Journal of Computer-Aided Molecular Design, 10 (1996) 427-440 ESCOM

J-CAMD 351

非共价蛋白质 - 配体相互作用的评分：一种用于计算结合亲和力的连续可微函数

Ajay N. Jain

美国加利福尼亚州旧金山市牡蛎角大道 385 号，阿瑞斯制药公司，邮编 94080

1996 年 2 月 7 日收到
1996 年 4 月 28 日接受

关键词：分子对接；蛋白质配体相互作用；打分函数

摘要

通过分子对接来利用蛋白质结构以寻找潜在药物先导化合物，这在很大程度上取决于对假定蛋白质 - 配体相互作用进行评分的方法。理想的评分函数必须具有预测准确性高和计算速度快的特点，并且必须能够容忍蛋白质 - 配体分子相对排列和构象的变化。本文描述了一种基于蛋白质 - 配体复合物的结合亲和力及其晶体结构所推导出的经验评分函数的开发。该函数的主要项涉及疏水性和极性互补性，另外还有熵和溶剂化效应的项。通过构建一个连续可微的非线性函数解决了排列/构象依赖性问题，该函数要求配体构象/排列空间中的最大值与晶体结构紧密对应。基于交叉验证，预测亲和力的预期误差为 1.0 个对数单位。该函数足够快速和准确，可用作分子对接搜索引擎的目标函数。该函数特别适合对接问题，因为它具有空间上狭窄的极大值，且可通过梯度下降法广泛地到达这些极大值。

介绍

随着已知原子分辨率结构的蛋白质数量不断增加，针对小分子先导化合物的三维数据库搜索方法变得愈发重要。已有大量通过数据库筛选发现先导化合物的报道[1-4]。对接系统的一个关键改进领域在于评估假定蛋白质 - 配体相互作用的方法。评分是对接问题的核心。分子对接程序的目标是在给定蛋白质和配体的情况下，通过改变配体相对于蛋白质的排列和构象（本文余下部分将分子的构象/排列称为构象）来使评分函数的值达到极值。本文提出了一种经经验调整的评分函数，用于预测蛋白质 - 配体相互作用的结合亲和力，其具体目标是用于灵活分子对接系统。

理想的评分函数具有三个关键特性：（i）预测结合亲和力的准确性；（ii）速度；（iii）对不准确配体构象的容忍度。前两个特性显而易见，但第三个特性需要进一步讨论。一个几何搜索引擎在给定蛋白质结合位点的情况下尝试为配体生成合适的构象，会在一定时间内以一定概率生成在一定容差范围内的“正确”构象。随着容差的增加，快速生成足够好的构象的概率也会增加。因此，评分函数在给定不准确构象的情况下计算准确结合亲和力的能力，将对其在该系统中的实用性产生重大影响。

实现对姿态不敏感这一目标的一种方法是构建一个空间上较为粗糙的函数——即姿态的小幅变动不会导致预测亲和力出现显著变化。然而，这必然会导致函数特异性的丧失。函数中“最佳结合区域”越宽泛，就会有越多的配体（错误地）显示出良好的得分。对接中的假阳性问题要比假阴性问题严重得多。假设在一个包含 100000 种化合物的数据库中，只有 20 种化合物能与目标以高于某个阈值的亲和力结合。即便假阴性率很高（比如 50%），找到 10 种真正的有效化合物也可能会让人感到满意。但是——

0920-654X/s 6.00 + 1.00 0 1996 ESCOM Science Publishrs B.V.

然而，根据实际能够获得（通过购买或合成）并进行检测的化合物数量，假阳性率对发现真正有效的化合物的可能性有着重大影响。如果假阳性率仅为 1%，那么在这次筛选中就会有 1000 个假阳性“命中”。如果只能获取并检测 1010 个命中中的 50 个，且没有原则性方法区分真阳性与假阳性，那么有 60% 的可能性一个真阳性都找不到。为了有 90% 的可能性至少找到一个真阳性，绝对假阳性率必须低于 0.2%。

尽管上述分析具有推测性，但从定性角度而言，它仅基于这样一个观察结果：在大型化合物数据库中，真正具有特异性结合的配体数量远少于非结合配体的数量，这一点是毋庸置疑的。显然，构建尽可能特异性的函数至关重要。这里所采用的方法是构建具有明确狭窄峰值的函数，以尽量降低假阳性率。然而，这使得对接程序难以找到对应于峰值的配体构象。对函数施加一个额外的约束条件可以解决这个问题。该函数是连续且可微的，因此对接程序只需找到接近函数最大值的构象，然后依靠基于梯度的搜索来找到最大值。因此，不准确的初始配体构象由评分函数进行校正，而不是依赖对接程序生成完美的构象。为了适应这一点，函数与校准数据的拟合度被定义为在基于梯度的配体构象变化下函数的最大值。该函数的理想表现是：蛋白质与配体相互作用的最高分接近正确的结合亲和力，且对应的构象接近真实结构。

这里所采用的函数形式遵循了 Compass 技术的模式，结合了非线性组合。

表 1
校准套装中使用的配合物

Number	Protein	Ligand	Affinity[a]	Source
1	Carboxypeptidase	ZFV^{P}(O)F	14.0	7CPA
2	Streptavidin	Biotin	13.4	1STP
3	Carboxypeptidase	A-ZAA^{P}(O)F	11.52	6CPA
4	Thermolysin	ZF^{P}LA	10.19	4TMN
5	DHFR	Methotrexate	9.70	4DFR
6	HIV Protease	L700,417	9.15	4PHV
7	Thrombin	NAPAP	8.52	1DWD
8	Thermolysin	ZG^{P}LL	8.04	5TMN
9	Galactose BP	Galactose	7.60	2GBP
10	Thermolysin	Phosphoramidon	7.55	1TLP
11	Thrombin	MQPA	7.40	1ETR
12	Thermolysin	CLT	7.30	1TMN
13	Retinol BP	Retinol	6.72	1RBP
14	Trypsin	NAPAP	6.46	1PPC
15	Thermolysin	HONH-BAGN	6.37	5TLN
16	Trypsin	3-TAPAP	6.22	1PPH
17	Thrombin	TAPAP	6.19	1ETT
18	Cytochrome P450	4-Phe-imidazole	6.07	1PHF
19	DHFR	2,4-Diaminopteridine	6.00	4DFR[b]
20	Cytochrome P450	Adamantone	5.88	5CPP
21	Xylose isomerase	Xylose	5.82	1XIS
22	Fatty acid BP	C_{15}COOH	5.43	2IFB
23	PNP	Guanine	5.30	1ULB
24	TIM	Phosphoglycilic acid	4.82	2YPI
25	Trypsin	Benzamidine	4.74	3PTB
26	PHBH	*p*-Hydroxybenzoate	4.68	2PHH
27	Thermolysin	PLN	4.67	2TMN
28	Trypsin	Phenylguanidine	4.14	3PTB[c]
29	Thrombin	Amidinopiperidine	3.82	1DWD[c]
30	Thermolysin	Leu-NHOH	3.72	4TLN
31	Trypsin	Benzylamine	3.42	3PTB[c]
32	Chymotrypsin	Indole	3.10	4CHA[c]
33	Thrombin	Benzamidine	2.92	1DWB
34	Trypsin	Butylamine	2.82	3PTB[c]

单位为 -log（K.）。

b 从 5 中获取的蛋白质结构是通过删除甲氨蝶呤中除蝶啶片段以外的所有部分而得到的。

配体通过 Hammerhead [11] 软件对接到结合位点。

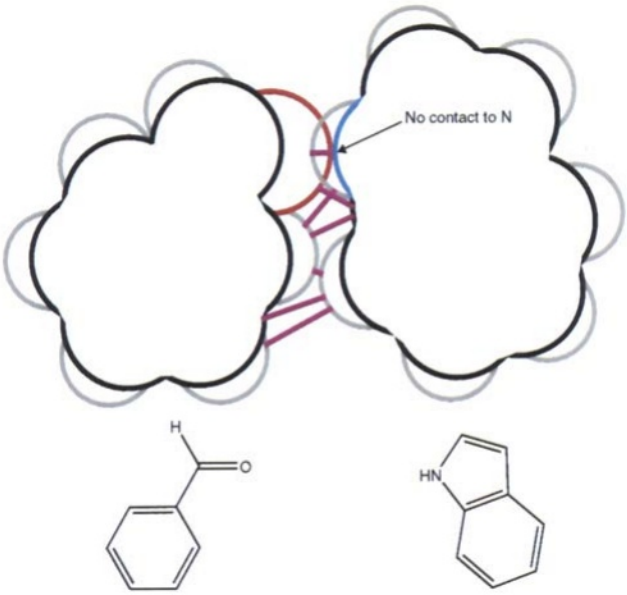

图 1.参与计算相互作用的蛋白质 - 配体原子对。紫色线条表示未被遮挡的原子间接触，因此构成了相互作用能总和的基础。

具有可调参数的基函数通过迭代梯度下降法进行估计[5-7]。除了 Compass 功能方法外，这项工作最直接借鉴了 Bohm [8] 以及 Bohacek 和 McMartin [9] 的成果。Bohm 表明，通过考虑疏水表面接触面积、有利的极性接触以及熵校正，可以构建一个简单且准确的经验调优评分函数。此前，Bohacek 已经证明，互补疏水和极性蛋白质配体接触数量的加权线性和能极好地拟合一系列热溶菌素抑制剂的 K 值。这里报告的类似 Compass 的评分函数的主要项是疏水性和极性互补性的度量，另外还有熵和溶剂化效应的项。

校准数据集由 34 个具有已知亲和力和几何结构的蛋白质 - 配体复合物组成。在留一法交叉验证下，预测结合亲和力的平均误差为 1.0 个对数单位。在配体构象变化条件下，评分函数的最大值与实验确定的结构吻合良好。这些最大值非常狭窄：与最大构象偏差 0.5 埃的均方根距离会导致计算结合亲和力下降 3 个对数单位。然而，该函数的梯度表现良好，以至于最大 2.0 埃的位移不会导致无法通过梯度下降法优化的构象。该函数计算速度很快（在胰蛋白酶 - 苯甲脒相互作用中，每次评估仅需 0.03 秒），这使其适用于分子对接。该技术构成了自动识别和表征蛋白质结合位点方法的基础[10]。它还被用于一个灵活的分子对接系统[11]。

方法

校准数据集

用于估计该函数参数的复合物是 Bohm [8] 所给出复合物的一个子集。未能满足以下条件的复合物被排除在外：

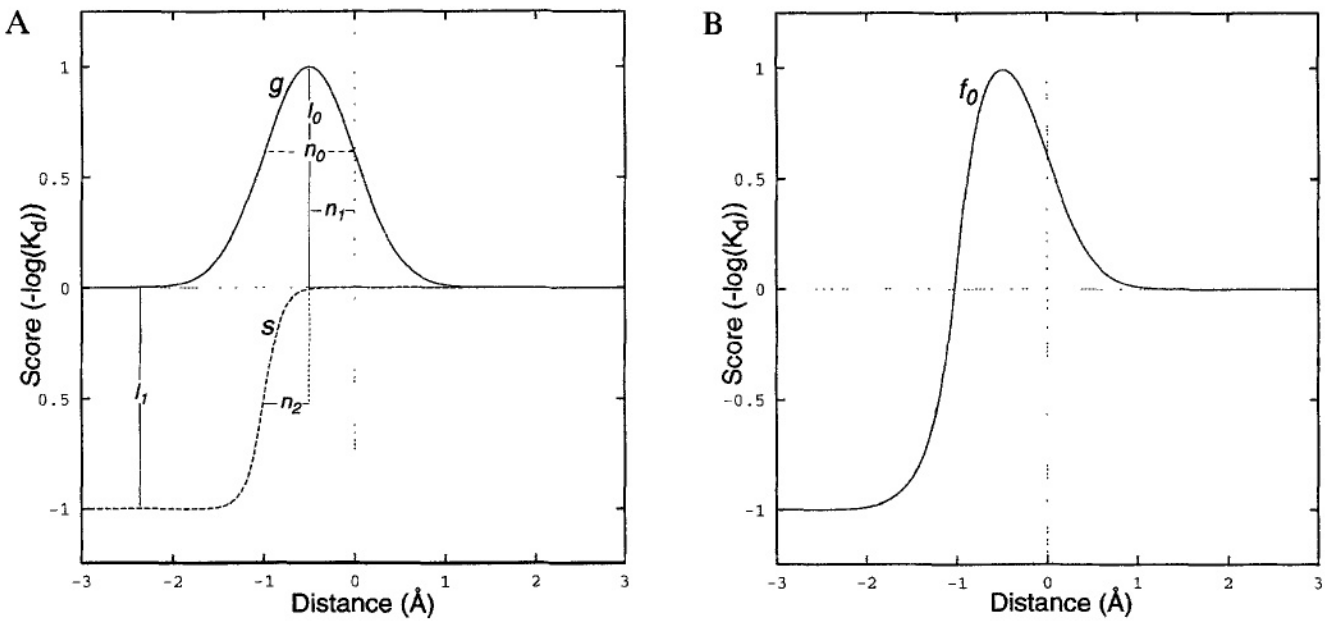

图 2. 基函数的图示。（A）高斯型和 S 型基函数具有可调参数。（B）两者结合形成一个关于原子间距离的函数，该函数具有一个奇异的最大值，原子间距离较大时贡献逐渐减小趋于零，而原子间深度重叠时贡献为负。

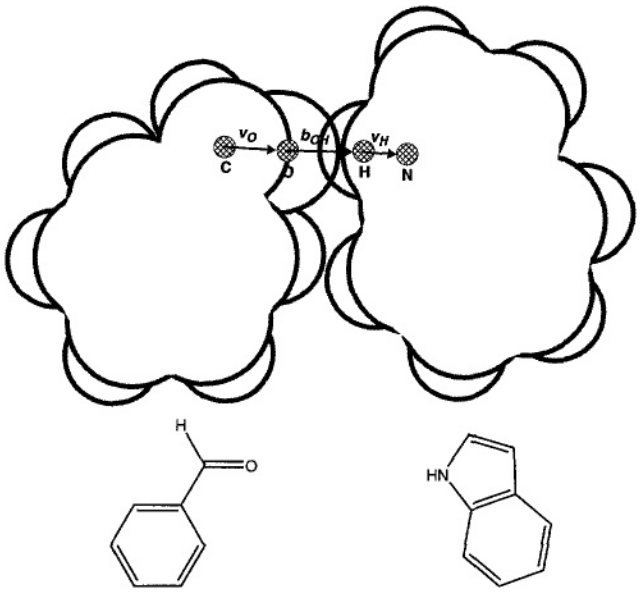

图 3 极性相互作用的方向性。在计算氢键的方向性贡献时，使用了定义 C-O 键、O-H 氢键和 H-N 键方向的矢量。

以下标准被剔除：可得或可推导的结构、解离常数小于 10-3 M 以及质子化后蛋白质配体无显著空间重叠。在评估蛋白质配体重叠之前，系统地重新定位蛋白质和配体上的羟基和巯基，以使不利的相互渗透最小化。这产生了 34 个复合物，其 Ky 值范围为 10-3 至 10-14 M。表 1 列出了复合物、亲和力和数据来源。所有情况下的结合相互作用均被认为是非共价的。请注意，少数复合物是通过将小而相对刚性的配体对接到蛋白质的活性位点中获得的。在这些情况下，其他非常相似的分子或其片段的复合物已通过实验解决。

函数形式

评分函数是暴露的蛋白质配体原子范德华表面距离的非线性函数的加权和。它考虑了疏水接触、极性接触和方向性、溶剂化效应以及熵效应。该函数分段连续且可微。在整篇论文中，结合亲和力均以 -log⊠⊠(K) 单位表示。

图 1 展示了用于计算相互作用能的原子对的二维卡通图。对于每对原子，计算其最近表面距离（相互穿透视为负距离）。如果距离大于 2.0 埃，或者路径穿透了另一个原子，则忽略该对原子。虽然可以计算最近的非遮挡距离，但这涉及更复杂的计算。假设在遮挡阻止原子对相互作用的情况下，导致遮挡的原子间的相互作用往往会占主导地位。通过这些测试的蛋白质 - 配体原子对将被称为一对。

蛋白质和配体上的每个原子都被标记为非极性（例如，CH 中的 H）或极性（例如，N-H 中的 H 或 C-O 中的 O），并且极性原子还被赋予电荷。关于评分函数形式的讨论将大致按照重要性顺序进行：疏水互补性、极性互补性、溶剂化项和熵项。完整的评分函数是这些各项的总和。对评分函数实用性的细节不感兴趣的读者可以跳过本节的其余部分。

疏水互补性

疏水作用通过成对表面距离的高斯函数 g 和 S 型函数 s 的加权和来体现，如 Compass [5-7] 中所述：

$$g(x,\mu,\sigma) = e^{-(x+\mu)^2/\sigma} \quad (1)$$

$$s(x,\mu) = 1/(1+e^{10(x+\mu)}) \quad (2)$$

函数 g 捕获原子接触的正向部分，而 s 捕获由于过度嵌入产生的部分。在接下来的方程中，可调的线性参数将用 1 表示，与结合亲和力具有非线性关系的可调参数将用 n 表示：

$$f_0(x) = l_0\, g(x,n_0,n_1) + l_1\, s(x,n_2+n_1) \quad (3)$$

函数 fo 定义了疏水性对结合亲和力的贡献。图 2 展示了 lo = 1.0、1 = -1.0、n, = 0.5、n = 0.5 以及 n， = 0.5 时 fo 的曲线图（这些参数仅作说明之用）。图 A 展示了 g

表 2
定义F的方程总结

Hydrophobic term	$\sum_{i,j} f_0(d(i,j))$
Polar term	$\sum_{i,j} f_1(d(i,j),i,j)$
Repulsive term	$\sum_{i,j} f_2(d(i,j),i,j)$
Solvation term	$(l_5 \cdot phbe) + (l_6 \cdot lhbe)$
Entropic term	$(l_7 \cdot n\_rot) + (l_8 \log(mol\ weight))$
	$f_0(x) = l_0\, g(x,n_0,n_1) + l_1\, s(x,n_2+n_1)$
	$f_1(x,i,j) = f_{1a}(x)\, f_{1b}(i,j)\, (1+n_6c_i)\, (1+n_6c_j)$
	$f_2(x) = l_5\, g(x,n_7,n_8)\, f_{1a}(x)\, f_{1b}(i,j)\, (1+n_6c_i)\, (1+n_6c_j)$
	$f_{1a}(x) = l_2\, g(x,n_3,n_4) + l_3\, s(x,n_2,n_4)$
	$f_{1b}(i,j) = s(-(b_{ij} \cdot v_i)\, (b_{ij} \cdot v_j), -n_5$
	$g(x,\mu,\sigma) = e^{-(x+\mu)^2/\sigma}$
	$s(x,\mu) = 1/(1+e^{10(x+\mu)})$
	$d(i,j) = ((x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2)^{1/2} - r_i - r_j$

详情请见正文。

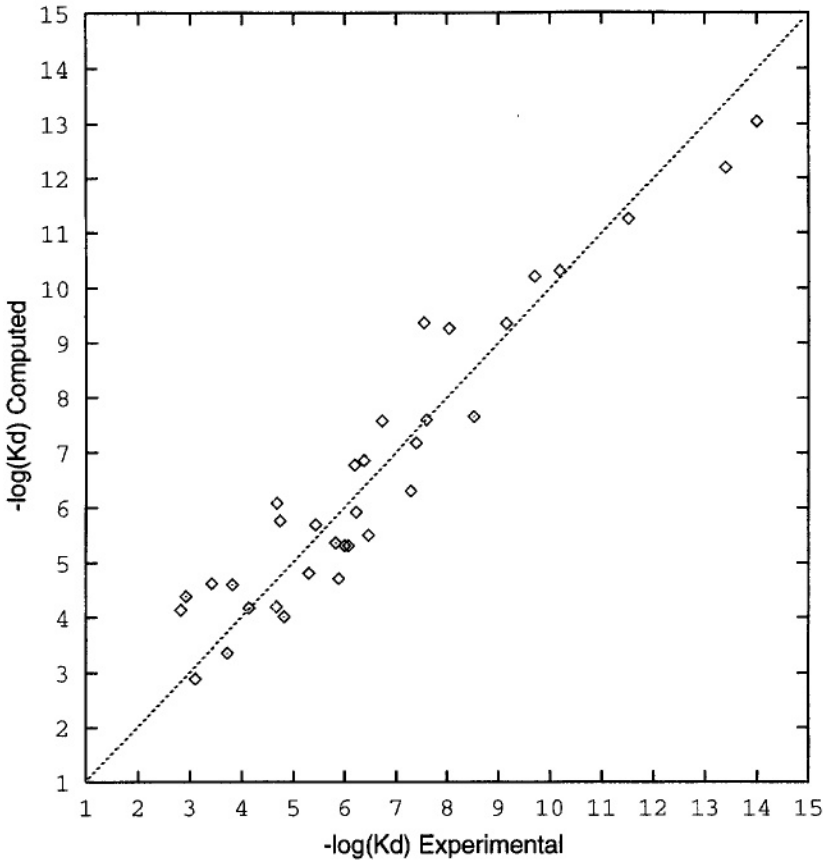

图 4. 计算亲和力与实验亲和力的对比图。

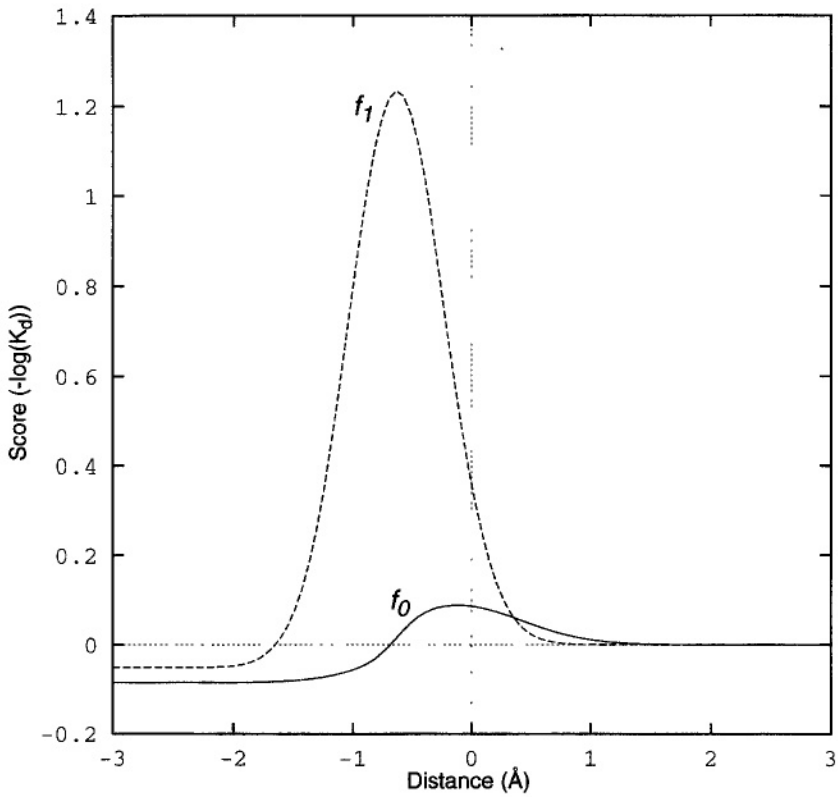

图 5. F 的 f⊠ 和 f⊠ 曲线图。

图 A 分别展示了各项贡献，图 B 则展示了综合函数。该函数在特定的表面距离处达到正值最大值；随着距离的减小，函数值下降至负值最小值。

线性参数控制着对结合亲和力的总体贡献，而非线性参数则控制着基函数的内在形状。参数 I 表示单个理想非极性接触的正疏水性贡献，l 表示过量空间重叠的最大代价。参数 no 对应疏水相互作用的狭窄程度，n 表示实现最大相互作用所需的相互渗透程度，nz 则表示达到最大相互渗透惩罚一半所需的过量相互渗透量。

相互作用能的总疏水成分是：

哪里

$$\sum_{n,j} f_0(d(i,j)) \qquad (4)$$

$$d(i,j) = ((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2)^{1/2} - r_i - r_j \qquad (5)$$

原子 i 的原子坐标用 x、y 和 z 表示。原子 i 的范德华半径用 r 表示。求和是对所有不全是极性原子的原子对进行的。

极性互补性

氢键和盐桥的作用力通过与疏水项类似的求和方式来计算。极性基函数对所有互补极性原子对进行求和。其形式与 fo 相同，但参数不同：

$$f_{1a}(x) = l_2\, g(x,n_3,n_4) + l_3\, s(x,n_2,n_4) \qquad (6)$$

函数 fia 定义了方向性最佳的中性氢键的相互作用贡献。由于氢键的形成除了受纯距离的影响外，还受其他因素的影响，因此还需要一个方向性项：

$$f_{1b}(i,j) = s(-(b_{ij}\cdot v_i)\,(b_{ij}\cdot v_j), -n_5 \qquad (7)$$

向量 b 是从原子 i 到原子 j 的归一化向量，v 是原子 i 的“出”方向，v 是原子 j 的“入”方向。图 3 说明了这一点（O 为原子 i，H 为原子 j）。请注意，方向因子是对称的：交换原子 i 和 j 会得到相同的值。当向量重合时，该函数达到最大值 1.0，而当方向相异时则降至 0.0。强度达到最大值一半时的点由 ng 决定。该函数实质上定义了围绕每个极性接触参与者相对于其相互接触方向的有利锥体。入/出方向对于

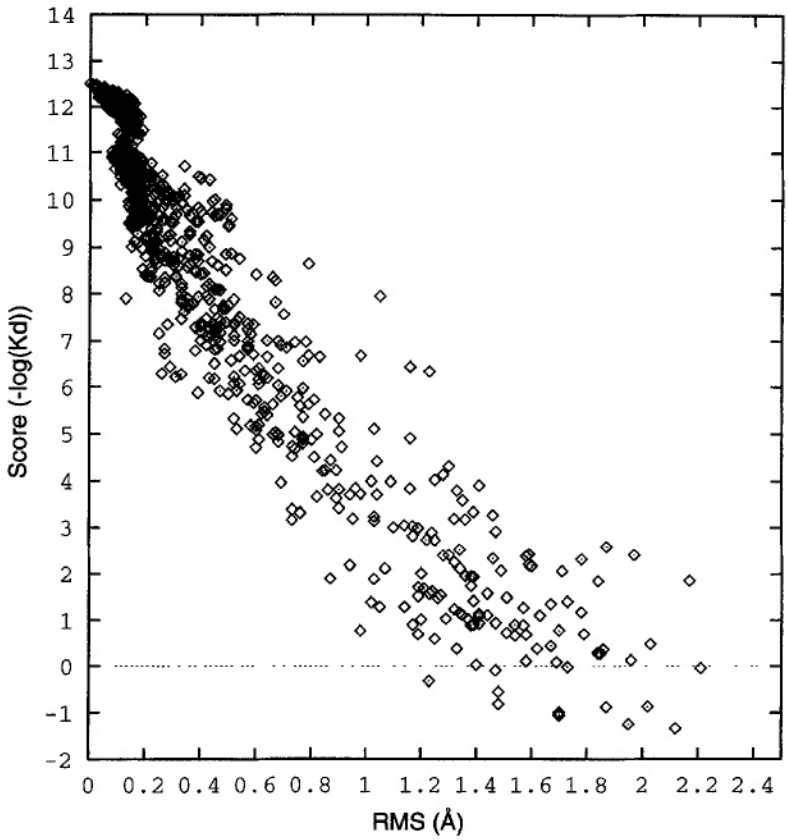

图 6. 最大构象偏差的均方根值与计算亲和力的关系图。

每个极性原子的计算均基于与其相连的原子的质心以及该极性原子的位置。对于羰基，还会生成两个替代向量，分别从氧原子发出，且与从 sp' 碳原子到其其他取代基的向量平行。在计算中使用由最有利向量组产生的相互作用强度。这近似于在晶体学测定结构中观察到的氢键几何分布情况[12]。

还需要做最后的补充，以考虑正式电荷相互作用的影响。定义单个极性接触强度的最终方程为

$$f_1(x,i,j) = f_{1a}(x)\, f_{1b}(i,j)\, (1 + n_6 c_i)\, (1 + n_6 c_j) \qquad (8)$$

原子 i 的形式电荷用 c 表示。因此，随着任一原子的形式电荷增加，相互作用的强度会像库仑定律那样呈线性增加。这里不使用部分电荷。

为了考虑不利的极性接触，对同号极性原子对求和采用以下函数：

$$f_2(x) = l_5\, g(x,n_7,n_8)\, f_{1a}(x)\, f_{1b}(i,j)\, (1 + n_6 c_i)\, (1 + n_6 c_j) \quad (9)$$

此函数只有一个峰值，并且随着距离的增加而下降。

溶剂化效应

可以将溶剂化效应视为打破与水的有利相互作用以及未能与配体或蛋白质重新形成这些相互作用所产生的代价。因此，将配体的疏水部分与蛋白质的极性基团紧密接触，或者反之亦然，都会产生一定的代价。蛋白质和配体的溶剂化效应是同等对待的。其目的是计算蛋白质和配体在相互作用中可能形成的氢键等效数与实际形成的氢键数之间的差值。所以，如果蛋白质的羰基直接指向配体且距离较近，则会获得一个氢键等效数。如果蛋白质的极性原子朝向远离配体的方向，或者距离过远，则其值接近于零。在这两种情况下，都假定这些原子能够与溶剂分子或其他蛋白质原子相互作用。

对于蛋白质而言，会识别出蛋白质中每个包含极性原子的成对组合。对于每一对，都会计算出由任何配体原子产生的最佳潜在极性相互作用的强度（方向性和电荷效应）（每个配体原子都被视为具有正确极性的中性氢键参与者）。此值会通过 S 型距离函数 s（x,n） 进行衰减。参数 n 决定了衰减函数达到半最大值的点。

图 7.生物素（灰色）的梯度优化构象与晶体学测定结构的对比。

潜在蛋白质氢键等效物总数与实际极性相互作用量（phbe）之间的差值乘以 1s。同样，配体氢键等效物总数与实际极性相互作用量（Ihbe）之间的差值乘以 1。

熵成本

由于可旋转配体键的固定以及旋转和平动熵的损失而产生的熵成本构成了该函数的其余部分。这些成本与蛋白质无关，由两个简单的项来表示：第一项只是配体中自由旋转键的数量（n_rot）乘以一个比例因子 l，第二项是配体分子量的对数乘以比例因子 l（这是基于平动/旋转熵对质量的大致依赖性[13]）。对于蛋白质不计算固定成本，因为该评分函数的目标应用是在分子对接系统中，其中蛋白质被视为刚性。对于特定的结合位点，不同配体对蛋白质固定成本的影响变化很小。

本节中给出的所有方程均总结于表 2 中。完整的评分函数是疏水项、极性项、排斥项和熵项在适当原子对上的总和。

实验结果

回想一下，该评分函数将由一个.....来使用。

分子对接程序的目标是通过改变配体的构象来最大化其得分。因此，配体在其最佳构象下的得分应被视为其有效得分。理想的评分函数应在其全局最大值附近接近晶体学确定的构象。如果一个评分函数的准确性依赖于已知晶体学确定的构象，那么它的实用性就很小。这个问题最好通过考虑使用静态构象构建的函数，然后观察构象变化的影响来说明。

利用 34 种天然晶体结构构建了一个函数。通过从线性参数值较小的随机初始点开始进行梯度下降，计算出了该函数的可调参数。表 3 展示了此函数（称为 F.）与数据的拟合情况。所有误差值均以 -log(K) 单位报告。作为对计算出的复合物秩的非参数度量，还给出了配对秩相关系数（PRCC）[7]。PRCC 是对两个预测正确排序可能性的估计。性能看起来还算合理：平均误差为 1.2 logs err，均方根误差为 1.4。然而，如果通过梯度下降优化配体的构象，许多得分会显著提高。表 3 中 F 的第二组值是通过从天然构象开始进行构象优化得到的。虽然没有配体在均方根（rms）方面与天然构象相差很大，但根据计算出的亲和力进行拟合的效果却显著变差：平均误差为 1.5 logs err，均方根误差为 2.1。显然，在参数优化过程中考虑这种影响是很重要的，否则该函数的在线性能将会受到影响。

这个问题的处理方式与 Compass 算法处理问题的方式相同[5-7]。训练算法迭代参数估计和配体构象优化。在给定静态构象的情况下优化初始随机参数值，然后使用当前函数优化构象，如此反复。在参数变化过程中，要求复合物上评分函数的最大值接近实验值。在五次迭代内即可收敛到一组稳定的参数。

这种方法在处理配体与蛋白质之间的不利接触时又引发了一个问题。从一个“一无所知”的函数开始，然后逐步开发出一个经过调整的函数，这种做法存在一定的风险。因为训练集中没有任何信息表明空间重叠是不利的，因为该数据集仅包含未表现出显著蛋白质穿透的配体实例。所以，在中间构象优化过程中，可能会忽略诸如空间重叠之类的硬物理约束，从而将产生的构象视为完全合法。为避免此问题，在构象优化过程中，向函数中添加了一个空间重叠项（在上述 F 的构象变化实验中也采用了这种方法）。对于原子对 ij 的空间重叠项为 -10.0d + 8（d + 8），其中 8 对于互补极性接触为 0.7，对于其他情况为 0.1（这些值并非系统性选择）。因此，0.5 Å 的空间重叠会产生 1.6 个对数单位的惩罚。为了使评分函数最大化，必须考虑空间重叠项。理想情况下，系统应通过明确建模负数据来适当地学习空间重叠的负惩罚。然而，在缺乏负数据的情况下，施加额外的惩罚会使函数更紧致，因为函数受到的约束更严格。附加项迫使系统学习一组参数，这些参数需满足避免空间重叠的约束条件。

空间重叠项仅在构象优化过程中使用；它不会计入最终得分。人们可能会认为，即使是很小的相互穿透也应该在最终得分中受到显著惩罚。然而，即使是在紧密结合配体的高分辨率结构中，也常常会发现一些空间重叠。这种重叠与结合亲和力之间几乎没有相关性。当然，非常显著的重叠是不能被容忍的，但确定可容忍的阈值是一个与对接问题的筛选方面相关的经验性问题。

Fz 是在参数估计过程中从初始姿态开始通过在线姿态优化构建而成的。在姿态优化条件下，其性能明显优于 F，对数误差为 1.0，均方根误差为 1.2（而 F 分别为 1.5 和 2.1）。姿态优化是终端应用所必需的，但也有助于改善存在的问题。

表 3

不同函数的拟合优度

Function	Mean error	Rmse[a]	PRCC[b]	Conditions of test
F_1	1.15	1.39	0.90	Static native poses
F_1	1.51	2.09	0.90	Native poses, pose optimization
F_2	0.97	1.17	0.92	Native poses, pose optimization
F	0.72	0.85	0.95	Minimized ligands, pose optimization

均方根误差。

成对等级相关系数（1.0 表示完全一致的排序，0.5 表示随机排序，0.0 表示完全相反的排序）[7.]

表 4 参数值（缺失）

Parameter	Value	Parameter	Value
l_0	0.0898	n_0	0.6213
l_1	–0.0841	n_1	0.1339
l_2	1.2338	n_2	0.4880
l_3	–0.1796	n_3	0.3234
l_4	–0.0500	n_4	0.6313
l_5	–0.1539	n_5	0.6139
l_6	0.0000	n_6	0.5000
l_7	–0.2137	n_7	0.5010
l_8	–1.0406		

在复合物中详细原子位置的实验不确定性。这一点至关重要，因为评分函数完全依赖于原子间的距离。在 X 射线数据的分辨率范围内，从一个结构到另一个结构氢键距离和角度的微小变化本质上只是噪声。这种变化可能会对估计的参数产生重大影响。通过允许该函数对这种几何形状进行“正则化”，可以推导出具有更尖锐最大值的函数。

最后一个问题是配体和蛋白质的构象应变。在许多复合物中，配体（较少情况下是蛋白质）表现出不良的键角和/或二面角。这种能量未被评分函数所考虑，因此存在显著构象应变的复合物往往具有比实验测定值更强的计算亲和力。对于具有一定柔性的配体，可能只需要考虑配体的应变，因为配体和蛋白质都会分担应变（就像两个相互挤压的弹簧）。最终函数 F 的构建旨在帮助解决这种构象应变问题。对于每个复合物，将配体从结合位点移除，并在真空中进行最小化处理，然后再放回结合位点。在大多数情况下，最小化后的配体结构仅与原始结构略有偏差，但在某些情况下构象应变较为显著。F 和 Fz 的区别仅在于用于 F 的配体是预先最小化的。

最终评分函数 F 对数据的拟合效果明显优于任何中间函数。计算出的亲和力平均误差为 0.72 个对数单位（均方根误差为 0.85），复合物的亲和力排序近乎完美（成对排序相关系数为 0.95）。这一结果相当惊人，因为应用于配体构象的最小化过程对其的扰动与复合物及评分函数无关。图 4 展示了 F 计算出的结合亲和力与实验值的对比图。F 的参数值见表 4。与原始晶体结构相比，最终梯度优化后的配体构象的均方根偏差较小（平均值为 0.70 埃，标准偏差为 0.37）。

讨论

考虑各个项对结合亲和力的相对贡献是很有趣的。在利用 E 计算出所有亲和力之后，将各项的绝对值作为总和的百分比进行计算。对于 34 个复合物在 F 的不同项上的得分分布情况为：疏水性占 4%，极性占 26%，熵占 25%，溶剂化占 5%。在这些项中，特定于复合物的项（熵项不依赖于蛋白质）有 93%来自疏水性和极性项 fo 和 f。图 5 展示了 F 的这些项的图。请注意，每个函数的 S 形部分权重不大。原因有两方面。首先，由于所有配体都适配于蛋白质结合位点，复合物表现出的过度重叠非常少。其次，在构象优化过程中使用的空间重叠项可防止在训练过程中出现此类重叠。从筛选的角度来看，这是有问题的，因为大多数配体需要基于无法适配于目标位点而被排除。理想情况下，计算出的得分应适当考虑穿透效应的权重。实际上，采用了一个阈值，仅允许少量的相互渗透，以考虑蛋白质的某些移动性。

疏水成分的峰值位于范德华表面轻微相互渗透处（此处所用的范德华半径略大于从晶体学研究中推导出的值[5]）。单个理想的疏水接触可产生 0.09 个对数单位的结合亲和力。虽然这看似微不足道，但由于大多数接触都属于此类，因此这一项最为重要。有些特定的基序使用少量原子就能产生显著的疏水得分。例如，一个非极性质子与六元环的一个面紧密堆积，就能产生超过 0.5 个对数单位的亲和力。这将在链霉亲和素 - 生物素相互作用中进一步说明，其中关键的色氨酸侧链提供了理想的疏水堆积表面。

极性成分的峰值位于氢键距离为 2.0 埃的相互插入处。极性强度减半时与 HO 向量的夹角为 52°。理想的中性氢键可产生 1.2 个对数单位的结合亲和力。电荷的影响（ng = 0.5）通过一个例子来说明最为恰当。一个带正 0.25 电荷的脒基质子与带负 0.5 电荷的羧酸氧相互作用可产生 1.7 个对数单位的亲和力，而中性氢键则为 1.2 个对数单位。排斥极性项的权重很小，其原因与相互插入项权重小的原因基本相同。

溶剂化项总体上并非特别重要，但在某些情况下（例如配合物 12、热溶菌酶/CLT）却能造成高达 2 个对数单位的差异。尽管如此

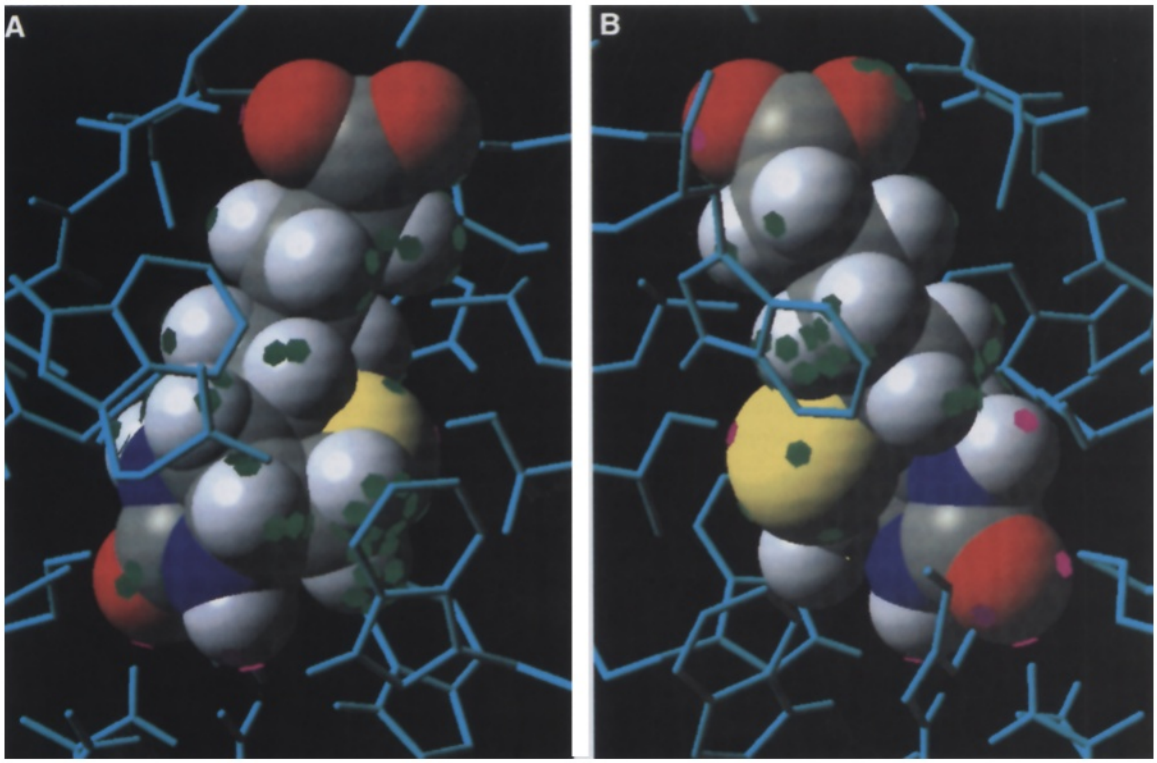

图 8 生物素与链霉亲和素结合的相互作用（两种视角）。紫色区域表示有利的极性相互作用；绿色区域表示有利的疏水相互作用。

由于对称定义，只有蛋白质溶剂化成分获得了显著的权重。其原因似乎在于蛋白质溶剂化效应的动态范围相当大，而配体的动态范围则较小。

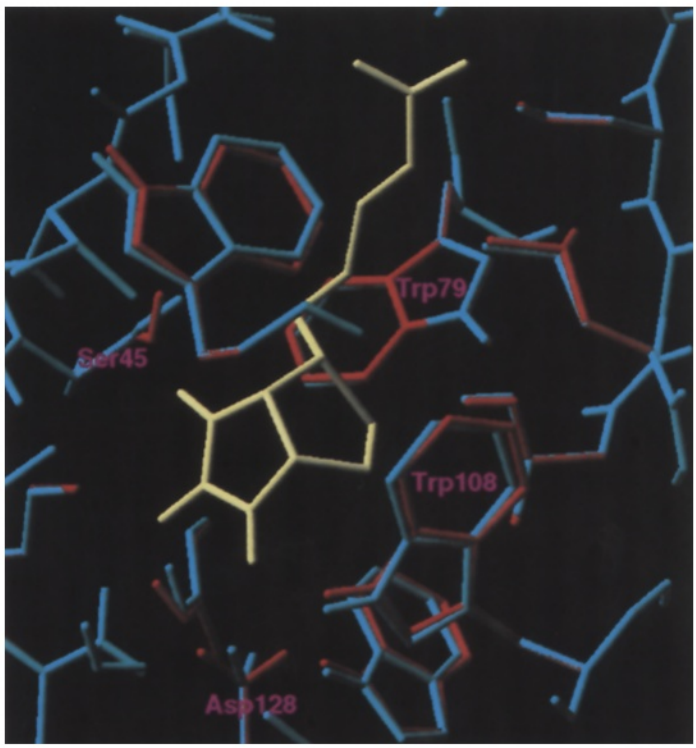

图 9. 利用 F 对蛋白质侧链的改变（两个视角）。链霉亲和素的原始结构为青色；修改后的结构为红色；生物素为黄色。

由于配体中极性基团较少，溶剂化效应要小得多。为了使溶剂化计算具有对称性，可以将 I 和 l 合并为一个参数，但这对函数的性能不会产生显著影响。

该函数并未明确处理一些已知会影响结合的因素。例如，长程极性效应和芳香环的极性效应。虽然可以为每个因素添加相应的项，但使用当前的函数集就能很好地解释这些数据。这在一定程度上是由于数据集的规模有限。随着更多数据的出现，以及计算亲和力不佳的实例增多，可能就需要添加这些项了。

与玻姆函数的比较

该函数与玻姆推导出的函数非常相似。由于两个函数中平动/转动熵损失项的表达式不同，其他线性参数的动态范围略有差异。对于特定示例进行直接比较最为简便，由于有玻姆函数的原始数据，因此可以对胰蛋白酶 - 苯甲脒的情况进行分析。玻姆计算出的该情况的亲和力为 5.49，具体分解如下：3.52 为疏水性，3.16 为极性，-0.25 为可旋转键熵，-0.95 为刚性体固定熵。对于 F 函数，计算出的亲和力为 5.69，具体分解如下：5.21 为疏水性，3.50 为极性，-0.64 为蛋白质溶剂化，-0.21 为可旋转键熵，-2.17 为刚性体固定熵。为了直接比较这些数值，将它们重新缩放，使两者使用相同的常数：4.38、2.94、-0.54、-0.18 和 -0.95。最后一步是将 F 函数的疏水性和蛋白质溶剂化项合并，以匹配玻姆的单个亲脂性项：3.84（3.52）疏水性，2.94（3.16）极性，-0.18（-0.25）熵，-0.95（-0.95）固定。两个函数的分解结果非常接近。考虑到其潜在的功能方法大不相同，这种匹配令人称奇。

Bohm [8] 对这些推导值与实验估计值之间的关系进行了详尽的讨论，此处不再重复那些函数一致的值的讨论。F 函数与 Bohm 函数的一个显著差异在于，由于平动和转动熵的损失而产生的固定成本更高。F 函数的成本范围为 -1.95 至 -2.9（11.2 - 16.5 千焦），这取决于分子量的对数，而 Bohm 的函数则为常数 -0.95（5.4 千焦）。F 函数的值更接近 Searle 和 Williams [13] 所估计的“9 至 45 千焦/摩尔”这一范围。

内部验证测试

鉴于所使用的非线性函数的表达能力很强，因此与校准数据的紧密拟合并不令人特别惊讶。然而，值得注意的是，可调参数的数量远远少于校准集中的复合物数量。此外，定义疏水性、极性和熵项的 13 个参数占总分的 95%。另外，构象变化增加了数据集的有效大小，因为该函数被约束为拟合每个配体的许多可能构象中的最大值。

为了进一步验证该函数的准确性，对 F 进行了留一法交叉验证。在与 F 相同的条件下构建了 34 个模型，每次将校准集中的一个复合物排除在外。交叉验证得出的平均预测亲和力误差为 1.00 个对数单位（1.26 均方根误差），交叉验证的 r 值为 0.79，PRCC 为 0.91。当然，这比对数据的拟合要稍差一些，但交叉验证实验中假阳性和假阴性的模式与校准实验中的模式一致。这表明函数参数由数据过度确定。交叉验证实验中最大的偏差与上述拟合实验中的最大偏差相对应，但偏差的幅度更大。由于缺乏足够数量的适当复合物，未对“盲集”进行正式实验。使用 F 进行柔性对接到几个蛋白质位点的实验正在进行中，这将提供有关其有用性和预测准确性的明确数据。

对精确姿态的不敏感性

分子对接中 F 的效用取决于其从几何搜索引擎生成的不准确初始构象中恢复的能力。为了评估这一特性，对生物素的最小化构象进行了随机变化，通过扰动其扭转角和对齐参数来实现。从这些扰动后的构象开始，应用梯度下降法以最大化评分函数的值。图 6 展示了在多次梯度下降过程中所考虑的所有构象与生物素最大构象的均方根偏差（rms）与评分值的关系图。有两个重要特征。首先，该函数在 rms 方向上最大值的宽度较窄。如前所述，构象空间中窄的最大值应能降低筛选过程中出现的假阳性率。在最大构象的 0.2 埃范围内存在一个评分值非常接近的平台，但随后迅速下降。在 0.5 埃 rms 时，评分值下降了约 3 个数量级，而在 1.0 埃 rms 时，评分值下降了约 8 个数量级。其次，该函数相对于构象的梯度表现良好，所有扰动后的构象都被拉到了一个可接受的最终构象（评分值接近最大值且 rms 较小）。图 7 展示了生物素的最大构象（得分为 12.5）以灰色显示与晶体的叠加情况。

黄色结构的构象（得分 9.7）。这些构象几乎无法区分，均方根偏差为 0.45 埃。

速度

在需要反复计算成千上万种分子对接的内循环中，速度至关重要。计算过程中最耗时的部分之一是确定参与相互作用的蛋白质 - 配体原子对。为避免计算所有成对距离，预先计算了附近蛋白质原子的三维网格。给定配体原子的位置，可以在网格中进行常量时间查找，以确定可能足够接近从而影响函数的所有蛋白质原子。通过这种简单的优化，在 SGI R4400 150 MHz 机器上计算苯甲脒与胰蛋白酶的亲和力平均时间为 0.03 秒。这已经足够快，使得在利用 F（111） 的柔性分子对接系统中，评分函数的计算不再是瓶颈。

训练集偏差

校准集所用复合物的偏差是一个令人担忧的问题。一个相对较小的担忧是某些类型复合物（例如蛋白酶）的过度代表。仅由一个代表物代表的蛋白质的预测亲和力平均误差为 1.0 个对数单位，与由多个代表物代表的复合物相同。每种蛋白质的配体变化性似乎足以使每种复合物都有很大的不同。一个更严重的问题是，所有数据基本上都是从正例中得出的。所有配体都已知能与蛋白质结合。此外，它们都表现良好，可以结晶。缺乏负数据导致对诸如空间重叠之类的排斥相互作用赋予的权重较小。可以通过设置保守的允许重叠阈值来避免由此产生的假阳性。然而，在溶剂化效应的情况下，其重要性可能远大于其权重，却没有简单的方法来消除假阳性。

人们可能会想利用筛选得到的阴性结果，通过为非结合物生成假定构象并赋予其极低的得分来将其作为阴性数据。然而，一条表明某种配体与蛋白质的结合能力低于某个阈值的数据可能意味着很多情况。理想的解释是该配体无法与已确定的结合位点结合。另一种解释是它与另一个位点结合得更好。所以，该配体可能适合对接位点，但检测方法却无法检测到。还有其他与检测方法假阴性相关的解释也可能说明结合结果为阴性。相对安全的阴性数据来源是那些预测结合能力高于实际结合能力的分子。如果预测某个分子以特定构象结合的浓度为 1 微摩尔，而实际检测结果为 100 微摩尔，那么可以安全地重新调整函数，使其在任何构象下预测的最大值为 4.0（-log(K)）。

链霉亲和素 - 生物素复合物

链霉亲和素 - 生物素复合物是一个研究得相当透彻的体系，这得益于其极高的结合亲和力（K = 10-134 M）[14]。图 8 展示了生物素与链霉亲和素结合的最大构象，图中在生物素表面标出了疏水和极性相互作用。紫色区域表示极性相互作用，绿色区域表示疏水相互作用。生物素与蛋白质形成了八个强的特异性极性相互作用：三个来自脲基氧，其余极性原子各一个。值得注意的是疏水堆积的模式。图 A 下方右侧标记的亚甲基与 Trp' 的吲哚形成了 13 个接触点。图 B 中间的亚甲基与 Trp" 的吲哚形成了 10 个接触点。得分（12.5）的分解情况为：疏水 7.61，极性 8.99，蛋白质溶剂化 0.53，可旋转键熵 -1.07，旋转/平移熵 -2.48。

很难将这些值与其他关于该复合物的研究联系起来。韦伯等人所做的热力学研究将自由能分为 ΔH 和 ΔS 两项[14]。然而，F 中的关键疏水项和极性项各自都包含了焓效应和熵效应。疏水项综合了焓的色散 - 吸引作用力、释放水分子的熵以及水分子在很大程度上疏水的蛋白质空腔外形成更优氢键的焓。极性项综合了库仑吸引力以及置换固定水分子的混合效应。更复杂的是，实验中的 ΔH 和 ΔS 可能会受到此处未考虑的蛋白质变化的显著影响。例如，生物素与链霉亲和素结合时，链霉亲和素的一个柔性环会固定下来[14]。

然而，尽管此前针对这一复合物的研究试图通过假设生物素脲基头部的部分极化来部分解释其高亲和力[8,14]，但此处报告的计算并未对这一部分做任何特殊假设。头部的极性原子被视为中性的氢键供体。这是本研究与 Bohm 研究的另一个不同之处。即便假设了部分极化，计算得出的生物素 - 链霉亲和素亲和力仍低了 2.7 个对数单位。这种差异可能在于对结合亲和力中疏水成分的精确计算。Bohm 的方法将亲脂性表面接触面积与结合亲和力线性关联。甲基与 Trplos 和 Trp 相互作用的情况检验了每个亲脂性接触面积 A2 均相等这一隐含假设。这些区域的原子对接触密度非常高。

实际上，在整个结合界面上，非极性原子对接触的密度相当高。如果接触密度对每平方埃的结合能有显著影响，那么链霉亲和素/生物素复合物对于假定表面接触面积与结合亲和力之间呈线性关系的函数来说，就是一个棘手的案例。

Miyamoto 和 Kolman 在文献 15 中对链霉亲和素 - 生物素复合物进行了大量的自由能微扰（FE）计算。他们通过将生物素突变为“假分子”，在两种环境中计算结合自由能：链霉亲和素内部和溶剂中。这两种环境下的能量差即为结合能。尽管计算出的极性相互作用能相当大，但他们发现生物素在溶剂中也形成了类似的相互作用，并得出结论认为疏水作用是主要的相互作用。他们认为 Trp'00 和 Trp 可能尤为重要。该研究结果表明疏水作用在解释极高的亲和力方面具有重要意义，但在极性相互作用的净重要性方面存在较大分歧。在自由能微扰计算中，复合物中存在的极性相互作用的能贡献在很大程度上被溶剂化效应所抵消。

虽然看起来 F 可能低估了溶剂化效应，但很难将宫本和科尔曼的自由能计算与微量热法数据[14]相协调。韦伯等人[14]表明，生物素与链霉亲和素的结合自由能主要由焓驱动（ΔH = -32.0），而熵则起不利作用（TAS = -13.7）。宫本和科尔曼解释说，焓和熵的符号及大小对疏水作用的预测效果不佳。这里给出的计算结果将该复合物的结合亲和力协调在韦伯等人的结果（侧重于极性成分）和宫本及科尔曼的结果（侧重于疏水成分）之间。

HABA 与链霉亲和素结合的案例提供了另一个例子。结合亲和力为 10⊠⊠ M，由熵效应驱动[14]。使用来自天然结构（PDB 参考号 ISRE）的 F 计算得到的 HABA 亲和力为 10⊠⊠⊠ M。疏水项主导了极性项（分别为 7.8 和 1.7 个对数单位）。这与量热数据定性一致。与生物素相比，极性贡献的大幅降低同时降低了结合的焓以及蛋白质和配体的固定程度。晶体学观察表明，复合物中链霉亲和素的柔性环未完全固定，这支持了这一解释 [14]。

蛋白质构象的影响

几乎所有用于校准的复合物均取自晶体结构中存在配体的情况。因此，蛋白质的构象是针对特定配体进行优化的。显然，这种情况并不能反映对接问题中蛋白质保持刚性的情况。在对接问题中，成千上万的假定配体要对接到固定不变的蛋白质结合位点上。为了评估这种效应在具有良好亲和力的小分子上可能产生的误差大小，将生物素对接到未结合的链霉亲和素结构中[16]。计算出的亲和力为 9.9（在配体构象优化之后），比正确的亲和力低 3.5 个对数单位，比从共晶结构计算出的亲和力低 2.6 个对数单位。这可能比交叉验证的结果更能表明在真正的对接应用中可能出现的误差程度。幸运的是，假设蛋白质处于相对低能的构象，大多数此类误差似乎都应是负向的（计算出的亲和力偏低）。这应当会提高假阴性率，但正如前面所讨论的，这不像对假阳性率产生不利影响那么严重。在这种特定情况下，从未结合的结构开始，人们仍会发现生物素是真正的阳性结果，因为它在亚纳摩尔范围内得分。

校正这种效应需要明确的蛋白质构象变化。这是通过使用 F 来改变蛋白质侧链的构象来实现的，同时保持蛋白质的其余部分以及配体不变。在该实验中，侧链无法区分蛋白质和配体，因此有可能最终得到的蛋白质构象会以牺牲蛋白质 - 配体相互作用为代价来优化蛋白质内部的相互作用。这会导致计算得出的蛋白质 - 配体结合亲和力相对于天然未复合结构有所下降。实验是迭代进行的，从生物素与未复合的链霉亲和素结构的近似对接开始（通过将天然结构与训练中使用的最小化生物素进行叠加生成）。首先调整生物素的构象，然后优化与生物素接触的每个侧链，仅改变侧链的扭转角，重复此过程直至收敛。蛋白质侧链能够使其对蛋白质和配体组合的“亲和力”提高 4.0 个对数单位。侧链所获得的大部分增益来自蛋白质内部的相互作用。最终计算得出的生物素与修饰后的链霉亲和素的亲和力为 11.4，比原始未复合结构提高了 1.5 个对数单位（仅比天然结构的计算亲和力低 1 个对数单位）。

图 9 展示了修改后的结构与原始结构在构象上的差异。侧链移动幅度相对较小，最大的原子位移为 0.76 埃（整体侧链均方根为 0.27 埃）。最大的移动对应于 Asp12（位于底部），其重新定向以形成两个氢键。

在与生物素的 N-H 保持氢键的同时，与其他蛋白质原子形成键。Trpl0 和 Trp" 在优化过程中均发生重新定向。此处的相互作用与天然结构中的相互作用之间最显著的剩余差异在于，生物素的另一个 N-H 与 Ser'5 之间的氢键。前者的距离为 2.34 埃，而在天然结构中为 2.06 埃。仅侧链运动就将初始生物素构象优化后原结构中的距离从 2.56 埃改善，但无法使相互作用达到最佳距离。为了在对接搜索中使侧链变化计算足够快速而实用，还需要进行更多的工作。主链蛋白质的变化则需要大量的工作。

筛选

要全面评估 F 的效用，唯一的方法是利用 F 作为目标函数，对大型数据库进行筛选，以寻找新的小分子配体。目前，我们正在开展相关实验，将 F 与一种自动识别和表征结合位点的方法以及一个灵活的搜索引擎相结合[10,11]。在针对链霉亲和素系统的初步实验中，结果令人鼓舞。特别是，在 80000 种化合物对接到链霉亲和素结合位点后，F 预测生物素将是亲和力最高的配体，高出 2 个对数单位。这对于解决假阳性问题而言是一个令人鼓舞的结果。对预测的强结合配体的测定正在进行中，结果将在后续论文中报告。正在筛选的其他系统包括细胞因子受体和丝氨酸蛋白酶。

结论

分子对接系统中对评分函数的三个关键要求是准确性、速度以及对蛋白质结合位点中假定配体不准确构象的容忍度。这里定义的函数 F 满足这些要求。通过在一组多样化的结合位点和配体上进行交叉验证，预测亲和力的预期平均误差估计为 1.0 个对数单位。通过一种简单的优化方法加快识别配体附近的蛋白质原子，计算苯甲脒与胰蛋白酶亲和力的时间为 0.03 秒，这足以用于对接搜索引擎。F 是一个连续且可微的函数，其最大值与晶体学确定的结构非常接近。因此，不精确的假定配体构象可以通过梯度下降法进行高效优化。

在数据库筛选问题中，重要的是要将评分函数的假阳性率降至最低，因为在数千种化合物中，实际上只有几十种会与特定靶点结合。F 的参数估计方案产生了一个具有空间狭窄最大值的函数。配体的最大构象偏差 0.5 埃均方根会导致计算亲和力降低 3 个对数单位。与涉及设计为空间粗略的函数的其他方法相比，F 的高空间特异性可能会降低筛选中观察到的假阳性率。

该函数考虑了疏水相互作用、极性相互作用（包括方向性和电荷）、熵效应以及溶剂化效应。在未添加任何专门术语、参数或局部极化效应的情况下，计算出的亲和力分解为这些成分的结果与实验观察到的生物素与链霉亲和素结合的焓和熵在定性上是一致的。使用 F 来改变蛋白质构象的初步实验为分子对接中更普遍地处理柔性问题带来了希望。

致谢

我感谢威尔·韦尔奇、吉姆·拉珀特和特里·克莱因的合作努力以及对书稿提出的宝贵意见，还要感谢迈克·罗斯和迈克·韦努蒂给予的支持。

参考文献

1 Rutenber, E., Fauman, E.B, Keenan, R.JI, Fong, S, Furth, P.S., Ortiz de Montellano, PR, Meng, E., Kuntz, I.D, DeCamp, D.L., Salto, R., Ros, J.R., Craik, C.S. and Stroud, R.M. J. Biol. Chem., 268 (1993) 15343.

2 Shoicet, B.K., Stroud, R.M., Santi, D.V, Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445.

3 Bodian, D.L., Yamasaki, R.B., Buswell, R.L., Stearns, J.F., White, J.M. and Kuntz, I.D., Biochemistry, 32 (1993) 2967.

4 Ring, C.S, Sun, E., McKerrow, J.H., Lee, G.K, Rosenthal, P., Kuntz, I.D. and Cohen, F.E., Proc. Nat⊠. Acad. Sci. USA, 90 (1993) 3583.

5 Jain, A.N., Harris, N.L. and Park, J.Y, I. Med. Chem., 38 (1995) 1295.

6 Jain, A.N., Dietterich, T.G., Lathrop, R.L., Chapman, D., Critchlow, R.E., Baur, B.E.., Webster, T.A. and Lozano-Perez, T., J. Comput.-Aided Mol. Design, 8 (1994) 635.

7 Jain, A.N, Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.

8 Böhm, H.J., J. Comput.-Aided Mol. Design, 8 (1994) 243.

9 Bohacek, R.S. and MeMartin, C., J. Med. Chem., 35 (1992) 1671.

10 Ruppert, J., Welch, W. and Jain, A.N., manuscript submited for publication.

1 Welch, W. Ruppert, J. and Jain, A.N, Chem. Biol, 3 (199) 449.

12 Murray-Rust, P. and Glusker, J.P, I. Am. Chem. Soc, 106 (1984) 1018.

13 Searle, M.S. and Williams, D.H. J. Am. Chem. Soc, 114 (1992) 10690.

14 Weber, P.C., Wendoloski, JJ., Pantoliano, M.W. and Salemme, FR., J. Am. Chem. Soc., 114 (1992) 3197.

15 Miyamoto, S. and Kollman, P.A., Proteins, 16 (1993) 226. 16 Katz, B.A., Biochemistry, 34 (1995) 1421.