arXiv:2611.29224v1 [cs.LG] 22 Nov 2026

星系属性估计中的曲率失配与局部误差

Austin Lutterbach
EleutherAI SOAR
UniverseTBD
rmhslutterbach@gmail.com

Angus Hoy EleutherAI SOAR
UniverseTBD
angusjwh@gmail.com

Jasper Chong昆士兰大学 EleutherAI SOAR
UniverseTBD
jasper.chong@uq.edu.au

Aman Kumar
UniverseTBD EleutherAI
IUCAA
kumaramanasci@gmail.com

Michael J. Smith AstroAI 哈佛-史密松天体物理中心 UniverseTBD EleutherAI
mike@mjjsmith.com

摘要

线性探针广泛用于评估学习到的表示，并从无标注数据中推断物理量。探针的预测来自环境空间中的一个仿射函数，该函数受限于底层表征；因此，即使某种表征包含预测目标所需的全部信息，如果这些信息以线性探针无法访问的形式编码，也可能产生系统性探针误差。在此，我们指出，这类探针误差有一部分源于流形固有的几何性质。在线性探针被应用于弯曲流形时，通常会产生非零的内在 Hessian（我们通俗地称之为“面向探针的曲率”），其表达式为 $\langle w_N, \mathrm{II} \rangle$，即流形的第二基本形式与探针的法向分量的点积。我们将线性探针拟合到来自五种预训练编码器的星系嵌入，并对这一曲率项进行干预，以评估其对探针性能的贡献。我们发现，当继承曲率与标签曲率相匹配时，保留它能够改善局部拟合。若探针的曲率分量与标签曲率不匹配，移除该分量同样能带来改进；在每个邻域中对其进行校正，可使星系数据上的留出局部 R^2 提升 0.006 至 0.010，且两种曲率差异越大，提升越明显。我们证实，这些收益源于底层几何，而非最小二乘训练必然带来的结果。

1 引言

线性探针通过衡量能否从冻结嵌入中准确恢复目标，来评估学习到的表示 [Alain 和 Bengio, 2016]。在物理科学中，局部可靠性至关重要，因为探针要在没有直接测量的情况下推断物理量。对于星系图像，线性探针可预测星表量，例如星等、红移和恒星质量 [Smith 等人, 2024], 但仅凭整体评分，既无法刻画局部预测误差，也无法解释其几何来源 [Eyuboglu 等人, 2022]。我们表明，被评估表示所继承的曲率会影响线性探针的性能，并捕获预测误差的一个局部来源。

我们贡献了：（i）实现并通过合成数据验证了基于解码器的第二基本形式估计，用于局部探针分析；（ii）将面向探针的曲率 $\mathrm{Hess}_M(w \cdot x) =$

第九届机器学习与物理科学研讨会（ML4PS 2026）。

$\langle w_N$，II$\rangle$ 应用于拟合探针，并通过残差展开识别其二阶近似误差；（iii）来自 86,471 个星系和五种编码器的证据表明，在“幸运”的几何对齐情形下，嵌入曲率能够提升探针性能，而在发生错位时则会产生阻碍；（iv）提出一种弯曲校正，使探针权重能够在各个邻域中顺应流形的弯曲，从而在这种弯曲与标签未对齐的区域提高留出集局部 R^2。代码、记录和笔记本见 https://github.com/alutterb/galaxy-curvature。

2 解码器的逐点曲率

数据。 来自 AstroPT 星系数据集的 86,471 个星系 [Smith 等人， 2024] 及其目录标签（r波段星等、光度红移、平滑成分比例、恒星质量），由一个在 ImageNet-21k 上监督训练的 ViT-B/16（下文简称“ViT-B”）和另外四个编码器（附录 A）生成嵌入；Platonic Universe 发布版本中的嵌入 [UniverseTBD 等人， 2025]。我们对嵌入进行单位归一化（数据位于 S^{D-1}； $D = 768$适用于 ViT-B）。

方法。 令 F： $\mathbb{R}^d \to \mathbb{R}^D$ 为普通自编码器的解码器（3 个宽度为 250 的隐藏层，使用 SiLU），在 80% 的样本行上拟合；我们在留出样本行上评估曲率。我们通过自动微分计算 $J = DF(z)$、度量 $g = J^\top J$、法向投影算子 $P_N = I - Jg^{-1}J^\top$ 和第二基本形式 II $= P_N D^2 F(z)$；参见 Lee 和 Park, 2023，Acosta 等人， 2023]，计算点满足 DF 具有满列秩 d。我们对投影后的解码器 $F/\|F\|$求导，其像位于单位球面上，因此 II 精确分解为球面的普适径向部分和球内形状部分（式（2））。

验证。 在具有闭式第二基本形式（附录 A）的合成球内曲面上，当 $D = 768$ 解码器估计的 II 的真实球内迹 H^S 在方向（余弦相似度 0.999 ）、星等（比率 1.00 ）和各点的秩（ 0.94 ）方面均与之吻合； K_S 完整张量的余弦中位数为 0.96（附录 A）。

3 设置与符号

We a设流形假设 [Fefferman 等人， 2016], 成立，即嵌入位于某个光滑
流形附近。

符号。 本文其余部分使用的符号：

• 流形。 $M \subset S^{D-1} \subset \mathbb{R}^D$ 是一个光滑的 d维**流形** ($d < D$)，具有**锚点** x_0、正交归一**切向标架** J u x u $= x_0 + Ju +$和**法向坐标** （图表），$\frac{1}{2}$ u u $+ O$ $\|u\|^3$ [2018],
因此（ ）II(,)（ ）Lee，其中 II 是取值于法向的**第二基本形式**。在球体上，II $=$ II$^S - g\,x$，其中球内**形状部分** II$^S \perp x$，且**迹** H^S。数值上，$J = DF$、 $g = J^\top J$（$g = I$在正交归一标架中）以及 $u = g^{-1}J^\top$（$x - x_0$）。

• 探针。 $\hat{}y = w \cdot x + b_0$，带有 **岭惩罚** α ($\alpha = 100$ **预注册**， α^* **调优的**）； $w = w_T + w_N$(**切向**， **法向**）以及 $w_N = w_S + w_{\mathrm{rad}}$，其中 $w_S \perp x$ 和 $w_{\mathrm{rad}} = (w \cdot x)\,x$。
• 曲率。 H $=$Hess$_M\,y$对于光滑目标 y； K $=$Hess$_M\hat{}y$ $= \langle w_N$、II$\rangle$，即探针曲率（公式（1）），可按K $= K_S + K$sph（公式（2））分解为形状项 K_S $=\langle w_S$、II$^S\rangle$，由 M在球体内部的弯曲方式决定；以及球面项 Ksph $= -(\hat{}y - b_0)\,g$，球体上的每个探针都具有该项，且它仅取决于预测值；失配 $\Delta = H - K$； $R = H - K$sph，即 H需要复现的部分：在 Ksph固定的情况下， $\Delta = R - K_S$；以及增益 $\Gamma = 2\langle H\,K\rangle_g - \|K\|_g^2 = \|H\|_g^2 - \|\Delta\|_g^2$，当探针的弯曲降低局部误差的曲率部分时为正。• 内积。对于对称二阶张量 $\langle A\,B\rangle_g$ $= g^{ia}g^{jb}A_{ij}B_{ab}$和 $\|A\|_g^2$ $= \langle AA\rangle_g$，这是度量对应的弗罗贝尼乌斯内积版本。

• 局部区域与干预。在**局部区域** $u \sim N(0,\ s^2 I_d)$上， $\bar{c}$是 $y - \hat{}y$与 $\delta = (dy - d\hat{}y)_J$的平均值，即**斜率失配**。$\hat{}y_t$ $= (w_T + w_{\text{弧度}})$ $\cdot\, x + \frac{t}{2}K_S(u,\ u)$ $+ c$按 t进行缩放，并重新拟合 c；在邻域点上， q_c是中心化后的 $\frac{1}{2}K_S(u_i,\ u_i)$， r_0是 $t = 0$处中心化后的**残差**，而 t^* $=\langle r_0,\ q_c\rangle/\|q_c\|^2$。

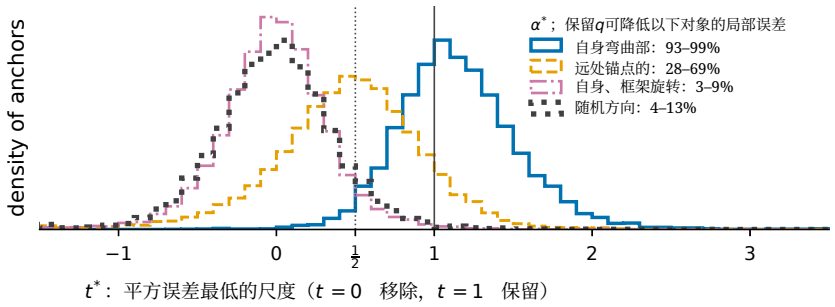

图1：解码器在每个星系处测得的曲率，就是拟合探针所使用的曲率。固定探针的权重后，我们用其切向读出和径向读出以及 $t\,q$，对每个锚点的 2,048 个邻域点进行评分，并针对四种大小相等的二次型 q 进行评估。 t^* 是这些邻域点上平方误差最低的尺度（局部 R^2最高），且仅当$t^* > \frac{1}{2}$（竖直点线）时，保留 q 比移除它拟合得更好。自身弯曲 $\frac{1}{2}\langle w_S,\ \mathrm{II}^S\rangle$在 $t^* = 1$处达到峰值：无标签解码器给出的弯曲部，恰好处于探针所使用的尺度。远处锚点的 II^S 在 $_1{}_2$附近达到峰值：曲率具有局部性。将自身形式的切向标架旋转后（主曲率不变），其峰值出现在 0：方向至关重要。灰色表示随机法向方向，供参考。在 $d = 16$处调优的探针、五个编码器、四个标签。

4 弯曲流形上的线性探针

通过这一设定，我们可以确定三点。首先，尽管探针在环境空间中是仿射的，但它限制在 M 上时，通常会在内禀坐标中呈非线性：

$$(d\hat{y})_J = J^\top w, \qquad \mathrm{Hess}_M\,\hat{y} = \langle w_N, \mathrm{II}\rangle =: K, \tag{1}$$

其中， K 是面向探针的曲率，这是一个经典恒等式（ [Milnor，1963年；Chern 和 Lashof， 1957]）。其次，在单位球面上，它可以分解为球内形状项和径向项：

$$K = K_S + K_{\mathrm{sph}}, \qquad K_S = \langle w_S, \mathrm{II}^S\rangle, \qquad K_{\mathrm{sph}} = -(\hat{y} - b_0)g. \tag{2}$$

因此，探针从表征的几何结构中继承了内在曲率。

第三，在尺度为 s 的局部区域上，对于足够平滑的嵌入以及高阶导数受控的目标函数（附录 A），

$$\mathbb{E}_{\mathrm{patch}}[(y - \hat{y})^2] = \bar{c}^{\,2} + s^2\|\delta\|^2 + \tfrac{1}{2}s^4\|\Delta\|_g^2 + O(s^4\|\delta\|) + O(s^6). \tag{3}$$

因此，在该模型下，海森失配决定了二次残差对中心化局部误差的贡献。相对于零海森基线 $K = 0$，当且仅当 $\Gamma > 0$ 时，探针会降低二次项对局部误差的贡献。因此， K 必须朝目标的曲率方向弯曲，也就是说，比 $K = 0$ 更接近 H。在固定 K_{sph} 的情况下， K_S 的相应条件为 $2\langle R,\ K_S\rangle_g > \|K_S\|_g^2$。第5节将检验这一关系。由于二次项的贡献按 s^4 缩放，我们的实证分析控制了局部邻域尺度的变化。

我们研究曲率项 $\frac{1}{2}s^4\|\Delta\|_g^2$，因为它是唯一能反映完整张量 K 的项： δ 是一阶导数失配，而弯曲部只能通过 Δ的迹作用于 $\bar{c}$。

5 实验

我们拟合五折样本外岭探针 [Alain 和 Bengio， 2016] 在预注册的 $\alpha = 100$处建立从嵌入到标签的映射，在 512 个锚点各自的 2,048 个邻域内对局部 R^2 评分，并使用 Freedman–Lane$p$的 [值 1983], ，计算每种曲率量与局部 R^2 之间的秩偏 Spearman 相关（Freedman 和 Lane， 1983], ），控制对数补丁半径、局部标签方差和计数；更严格时，还控制五种尺度下的对数半径。我们还报告调优的 α^* （附录 A）。我们使用百分比范围（x%–y%）表示编码器与标签配对中观测到的最小值和最大值。

探针曲率与标签曲率的失配。 在各个锚点处，较大的失配 $\|\Delta\|_g$ 与较低的局部 R^2 星等和红移相关，但其中大部分是对 H估计的噪声；去除这部分噪声后，只有红移仍保持这种关联，而且仅在 $\alpha = 100$ 时如此（附录 B）。

保留、去除或校正探针的弯曲。 接下来，我们考察探针的弯曲对其局部拟合有何影响。在固定流形和（$w_T + w_{\mathrm{rad}}$）$\cdot x$ 的情况下，我们使用 $\hat{}y_t$ 为每个锚点的 2,048 个邻域点评分（第3节）。因此，$t = 1$ 保留探针从流形继承的 K_S，而$t = 0$将其去除。

由于 c经过重新拟合，局部误差在 t中呈抛物线形式：SSE(t) $= \|r_0 - t\, q_c\|^2$，在 t^*处达到最小值，并且仅当 $2\langle r_0,\ q_c\rangle \le \|q_c\|^2$时，$t = 0$ 优于 $t = 1$（即进行干预）。在不同编码器和标签中，$t = 1$ 在 79–99% 的锚点上优于 $t = 0$；使用调优探针后，该范围提升至 93–100%。
随后，我们考察对曲率进行干预是否能改善流形与标签曲率之间发生错位的位置。由于无法直接比较 II 和 H，我们将 K_S与 $R = H - K_{\mathrm{sph}}$ 进行比较，以去除球面部分，然后计算K_S 与 R之间的余弦相似度。在调优的 α^*下，当存在错位而非对齐时，干预 $t = 0$ 相较于 $t = 1$ 更常带来改善。在流形中错位程度最高的五分位组区域里，干预平均有 5.81% 的概率改善探针误差，而在对齐程度最高的五分位组区域中，这一比例为 2.03%。（完整数值范围请参见附录 C。）

调优探针仍可受益于一种随邻域点变化的权重。在每个邻域中，我们使用一半的邻域点拟合弯曲校正：进行小型岭拟合，使调优探针的权重沿解码器的 136 个弯曲方向移动（来自 II^S；附录D），同时固定斜率；然后在另一半邻域点上对其评分。在 ViT-B 上，留出的局部 R^2 从 0.0059 上升至 0.0095。其中不到一半的增益来自曲率，而当探针的弯曲与标签的曲率对齐程度较低时，这部分增益更大（四个标签中有三个的秩偏 Spearman -0.17 至 -0.28）。

为将几何因素与最小二乘区分开来（图1），我们保留拟合权重 w_S，仅改变它所读取的弯曲，并将其重新缩放至相同幅度 $\|q_c\|$。在拟合完全相同的情况下，最小二乘没有理由偏好某一种弯曲。首先，我们从一个遥远的锚点取出 II^S，该锚点的邻域与当前邻域不共享任何星系：在 82–99% 的锚点中，该锚点自身的弯曲优于它。其次，我们保留该锚点自身的形态，但旋转其切向方向，使曲面沿错误方向以相同幅度弯曲：这样在 17–34% 的锚点中会有所帮助（在 α^*处为 3–9%），其发生频率不高于随机法向方向。因此，增益取决于流形在锚点处如何弯曲以及沿哪些方向弯曲，而不仅仅取决于拟合权重。

应用于物理科学。曲率可以通过邻域定位星表中的薄弱环节。对于星表研究，我们提出三条规则。第一，调整岭惩罚。流形弯曲与标签曲率失配的区域，需要探针权重补偿二者之间的差异，而过强的惩罚项可能会妨碍这种修正。第二，在读取标签的局部曲率之前，先去除噪声底。第三，按邻域而非仅按整体读取准确率。按邻域点的平均 r 波段星等对锚点分组后，星等最暗的三分之一星系在四个标签中的三个标签上，其局部 R^2 准确率比最亮的三分之一低至多 0.22；而这一下降的大部分源自标签本身的噪声：在这些区域，标签噪声占局部方差的比例高出 0.11 到 0.19（附录 B）。因此，暗淡星系上的低准确率主要反映的是星表的问题，而非探针的问题：这意味着需要改进标签，或为这些区域提供更宽的误差条。

6 讨论

弯曲嵌入上的线性探针会继承流形的弯曲；我们已证明，当这种曲率与估计的标签曲率一致时，它可以提升探针的性能，否则则会损害性能。我们还证明，当探针在全局范围内训练时，流形会决定探针在每个局部邻域中的弯曲；当标签的曲率与这种弯曲不同时，就会导致探针产生局部误差。

这些发现带来了三个具体的应用场景：确保岭惩罚经过调优，因为较强的惩罚项会最大程度地压缩探针的弯曲部，使其无法匹配标签的曲率；按邻域而不是全局读取准确率，因为探针对暗星系邻域的拟合效果最差；如果有可用的解码器，则使用弯曲校正在局部细化探针的估计结果。

局限性。 我们研究了一个星系样本、四个目标和五个编码器。估计值$\|H\|_g^2$ 中约有一半属于噪声底（附录 B），而每个锚点的残差曲率无法对单个对象进行秩排序（拆分-一半 $\rho \approx 0.33$）。干预改变的是局部代理模型，而不是编码器。失配和对齐是在预注册平均曲率分析之后定义的，干预属于探索性分析；

置换 p-值（表格中的*）以及干预中各锚点所占的比例，均将重叠锚点视为相互独立。

致谢

我们谨向 Stella Biderman 和 EleutherAI 致谢，感谢他们提供计算资源和基础设施支持。

参考文献

Guillaume Alain 和 Yoshua Bengio．使用线性分类器探针理解中间层，2016年 arXiv:1610.01644。

Michael Arcuschin、Jett Janiak、Robert Krzyzanowski、Senthooran Rajamanoharan、Neel Nanda 和 Arthur Conmy。《思维链推理在开放权重语言模型中并非忠实》，2025年。doi: 10.48550/arXiv.2503.08679。URL https://arxiv.org/abs/2503.08679。

Adrien Bardes、Jean Ponce 和 Yann LeCun。VICReg：用于自监督学习的方差-不变性-协方差正则化，2022年。

Sabri Eyuboglu、Maya Varma、Khaled Saab、Jean-Benoit Delbrouck、Christopher Lee-Messer、Jared Dunnmon、James Zou 和 Christopher Ré。Domino：利用跨模态嵌入发现系统性错误。载于国际学习表征会议，2022年。

UniverseTBD、Trinidad Borrell、Steven Dillmann、Kshitij Duraphe、Furkan Eris、Kartheik Iyer、Ashod Khederlarian、Aman Kumar、Giovanni Marraffini、Michael J. Smith、Shashwat Sourav、Rocco Di Tella 和 John F. Wu。《柏拉图式宇宙：基础模型看到的是同一片天空吗？》载于*NeurIPS 2025*机器学习与物理科学研讨会，2025年。URL https://arxiv.org/abs/2509.19453。

YonghyeonLeeandFrankC. Park. 深度生成模型中的显式曲率正则化。载于 机器学习中的拓扑、代数与几何研讨会（*TAG-ML*）第二届年会论文集，第 221 卷，机器学习研究论文集，第 505–518 页。PMLR，2023年。

Francisco Acosta、Sophia Sanborn、Khanh Dao Duc、Manu Madhav 和 Nina Miolane。量化神经流形中的外在曲率。载于 *2023*年 *IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*），第 610–619 页，2023年。doi: 10.1109/CVPRW59228.2023.00068。

Charles Fefferman、Sanjoy Mitter 和 Hariharan Narayanan。检验流形假设。美国数学会杂志，29(4): 983–1049，2016年。doi：10.1090/jams/852。

John M. Lee.《黎曼流形导论》，第176卷，属于《数学研究生教材》，2018年，第二版。doi：10.1007/978-3-319-91755-9。

约翰·米尔诺和詹姆斯·斯塔希夫，《特征类》，普林斯顿，1953《数学年刊研究》doi: 10.1515/9781400881802。根据 M. Spivak 和 R. Wells 的讲义编写。

陈省身和Lashof，《美国数学杂志》，79(2):306–318，1957，流形的总曲率。doi: 10.2307/2372684。

David Freedman、Robert Pisani 和 Roger Purves，《商业与经济统计》，报告的S.E.系数的非随机解释。doi: 10.1080/07350015.1983.10509354。

Peter Hall 和 J. S. Marron，《统计与概率快报》，6(2):109–115，1987，估计。doi: 10.1016/0167-7152(87)90083-6。

Alexander Walther、Hamed Nili、Naveed Ejaz、Arjen Alink、Nikolaus Kriegeskorte 和 Jörn Diedrichsen。多体素模式分析中差异性度量的可靠性。*NeuroImage*，137:188–200，2016年。doi: 10.1016/j.neuroimage.2015.12.012。

James G. MacKinnon 和 Halbert White，《计量经济学杂志》，29(3):305–325，1985年。doi: 10.1016/ 0304-4076(85)90158-7。

Raghunathan Ramakrishnan、Pavlo O. Dral、Matthias Rupp 和 O. Anatole von Lilienfeld。134,000 个分子的量子化学结构与性质。*Scientific Data*，1:140022，2014年。doi: 10.1038/sdata.2014.22。

Jerret Ross、Brian Belgodere、Vijil Chenthamarakshan、Inkit Padhi、Youssef Mroueh 和 Payel Das。《大规模化学语言表征捕获分子结构和性质》。*Nature Machine Intelligence*，4（12）：1256–1264，2022年。doi: 10.1038/s42256-022-00580-7。

附录 A：ViT-B 及跨编码器的关联结果

邻域与标签海森矩阵。我们从解码器未见过的行中抽取 512 个锚点；每个锚点的 $k = 2,048$ 个邻域点，是根据单位归一化嵌入之间的欧几里得距离（即余弦相似度），从全部 86,471 个星系中选出的最近邻。我们通过在切空间投影坐标 $u = g^{-1}J^{\top}(x - x_0)$ 中，对邻域点的标签拟合最小二乘二次函数，以估计每个锚点处的$H = \mathrm{Hess}_M\, y$。这些坐标在锚点处的克里斯托费尔符号为零，因此在小邻域极限下，二次函数系数可估计协变海森矩阵。

已知曲面上的验证。在 $D = 768$ 球内曲面 $G(z) = Q\mathrm{normalize}([\phi(z); 0.8\,h(z); 0\ldots])$（$\phi$ 逆立体投影 $\mathbb{R}^{16} \to S^{16}$、$h$ 四个高斯隆起、Q 一次旋转）具有闭式曲率。解码器的球内迹 H^S II 的秩与真实 $\|H^S\|$之间的秩为 0.94，方向余弦为 0.999，幅值比为 1.00。在同一曲面上使用六个已知海森矩阵的标签（$d = 16$、$n = 86,471$）时，解码器的 K_S 与真实值之间的张量余弦，在无噪声时的中位数为 0.963–0.965；在各向同性曲面外噪声下为 0.67–0.84，该噪声的大小为斑块半径中位数的四分之一。平均曲率这一通常报告的标量仅保留 II 的迹，因此表现不佳；这是因为它与读出准确率的关联会随着已知曲面上的采样方式以及星系中的潜在维度而改变符号。

惩罚项与编码器。表 1 和表 2 展示了预注册的 $\alpha = 100$ 探针和调优后的 α^*探针；这些探针针对每个编码器和标签，通过嵌套五折交叉验证（一倍标准误规则）选定，之后才在其上计算任何几何量；每个单元格中的 α^* 均为 0.1 或 0.32。表 1 展示了在 $d = 16$下对 ViT-B 的分析，并将其扩展至 Platonic Universe 发布版本中的另外四种编码器：DINOv3 ViT-B/16、CLIP ViT-B、ConvNeXt-B 和 ViT-L（宽度分别为 768、512、1,024 和 1,024；均经过单位归一化），每种编码器都搭配了架构和训练方案相同的解码器（种子 0）。对于星等和红移，在所有五种编码器和两种惩罚项下，失配都与局部 R^2 呈负相关；这种关联的大部分来自标签 Hessian 噪声底（附录 B）。

		ViT-B		DINOv3		CLIP-B		ConvNeXt-B		ViT-L		† 计数	
标签	数量	100	α^*	100	α^*	100	α^*	100	α^*	100	α^*	100	α^*
mag_r	失配	$-0.30^{\dagger}$	-0.24	$-0.58^{\dagger}$	$-0.58^{\dagger}$	-0.29	$-0.34^{\dagger}$	-0.33	-0.16	$-0.45^{\dagger}$	$-0.49^{\dagger}$	3/0	3/0
	对齐	$+0.35$	$+0.34$	$+0.53$	$+0.54$	$+0.32$	$+0.53$	$+0.21$	$+0.31$	$+0.13$	$+0.33$		
	球体	-0.26	-0.23	-0.23	-0.31	-0.40	-0.26	-0.24	-0.34	-0.08^{*}	$+0.17$		
照片_z	失配	$-0.39^{\dagger}$	$-0.45^{\dagger}$	$-0.60^{\dagger}$	$-0.62^{\dagger}$	$-0.42^{\dagger}$	$-0.41^{\dagger}$	$-0.46^{\dagger}$		-0.34	$-0.44^{\dagger}$	$-0.49^{\dagger}$	5/0	4/0
	对齐	$+0.26$	$+0.25$	$+0.33$	$+0.39$	$+0.06^{*}$	$+0.09$	$+0.22$	$+0.50$	$+0.14$	$+0.41$		
	球体	-0.33	$+0.04^{*}$	-0.21	$+0.01^{*}$	-0.29	$+0.21$	-0.46	$+0.03^{*}$	-0.44	$+0.12$		
平滑 _不匹配比例		-0.09^{*}		-0.13	-0.29	$-0.38^{\dagger}$	-0.24	-0.34	-0.19	-0.11	$+0.01^{*}$	-0.20	0/0	1/0
	对齐	$+0.19$	$+0.27$	-0.02^{*}	$+0.22$	$+0.24$	$+0.27$	$+0.04^{*}$	$+0.26$	$+0.14$	$+0.38$		
	球体	-0.13	-0.10	-0.06^{*}	-0.22	$+0.05^{*}$	-0.24	-0.15	-0.19	-0.11	-0.03^{*}		
恒星 _mass	失配	$+0.00^{*}$	$+0.03^{*}$	$+0.04^{*}$		-0.09	-0.12	-0.03^{*}	-0.03^{*}	-0.12	-0.02^{*}	-0.19	0/0	0/0
	对齐	$+0.01^{*}$	$+0.12$	-0.05^{*}	$+0.07^{*}$	-0.01^{*}	-0.06^{*}	$+0.07^{*}$	$+0.33$	-0.03^{*}	$+0.14$		
	球体	-0.35	-0.01^{*}	-0.26	-0.11	-0.38	$+0.06^{*}$	-0.40	$+0.02^{*}$	-0.38	-0.04^{*}		

表 1：在各个编码器上，$d = 16$，512 个锚点：在多尺度密度控制下，与局部 R^2 的秩偏斯皮尔曼相关，包括失配 $\|\Delta\|_g$、对齐 $\langle H,\ K_S\rangle_g/(\|H\|_g\|K_S\|_g)$ 以及球面项的大小 $\|K_{\mathrm{sph}}\|_g = \sqrt{d}\,\hat{}y - b_0|$。每对列先给出预注册 $\alpha = 100$ 时的探针结果，再给出调优的 α^* 时的结果。* 在 0.05 水平上不显著。† 对于失配，聚类自助法的 95% 百分位区间不包含零（根据邻域重叠进行平均连接聚类，将锚点分为 32 个区块，重采样 2,000 次）。最后一对列统计失配偏相关为负的编码器数量（–）和为正的编码器数量（+），其中 †。在计算任何失配之前，我们便将支持某项主张的门槛定为五个中至少三个为负。

弱岭。简单的关联无需欠拟合探针：在 $\alpha = 1$（ViT-B；全局 R^2从 0.51 升至 0.66）下，星等和红移的符号与显著性保持不变（失配 -0.25 至 -0.50，对齐 $+0.26$至 $+0.47$），且调优的 α^* 结果一致（表 1）。

公式（3）的推导（3）。在正交归一标架（$g = I$处的 x_0）中，将沿 M方向的残差 $e = y - \hat{}y$写作 $e(u) = e_0 + \delta^{\top}u + \frac{1}{2}u^{\top}\Delta u + \rho(u)$。二次项系数为 Δ，因为法坐标在 x_0 处的克里斯托费尔符号为零，且 $\mathrm{Hess}_M\,\hat{}y = K$（公式（1））；受控的高阶导数

给出 $\rho(u) = O(\|u\|^3)$。确切地说，$\mathbb{E}[e^2] = \bar{c}^2 + \mathrm{Var}(e)$。对于 $u \sim N(0,\ s^2 I_d)$，线性项与二次项不相关（奇阶矩为零），$\mathrm{Var}(\delta^\top u) = s^2\|\delta\|^2$，且 $\mathrm{Var}(\frac{1}{2}u^\top \Delta u) = \frac{1}{2}s^4\|\Delta\|^2$。$\rho$ 的线性项与三次项的协方差为 $O(s^4\|\delta\|)$，其他每一项均为 $O(s^6)$，由此得到公式（3）。

附录 B：标签 Hessian 噪声底（其他分析）

由于 Δ 是残差 $y - \hat{}y$ 的海森矩阵，按照构造方式，预计它与局部准确率存在一定关联（式 (3)），但其中很大一部分是拟合噪声。探针的 K 几乎没有参与其中：在 $d = 16$处的 40 ˆ 编码器、标签和惩罚单元中，$\|\Delta\|_g$ 的偏导数与标签 Hessian 范数 $\|H\|_g$的偏导数的中位数差异为 0.016，相关系数为 0.988。

ˆ噪声底。标签海森矩阵 H 是关于 n 含噪标签的最小二乘二次函数。在给定邻域点的条件下，设标签噪声 ε 的均值为零、对角协方差为 Σ，并设 A 为 $n \times p$ 局部二次设计矩阵，$C = (A^\top A)^{-1}$ 带有二次块 C_{hh}，W 为关于二次系数 $\|\cdot\|_g$ 的 Gram 矩阵，H 为无噪声海森矩阵。拟合得到的二次系数为 $\hat{}h = \bar{h} + [CA^\top\varepsilon]_h$，因此 $\mathbb{E}\,\hat{}H = \bar{H}$ 但

$$\mathbb{E}\|\hat{H}\|_g^2 = \|\bar{H}\|_g^2 + \mathrm{tr}(WV), \qquad V = [CA^\top \Sigma AC]_{hh} \tag{4}$$

（$\sigma^2\,\mathrm{tr}(WC_{hh})$，针对同方差噪声）。对于给定邻域点的 K，$\|\hat{}H - K\|_g^2$ 含有相同的项，而 $\langle \hat{}H,\ K\rangle_g$ 则不含，因此增益 $\Gamma = 2\langle \hat{}H,\ K\rangle_g - \|K\|_g^2$不存在噪声底。平方估计量的向上偏差及通过独立两半样本间的乘积消除该偏差，都是标准结果 [Hall and Marron，1987，Walther 等人，2016]。我们使用 MacKinnon and White 的 HC2 三明治估计量 [1985],，根据单次拟合的残差估计噪声底；或者通过对局部区域的两个随机两半样本分别拟合，再取交叉拟合乘积 ˆ ˆ$\langle H_A,\ H_B\rangle_g$，来消除噪声底。相同的噪声也会降低局部准确率。令$P_A$ 为投影到 A各列所张成空间上的投影算子，则 $\hat{}\sigma_q^2 = \|(I - P_A)y\|^2/n$ 和 lv 为局部区域的局部块方差。局部区域图表中的任何探针二次函数都满足 $1 - R^2 \geq \hat{}\sigma_q^2/\mathrm{lv}$；当且仅当它等于标签的局部二次拟合时，等号成立，因为 $y - \hat{}y = (I - P_A)y + (P_A y - \hat{}y)$，且两部分相互正交。因此，在引入任何几何因素之前，噪声更大的标签就会增大 $\|\hat{}H\|$ 和 $\|\hat{}H - K\|$，并降低 R^2。

关于星系。交叉拟合乘积将误差下限估计为 $\|\hat{}H\|_g^2$ 的 45–66%（在锚点上对 $1 - \langle \hat{}H_A,\ \hat{}H_B\rangle_g / \|\hat{}H\|_g^2$ 取中位数；20 个编码器-标签单元的中位数为 0.51），HC2 校正与之吻合（中位数比值为 0.95–1.04）。在 40 个单元中，$\hat{}\sigma_q^2$ 均与局部 R^2 呈负相关，且区间不包含零；在 $\alpha = 100$ 处，98–100% 的锚点满足该界限，此时探针的局部二次近似误差仅为其误差的 1–4%（在 α^* 处则不满足）。去除误差下限后（表 2），星等在所有编码器中均不再保持相关性；红移在 $\alpha = 100$ 处的五个编码器中有三个保持相关性，在 α^* 处则至多有两个。交叉拟合失配 $\langle \hat{}H_A - K,\ \hat{}H_B - K\rangle_g$ 也呈现相同情况：星等在两个探针处的五个编码器中均为零个保持相关性；红移在 $\alpha = 100$ 处有四个，在 α^* 处有两个。这些零值是误差下限相互抵消所致，而非统计功效降低，因为在 40 个单元中，半块范数的偏导数比完整块的更负。两种估计量都将图表二次项无法拟合的部分计作误差下限。

将探针的局部误差精确拆分为 $\|y - y\|^2 = \|\ I - P_A\ y\|^2 + \|\ I - P_A\ y\|^2 - 2\langle\ I - P_A\ y\ \ I - P_A\ y\rangle + \|P_A\ y - y\ \|^2$ ˆ()()ˆ(), ()ˆ(ˆ)，其中最后一项按 A 的常数块、线性块和二次块进行拆分，在 $\alpha = 100$ 处的中位数占比分别为：标签在其局部二次项之后的残差占 0.69–0.79，探针的非二次部分占 0.01–0.04，-0.14 至 -0.04 为它们的交叉项，以及失配的常数、线性和二次部分分别占 0.01–0.05、0.09–0.22 和 0.10–0.15；最后一项是一个上界，因为拟合得到的标签二次项带有噪声。去除基线后，K 对修正后的 $\|\hat{}H\|_g^2$ 的中位数捕获率为 0.11（$\alpha = 100$）和 0.28（α^*）；我们将这一比率 $\Gamma/(\|\hat{}H\|_g^2 - \widehat{\ \ }$界限）称为捕获率。在这里它不可能达到 1，因为它将解码器的 K 设定为基于数据拟合的海森矩阵。

ViT-B 上的线性下限。在 $d = 16$的 ViT-B 上，我们将每个探针的留出误差拆分为三部分：重新拟合仿射读出所消除的误差、灵活残差读出（近似普通最小二乘加上对其残差进行核岭拟合）在此基础上进一步消除的误差，以及剩余误差。重新拟合消除了 $\alpha = 100$ 探针误差的 0.303–0.323；对于通过留一交叉验证调优的探针，则消除了其误差的 -0.000 至 $+0.001$（$\alpha = 0.1$ 即按标签计为 0.001，而非附录 A 中嵌套的 α^*）。灵活读出

		ViT-B		DINOv3		CLIP-B		ConvNeXt-B		ViT-L		† 计数	
标签	数量	100	α^*	100	α^*	100	α^*	100	α^*	100	α^*	100	α^*
mag_r	$\|H\|^2$–floor	$+0.06^*$	$+0.16^\dagger$	-0.05^*	$+0.02^*$	-0.07^*	$+0.14$	-0.08^*	$+0.14$	-0.08^*	$+0.01^*$	0/0	0/1
	$\langle H_A, H_B\rangle$	$+0.09$	$+0.17^\dagger$	-0.06^*	$+0.02^*$	-0.05^*	$+0.13^\dagger$	-0.00^*	$+0.15^\dagger$	-0.08^*	$+0.01^*$	0/0	0/3
	分割余弦	$+0.35$		$+0.35$		$+0.31$		$+0.29$		$+0.30$			
光度_z	$\|H\|^2$–下限	-0.22	-0.24	$-0.35^\dagger$	$-0.36^\dagger$	$-0.20^\dagger$	-0.16	$-0.29^\dagger$	-0.12	-0.09	-0.02^*	3/0	1/0
	$\langle H_A,\ H_B\rangle$	$-0.22^\dagger$	$-0.26^\dagger$	$-0.27^\dagger$	$-0.27^\dagger$	-0.14	-0.13	$-0.27^\dagger$	-0.12	-0.06^*	$+0.02^*$	3/0	2/0
	分割余弦	$+0.32$		$+0.24$		$+0.25$		$+0.30$		$+0.24$			
平滑_分数	$\|H\|^2$–下限	$+0.14$	$+0.22$	-0.16	-0.13	-0.11	-0.05^*	-0.14	$+0.06^*$	$+0.26$	$+0.36$	0/0	0/0
	$\langle H_A, H_B\rangle$	$+0.11$	$+0.20$	-0.13	-0.09^*	-0.06^*	-0.02^*	-0.13	$+0.04^*$	$+0.21$	$+0.28$	0/0	0/0
	分割余弦	$+0.33$		$+0.31$		$+0.28$		$+0.32$		$+0.30$			
恒星质量_	$\|H\|^2$–下限	$+0.13$	$+0.25^\dagger$	-0.00^*	-0.03^*	-0.05^*	$+0.09$	$+0.03^*$	$+0.08^*$	$+0.05^*$	$+0.02^*$	0/0	0/1
	$\langle H_A, H_B\rangle$	$+0.14^\dagger$	$+0.22^\dagger$	$+0.04^*$	$+0.01^*$	-0.02^*	$+0.11^\dagger$	-0.01^*	$+0.08^*$	$+0.08^*$	$+0.04^*$	0/1	0/2
	分割余弦	$+0.31$		$+0.22$		$+0.19$		$+0.29$		$+0.21$			

表 2：去除噪声底后的局部 R^2 关联，$d = 16$，512 个锚点，多尺度密度控制：标签 Hessian 范数减去其 HC2 下限后的秩偏 Spearman、$\|H\|_g^2 -$ $\widehat{}$下限，以及两个半补丁拟合的交叉拟合乘积 $\langle H_A, H_B\rangle_g$。分割余弦是 H_A 与 H_B之间的余弦相似度中位数；它与探针无关。列对、标记、计数和横线均同表 1。

可进一步减少调优探针误差的 0.124–0.179，这是对线性下限的估计，而非界限。预注册检验旨在探究探针曲率能否预测该下限最大的区域：在基于局部标签方差、对数补丁半径、计数和密度构建的秩模型中，最佳仿射读出的白化 $\|K\|_g$ 对星等增加了 $\Delta\rho = -0.011$ $[-0.032, +0.011]$（没有额外价值），对红移增加了-0.002 $[-0.078, +0.025]$（结果不确定）；区间取自 32 块自助法的第 1.25 至第 98.75 百分位数。

错位，留出检验。在 ViT-B 上，我们在每个邻域的一半样本中测量 R 与 K_S之间的余弦相似度，并扣除 $\|R\|_g^2$ 的噪声底；然后在另一半样本中检验 $t = 0$ 是否胜出。在余弦相似度最低的五分位组锚点中，移除操作比在余弦相似度最高的五分位组中更常胜出：合并各标签后，差值在 $\alpha = 100$时为 $+0.066$（区间为 -0.001 至 $+0.122$），在 α^*时为 $+0.038$（区间为$+0.010$ 至 $+0.070$）；对于星等，差值为 $+0.070$ 和 $+0.100$（在 α^* 时，只有该标签的区间不包含零）。同等幅度的随机弯曲未显示出这种差异（-0.025 和-0.005）。

邻域类型。对于每个 ViT-B 锚点，我们对其邻域点中的星表星等、红移和和平成分比例取平均值，并在事后分析中比较锚点中排名最高和最低的三分之一。邻域彼此混合，因此两组三分位数差异很小：星等的截点为 18.16 和 18.37，红移为 0.180 和 0.206，和平成分比例为 0.724 和 0.786。探针对较暗邻域中的星等曲率捕获得更多（$+0.104$ 位于 $\alpha = 100$），对更平滑邻域中的星等、和平成分比例和恒星质量曲率捕获得更多（$+0.067$ 至 $+0.098$ 位于 $\alpha = 100$），而在高红移处对红移曲率捕获得更少（-0.036 位于 $\alpha = 100$，-0.057 位于 α^*）。较暗和红移较高的邻域噪声也更大，并且对于每个数量进行 24 次比较时，一两个区间可能会因偶然而排除零值。探针对较暗邻域的拟合最差。在最暗的三分之一中，其局部 R^2 相比最亮的三分之一，在 $\alpha = 100$处的星等、红移和和平成分比例分别低 0.13、0.14 和 0.22，在 α^* 处分别低 0.07、0.10 和 0.22（除星等在 α^*处的区间外，每个区间都排除了零值）；恒星质量并不遵循这一趋势（$+0.04$ 和 $+0.00$）。相对于自身的离散程度，那里标签的噪声更大：在每个邻居云的 16 个主成分中读取、且不使用解码器时，残差占比 $\hat{}\sigma_q^2$/lv 高出 0.19、0.11 和 0.18；根据该界限，局部 R^2 仅在 $\alpha = 100$处降低。弯曲强度 $\|\mathrm{II}^S\|$是锚点处内球第二基本形式的范数，无需标签。它不随邻域的星等或红移变化（跨锚点的秩相关为 -0.03 和 -0.06），并且在更平滑的邻域中更大（$+0.30$，区间为 $+0.05$ 至 $+0.51$）。

附录 C：移除曲率不一致处的弯曲部

第5节比较了在 R 和 K_S 最不对齐的锚点五分位组中，移除弯曲部获胜的频率与二者最对齐的五分位组中的频率。在 ViT-B 上，对于 α^*，我们按照附录 B 的方法，根据每个邻域一半部分上的余弦相似度对锚点排序，并在另一半上评分。移除弯曲部获胜

在某个锚点处，当 $t = 0$ 在该处的平方误差较低时，移除操作便胜出，其优势为 $R^2(t{=}0) - R^2(t{=}1)$。表3列出了每个五分位组中移除操作胜出的锚点数量及其优势幅度。在四个标签中，有三个标签的未对齐五分位组里移除操作更常胜出，但优势幅度很小：中位数为 R^2 的 0.5 至 1.6 个百分点（红移为 6.9 个百分点，基于两个锚点）。这项分析属于事后分析。

	对齐的五分位组		未对齐五分位组	
标签	移除操作胜出	R^2 优势	移除胜出	R^2 增益
星等_r	100 个中的 1 个	$+3.2$	100 个中的 11 个	$+1.6$ (12.4)
照片_z	99 中的 0	–	99 中的 2	$+6.9$ (13.6)
平滑成分比例_	99 中的 4	$+0.8$ (1.3)	99 页中的第 3 页	$+1.3$ (1.4)
恒星质量_	97 页中的第 3 页	$+0.9$ (1.6)	97 页中的第 7 页	$+0.5$ （0.7）

表3：ViT-B、调优探针（α^*）、512 个锚点；事后分析。根据每个邻域 A 半部分中R 和 K_S 的余弦相似度，将锚点分为五分位组：对齐（余弦相似度最高）和未对齐（最低）。移除获胜项：其中 $t = 0$ 对留出的一半 B 的拟合优于 $t = 1$的锚点。 R^2 增益：在这些锚点上，R^2（$t{=}0$） $- R^2$（$t{=}1$）的中位数，以局部 R^2的点数表示；当存在两个或更多获胜项时，括号内给出最大值。

附录D：分子嵌入上的弯曲校正（补充分析）

查看星系数据的结果后，我们想知道这种方法是否能推广到其他介质。我们在 QM9 [上重复了第5节中的弯曲校正。Ramakrishnan 等人（ 2014]）的研究使用 MoLFormer-XL [（Ross 等人， 2022], ）嵌入了 130,744 个分子，并提供了四种量子化学标签（HOMO–LUMO能隙、偶极矩 μ、各向同性极化率和热容 c_v）。实验设置与 ViT-B 相同：使用 $d = 16$处的种子0解码器、512 个锚点（每个锚点有 2,048 个邻域点），并采用处于调优惩罚项下的探针。在每个邻域中，我们用一半的邻域点重新拟合探针曲率，并用另一半进行评分。表4并列展示了这两个数据集。

弯曲校正。 在 $d = 16$ 中，图表中的一个二次形状具有 136 个独立分量，而解码器在锚点处的 II^S 会将每个分量映射到嵌入空间中的一个方向。我们取其 136 个奇异方向 n_k：沿 n_k 移动探针的权重，会使探针在第 k种形状上发生弯曲，而邻域点沿该方向的坐标为 $n_k \cdot (x - x_0)$ 。在邻域点的一半上，我们根据调优探针在这 136 个坐标上的残差拟合岭校正；在此之前，先移除其中恒定或在图表中呈嵌入空间中的线性函数的部分，使校正改变探针曲率而不改变其斜率。其惩罚项通过在该半部分上采用留一法来选择，并将不进行校正作为一种选项，然后在另一半上评估校正后的探针。校正后的探针在嵌入空间中仍保持线性，其权重会随邻域而变化。

	QM9、MoLFormer-XL	星系、ViT-B
邻域留出方差的弯曲部分占比	0.12–0.15	0.06–0.08
调优探针未读取的曲率（读取较低）	0.025–0.080	0.006–0.016
图表中自由二次函数的增益	0.023–0.041	0.0015–0.0039
弯曲校正的增益	0.044–0.049	0.0059–0.0095

Table 4：第5节在两个数据集上的弯曲校正，调优探针， $d = 16$，512 个锚点。Share s and 增益是在留出的局部 R^2 上计算的；每个范围覆盖四个标签。

弯曲校正使调优探针对间隙、极化率和 c_v 的局部误差降低了 18% 到 22%，对 μ 则降低了 8%；该探针对后者的拟合效果最差（局部 R^2 0.35，而其他项为 0.74 到 0.80）。在两个数据集上，调优探针读出的各标签曲率占比大致相同：这里为 0.40 到 0.69，在 ViT-B 上为 0.58 到 0.64。这个占比是探针拟合的图表海森矩阵对标签拟合的图表海森矩阵的回归系数；两者均由每个邻域两半的数据拟合而得，再将两半的系数相乘，并对所有锚点汇总。因此，它不包含附录 B 中比率所包含的解码器与数据之间的差距。分子标签所需读出的曲率约为其两倍，而探针遗漏的部分则大三到五倍；在半个邻域上重新拟合能恢复其中更多部分（这里可恢复未读出曲率的 0.56 到 0.99，而 ViT-B 上为 0.26 到 0.33）。我们的估计量在两个数据集上都低估了这些未读出的量

。对于每个标签自身的线性拟合（探针能够复现该拟合），它报告这里有 -0.11 到 -0.15 的曲率未被读出，而 ViT-B 上为 -0.21 到 -0.26；此处真实值接近 0。因此，表4中未读出这一行以及据此得出的比率都只是粗略估计。增益是两个拟合之间的差值，不受这种偏差影响。

有两项限制。弯曲校正在一个编码器和一个解码器上运行。而在这里，解码器遗漏了一些结构。在留出数据上，图表中的自由二次函数可解释各法向坐标的 0.19 至 0.36，而解码器自身的弯曲可解释 0.005 至 0.16；在 ViT-B 上，解码器的解释量超过了自由二次函数。因此，部分分子增益可能来自解码器在$d = 16$处的拟合方式，高于编码器估计的 11 维内在维度。弯曲校正带来的增益也大约一半来自曲率（0.020 至 0.029），一半来自图表之外的结构（0.019 至 0.026）；其弯曲部分达到自由二次函数所恢复部分的 0.70 至 0.88，而在 ViT-B 上两者相当。在 ViT-B 上，弯曲部分为 0.0022 至 0.0039，占总量 0.0059 至 0.0095；其余部分来自图表之外的结构。