, 2005, 21(8): 1603-1609.

[42] HAFNER J, PAYNE J, MOHAMMADIPEYHANI H, et al. A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives[J]. Nature Communications, 2021, 12: 1760.

[43] TIAN Y, WU L, YUAN L, et al. BCSExplorer: a customized biosynthetic chemical space explorer with multifunctional objective function analysis[J]. Bioinformatics, 2020, 36(5): 1642-1643.

[44] TU W Z, ZHANG H R, LIU J, et al. BioSynther: a customized biosynthetic potential explorer[J]. Bioinformatics, 2016, 32(3): 472-473.

[45] KREUTTER D, SCHWALLER P, REYMOND J L. Predicting enzymatic reactions with a molecular transformer[J]. Chemical Science, 2021, 12(25): 8648-8659.

[46] LITSA E E, DAS P, KAVRAKI L E. Prediction of drug metabolites using neural machine translation[J]. Chemical Science, 2020, 11(47): 12777-12788.

[47] YUAN L, TIAN Y, DING S Z, et al. PrecursorFinder: a customized biosynthetic precursor explorer[J]. Bioinformatics, 2019, 35(9): 1603-1604.

[48] KOCH M, DUIGOU T, FAULON J L. Reinforcement learning for bioretrosynthesis[J]. ACS Synthetic Biology, 2020, 9(1): 157-168.

[49] FINNIGAN W, HEPWORTH L J, FLITSCH S L, et al. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades[J]. Nature Catalysis, 2021, 4(2): 98-104.

[50] ZHENG S J, ZENG T, LI C T, et al. Deep learning driven bio-

synthetic pathways navigation for natural products with BioNavi-NP[J]. Nature Communications, 2022, 13: 3342.

[51] PROBST D, MANICA M, NANA TEUKAM Y G, et al. Biocatalysed synthesis planning using data-driven learning[J]. Nature Communications, 2022, 13: 964.

[52] RAHMAN S A, CUESTA S M, FURNHAM N, et al. EC-BLAST: a tool to automatically search and compare enzyme reactions[J]. Nature Methods, 2014, 11(2): 171-174.

[53] CARBONELL P, WONG J, SWAINSTON N, et al. Selenzyme: enzyme selection tool for pathway design[J]. Bioinformatics, 2018, 34(12): 2153-2154.

[54] HADADI N, MOHAMMADIPEYHANI H, MISKOVIC L, et al. Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(15): 7298-7307.

[55] MORIYA Y, YAMADA T, OKUDA S, et al. Identification of enzyme genes using chemical structure alignments of substrate-product pairs[J]. Journal of Chemical Information and Modeling, 2016, 56(3): 510-516.

[56] MELLOR J, GRIGORAS I, CARBONELL P, et al. Semisupervised Gaussian process for automated enzyme search[J]. ACS Synthetic Biology, 2016, 5(6): 518-528.

[57] RUSS W P, FIGLIUZZI M, STOCKER C, et al. An evolution-based model for designing chorismate mutase enzymes[J]. Science, 2020, 369(6502): 440-445.

[58] RYU J Y, KIM H U, LEE S Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(28): 13996-14001.

[59] WATANABE N, MURATA M, OGAWA T, et al. Exploration and evaluation of machine learning-based models for predicting enzymatic reactions[J]. Journal of Chemical Information and Modeling, 2020, 60(3): 1833-1843.

[60] FA R, COZZETTO D, WAN C, et al. Predicting human protein function with multi-task deep neural networks[J]. PLoS One, 2018, 13(6): e0198216.

[61] LUO Y N, JIANG G D, YU T H, et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering[J]. Nature Communications, 2021, 12: 5743.

[62] GELMAN S, FAHLBERG S A, HEINZELMAN P, et al. Neural networks to learn protein sequence-function relationships from deep mutational scanning data[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(48): e2104878118.

[63] KROLL A, ENGQVIST M K M, HECKMANN D, et al. Deep learning allows genome-scale prediction of Michaelis constants from structural features[J]. PLoS Biology, 2021, 19(10): e3001402.

[64] HECKMANN D, LLOYD C J, MIH N, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models[J]. Nature Communications, 2018, 9: 5252.

[65] LI F R, YUAN L, LU H Z, et al. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction[J]. Nature Catalysis, 2022, 5(8): 662-672.

[66] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 6000-6010.

[67] COREY E J. General methods for the construction of complex molecules[J]. Pure and Applied Chemistry, 1967, 14(1): 19-38.

[68] COREY E J, WIPKE W T. Computer-assisted design of complex organic syntheses[J]. Science, 1969, 166(3902): 178-192.

[69] DELÉPINE B, DUIGOU T, CARBONELL P, et al. RetroPath2.0: a retrosynthesis workflow for metabolic engineers[J]. Metabolic Engineering, 2018, 45: 158-170.

[70] SEGLER M H S, PREUSS M, WALLER M P. Planning chemical syntheses with deep neural networks and symbolic AI[J]. Nature, 2018, 555(7698): 604-610.

[71] LIU B W, RAMSUNDAR B, KAWTHEKAR P, et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models[J]. ACS Central Science, 2017, 3(10): 1103-1113.

[72] CHEN B H, LI C T, DAI H J, et al. Retro*: learning retrosynthetic planning with neural guided a* search[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 1608-1616.

[73] 张建志, 付立豪, 唐婷, 等. 基于合成生物学策略的酶蛋白元件规模化挖掘[J]. 合成生物学, 2020, 1(3): 319-336.
ZHANG J Z, FU L H, TANG T, et al. Scalable mining of proteins for biocatalysis via synthetic biology[J]. Synthetic Biology Journal, 2020, 1(3): 319-336.

[74] FIGLIUZZI M, BARRAT-CHARLAIX P, WEIGT M. How pairwise coevolutionary models capture the collective residue variability in proteins? [J]. Molecular Biology and Evolution, 2018, 35(4): 1018-1027.

[75] HUANG P S, BOYKEN S E, BAKER D. The coming of age of *de novo* protein design[J]. Nature, 2016, 537(7620): 320-327.

[76] DAUPARAS J, ANISHCHENKO I, BENNETT N, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. Science, 2022, 378(6615): 49-56.

[77] LIU Y F, ZHANG L, WANG W L, et al. Rotamer-free protein sequence design based on deep learning and self-consistency[J]. Nature Computational Science, 2022, 2(7): 451-462.

[78] JIANG L, ALTHOFF E A, CLEMENTE F R, et al. *De novo* computational design of Retro-Oldol enzymes[J]. Science,

2008, 319(5868): 1387-1391.

[79] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215 (3): 403-410.

[80] LI Z R, LIN H H, HAN L Y, et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence[J]. Nucleic Acids Research, 2006, 34(suppl_2): W32-W37.

[81] VAVRICKA C J, TAKAHASHI S, WATANABE N, et al. Machine learning discovery of missing links that mediate alternative branches to plant alkaloids[J]. Nature Communications, 2022, 13(1): 1405.

[82] MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: the protein families database in 2021[J]. Nucleic Acids Research, 2021, 49(D1): D412-D419.

[83] YANG K K, WU Z, ARNOLD F H. Machine-learning-guided directed evolution for protein engineering[J]. Nature Methods, 2019, 16(8): 687-694.

[84] WITTMANN B J, JOHNSTON K E, WU Z, et al. Advances in machine learning for directed evolution[J]. Current Opinion in Structural Biology, 2021, 69: 11-18.

[85] KLUMPP S, SCOTT M, PEDERSEN S, et al. Molecular crowding limits translation and cell growth[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(42): 16754-16759.

[86] CHEN Y, NIELSEN J. Energy metabolism controls phenotypes by protein efficiency and allocation[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(35): 17592-17597.

[87] ALLEY E C, KHIMULYA G, BISWAS S, et al. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature Methods, 2019, 16(12): 1315-1322.

[88] BORGER S, LIEBERMEISTER W, KLIPP E. Prediction of enzyme kinetic parameters based on statistical learning[J]. Genome Informatics International Conference on Genome Informatics, 2006, 17(1): 80-87.

[89] YAN S M, SHI D Q, NONG H, et al. Predicting K_m values of beta-glucosidases using cellobiose as substrate[J]. Interdisciplinary Sciences: Computational Life Sciences, 2012, 4(1): 46-53.

[90] LI F R, CHEN Y, ANTON M, et al. GotEnzymes: an extensive database of enzyme parameter predictions[J]. Nucleic Acids Research, 2023, 51(D1): D583-D586.

[91] MACKLIN D N, AHN-HORST T A, CHOI H, et al. Simultaneous cross-evaluation of heterogeneous *E. coli* datasets *via* mechanistic simulation[J]. Science, 2020, 369(6502): eaav3751.

[92] THORNBURG Z R, BIANCHI D M, BRIER T A, et al. Fundamental behaviors emerge from simulations of a living minimal cell[J]. Cell, 2022, 185(2): 345-360.e28.

[93] ENGQVIST M K M. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures[J]. BMC Microbiology, 2018, 18(1): 177.

[94] LI G, HU Y T, ZRIMEC J, et al. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism[J]. Nature Communications, 2021, 12: 190.

[95] REMBEZA E, ENGQVIST M K M. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class[J]. PLoS Computational Biology, 2021, 17(9): e1009446.

[96] RAO R, BHATTACHARYA N, THOMAS N, et al. Evaluating protein transfer learning with TAPE[J]. Advances in Neural Information Processing Systems, 2019, 32: 9689-9701.

[97] RIVES A, MEIER J, SERCU T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[98] JAEGER S, FULLE S, TURK S. Mol2vec: unsupervised machine learning approach with chemical intuition[J]. Journal of Chemical Information and Modeling, 2018, 58(1): 27-35.

[99] UNSAL S, ATAS H, ALBAYRAK M, et al. Learning functional properties of proteins with language models[J]. Nature Machine Intelligence, 2022, 4(3): 227-245.

[100] COLEY C W, ROGERS L, GREEN W H, et al. Computer-assisted retrosynthesis based on molecular similarity[J]. ACS Central Science, 2017, 3(12): 1237-1245.

[101] FERRUZ N, SCHMIDT S, HÖCKER B. ProtGPT2 is a deep unsupervised language model for protein design[J]. Nature Communications, 2022, 13(1): 4348.

[102] SHIN J E, RIESSELMAN A J, KOLLASCH A W, et al. Protein design and variant prediction using autoregressive generative models[J]. Nature Communications, 2021, 12: 2403.

[103] REPECKA D, JAUNISKIS V, KARPUS L, et al. Expanding functional protein sequence spaces using generative adversarial networks[J]. Nature Machine Intelligence, 2021, 3(4): 324-333.

[104] BISWAS S, KHIMULYA G, ALLEY E C, et al. Low-N protein engineering with data-efficient deep learning[J]. Nature Methods, 2021, 18(4): 389-396.

[105] LUO S T, SU Y F, PENG X G, et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures[EB/OL]. bioRxiv, 2022[2022-12-01]. https://www.biorxiv.org/content/10.1101/2022.07.10.499510v5.

[106] WANG L, TITOV A, MCGIBBON R, et al. Discovering chemistry with an *ab initio* nanoreactor[J]. Nature Chemistry, 2014, 6(12): 1044-1048.

[107] SIMM G N, VAUCHER A C, REIHER M. Exploration of reaction pathways and chemical transformation networks[J]. The Journal of Physical Chemistry A, 2019, 123(2): 385-399.

[108] WANG Y H, XU H C, ZOU J, et al. Catalytic role of carbonyl oxygens and water in selinadiene synthase[J]. Nature Catalysis, 2022, 5(2): 128-135.

[109] ZANGHELLINI A, JIANG L, WOLLACOTT A M, et al. New algorithms and an *in silico* benchmark for computational enzyme design[J]. Protein Science, 2006, 15(12): 2785-2794.

[110] LEAVER-FAY A, TYKA M, LEWIS S M, et al. Chapter nineteen-Rosetta3: an object-oriented software suite for the simulation and design of macromolecules[M]// Methods in enzymology. Pittsburgh, PA, USA: Academic Press, 2011, 487: 545-574.

[111] SIEGEL J B, SMITH A L, POUST S, et al. Computational protein design enables a novel one-carbon assimilation pathway[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(12): 3704-3709.

[112] VERGES A, CAMBON E, BARBE S, et al. Computer-aided engineering of a transglycosylase for the glucosylation of an unnatural disaccharide of relevance for bacterial antigen synthesis[J]. ACS Catalysis, 2015, 5(2): 1186-1198.

[113] CUI Y L, WANG Y H, TIAN W Y, et al. Development of a versatile and efficient C-N lyase platform for asymmetric hydroamination via computational enzyme redesign[J]. Nature Catalysis, 2021, 4(5): 364-373.

[114] ZENG T, HESS B A, ZHANG F, et al. Bio-inspired chemical space exploration of terpenoids[J]. Briefings in Bioinformatics, 2022, 23(5): bbac197.

[115] ZHANG L F, HAN J Q, WANG H, et al. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics[J]. Physical Review Letters, 2018, 120(14): 143001.

[116] HU Q N, DENG Z, HU H N, et al. RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity[J]. Bioinformatics, 2011, 27 (17): 2465-2467.

通讯作者：巫瑞波（1984—），男，教授，博士生导师。研究方向为基于多尺度模拟的萜类天然产物生物智造与药效挖掘。

E-mail：wurb3@mail.sysu.edu.cn

第一作者：曾涛（1995—），男，博士研究生。研究方向为计算驱动的生物合成路线设计与优化。

E-mail：zengt28@mail2.sysu.edu.cn