Brief Communication
简短的沟通
nature biotechnology
Https://doi.org/10.1038/s41587-019-0224-x

Deep learning enables rapid identification of potent DDR1 kinase inhibitors

深度学习使快速识别潜在的 DDR1 激酶抑制剂成为可能

Alex Zhavoronkov[1]*, Yan A. Ivanenkov[1], Alex Aliper[1], Mark S. Veselov[1], Vladimir A. Aladinskiy[1],
1 *， Yan a. Ivanenkov1， Alex Aliper1， Mark s. veselov 1， Vladimir a. Aladinskiy1,

Anastasiya V. Aladinskaya[1], Victor A. Terentiev[1], Daniil A. Polykovskiy[1], Maksim D. Kuznetsov[1],
1， Daniil a. Polykovskiy1， Maksim d. Kuznetsov1,

Arip Asadulaev[1], Yury Volkov[1], Artem Zholus[1], Rim R. Shayakhmetov[1], Alexander Zhebrak[1],
1， Yury Volkov1， Artem zhoulus1， Rim r. shaykhmetov1， Alexander zebrak1,

Lidiya I. Minaeva[1], Bogdan A. Zagribelnyy[1], Lennart H. Lee[2], Richard Soll[2], David Madge[2], Li Xing[2],
1， Bogdan a. Zagribelnyy1， Lennart h. Lee 2， Richard Soll2， David Madge2， Li Xing2,

Tao Guo[2] and Alán Aspuru-Guzik[3,4,5,6]
陶果 2 和阿尔 阿斯普鲁-古兹克 3,4,5,6

We have developed a deep generative model, generative tensorial reinforcement learning (GENTRL), for de novo small-molecule design. GENTRL optimizes synthetic feasibility, novelty, and biological activity. We used GENTRL to discover potent inhibitors of discoidin domain receptor 1 (DDR1), a kinase tar-get implicated in fibrosis and other diseases, in 21 days. Four compounds were active in biochemical assays, and two were validated in cell-based assays. One lead candidate was tested and demonstrated favorable pharmacokinetics in mice.

我们已经开发了一种深生成模型，生成微粒强化学习(GENTRL)，用于从头设计小摩尔微粒。Gentrl 优化了合成的可行性、新颖性和生物活性。我们使用 GENTRL 在 21 天内发现了强效的椎间盘蛋白结构域受体 1(DDR1)抑制剂，这是一种与纤维化和其他疾病有关的激酶。四种化合物在生化检测中有活性，其中两种在细胞检测中得到验证。其中一个领先候选者在小鼠身上进行了测试并证明了良好的药代动力学。

Drug discovery is resource intensive, and involves typical timelines of 10–20 years and costs that range from US$0.5 billion to US$2.6 billion[1,2]. Artificial intelligence promises to accelerate this process and reduce costs by facilitating the rapid identification of compounds[3,4]. Deep generative models are machine learning techniques that use neural networks to produce new data objects. These techniques can generate objects with certain properties, such as activity against a given target, that make them well suited for the discovery of drug candidates. However, few examples of generative drug design have achieved experimental validation involving synthesis of novel compounds for in vitro and in vivo investigation[5–16].

药物开发是资源密集型的，通常需要 10-20 年的时间，成本从 5 亿美元到 26 亿美元不等。人工智能有望通过促进快速识别化合物 3、4 来加速这一过程并降低成本。深层生成模型是一种利用神经网络产生新数据对象的机器学习技术。这些技术可以生成具有特定性质的物体，例如针对给定目标的活动，这使得它们非常适合于发现候选药物。然而，很少有生殖药物设计的例子实现了包括合成新化合物用于体外和体内研究的实验验证 5-16。

Discoidin domain receptor 1 (DDR1) is a collagen-activated pro-inflammatory receptor tyrosine kinase that is expressed in epithelial cells and involved in fibrosis[17]. However, it is not clear whether DDR1 directly regulates fibrotic processes, such as myofibroblast activa-tion and collagen deposition, or earlier inflammatory events that are associated with reduced macrophage infiltration. Since 2013, at least eight chemotypes have been published as selective DDR1 (or DDR1 and DDR2) small-molecule inhibitors (Supplementary Table 1). Recently, a series of highly selective, spiro-indoline-based DDR1 inhibitors were shown to have potential therapeutic efficacy against renal fibrosis in a *Col4a3*$^{-/-}$ mice model of Alport syndrome[18]. A wider diversity of DDR1 inhibitors would therefore enable further basic understanding and therapeutic intervention.

盘状蛋白结构域受体 1(DDR1)是一种胶原激活的促炎症受体酪氨酸激酶，在上皮细胞中表达，参与纤维化 17。然而，还不清楚是否 DDR1 直接调节纤维化过程，如肌成纤维细胞活化和胶原沉积，或早期炎症事件与减少巨噬细胞浸润有关。自 2013 年以来，至少有 8 种化学类型被公布为选择性 DDR1(或 DDR1 和 DDR2)小分子抑制剂(补充表 1)。最近，一系列高选择性的、以螺旋吲哚啉为基础的 DDR1 抑制剂被证明对 Alport 综合征的 Col4a3/-小鼠模型的肾纤维化有潜在的治疗效果。因此，更广泛的 mdr1 抑制剂的多样性将有助于进一步的基础理解和治疗干预。

We developed generative tensorial reinforcement learning (GENTRL), a machine learning approach for de novo drug design. GENTRL prioritizes the synthetic feasibility of a compound, its

effectiveness against a given biological target, and how distinct it is from other molecules in the literature and patent space. In this work, GENTRL was used to rapidly design novel compounds that are active against DDR1 kinase. Six of these compounds, each complying with Lipinski's rules[1], were designed, synthesized, and experimentally tested in 46 days, which demonstrates the potential of this approach to provide rapid and effective molecular design (Fig. 1a).

我们开发了生殖张量强化学习(GENTRL)，这是一种用于新药设计的机器学习方法。Gentrl 优先考虑一种化合物的合成可行性，它对特定生物目标的有效性，以及它与文献和专利空间中的其他分子的区别。在这项工作中，GENTRL 被用来快速设计新的化合物是活性对 DDR1 激酶。其中六个化合物，每个符合 Lipinski 的规则 1，设计，合成，和在 46 天的实验测试，这表明这种方法的潜力，以提供快速和有效的分子设计(图 1a)。

To create GENTRL, we combined reinforcement learning, variational inference, and tensor decompositions into a generative two-step machine learning algorithm (Supplementary Fig. 1)[19]. First, we learned a mapping of chemical space, a set of discrete molecular graphs, to a continuous space of 50 dimensions. We parameterized the structure of the learned manifold in the tensor train format to use par-tially known properties. Our auto-encoder-based model compresses the space of structures onto a distribution that parameterizes the latent space in a high-dimensional lattice with an exponentially large number of multidimensional Gaussians in its nodes. This parameter-ization ties latent codes and properties, and works with missing values without their explicit input. In the second step, we explored this space with reinforcement learning to discover new compounds.

为了创建 GENTRL，我们将强化学习、变量推理和张量分解结合到一个产生式两步机器学习算法中。首先，我们学习了化学空间的映射，一组离散的分子图，到一个 50 维的连续空间。我们将学习流形的结构参数化为张量列车格式，以便利用已知的特性。我们的基于自动编码器的模型将结构空间压缩到一个分布上，这个分布参数化了高维格子中的潜在空间，其节点中有大量的多维高斯数。这个参数化关系到潜在的代码和属性，并且在没有显式输入的情况下处理缺失的值。在第二步，我们和强化学习一起探索这个空间，以发现新的化合物。

GENTRL uses three distinct self-organizing maps (SOMs) as reward functions: the trending SOM, the general kinase SOM, and the specific kinase SOM. The trending SOM is a Kohonen-based reward function that scores compound novelty using the applica-tion priority date of structures that have been disclosed in patents. Neurons that are abundantly populated with novel chemical entities reward the generative model. The general kinase SOM is a Kohonen map that distinguishes kinase inhibitors from other classes of molecules. The specific kinase SOM isolates DDR1 inhibitors from the total pool of kinase-targeted molecules. GENTRL prioritizes the structures it generates by using these three SOMs in sequence.

Gentrl 使用三种不同的自组织映射(SOMs)作为奖赏功能:趋势 SOM、一般激酶 SOM 和特异性激酶 SOM。该方法是一种基于 kohonen 的奖励函数，利用专利中已公开的结构的应用优先级数据对复合新颖性进行评分。神经元中充满了新奇的化学物质，这是对生成模型的奖赏。一般激酶 SOM 是一个 Kohonen 地图，区分激酶抑制剂从其他类分子。特异性激酶 SOM 从靶向激酶的分子总库中分离出 dr1 抑制剂。通过按顺序使用这三个 soma，GENTRL 对其生成的结构进行优先排序。

We used six data sets to build the model: (1) a large set of molecules derived from a ZINC data set, (2) known DDR1 kinase inhibitors, (3) common kinase inhibitors (positive set), (4) molecules that act on non-kinase targets (negative set), (5) patent data for biologically active molecules that have been claimed by pharmaceutical companies, and (6) three-dimensional (3D) structures for DDR1 inhibitors (Supplementary Table 1). Data sets were preprocessed to exclude gross outliers and to reduce the number of compounds that contained similar structures (see Methods).

我们使用了六个数据集来建立模型:(1)一大组来自锌的数据集的摩尔筛选，(2)已知的 DDR1 激酶抑制剂，(3)普通激酶抑制剂(阳性组)，(4)作用于非激酶靶标的分子(阴性组)，(5)制药公司声称的生物学活性分子的专利数据，和(6)DDR1 抑制剂的三维结构(补充表 1)。对数据集进行了预处理，以排除粗略的异常值，并减少含有类似结构的化合物的数量(见方法)。

We started to train GENTRL (pretraining) on a filtered ZINC database (data set 1, described earlier), and then continued train-ing using the DDR1 and common kinase inhibitors (data set 2 and data set 3). We then launched the reinforcement learning stage with the reward described earlier. We obtained an initial output of 30,000 structures (Supplementary Data Set), which were then

我们开始在一个过滤锌数据库(数据集 1，前面描述过)上训练 GENTRL(预训练)，然后继续使用 DDR1 和普通激酶抑制剂(数据集 2 和数据集 3)进行训练。然后我们用之前提到的奖励启动了强化学习。我们获得了 30,000 个结构(补充数据集)的初始输出

[1]Insilico Medicine Hong Kong Ltd, Pak Shek Kok, New Territories, Hong Kong. [2]WuXi AppTec Co., Ltd, Shanghai, China. [3]Department of Chemistry, University of Toronto, Toronto, Ontario, Canada. [4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [5]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. [6]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. *e-mail: 1038 Nature Biotechnology | VOL 37 | SEPTEMBER 2019 | 1038–1040 | alex@insilico.com

香港新界白石角一硅医药香港有限公司。上海二无锡应用技术有限公司。加拿大安大略省多伦多市多伦多大学化学系。加拿大安大略省多伦多市多伦多大学计算机科学系。5.加拿大安大略省多伦多市 vector 人工智能研究所。6、加拿大安大略省多伦多市加拿大高级研究院。自然-生物技术:1038|VOL37|SEPTEMBER2019|1038-1040|

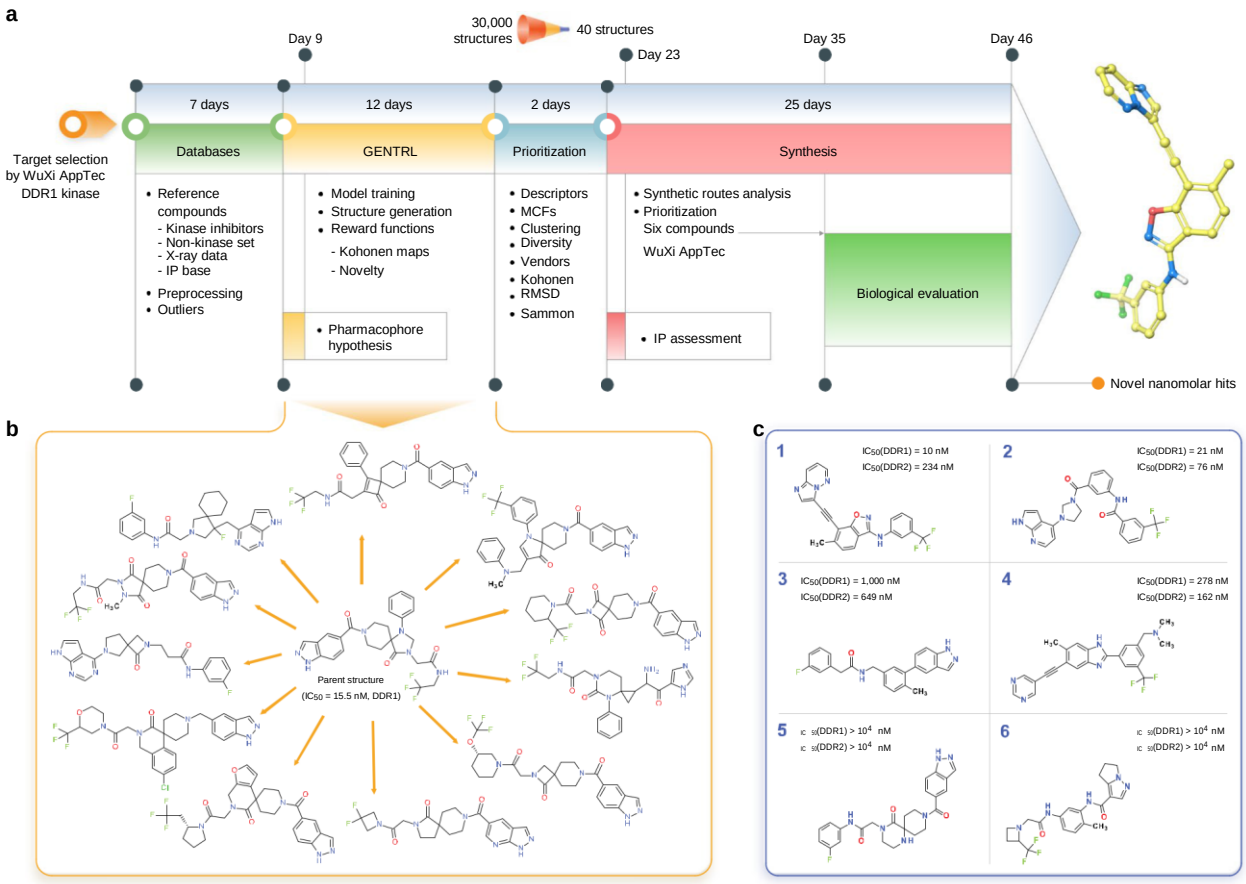

Fig. 1 | GENTRL model design, workflow, and nanomolar hits. a, The general workflow and timeline for the design of lead candidates using GENTRL. IP, intellectual property. **b**, Representative examples of generated structures compared to the parent DDR1 kinase inhibitor. **c**, Generated compounds with the highest inhibition activity against human DDR1 kinase.

图 1|GENTRL 模型设计、工作流程和纳摩尔点击率。使用 GENTRL 设计主要候选人的一般工作流程和时间表。知识产权、知识产权。生成的结构的典型例子相比，母体 DDR1 激酶抑制剂。生成的对人类 DDR1 激酶具有最高抑制活性的化合物。

automatically filtered to remove molecules bearing structural alerts or reactive groups, and the resulting chemical space was reduced by clustering and diversity sorting (Supplementary Table 2). We then evaluated structures using (1) the general and specific kinase SOMs, and (2) pharmacophore modeling on the basis of crystal structures of compounds in complex with DDR1 (Supplementary Figs. 2 and 3). On the basis of the values of molecular descriptors and root-mean-square deviation (RMSD) calculated in two previous steps (steps 6 and 7), we used Sammon mapping to assess the distribution of the remaining structures.

自动过滤去除带有结构警告或活性基团的分子，由此产生的化学空间通过聚类和多样性分类而减少(补充表 2)。然后我们使用(1)一般和特异性激酶 SOMs 评价结构，(2)药效团模型的基础上化合物的晶体结构与 DDR1 的复合物(补充图 2 和 3)。基于前两步计算的分子描述符和均方根差，我们使用 Sammon 映射来评估剩余结构的分布。

To narrow our focus to a smaller set of molecules for analysis, we randomly selected 40 structures that smoothly covered the resulting chemical space and distribution of RMSD values (Supplementary Fig. 4 and Supplementary Table 3). Of the 40 selected structures, 39 were likely to fall outside the scope of any published patents or applications (Supplementary Table 4) . Six of these were chosen for experimental validation on the basis of syn-thetic accessibility. Of note, our approach led to several examples of nontrivial potentially bioisosteric replacements and topological modifications (Fig. 1b).

为了将我们的注意力集中在一组较小的分子上进行分析，我们随机选择了 40 个结构，顺利地覆盖了最终的化学空间和 RMSD 值的分布(补充图 4 和补充表 3)。在 40 个选定的结构中，有 39 个可能不属于任何已发表的专利或应用的范围(补充表 4)。其中六个是选择实验验证的基础上综合易达性。值得注意的是，我们的方法导致了几个潜在的非平凡的生物同位素置换和拓扑修改的例子(图 1b)。

By day 23 after target selection, we had identified six lead candidates, and by day 35, these molecules had been successfully synthesized (Fig. 1c). They were then tested for in vitro inhibitory activity in an enzymatic kinase assay (Supplementary Fig. 5). Compounds 1 and 2 strongly inhibited DDR1 activity (half-maximum inhibi-tory concentration (IC_{50}) values of 10 and 21 nM, respectively), compounds 3 and 4 demonstrated moderate potency (IC_{50} values of 1 μM and 278 nM, respectively), and compounds 5 and 6 were inactive. Compounds 1 and 2 both exhibited selectivity towards DDR1

到目标选择后的第 23 天，我们已经确定了 6 个铅的大小，到第 35 天，这些分子已经成功地合成了大小(图 1c)。然后他们在体外酶激酶抑制活性测试(补

充图 5)。化合物 1 和 2 对 DDR1 活性有强烈的抑制作用(半最大抑制浓度(IC50)分别为 10 和 21nm)，化合物 3 和 4 具有中等效力(IC50 值分别为 1m 和 278nm)，化合物 5 和 6 具有抑制作用。化合物 1 和 2 都对 DDR1 有选择性

over DDR2 (Fig. 1c). Furthermore, compound 1 exhibited a relatively high selectivity index compared to those of 44 diverse kinases (Supplementary Fig. 6).

超过 DDR2(图 1c)。此外，与 44 种不同的激酶相比，化合物 1 具有相对较高的选择性指数(补充图 6)。

Next, we investigated the DDR1 inhibitory activity of compound 1 and compound 2 as measured by autophosphorylation in U2OS cells. The compounds showed IC_{50} values of 10.3 and 5.8 nM, respectively (Supplementary Fig. 7). Both molecules inhibited the induction of fibrotic markers α-actin and CCN2 in MRC-5 lung fibroblasts (Supplementary Fig. 8). These molecules also inhibited the expression of collagen (a hallmark of fibrosis) in LX-2 hepatic stellate cells, with compound 1 showing potent activity at 13 nM (Supplementary Fig. 9).

接下来，我们研究了自体磷酸化在 U2OS 细胞中测定的 compound1 和 compound2 的 DDR1 抑制活性。化合物的 IC50 值分别为 10.3 和 5.8nM(补充图 7)。这两种分子都能抑制 MRC-5 肺成纤维细胞中肌动蛋白和 CCN2 的表达(图 8)。这些分子也抑制 LX-2 肝星状细胞胶原(纤维化的标志)的表达，化合物 1 在 13nM 处显示强大的活性(补充图 9)。

We then performed in vitro microsomal stability studies to characterize the metabolic stability of compounds 1 and 2 in human, rat, mouse, and dog liver microsomes. Compounds 1 and 2 had half-life and clearance values that were similar to or more favor-able than those of routinely used control molecules (Supplementary Table 5). Compound 2 was also found to be very stable in buffer conditions (Supplementary Table 6). Neither compound strongly inhibited cytochrome P450, and both compounds showed favor-able physiochemical properties, including satisfying Lipinski's rules (Supplementary Tables 7 and 8).

然后，我们进行了体外微粒体稳定性研究，以确定化合物 1 和 2 在人、大鼠、小鼠和狗肝微粒体中的代谢稳定性。化合物 1 和 2 的半衰期和清除值与常规使用的对照分子相似或更有利(补充表 5)。化合物 2 也被发现是非常稳定的缓冲条件(补充表 6)。两种化合物对细胞色素 P450 都没有明显的抑制作用，两种化合物都具有良好的理化性质，包括符合 Lipinski 定律(补充表 7 和 8)。

Finally, we tested compound 1 in a rodent model. Compound 1 was delivered to mice intravenously (i.v.) (10 mg kg^{-1}) and orally (p.o., 15 mg kg^{-1}). The two administrations resulted in similar half-lives, ~3.5 h (Fig. 2a and Supplementary Tables 9 and 10). I.v. administration conferred a peak plasma concentration of 2,357 ng ml^{-1} on initial delivery, whereas p.o. administration resulted in a lower maximum of 266 ng ml^{-1}, which peaked 1 h after delivery.

最后，我们在一个啮齿动物模型中测试了化合物 1。化合物 1 静脉注射(10mgkg-1)，口服(15mgkg-1)。两种药物导致相似的半衰期，约 3.5 小时(图 2a 和补充表 9 和 10)。静脉注射给药初次分娩血浆浓度峰值为 2,357ngml-1，而静脉注射给药的最高峰值为 266ngml-1，在分娩后 1 小时达到峰值。

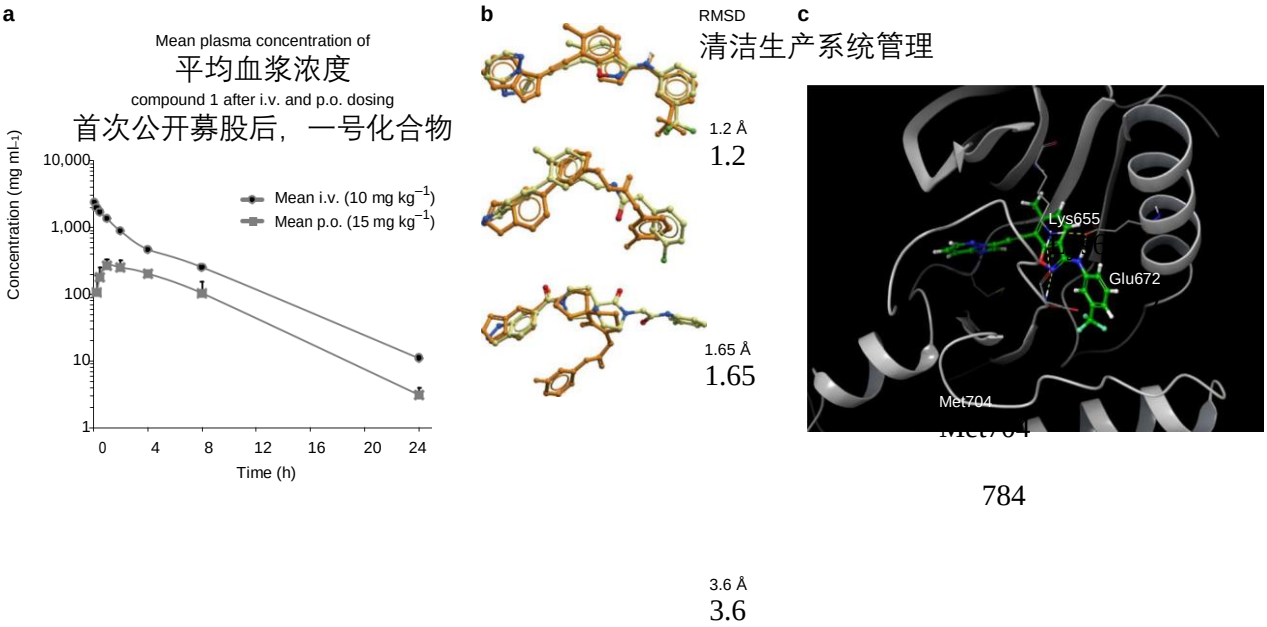

Fig. 2 | Pharmacokinetic characterization and structural basis of hit activity. a, Plasma concentrations of compound 1 in mouse pharmacokinetic study at doses of 10 and 15 mg kg^{-1} for i.v. and p.o. treatment, respectively. Measure of center is mean; error bars are s.d.; $n = 3$ biologically independent animals used for each route of administration. **b**, The rigid alignment of a conformation that best fit the pharmacophore hypothesis and a conformation predicted by quantum mechanical calculations. Superpositions are presented for compound 1, compound 3, and compound 5. Orange, quantum mechanical calculation; yellow, pharmacophore modeling. **c**, The putative binding mode of compound 1 (IC_{50} = 10 nM) in DDR1 kinase (PDB code: 3ZOS) derived from docking simulations. The receptor is shown in gray; compound 1 is shown as sticks and balls, and key receptor residues that are involved in ligand binding are shown as sticks. Hydrogen bonds are shown as yellow dashed lines.

图 2|药代动力学角色塑造和撞击活性的结构基础。A，药物 1 在小鼠药代动力学研究中的血浆浓度，剂量分别为 10 和 15 毫克 kg-1，分别用于静脉注射和缓解治疗。中心测量值是平均值;误差线是 s.d.;每种给药途径使用 3 种生物学上独立的动物。最符合药效团假说的构象的刚性排列和量子力学计算所预测的构象。化合物 1、化合物 3 和化合物 5 的叠加位置。橙色，量子力学计算;黄色，药效团模型。化合物 1(ic5010nm)在 DDR1 激酶(PDB 代码:)中的假定结合模式。受体以灰色显示;化合物 1 以棍状和球状显示，参与配体结合的关键受体残基以棍状显示。氢键显示为黄色虚线。

Quantum mechanical analysis was used to explore the mechanistic basis of the activity of compound 1. The predicted conformation of compound 1 according to pharmacophore modeling was very similar to the conformation predicted to be preferred and stable by quantum mechanical calculations (Fig. 2b). We proposed a 'lock and key' entropy-driven binding mechanism between compound 1 and DDR1, and further characterized this binding via molecular docking. The putative binding mode suggests a type II inhibition mechanism (Fig. 2c). In summary, compound 1 forms multiple hydrogen bonds and has favorable charge and hydrophobic inter-actions with the active site residues of DDR1 kinase. The comple-mentarity of compound 1 to the ATP site may help to explain its inhibitory activity against DDR1.

用量子力学方法研究了化合物 1 活性的机理基础。根据药效团模型预测化合物 1 的构象与量子力学计算预测的首选构象和稳定构象非常相似(图 2b)。我们提出了化合物 1 和 DDR1 之间由熵驱动的"锁和钥匙"结合机制，并通过分子对接进一步表征了这种结合。假定的结合模式表明了一种 II 型抑制机制(图 2c)。综上所述，化合物 1 与 DDR1 激酶的活性位点残基形成多重氢键，具有良好的电荷和疏水作用。化合物 1 对 ATP 位点的复合作用可能有助于解释其对 DDR1 的抑制活性。

Despite reasonable microsomal stability and pharmacokinetic properties, the compounds that have been identified here may require further optimization in terms of selectivity, specificity, and other medicinal chemistry properties.

尽管有合理的微粒体稳定性和药代动力学性质，这里已经确定的化合物可能需要在选择性、特异性和其他药物化学属性方面进一步优化。

In this work, we designed, synthesized, and experimentally vali-dated molecules targeting DDR1 kinase in less than 2 months and for a fraction of the cost associated with a traditional drug discovery approach[1]. This illustrates the utility of our deep generative model for the successful, rapid design of compounds that are synthetically feasible, active against a target of interest, and potentially innovative with respect to existing intellectual properties. We anticipate that this technology will be improved further as a useful tool to identify drug candidates.

在这项工作中，我们在不到两个月的时间里设计、合成和实验性地测定了靶向 DDR1 激酶的分子，而且成本只是传统药物研发成本的一小部分。这说明

了我们的深生成模型在成功的快速设计化合物方面的效用，这些化合物是综合可行的，对感兴趣的目标具有活性，并且在现有的知识产权方面具有潜在的创新性。我们期望这项技术将得到进一步改进，成为确定候选药物的有用工具。

Online content

在线内容

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/ s41587-019-0224-x.

任何方法、附加参考文献、自然研究报告摘要、源数据、代码声明和数据可用性及相关的加入代码均可在。

Received: 1 November 2018; Accepted: 12 July 2019;
收到:2018 年 11 月 1 日;接受:2019 年 7 月 12 日;
Published online: 2 September 2019
发表于 2019 年 9 月 2 日

References

参考资料

\1.\ Paul, S. M. et al. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
1.1997 年，他在《经济学人》杂志上发表了《经济学人》一文。纳特。Rev.Drug 的斯科夫。9,203-214(2010).
2. Avorn, J. *N. Engl. J. Med.* **372**, 1877–1879 (2015).
图 2。艾沃恩，j.n.Engl。J.Med.372,1877-1879(2015).
\3.\Goodfellow, I. et al. Generative adversarial nets. in *Advances in Neural Information Processing Systems* 2672–2680 (2014).
3.“好家伙，我等人。生殖对抗性网。《神经信息处理系统的进展》2672-2680(2014)。
\4.\ Mamoshina, P. et al. *Mol. Pharm.* **13**, 1445–1454 (2016).
4.等人。分子。Pharm.13,1445-1454(2016).
\5.\ Sanchez-Lengeling, B. & Aspuru-Guzik, A. *Science* **361**, 360–365 (2018).
5. Sanchez-lengeling，b. & Aspuru-Guzik，a. Science 361,360-365(2018).
\6.\ Kadurin, A. et al. *Oncotarget* **8**, 10883–10890 (2016).
6.1997 年，在中国南方广播电视台播出的《广播电视》节目中，广播电视台播出了广播电视节目《广播电视节目》。Oncotarget8,10883-10890(2016).
\7.\ Kadurin, A. et al. *Mol. Pharm.* **14**, 3098–3104 (2017).
7.1997 年，在中国南方广播电视台播出的《广播电视》节目中，广播电视台播出了广播电视节目《广播电视节目》。分子。Pharm.14,3098-3104(2017).
\8.\ Gómez-Bombarelli, R. et al. *ACS Cent. Sci.* **4**, 268–276 (2018).
8.等人。美国化学学会美分。科学。4,268-276(2018).
\9.\ Putin, E. et al. *Mol. Pharm.* **15**, 4386–4397 (2018).
9.普京等人。分子。Pharm.15,4386-4397(2018).
\10.\Putin, E. et al. *J. Chem. Inf. Model.* **58**, 1194–1204 (2018).
10.普京等人。北京杰凯化工技术有限公司。Inf.模型。58,1194-1204(2018).
\11.\Harel, S. & Radinsky, K. *Mol. Pharm.* **15**, 4406–4416 (2018).
11. Harel，s. & Radinsky，k. Mol. Pharm. 15,4406-4416(2018).
\12.\Polykovskiy, D. et al. *Mol. Pharm.* **15**, 4398–4405 (2018).
12.波利可夫斯基等人。分子。Pharm.15,4398-4405(2018).
\13.\Kuzminykh, D. et al. *Mol. Pharm.* **15**, 4378–4385 (2018).
13.等人。分子。Pharm.15,4378-4385(2018).
\14.\Segler, M. H. S. et al. *Nature* **555**, 604–610 (2018).
14. Segler，M.H.S.et al. . Nature 555,604-610(2018).
\15.\Merk, D. et al. *Mol. Inform.* **37**, 1–2 (2018).
15.等人。分子。通知。37,1-2(2018).
\16.\Merk, D. et al. *Commun. Chem.* **1.1**, 68 (2018).
16.等人。同“Commun”。化学。1.1,68(2018).
\17.\Moll, S. et al. *Biochim. Biophys. Acta Mol. Cell Res.* https://doi.org/10.1016/j.bbamcr.2019.04.004 (2019).
17.莫尔等人。Biochim.Biophys.ActaMol.CellRes.(2019 年)。
\18.\Richter, H. et al. *ACS Chem. Biol.* **14**, 37–49 (2019).
18.里克特等人。美国化学学会。Biol.14,37-49(2019).
\19.\Elton, D. C. et al. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
19.埃尔顿，特区等人。分子。Syst.女名女子名。英国。4,828-849(2019).

Acknowledgements

鸣谢

The authors thank T. Oprea (University of New Mexico School of Medicine) for the valuable contributions, review, and assessment of the novelty of the intellectual property generated by GENTRL. The authors would like to thank NVIDIA Corporation and
作者感谢 T.Oprea(新墨西哥大学医学院)的宝贵贡献，审查和评估的新颖性的知识产权产生的 GENTRL。作者想要感谢英伟达和
M. Berger for providing early access to the graphics processing equipment used for deep learning applications by Insilico Medicine. The authors acknowledge T. Lu, L. Duan, Y. Hu, and the WuXi AppTec chemistry team for providing chemical synthesis of the presented compounds. The authors thank S. Djuric, whose valuable comments informed further experiments.
为早期使用 Insilico 医学公司用于深度学习应用程序的图形处理设备提供了机会。作者承认 t.Lu，l.Duan，y.Hu 和无锡 AppTec 化学团队提供了化学合成化合物。作者感谢 s.Djuric，他的宝贵意见为进一步的实验提供了参考。

Author contributions

作者贡献

A. Zhavoronkov, Y.A.I., and A.A. led the project, designed and planned the experiments, and wrote the manuscript. M.S.V., V.A.A., A.V.A., and V.A.T. planned and performed computational chemistry experiments. D.A.P., M.D.K., A. Zholus, A.A., Y.V., R.R.S., and A. Zhebrak developed and implemented the GENTRL. L.I.M. curated chemical synthesis, and B.A.Z. collected and prepared the data. L.H.L., R.S., D.M., L.X., and T.G. helped write the manuscript. A.A.-G. provided manuscript and methodological feedback.

和 A.a.领导了这个项目，设计和规划了实验，并写下了手稿。和 v.a.t 计划并实施了计算化学实验。和 a.zebrak 开发并实施了 GENTRL。和 b.a.z 收集并准备了这些数据化学合成。和 t.g.参与了手稿的撰写。A.a.-g.提供手稿和方法反馈。

Competing interests

相互竞争的利益

A. Zhavoronkov, Y.A.I., A. Aliper, M.S.V., V.A.A., A.V.A., V.A.T., D.A.P., M.D.K.,

作者:a.Zhavoronkov，y.A.i.，a.Aliper，m.s.v.，v.A.a.，A.v.a.，v.A.t.，d.A.p.，m.d.k.,

A. Zholus, A. Asadulaev, Y.V., A. Zhebrak, R.R.S., L.I.M., and B.A.Z. work for Insilico Medicine, a commercial artificial intelligence company. L.H.L., R.S., D.M., L.X., and T.G. work for WuXi AppTec, a commercial research organization. A.A.-G. is a cofounder and board member of, and consultant for, Kebotix, an artificial intelligence-driven molecular discovery company and a member of the science advisory board of Insilico Medicine.

A.Zholus，a.asadlaev，y.v.，a.zebrak，r.r.r.s.，l.i.m.，andb.A.z 在 insikoMedicine，一家商业人工智能公司工作。和 t.g.在商业研究机构无锡 AppTec 工作。A.a.-g.是 Kebotix 的联合创始人、董事会成员和顾问，Kebotix 是一家人工智能驱动的分子发现公司，也是 Insilico 医学科学咨询委员会的成员。

Additional information

补充资料

Supplementary information is available for this paper at https://doi.org/10.1038/ s41587-019-0224-x.

有关补充资料，请浏览。

Reprints and permissions information is available at www.nature.com/reprints.

重印和权限信息可在。

Correspondence and requests for materials should be addressed to A.Z.

信件和索取材料的要求应寄到 a.z。

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

出版者的说明:施普林格自然保持中立的管辖权主张在出版地图和机构附属机构。

Methods

方法

Pretraining data set. For the pretraining procedure, we have prepared a data set of structures using the Clean Leads set from the ZINC database[20] and proprietary databases from our partners. We have removed structures containing atoms other than carbon, nitrogen, oxygen, sulfur, uorine, chlorine, bromine, and hydrogen. Routine medicinal chemistry flters were applied to exclude compounds with potentially toxic and reactive groups.

训练前数据集。对于预培训程序，我们已经准备了一个数据集的结构使用的清洁线索集从锌数据库 20 和专有数据库从我们的合作伙伴。我们已经去除了碳、氮、氧、硫、铀、氯、溴和氢以外的原子结构。常规的药物化学用于排除具有潜在毒性和活性基团的化合物。

Kinase inhibitors and 'negative' data set. The data set of molecules that actively inhibit and do not inhibit various kinases was prepared using the data available in the Integrity and ChEMBL databases.

激酶抑制剂和阴性数据集。数据集的分子积极抑制和不抑制各种激酶是利用完整性和化学/生物学数据库的数据准备的。

Compounds from patent records by priority date. The Integrity database was used to collect the data set of structures claimed as new drug substances in patent records from 1950 to the present day by the top ten pharmaceutical companies (as ranked by market capitalization in 2017 according to https://www.globaldata.com). The final data set contained 17,000 records.

按优先权日期分列的专利记录化合物。完整性数据库用于收集 1950 年至今排名前 10 位的制药公司的专利记录中声称为新药物的结构的数据集。最终的数据集包含 17,000 条记录。

Model. Our generative pipeline was created using the GENTRL model, a variational auto-encoder with a rich prior distribution in the latent space (Supplementary Code and Supplementary Fig. 1). We used tensor decomposition to encode the relationships between molecular structures and their properties, and trained a model in a semisupervised fashion without imputing unknown biochemical properties of molecules.

模型。我们的生成流水线是使用 GENTRL 模型创建的，这是一个在潜在空间中具有丰富先验分布的变分自动编码器(补充代码和补充图 1)。我们使用张量分解对分子结构及其性质之间的关系进行编码，并以半监督的方式训练了一个模型，而没有插入未知的分子生物化学性质。

The tensor-train decomposition[21] approximates high-dimensional tensors using a relatively small number of parameters. A joint distribution $p(r_1, r_2, \ldots, r_n)$ of discrete random variables $r_i \in \{0, \ldots N_i - 1\}$ can be represented as elements of n-dimensional tensor:

张量列式分解 21 使用相对较少的参数来近似高维张量。离散随机变量 ri∈{0，...Ni-1}的联合分布 p(r1，r2，...，rn)可表示为 n 维张量的元素:

$$p(r_1, r_2, \ldots, r_n) = \frac{1}{Z} \mathbf{1}_m \left(\prod_{i=1}^{n} Q_i[r_i] \right) \mathbf{1}_m^T$$

where tensors $Q_i \in \mathbb{R}^{N_i \times m \times m}$ are cores, $\mathbf{1_m}$ is a vector of ones, and Z is a normalizing constant. With larger core sizes, the flexibility of the model improves, although the number of parameters grows quadratically with core size m. In tensor train, we can efficiently marginalize the distribution with respect to any variable,

其中张量 qir+n-im 为核心，1m 为 1 的向量，z 为正规化常数。随着岩心尺寸的增大，模型的参数数目随岩心尺寸的增大呈二次增长，但模型的灵活性有所提高。在张量列车中，我们可以对任何变量有效地边缘化分布，

as follows:

详情如下:

$$p(r_1, \ldots, r_k, r_{k+1}, \ldots, r_n) = \frac{1}{Z} \left(\prod_{i=1}^{k} Q_i[r_i] \right) \left(\prod_{i=k+1}^{n} \tilde{Q}_i[r_i] \right) \mathbf{1}^T$$

where $\tilde{Q}_k = \sum_{r_i} Q_k[r_i]$ can be computed efficiently. With marginal distributions, we can compute the conditional distributions and sample using a chain rule. The normalizing constant Z is given by

其中 QkriQk[ri 可以有效地计算。对于边际分布，我们可以用链式规则计算条件分布和样本。给出了正规化常数 z

$$Z = \mathbf{1}_m \prod_{i=1}^{n} Q_i \mathbf{1}_m$$

I 1

As generative auto-encoders use continuous latent codes, we use continuous tensor-train representation. For simplicity of notation, assume that latent codes z are continuous and properties y are discrete. We approximate distributions $p_\psi(z_i)$ as mixtures of Gaussians with component index s_i. The joint distribution on z and y is

由于生成式自动编码器使用连续的潜码，我们使用连续的张量训练表示。为了简单起见，假设潜码 z 是连续的，性质 y 是离散的。本文将 p(zi)近似分布作为组分指数为 1 的高斯混合物。Z 和 y 上的联合分布为

$$p(z, y) = \sum_{s_1, \ldots, s_d} p(z, y, s_1, \ldots, s_d) = \sum_{s_1, \ldots, s_d} p(z \mid y, s_1, \ldots, s_d) P(y \mid s_1, \ldots, s_d) \prod_{i=1}^{d} p(s_i)$$

P (z， y) p (， s z， y) p [s， y] p (z y， s)

S， ... ， s， ... ， s

1 d 1 d

For conditional distribution $p_\psi(\mathbf{z}|\mathbf{y},\mathbf{s})$, we select a fully factorized Gaussian that does not depend on y:

对于条件分布 p(z|y，s)，我们选择一个不依赖于 y 的完全分解高斯:

$$p(z \mid y, s) = p(z \mid s) = \prod_{k=1}^{d} \mathcal{N}(z_k \mid \mu_{k,s_k}, \sigma^2_{k,s_k})$$

P (z y， s) p (z s) n (zk， sk， k， 2 sk)

K1

The tunable parameters ψ of the distribution p_ψ are tensor-train cores Q_i, means $\mu_{k,sk}$, and variances $\sigma^2_{k,sk}$ of the Gaussian components. We store tensor $P[\mathbf{s},\mathbf{y}]$ in a tensor-train format. The resulting distribution becomes

分布的可调参数是张量列核 Qi，平均 k，sk 和高斯分量的方差 k2，sk。我们以张量列式格式存储张量 p[s，y]。由此产生的分布成为

$$p(z, y) = \sum_{s_1, \ldots, s_d} P[s_1, \ldots, s_d, y] \prod_{k=1}^{d} \mathcal{N}(z_k \mid \mu_{k,s_k}, \sigma^2_{k,s_k})$$

Our model is a variational auto-encoder with a prior distribution $p_\psi(\mathbf{z},\mathbf{y})$, encoder q_φ, and a decoder p_θ. Consider a training example $(\mathbf{x}, \mathbf{y}_{ob})$, where x is a molecule and $\mathbf{y}_{ob}$ are its known properties. The lower bound on a log-marginal likelihood (also known as the evidence lower bound) for our model is

我们的模型是一个具有先验分布 p(z，y)、编码器 q 和解码器 p 的变分自动编码器。考虑一个训练示例(x，yob)，其中 x 是一个分子，yob 是它的已知属性。我们的模型的对数边际似然(也称为证据下界)的下界是

$$L(\theta, \varphi, \psi) = \mathrm{E}_{q(z \mid x, y_{ob})} (\log p(x \mid z, \theta) + \log p(y_{ob} \mid z, \psi))$$

(， ,) Eq (zx， yob)(log p (xz， yob) + log p (yob z))

− KL(q (z x , yob) p (z yob))

- KL (q (zx， yob) p (zyob))

As the molecule determines its properties, we assume that $q_\varphi(\mathbf{z}|\mathbf{x},\mathbf{y}_{ob}) = q_\varphi(\mathbf{z}|\mathbf{x})$. We also assume that $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y}_{ob}) = p_\theta(\mathbf{x}|\mathbf{z})$, indicating that an object is fully defined by its latent code. The resulting evidence lower bound is

由于分子决定了它的性质，我们假设 q(z|x，yob)q(z|x)。我们还假设 p(x|z，yob)p(x|z)，表示对象完全由其潜在代码定义。由此得到的证据下限为

$$L(,,) = \mathrm{E}_{q(z\,x)}(\log p(x\,z) + \log p(y_{ob}\,z)) - \mathrm{KL}(q(z\,x)\ p(z\,y_{ob}))$$

$$\frac{1}{l}\sum_{i=1}^{l} \log p(x\,z)_i + \log p(y\,z)_{ob}^{\,i} - \log \frac{q(z\,x)_i}{p(z_i\,y)_{ob}}$$

where $\mathbf{z}_i \sim q_\varphi(\mathbf{z}|\mathbf{x})$. For the proposed joint distribution $p_\psi(\mathbf{z}, \mathbf{y})$, we can compute the density of the posterior distribution on the latent codes, given observed properties

其中 zi~q(z|x)。对于所提出的联合分布 p(z，y)，我们可以根据观察到的性质计算潜码上的后验概率密度

$p_\psi(\mathbf{z}|\mathbf{y}_{ob})$, analytically.

从分析的角度来说(z|yob)。

By maximizing the evidence lower bound, we trained an auto-encoder and a prior distribution on three data sets described above (pretraining, kinase and patent data sets): we sampled molecules in a simplified molecular input line entry system (SMILES) format from the data set along with their properties, including MCE-18, pIC_{50} (negative common logarithm of IC_{50}) and a binary feature that indicates whether a molecule passed medicinal chemistry filters (MCFs). We trained this model and obtained a mapping from the chemical space to the latent codes. This mapping was aware of the relationship between molecules and their biochemical properties.

通过最大化证据下限，我们训练了一个自动编码器和上述三个数据集(预训练，激酶和专利数据集)的先验分布:我们采样分子在一个简化的分子输入线输入系统(SMILES)格式的数据集及其属性，包括 MCE-18，pIC50(IC50 的负常用对数)和一个二进制特征，表明分子是否通过药物化学过滤器(MCFs)。对模型进行训练，得到从化学空间到潜在编码的映射关系。这个映射意识到了分子和它们的生化特性之间的关系。

In the next stage of training, we fine-tuned the model to preferentially generate DDR1 kinase inhibitors. We used reinforcement learning to expand the latent manifold towards discovering novel inhibitors with reward functions (general kinase SOM, specific kinase SOM, and trending SOM), which are described in the next section. We used the REINFORCE[22] algorithm (also known as a log-derivative trick) to directly optimize the model:

在接下来的训练阶段，我们对模型进行微调，以优先生成 DDR1 激酶抑制剂。我们使用强化学习扩展潜在的多方面发现新的具有奖赏功能的抑制剂(一般激酶 SOM，特异性激酶 SOM，和趋势 SOM)，这是在下一节描述。我们使用 REINFORCE22 算法(也称为对数导数技巧)直接优化模型:

$$max\ \mathrm{E}_{z \sim p(z)} R(z), \quad R(z) = \mathrm{E}_{x \sim p(x\,z)}[R_{general}(x) + R_{specific}(x) + R_{trending}(x)]$$

$$\mathrm{E}_{z \sim p(z)} R(z) = \mathrm{E}_{z \sim p(z)} \log p(z)\, R(z)$$

E z ~ p (z) r (z) Ez ~ p (z) logp (z) r (z)

We reduced the variance of the gradient using a standard variance reduction technique called a ‘baseline’. The rewards for each molecule in a batch are calculated and averaged, and the average reward is then subtracted from each individual reward:

我们使用一种称为基线的标准方差减少技术来减少梯度的方差。一批中每个分子的奖励被计算和平均，然后从每个单独的奖励中减去平均奖励:

$$E_{z\sim p(z)} R(z) \frac{1}{N}\sum_{i=1}^{N} \log p()_i zR z()-_i \frac{1}{N}\sum_{j=1}^{N} R(z)_j$$

To preserve the mapping of the chemical space, we fixed the parameters of the encoder and decoder, and trained only the manifold distribution $p_\psi(z)$. We combined exploration and exploitation approaches. For exploration, we sampled $z^{explore} \sim N(,(2)^2$ outside from the currently explored latent space, where μ and σ_2 are the mean and variance of $p_\psi(z)$ for all dimensions. If the reward $R(z_{explore})$ for a newly discovered area was high, the latent manifold expanded toward it (Supplementary Fig. 1).

为了保持化学空间的映射，我们固定了编码器和解码器的参数，只训练流形分布 p(z)。我们把勘探和开发结合起来。为了探索，我们从当前探索的潜在空间外部取样 zexplore~n(，(2)2，其中和 2 是 p(z)对所有维度的均值和方差。如果一个新发现的区域的奖励 r(zexplore)是高的，潜在的多方面向它扩大(补充图 1)。

The comparison of generative chemistry models is very important for the advancement of this emerging field, and there are several benchmarking platforms in development[12,23]. We successfully compared the performance of GENTRL with previous approaches, including objective-reinforced generative adversarial networks (ORGAN)[24,25], reinforced adversarial neural computer (RANC)[10], and adversarial threshold neural computer (ATNC)[9]. Training details are provided in the Supplementary Note.

生成化学模型的比较对于这一新兴领域的进展非常重要，在开发中有几个基准平台 12、23。我们成功地将 GENTRL 的性能与以前的方法进行了比较，包括客观强化生成对抗性网络(ORGAN)24、25、强化对抗性神经计算机(RANC)10 和对抗性阈值神经计算机(ATNC)9。培训详情载于补充说明。

Reward function. A reward function was developed on the basis of the Kohonen self-organizing maps (SOM)[26] (Supplementary Fig. 3). This algorithm was introduced by Teuvo Kohonen as a unique unsupervised machine- learning dimensionality reduction technique. It can effectively reproduce an intrinsic topology and patterns hidden in the input chemical space in a faithful and unbiased fashion. The input chemical space is usually described in terms of molecular descriptors (input vector), and the output typically includes a 2D or 3D feature map for convenient visual inspection. An ensemble of three SOMs was used as a reward function: the first SOM (general kinase SOM, Rgeneral) was trained to predict the activity of compounds against kinases, the second SOM (specific kinase

奖励功能。在 Kohonen 自组织映射图(SOM)26 的基础上建立奖励函数(补充图 3)。该算法由 TeuvoKohonen 提出，是一种独特的无监督机器学习降维技术。它可以有效地再现一个内在的拓扑结构和模式隐藏在输入化学空间的忠实和无偏见的方式。输入的化学空间通常用分子描述符(输入向量)来描述，输出通常包括一个二维或三维特征映射，以方便视觉检查。以三个 SOMs 为奖赏函数，第一个 SOM(一般激酶 SOM，Rgeneral)被训练用于预测化合物对激酶的活性，第二个 SOM(特异性激酶)被训练用于预测化合物对激酶的活性

SOM, $R_{specific}$) was developed to select compounds located in neurons associated with DDR1 inhibitors within the whole kinase map, and the last SOM (trending

在全激酶图谱中筛选与 DDR1 抑制剂相关的神经元内的化合物，并建立最后的 SOM(Rspecific)模型

SOM, $R_{trending}$) was trained to assess the novelty of chemical structures in terms of the current trends in medicinal chemistry. During learning, the generative model

受过训练，能够根据药物化学的当前趋势来评估化学结构的新颖性。在学习过程中，生成模型

was rewarded when the generated structures were classified as molecules acting on kinases, positioned in neurons attributed to DDR1 inhibitor. The model was also rewarded for generating novel structures.

当所产生的结构被归类为作用于激酶的分子，位于 cdr1 抑制剂所属的神经元中时，就会得到回报。该模型还因为生成新颖的结构而获奖。

Pharmacophore hypotheses. On the basis of X-ray data available in the Protein Data Bank (PDB) database (PDB codes 3ZOS, 4BKJ, 4CKR, 5BVN, 5BVO, 5FDP,

药效团假说。根据蛋白质数据库(PDB)数据库(PDB 代码，，，，，

5FDX, and 6GWR), we have developed three pharmacophore models describing DDR1 inhibitors. To obtain the superposition of the ligands, 3D alignment of the complexes was carried out. These three-, four- and five-centered pharmacophore hypotheses contain key features that are responsible for binding to the active site of DDR1 kinase, including a hydrogen bond acceptor at the hinge region, an aromatic or hydrophobic linker, and a hydrophobic center in the pocket located in proximity to the DFG motif. For detailed information on pharmacophore features and distances, see Supplementary Fig. 2.

我们已经建立了三个描述 DDR1 抑制剂的药效团模型。为了实现配体的叠加，对配合物进行了三维取向。这三个、四个和五个中心的药效团假说包含了与 DDR1 激酶活性位点结合的关键特征，包括铰链区域的氢键受体、芳香族或疏水性连接体以及位于 DFG 基序附近的口袋中的疏水性中心。有关药效团特征和距离的详细信息，请参阅附图 2。

Nonlinear Sammon mapping. To make the final selection, we used a Sammon-based mapping technique[27]. The main goal of this algorithm lies in the approximation of local geometric and topological relationships hidden in the input chemical space on a visually intelligible 2D or 3D plot. The fundamental idea of this method is to substantially reduce the high dimensionality of the initial data set into the low-dimensional feature space, and, in this aspect, it resembles an SOM approach with multidimensional scaling. However, in contrast to other algorithms, a classical Sammon-based method allows scientists to construct a projection that reflects global topographic relationships as pair-wise distances between all of the objects within the whole space of input vector samples. Structures that successfully passed all of the selection procedures described earlier were used as an input chemical space. For mapping, we used the same set of molecular descriptors that was applied for specific kinase SOM and added RMSD values obtained during pharmacophore modeling as additional inputs. Euclidean distances were used

非线性 Sammon 映射。为了进行最终的选择，我们使用了基于 sammon 的映射技术 27。该算法的主要目标是在一个直观易懂的二维或三维图形上逼近隐藏在输入化学空间中的局部几何和拓扑关系。这种方法的基本思想是将原始数据集的高维数据大幅度降低到低维特征空间，在这方面，它类似于一个带有多维标度的 SOM 方法。然而，与其他算法不同的是，经典的基于 sammon 的方法允许科学家构造一个投影，反映全球地形关系，即输入矢量样本整个空间内所有物体之间的成对距离。成功通过前面描述的所有选择程序的结构被用作输入化学空间。对于映射，我们使用相同的分子描述符集应用于特定的激酶 SOM 和添加 RMSD 值药效团模型作为额外的输入。使用欧氏距离

as a similarity metric. The stress threshold was 0.01, the interaction number was 300, the optimization step was 0.3 and the structural similarity factor was 0.5. The resulting map (Supplementary Fig. 4) demonstrates that structures are normally distributed within the Sammon plot.

作为一个相似度量。应力阈值为 0.01，相互作用次数为 300，优化步骤为 0.3，结构相似性因子为 0.5。结果图(补充图 4)表明，结构是正态分布的 Sammon 图。

Molecule generation and selection procedure. Using our model, we generated 30,000 unique valid structures by sampling latent codes from the learned manifold $p_\psi(z)$ and sampling structures from the decoder distribution $p_\theta(x|z)$. To select the batch of molecules for synthesis and biological studies, we developed a prioritization pipeline (for examples of rejected molecules, see Supplementary Fig. 10). At the initial step, the data set was reduced to 12,147 compounds using the following molecular descriptor thresholds: $-2 < \log P < 7$, $250 < MW < 750$, HBA

分子生成和选择过程。使用我们的模型，我们通过从学习流形 p(z)和从解码器分布 p(x|z)的抽样结构中抽取潜码生成 30,000 个唯一的有效结构。为了选择用于合成和生物学研究的一批分子，我们开发了一个优先级管道(对于排斥分子的例子，见附图 10)。在最初的步骤，数据集减少到 12,147 个化合物使用下列分子描述符阈值:-2logp7,250MW750，HBA

+ HBD < 10, TPSA < 150, and NRB < 10. After that, 150 in-house MCFs were applied to remove potentially toxic structures and compounds containing reactive and undesirable groups. These include substrates for 1,4-addition (Michael-bearing moieties) and other electrophilic species (for example, *para*- or *ortho*-halogen-substituted pyridines, 2-halogen-substituted furans and thiophenes, alkyl halides, and aldehydes and anhydrides), disulfides, isatins, barbiturates, strained heterocycles, fused polyaromatic systems, detergents, hydroxamic acids and diazo-compounds, peroxides, unstable fragments, and sulfonyl ester derivatives. In addition, we used more trivial filtering rules that excluded the following: <2 NO_2 groups, <3 Cl, <2 Br, <6 F, and <5 aromatic rings, and undesired atoms, such as silicon, cobalt or phosphorus. This reduced the number of structures spread within the entire chemical space to drug-like molecules without structural alerts. This procedure resulted in 7,912 structures. A clustering analysis was then performed using Tanimoto similarity as a metric and standard Morgan fingerprints implemented in the RDKit package. All compounds that satisfied a

10，TPSA150，andNRB10.在此之后，应用了 150 个内部 MCFs 去除具有潜在毒性的结构和含有反应性和不良基团的化合物。这些基质包括 1,4 加成物(含 michael 基团)和其他亲电性物种(例如，对位或邻位卤代吡啶，2-卤代呋喃和噻吩，烷基卤化物，醛和酐)，双硫醚，异硫氰酸盐，巴比妥酸盐，张力杂环，稠合多环芳烃体系，偶氮，羟基酸和二元化合物，过氧化物，不稳定碎片和磺酰基酯衍生物。此外，我们使用了更多的平凡过滤规则，排除了以下几点:2NO2 基团，3Cl，2Br，6f 和 5 个芳环，以及不想要的原子，如硅，钴或磷。这减少了扩散到整个化学空间的类药物分子的结构数量，而没有结构警报。这个过程产生了 7912 个结构。然后使用 Tanimoto 相似度作为 RDKit 包中实现的度量和标准 Morgan 指纹进行聚类分析。所有的化合物都能满足

0.6 similarity threshold were assigned to the same cluster, with a minimum value of five structures per cluster. Inside each cluster, the compounds were sorted according to their internal dissimilarity coefficient to output the top five items with the maximum diversity in structure. As a result, the data set was reduced to 5,542 molecules. Then, we performed a similarity search using vendors' collections (MolPort (https://www.molport.com) and ZINC[18]) and removed a further

0.6 的相似性阈值指定给同一聚类，每个聚类最小值为 5 个结构。在每个聚类中，根据化合物的内部差异系数进行分类，得到结构差异最大的前五个项目。结果，数据集减少到 5542 个分子。然后，我们使用供应商的集合(MolPort()和 ZINC18)执行了一个最近邻搜索，并进一步删除了

900 compounds with similarity >0.5 to increase the novelty of the generated structures. General kinase SOM and specific kinase SOM were used to prioritize the compounds by their potential activity against DDR1 kinase. Out of 2,570 molecules classified as kinase inhibitors by general kinase SOM, 1,951 molecules were classified as DDR1 inhibitors by specific kinase SOM and were used for pharmacophore-based virtual screening. For every molecule, ten conformations were generated and minimized using RDKit's implementation of the universal force field[28]. Using the developed hypotheses, the screening procedure was carried out, resulting in a set of RMSD values for 848 molecules matching at least one pharmacophore hypothesis. On the basis of Sammon mapping, we uniformly selected 20 molecules from ellipses corresponding to four- and five-centered pharmacophores (Supplementary Table 3 and Supplementary Fig. 4). Forty molecules were selected for synthesis and subsequent biological evaluation.

900 个化合物，相似度为 0.5，以增加产生的结构的新颖性。一般激酶 SOM 和特异性激酶 SOM 被用来优先化化合物的潜在活性对 DDR1 激酶。在

2,570 个被一般激酶 SOM 分类为激酶抑制剂的分子中，1,951 个被特异性激酶 SOM 分类为 DDR1 抑制剂，并用于基于药效团的虚拟筛选。对于每个分子，通过使用 RDKit 的通用力场 28 的实现，生成并最小化了 10 个构象。利用发展的假设，进行了筛选程序，导致了一组 848 个分子的 RMSD 值符合至少一个药效团假设。在 Sammon 映射的基础上，我们从椭圆中统一选择了 20 个分子，它们分别对应于四个和五个中心的药物载体(补充表 3 和补充图 4)。选择了四十个分子进行合成和随后的生物学评价。

Ab initio calculation details. We carried out first-principles calculations to the lowest conformer as predicted with the universal-force-field methodology presented earlier. Geometry optimization was performed using a local correlated coupled-cluster method that included single and double excitations (LCCSD) with the 6-31++G basis set. Final energies were calculated at the LCCSD(T) level of theory. The localized Pipek–Mezey procedure was used to obtain the initial molecular orbitals.

从头计算细节。我们进行了第一性原理计算的最低成员与预测的通用力场方法提出了前面。采用 6-31++g 基组的单激发和双激发局域相关耦合团簇方法(lcssd)进行了几何优化。最后的能量计算在 lcsd(t)的理论水平。采用局域化的 Pipek-Mezey 程序获得了初始分子轨道。

Docking simulations. Molecular modeling was performed in the Maestro suite

分子对接模拟在 Maestro 套件中进行

(https://www.schrodinger.com). PDB structure 3ZOS was preprocessed and energy

()。PDB 结构的预处理和能量

minimized using the Prep module. The binding site grid was generated around the ATP binding site with 20 Å buffer dimensions. Docking poses were generated by extra-precision (XP) Glide runs using the optimized ligand structure. The final model was selected on the basis of its docking score of –15 kcal mol^{-1}, which is lowest among all of the obtained models.

最小化使用准备模块。结合位点网格是围绕 ATP 结合位点生成的，具有 20 个缓冲尺寸。利用优化的配体结构，通过超精度滑翔运动生成对接姿态。最终模型的选择是基于其对接分数-15 千卡 mol-1，这是所有获得的模型中最低的。

In vitro activity assays. The activity of the molecules against human DDR1 and human DDR2 kinases was assessed using KinaseProfiler (Eurofins Scientific).

体外活性测定。用

KinaseProfiler(EurofinsScientific)检测了这些分子

对人类 DDR1 和人类 DDR2 激酶的活性。

Cell-culture activity assay. To measure autophosphorylation, the gene encoding human DDR1b with a hemagglutinin tag was cloned into pCMV Tet-On vector (Clontech), and stable inducible cell lines established in U2OS were used for the IC_{50} test. *DDR1* expression was induced for 48 h before DDR1 activation by rat tail collagen I (Sigma 11179179001). The cells were detached with trypsinization and transferred to a 15 ml tube. Then after pretreatment with the compound for 0.5 h, the cells were treated with compounds in the presence of 10 μg ml^{-1} rat tail collagen I for 1.5 h at 37 ℃.

细胞培养活性测定。为了检测自体磷酸化，将携带血凝素标签的人 cdr1b 基因克隆到 pCMVTet-On 载体中，并用 U2OS 中建立的稳定的诱导细胞系进行 IC50 检测。在 DDR1 被鼠尾胶原 i(Sigma11179179001)激活前 48 小时诱导 DDR1 的表达。细胞用胰蛋白酶消化分离，转移到 15 毫升试管中。经 0.5h 预处理后，用 10gml-1 鼠尾胶原 i 在 37c 条件下处理细胞 1.5h。

Cell-culture fibrosis assay. MRC-5 or human hepatic LX-2 cells were grown in reduced serum medium and treated with compounds for 30 minutes. Subsequently, the cells were stimulated with 10 ng ml^{-1} or 4 ng ml^{-1} TGF-β (R&D Systems, 240-B-002) for 48 or 72 h. The cells were lysed in radioimmunoprecipitation assay buffer and cell lysate of each sample was loaded onto a Wes automated western blot system (ProteinSimple, a Bio-Techne brand).

细胞培养纤维化试验。Mrc-5 或人肝 LX-2 细胞在减少的血清培养基中生长，并用化合物处理 30 分钟。随后，用 10ngml-1 或 4ngml-1 的 tgf-(r&dSystems，240-B-002)刺激细胞 48 或 72h。这些细胞在放射免疫沉淀分析缓冲液中溶解，每个样本的细胞溶解物被装载到韦斯自动免疫印迹系统(Bio-Techne 牌蛋白质简单)上。

Cytochrome inhibition. Water used in the assay and analysis was purified by ELGA Lab purification systems. Potassium phosphate buffer (PB, concentration of 100 mM) and $MgCl_2$ (concentration of 33 mM) were used. Test compounds (compound 1 and compound 2) and standard inhibitors (α-naphthoflavone, sulfaphenazole, (+)-*N*-3-benzylnirvanol, quinidine, and ketoconazole) working solutions (100×) were prepared. Microsomes were taken out of a freezer (–80 ℃) to thaw on ice, labeled with the date, and returned to the freezer immediately after use. Next, 20 μl of the substrate solutions was added to corresponding wells, 20 μl PB was added to blank wells, and 2 μl of the test compounds and positive control working solution was added to corresponding wells. We then prepared a working solution of human liver microsomes (HLM), and 158 μl of the HLM working solution was added to all wells of the incubation plate. The plate was prewarmed for approximately 10 minutes in a water bath at 37 ℃. Then, reduced nicotinamide adenine dinucleotide phosphate (NADPH) cofactor solution was prepared and

细胞色素抑制。分析中使用的水由 ELGA 实验室净化系统进行纯化。磷酸二氢钾缓冲液(PB，浓度 100mm)和 MgCl2(浓度 33mm)。制备了试验化合物(化合物 1 和化合物 2)和标准抑制剂(-萘黄酮、磺胺苯唑、(+)-n-3-苄基硝钒醇、奎尼丁和酮康唑)工作溶液(100)。微粒体从冰箱(-80 ℃)取出，在冰上解冻，贴上日期标签，使用后立即返回冰箱。其次，在相应的井中加入 20l 的基质溶液，在空白井中加入 20l 的 PB，在相应的井中加入 2l 的试验化合物和阳性对照工作液。然后制备人肝微粒体工作液(HLM)，并将 158l 的 HLM 工作液加入培养板的所有孔中。盘子在 37 摄氏度的水浴中预热了大约 10 分钟。制备了还原性烟酰胺腺嘌呤二核苷酸磷酸辅因子溶液，并对其进行了红外光谱、紫外-可见吸收光谱、紫外-可见吸收光谱和紫外-可见吸收光谱的测定

20 μl NADPH cofactor was added to all incubation wells. The solution was mixed and incubated for 10 minutes in a water bath at 37 ℃. At this point, the reaction was terminated by adding 400 μl cold stop solution (200 ng ml^{-1} tolbutamide and 200 ng ml^{-1} labetalol in acetonitrile (ACN)). The samples were centrifuged at

所有培养井均添加 20lNADPH 辅因子。溶液混合后在 37℃的水浴中培养 10 分钟。此时，在乙腈(ACN)中加入 400l 冷停止溶液(200ngml一1 甲苯磺丁脲和 200ngml一1 拉贝洛尔)，反应终止。样本经离心分离后，于

4,000 r.p.m. for 20 minutes to precipitate protein. Then, 200 µl supernatant was transferred to 100 µl HPLC water and shaken for 10 minutes. XLfit was used to plot the per cent of vehicle control versus the test compound concentrations, and for nonlinear regression analysis of the data. IC_{50} values were determined using three-or four-parameter logistic equation. IC_{50} values were reported as >50 µM when per cent inhibition at the highest concentration (50 µM) was less than 50%.

4000 转/分 20 分钟沉淀蛋白质。然后，将 200l 的上清液转移到 100l 的高效液相色谱水中，摇晃 10min。使用 XLfit 绘制了车辆控制百分比与测试化合物浓度的关系图，并对数据进行了非线性回归分析。采用三参数或四参数 logistic 方程确定 IC50 值。当最高浓度(50m)的抑菌率小于 50%时，IC50 值为 50m。

Microsomal stability. The microsomal stability of compound 2 was assessed as follows: working solutions of compound 2 and control compounds (testosterone, diclofenac, and propafenone) were prepared. The appropriate amount of NADPH powder (β-nicotinamide adenine dinucleotide phosphate reduced form, tetrasodium salt, NADPH 4Na, catalog no. 00616; Chem-Impex International) was weighed and diluted into $MgCl_2$ (10 mM) solution (working solution concentration, 10 units ml_{-1}; final concentration in reaction system, 1 unit ml_{-1}). The appropriate concentration of microsome working solutions (human: HLM, catalog no. 452117, Corning; SD rat: RLM, catalog no. R1000, Xenotech; CD-1 mouse: MLM, catalog no. M1000, Xenotech; Beagle dog: DLM, catalog no. D1000, Xenotech) was prepared with 100 mM PB. Cold ACN, including 100 ng ml_{-1} tolbutamide and 100 ng ml_{-1} labetalol as internal standard (IS), was used for

微粒体稳定性。化合物 2 的微粒体稳定性评价如下:制备了化合物 2 的工作液和对照化合物(睾酮、双氯芬酸和普罗帕酮)。将适量的 NADPH 粉末(-烟酰胺腺嘌呤二核苷酸磷酸还原形，四钠盐，NADPH4Na，编号 00616;Chem-ImpexInternational)称量，稀释成 MgCl2(10mm)溶液(工作液浓度 10unitml-1，反应体系终浓度 1unitml-1)。微粒体工作溶液的适当浓度(人类:HLM，目录编号 452117，康宁;SD 大鼠:RLM，目录编号。R1000，Xenotech;CD-1 鼠标:MLM，目录号。1000，Xenotech;Beagledog:DLM，catalogno.

以 100mmPB 为原料制备了 D1000，Xenotech)。以 100ng ml-1 甲苯磺丁脲和 100ng ml-1 拉贝洛尔为内标物，采用冷交联 ACN 法测定了 100ng ml-1 甲苯磺丁脲的含量

the stop solution. Compound or control working solution (10 µl per well) was added to all plates (T0, T5, T10, T20, T30, T60, and NCF60), except the matrix blank. Dispensed microsome solution (80 µl per well) was added to every plate by Apricot and the mixture of microsome solution and compound was incubated at 37 ℃ for approximately 10 minutes. After prewarming, dispensed NADPH regenerating system (10 µl per well) was added to every plate by Apricot to start a reaction. The solution was then incubated at 37 °C. Stop solution (300 µl per well, 4 °C) was then added to terminate the reaction. The sampling plates were shaken for approximately 10 minutes. The samples were centrifuged at 4,000 r.p.m. for

停止解决方案。除基体空白外，所有钢板(T0、T5、T10、T20、T30、T60、NCF60)均加入复合或对照工作液(每井 10l)。微粒体溶液(每孔 80l)加入杏仁中，微粒体溶液与化合物混合液在 37℃下培养约 10 分钟。杏经预热后，加入辅酶 NADPH 再生系统(每井 10l)，使之发生反应。然后溶液在 37℃温度下培养。然后加入停止溶液(每井 300l，4c)终止反应。取样板摇动约 10 分钟。样本以每分钟 4,000 转的速度离心分离

20 minutes at 4 ℃. While centrifuging, new 8 × 96-well plates were loaded with 300 µl HPLC water, and then 100 µl supernatant was transferred and mixed for liquid chromatography–tandem mass spectrometry (LC/MS/MS).

4 摄氏度 20 分钟。在离心过程中，用 300l 高效液相色谱水装载 8 个 96 孔板，然后转移 100l 上清液混合用于液相色谱-串联质谱法(lc/ms/ms)。

Buffer stability. The stability of compound 2 was assessed in phosphate buffer (pH 7.0 and 7.4). Test compounds (at 10 µM) were incubated at 25 °C with 50 mM phosphate buffer (pH 7.4), 8 mM MOPS (pH 7.0), and 0.2 mM EDTA (pH 7.0). Duplicate samples were used. Time samples (0, 120, 240, 360, and 1,440 minutes)

缓冲区稳定性。考察了化合物 2 在 ph7.0 和 7.4 的磷酸盐缓冲液中的稳定性。试验化合物(10m)在 25c 下，加入 50mm 磷酸盐缓冲液(ph7.4)、8mm 磷酸盐缓冲液(ph7.0)和 0.2mmEDTA(ph7.0)。我们使用了相同的样本。时间样本(0、120、240、360 和 1,440 分钟)

were removed and immediately mixed with cold 50% aqueous ACN solution containing IS. Curcumin was used as positive control in this assay at neutral–basic condition. The samples were analyzed by LC/MS/MS, and the disappearance of the test compound was assessed on the basis of peak area ratios of the analyte and IS (no standard curve).

去除后立即与含 IS 的 50%冷水 ACN 溶液混合。在中性-碱性条件下，以姜黄素作为阳性对照。样品用 lc/ms/ms 进行分析，并根据分析物和 IS 的峰面积比值(无标准曲线)评价试验化合物的消失情况。

Pharmacokinetic studies. A study was permitted by the Institutional Animal Care and Use Committee, Shanghai Site (IACUC-SH, WuXi AppTec (Shanghai) Co., Ltd.). The pharmacokinetic profiling of compound 1 was performed on male C57BL/6 mice (7–9 weeks old). Then, we performed i.v. (10 mg kg^{-1}) and p.o. (15 mg kg^{-1}) administration of compound 1. Each group consisted of three mice. *N*-Methyl-2-pyrrolidone:polyethylene glycol 400:H_2O = 1:7:2 solution was used as a vehicle at 5 and 3 ml kg^{-1} for i.v. and PO, respectively. All blood samples (approximately 25 µl blood per time point) were transferred into prechilled commercial K2-EDTA tubes, and then placed on wet ice. The blood samples were immediately processed for plasma by centrifugation at approximately 4 ℃, 3,200*g* for 10 minutes. The plasma was transferred into one prelabeled polypropylene microcentrifuge tube, quick frozen over dry ice, and kept at –70 ±10 ℃ until LC/ MS/MS analysis. Plasma concentration versus time data was analyzed by non-compartmental approaches using the Phoenix WinNonlin 6,3 software program.

药代动力学研究。这项研究得到了上海市动物保护和使用机构委员会(IACUC-SH，无锡市阿普泰克(上海)有限公司)的批准。用 c57bl/6 雄性小鼠(7-9 周龄)进行药代动力学研究。然后静脉滴注(10mgkg-1)和缓释(15mgkg-1)复方 1。每组由三只老鼠组成。以 n-甲基-2-吡咯烷酮:聚乙二醇 400:h2o1:7:2 溶液为载体，分别以 5 和 3mlkg-1 静脉滴注和 PO 滴注。所有的血液样本(大约每个时间点 25 升血液)都被转移到预冷的商用 K2-EDTA 试管中，然后置于湿冰上。血液样本立即处理血浆通过离心大约 4c，3,200 克 10 分钟。等离子体转移到一个预标记的聚丙烯微量离心管，在干冰上快速冷冻，并保持在-7010 ℃直到 LC/ms/ms 分析。血浆浓度随时间的变化数据采用非分室方法，使用 PhoenixWinNonlin6.3 软件进行分析。

Statistics and reproducibility. The sample sizes can be found in the figures and tables or corresponding legends. For microsomal stability experiments, R_2 values were calculated. The number of samples for each experiment can be found in the footnote to Supplementary Table 4. All western blot experiments were performed at least twice with similar results.

统计学和可重复性。样本大小可以在图表或相应的图例中找到。微粒体稳定性实验，计算 R2 值。每个实验的样本数可以在补充表 4 的脚注中找到。所有的免疫印迹实验至少进行两次，结果相似。

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

报告摘要。关于研究设计的更多信息可以在本文链接的自然研究报告摘要中找到。

Data availability

数据可用性

All data are available in the main text or the supplementary materials.

所有数据可在正文或补充材料中获得。

Code availability

代码可用性

The code for the GENTRL model is available at http://github.com/insilicomedicine/gentrl and in Supplementary Code.

Gentrl 模型的代码可在“补充代码”和“补充代码”中找到。

References

参考资料

\20.\Irwin, J. J. et al. *J. Chem. Inf. Model.* **52**, 1757–1768 (2012).

20.埃尔文，j.j.等。北京杰凯化工技术有限公司。Inf.模型。52,1757-1768(2012).

\21.\Oseledets, I. V. *SIAM J. Sci. Comput.* **33**, 2295–2317 (2011).

21.奥塞莱德茨，I.v.SIAMj.Sci。计算机。33,2295-2317(2011).

\22.\Williams, R. J. *Mach. Learn.* **8**, 229–256 (1992).

22.威廉姆斯，R.j.马赫。学习。8,229-256(1992).

\23.\Brown, N. et al. *J. Chem. Inf. Model.* **59**, 1096–1108 (2018).

23.布朗等人。北京杰凯化工技术有限公司。Inf.模型。59,1096-1108(2018).

\24.\Guimaraes, G. L. et al. Objective-Reinforced Generative Adversarial Networks (ORGAN) for sequence generation models. Preprint at https://arxiv.org/ abs/1705.10843 (2017).

24.1997 年，在中国南方科学院研究中心的基础上，通过对吉马良斯(Guimaraes)、g。用于序列生成模型的客观强化生成对抗网络(ORGAN)。预印在(2017)。

\25.\Sanchez-Lengeling, B. et al. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). Preprint at https://chemrxiv.org/articles/ORGANIC_1_pdf/5309668 (2017).

25.3.桑切斯-伦格林等人。分子空间上分布的优化。客观强化的逆设计化学(ORGANIC)生成对抗性网络。预印在(2017)。

\26.\Ritter, H. & Kohonen, T. *Biol. Cybern.* **61**, 241–254 (1989).

26. Ritter，h. & Kohonen，t. Biol. Cybern. 61,241-254(1989).

\27.\Sammon, J. W. *IEEE Trans. Comput.* **C-18**, 401–409 (1969).

27.Sammon，j.w.IEEETrans.计算机。C-18,401-409(1969).

\28.\Rappe, A. K. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).

28.Rappe，a.k.j.A..化学。Soc.114,10024-10035(1992).

自然-生物技术/盖蒂图片社/盖蒂图片社

nature research

Corresponding author(s): Alex Zhavoronkov
通讯作者:AlexZhavoronkov

Last updated by author(s): Jul 10, 2019
最后更新作者:2019 年 7 月 10 日

Reporting Summary
报告摘要

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.
《自然研究》希望提高我们发表的作品的可重复性。此表格为报告的一致性和透明度提供了结构。有关自然研究政策的进一步信息，请参阅作者和审稿人和编辑政策检查表。

Statistics
统计数字

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.
对于所有的统计分析，确认图例、表格图例、正文或方法部分中存在以下项目。

n/a | Confirmed
不确定

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
每个实验组/条件的精确样本容量(n)，以离散数和测量单位表示

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
关于是否从不同的样品进行测量或是否重复测量同一样品的说明

☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
使用的统计检验以及它们是单面还是双面的
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
只有常见的测试应该只用名称来描述;在方法部分描述更复杂的技术。

☒ ☐ A description of all covariates tested
所有被测协变量的描述

☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
对任何假设或更正的描述，如正态性检验和多重比较的调整

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient)
统计参数的详细说明，包括中央趋势(例如平均数)或其他基本估计(例如回归系数)
AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
与变异(例如标准差)或相关的不确定性估计(例如置信区间)

☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
对于零假设检验，检验统计量(例如 f，t，r)包含置信区间、效应大小、自由度和 p 值
Give P values as exact values whenever suitable.
只要合适，就给出准确的 p 值。

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
对于贝叶斯分析，信息的选择先行者和马尔科夫蒙特卡洛/设置

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
对于分级和复杂的设计，确定适当的测试水平和完整的结果报告

☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
效应大小的估计值(例如 Cohen 的 d，Pearson 的 r)，说明它们是如何计算的

Our web collection on statistics for biologists contains articles on many of the points above.
我们为生物学家收集的统计数据包含了上面许多观点的文章。

Software and code
软件和代码

Policy information about availability of computer code
关于计算机代码可用性的

策略信息

Data collection	The code for the GENerative adversarial Tensorial Reinforcement Learning (GENTRL) system is provided with the manuscript. It will be freely available at GitHub upon the acceptance of the manuscript.
Data analysis	Maestro Release 2018-3; Phoenix WinNonlin 6.3; custom python scripts using PyTorch 0.4.1 and RDKit 2018.03.4 libraries

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers.
对于使用自定义算法或软件的手稿，这些算法或软件是研究的中心，但尚未在已发表的文献中描述，软件必须提供给编辑/审稿人。

We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
我们强烈鼓励将代码存储在社区存储库中(例如 GitHub)。请参阅《自然研究》提交代码和软件的指导方针以获得更多信息。

Data
数据

Policy information about availability of data
关于数据可用性的策略信息

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
所有手稿必须包括一个数据可用性声明。本说明应酌情提供下列资料:

- Accession codes, unique identifiers, or web links for publicly available datasets
公开可用数据集的加入代码、唯一标识符或网络链接
- A list of figures that have associated raw data
与原始数据相关联的数字列表
- A description of any restrictions on data availability
对数据可用性的任何限制的说明

The 30,000 structures generated by GENTRL for the DDR1 kinase are available in supplementary materials.
由 GENTRL 为 DDR1 激酶产生的 30,000 个结构可以在辅助材料中获得。

Field-specific reporting
特定领域报告

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.
请选择下面最适合你研究的那一个。如果你不确定，在选择之前阅读适当的章节。

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

研究没有涉及野生动物
动物抵达无锡后，由兽医人员或其他授权人员对其进行一般健康状况评估。动物在被安置研究之前至少需要 3 天的时间来适应环境(在到达无锡之后)。
小鼠，c57bl/6，雄性，7-9 周

Life sciences study design
生命科学研究设计

All studies must disclose on these points even when the disclosure is negative.
所有的研究必须披露这些点，即使披露是负面的。

Sample size	No applicable. 样本容量不适用。
Data exclusions	Data outside 3xSD were excluded.
Replication	Results are verified by independent biological experiments.
Randomization	Not applicable.
Blinding	Not applicable.

Reporting for specific materials, systems and methods
具体材料、系统和方法的报告

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
我们需要从作者那里获得关于许多研究中使用的某些类型的材料、实验系统和方法的信息。在这里，指出是否每个材料，系统或方法列出是相关的研究。如果你不确定一个列表项目是否适用于你的研究，在选择回复之前阅读适当的部分。

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms 动物和其他生物
☒	☐ Human research participants 人类研究参与者
☒	☐ Clinical data 临床资料

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies
抗体

Antibodies used

Mouse anti- human COL1A1 antibody (3G3) (Santa Cruz, sc-293182); mouse anti-human CTGF antibody (E-5) (Santa Cruz, sc-365970); mouse anti human α-smooth muscle actin antibody (SPM332) (Santa Cruz, sc-56499); mouse anti-human GLYCERALDEHYDE-3-PDH, (Merck-Millipore, Merck-MAB374); rabbit anti-Phospho-DDR1 (Tyr513) (E1N8F) (Cell Signaling,

Mouse, C57BL/6, male, 7- 9 weeks

Study did not involve wild animals

Following arrival at WuXi AppTec animals were assessed as to their general health by a member of the veterinary staff or other authorized personnel. Animals were acclimated for at least 3 days (upon arrival at WuXi AppTec) before being placed on study.

ukaryotic cell lines
真核细胞系

Policy information about cell lines

Cell line source(s) 有关单元格线源的策略信息	MRC-5 was from ATCC (CCL-171); LX-2 was from Merck Millipore (SCC064); U-2 OS-DDR1 stable cell line was made at RSD Biology Department, WuXi Apptec, Inc. from U-2 OS parent cell line from ATCC® (HTB96™); Mrc-5 来源于 ATCC(CCL-171)，LX-2 来源于默克公司(SCC064)，U-2OS-DDR1 来源于 ATCC(HTB96TM)的 U-2OS 亲本细胞系;
Authentication 认证	The cells were authenticated by STR. 细胞经 STR 鉴定。
Mycoplasma contamination	All cell lines used have been tested free of mycoplasma contamination. 所有使用的细胞系已经检测没有支原体污染。
Commonly misidentified lines 支原体污染通常被错误识别的品系 (See ICLAC register) (见 ICLAC 登记册)	None. 没有。

nimals and other organisms
动物和其他生物

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research
有关动物研究的政策信息;关于动物研究报告的建议

Laboratory animals 实验室动物	
Wild animals 野生动物	
Field-collected samples 实地收集的样本	

	Animals were group housed during acclimation and individually housed during the study. The animal room environment ws controlled (target conditions: temperature 18 to 26°C, relative humidity 30 to 70%, 12 hours artificial light and 12 hours dark). Temperature and relative humidity was monitored daily. The animals were overnight fasted. They had access to Certified Rodent Diet ad libitum 4 hr post dose. The lot number and specifications of each lot used were archived at WuXi AppTec. Water was autoclaved before provided to the animals ad libitum. Periodic analyses of the water was performed and the results archived at WuXi AppTec. There are no known contaminants in the diet or water that, at the levels of detection, is expected to interfere with the purpose, conduct or outcome of the study.
Ethics oversight	Study was permitted by IACUC (Institutional Animal Care and Use Committee).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

注意，关于批准研究方案的完整信息也必须在手稿中提供。