Published in final edited form as:
*J Mol Biol*. 2012 June 8; 419(0): 211–222. doi:10.1016/j.jmb.2012.03.012.

# Fragmentation-Tree Density Representation for Crystallographic Modelling of Bound Ligands

**Gerrit G. Langer**, **Guillaume X. Evrard**, **Ciaran G. Carolan**, and **Victor S. Lamzin***
European Molecular Biology Laboratory c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

## Abstract

The identification and modelling of ligands into macromolecular models is important for understanding molecule's function and for designing inhibitors to modulate its activities. We describe new algorithms for the automated building of ligands into electron density maps in crystal structure determination. Location of the ligand-binding site is achieved by matching numerical shape features describing the ligand to those of density clusters using a “fragmentation-tree” density representation. The ligand molecule is built using two distinct algorithms exploiting free atoms with inter-atomic connectivity and Metropolis-based optimisation of the conformational state of the ligand, producing an ensemble of structures from which the final model is derived. The method was validated on several thousand entries from the Protein Data Bank. In the majority of cases, the ligand-binding site could be correctly located and the ligand model built with a coordinate accuracy of better than 1 Å. We anticipate that the method will be of routine use to anyone modelling ligands, lead compounds or even compound fragments as part of protein functional analyses or drug design efforts.

### Keywords

electron density map; small-molecule binders; shape; hybrid approach; drug design

## Introduction

The process of structure determination in macro-molecular crystallography (MX) is considered complete when all electron density in the unit cell has been modelled to the extent possible. The identification of small compounds—ligand molecules—and the appreciation of their binding mode are often keys to the understanding of chemical, biological or pharmaceutical processes in which the examined protein complex takes part. Highly automated procedures for the accurate identification of ligand-binding sites and placement of ligands in crystallographic maps are therefore highly desirable, permitting the rapid, convenient and unbiased completion of biological structures for subsequent analyses.

X-ray crystallography has proven very useful for the identification of bound low-affinity solutes arising from the crystallisation liquor—indeed it was one of the earliest methods capable of doing so.[1] The bound fragments located in the crystal structure allow the location of binding sites and directly indicate their binding modes, while their distributions on the protein surface may help guide the construction of new leads and drug-like molecules by the association of adjacent fragments in a “LEGO-like” manner. This formed the basis for a crystallography-led, fragment-based approach to drug discovery[2] successfully applied to a



*Corresponding author. victor@embl-hamburg.de.
Present address: G. X. Evrard, Bruker AXS, Östliche Rheinbrückenstrasse 49, 76187 Karlsruhe, Germany.

number of biological targets in recent years. [3,4] However, a large-scale application of this approach requires a large effort for data collection and modelling, and automated fragment location and building is of obvious benefit in such efforts.

Conventionally, crystallographic ligand building comprises a number of steps. Firstly, a suitable binding site is identified from the electron density and/or any other *a priori* available knowledge. Subsequently, the ligand molecule is placed in that binding site in a conformation that matches the electron density and is stereochemically sensible in itself as well as relative to the protein environment. Finally, the coordinates of the ligand molecule are refined, either in real or in reciprocal space. The thus obtained complex of the ligand with the protein is further refined against the measured diffraction data. A search for the next bound ligand in the new difference electron density can then be initiated. In all cases, subsequent final rounds of structure refinement and validation are necessary steps.

In MX, a variety of methods for automated modelling of bound ligands have been proposed over the last decade, each with emphasis on different aspects of the process. All of these approaches, in essence, fit the ligand model to a density blob, and the quality of the correspondence is evaluated by calculating some measure of the fit, typically a local map correlation coefficient. Most methods use variations of rotational and translational placement of the target ligand model as a whole, followed by its conformational optimisation. The AFITT method, drawing on the power of Monte Carlo techniques, was suggested by Wlodek *et al.*[5] Therein, an ensemble of low-energy plausible conformations is prepared, and each is in turn fit to the density using a force field. The molecular graphics application Coot[6] uses an adapted form of the X-LIGAND[7] method where potential binding sites are found using the analysis of unsatisfied electron density and where a Monte Carlo method is employed for optimisation of a ligand conformation to best fit the density blob.

A radically different approach—particularly suitable for the building of large ligands—was successfully implemented in Resolve[8] and then also provided within the Phenix suite. Therein, the modelling is achieved via the placement of a core fragment of the ligand with subsequent addition of the remaining parts according to the electron density and stereochemical considerations. An equally novel approach—based on the modelling of a density blob as free atoms—was implemented in the ARP/wARP modelling suite.[9,10] Atomic assignments of the free atoms are made using a graph-search approach leading to a full model, which is then optimised in real space. Graph search is also a key element of a method based on the medial axis transform,[11] where a set of points is computed from the surface of a density blob and is then thinned to follow the centre of the shape. The search ligand is matched to a “thinned” subset of the medial axis. A semiautomated procedure was proposed[12] to assist in the identification of bound ligands from unknown electron density by aligning the surface of the binding cleft and the representative set of ligands from the Protein Data Bank (PDB).[13]

A number of publications pointed to the potential benefits of using advanced mathematical features for the automatic recognition of molecular patterns. Such features included, for example, higher-order shape descriptors based on spherical harmonics for superposition of the surfaces of the ligand and the binding cavity through minimisation of the distance between their respective expansion coefficients.[14,15] More elaborate numerical descriptors that are less limited in their resolution of shapes, such as the Zernike moments, allow the recording of various structural properties in a concise way and were demonstrated to be applicable to ligand description.[16,17]

In this paper, we describe a number of novel techniques for shape recognition and ligand fitting. An innovative methodology for binding site identification based on the construction

of a “fragmentation tree” is introduced. New shape features that aid assignment of the appropriate site for an input ligand are also described. The “label-swapping” routine for ligand building previously described[9,10] has been developed and combined with a new Metropolis routine, and we demonstrate that their combination leads to better results than with either method alone.

## Results and Discussion

### Implementation of the software

The developed technologies have been implemented in the ARP/wARP ligand-building module,[18] providing a comprehensive and robust procedure for accurate building of ligands into crystallographic maps starting from separate ligand and protein structures provided in the PDB format. The full pipeline is shown in Fig. 1.

### Pre-processing of the input information

**Locating the binding site—**Electron density maps are typically presented in terms of iso-surfaces drawn/plotted at a given contour level. These closed surfaces define contiguous regions of density higher than the contouring level, which we hereafter refer to as density *clusters*.

We introduce a novel method for description of a three-dimensional electron density contoured at different density levels, which depicts the variation of volumes of density clusters and is extremely useful for discriminating ligand density clusters from background noise in a crystallographic map.

Let us consider a single atom whose electron density follows an isotropic three-dimensional normal distribution:

$$\rho(r)=\frac{Z}{(2\pi)^{3/2}s^3}\exp\left(-\frac{r^2}{2s^2}\right) \quad (1)$$

where $r$ is the distance from the centre of the atom, $s$ is the standard deviation of its density distribution and Z is a scaling factor. If this density is contoured at a threshold $\rho(r) = t$, the volume inside the contoured density cluster is:

$$V=\frac{4\pi}{3}r^3=\frac{4\pi}{3}s^3\left\{\ln\left(t^{-2}(2\pi)^{-3}s^{-6}Z^2\right)\right\}^{3/2} \text{ or}$$
$$V^{2/3}=\left(\frac{4\pi}{3}\right)^{2/3}s^2\left(\ln\left((2\pi)^{-3}s^{-6}Z^2\right)-2\ln(t)\right) \quad (2)$$

For a given cluster, the dependence of $V^{2/3}$ against $\ln(t)$—a *fragmentation tree*—should be linear. We note that an approximately linear dependence is also observed for clusters composed of more than one atomic Gaussian shape and computed from resolution-truncated data. In these cases, the scaling factor Z may be larger than the number of electrons in the atoms.[19,20] Due to overlaps of atomic density, the observed electron density clusters for bound ligands have somewhat smaller volumes than would be expected from Eq. (2). However, this does not adversely affect the use of the fragmentation tree for recognition of the density clusters.

When the bound ligand is fully occupied and is pronounced in the difference density, its density cluster has properties distinct from other clusters, as evident in the example fragmentation tree in Fig. 2. Upon increase of the density-contouring threshold, the clusters

of bound compounds reduce in their volume. Ligand density areas can be recognised from characteristic, approximately linear stretches. The slope of a stretch reflects the sharpness of the density inside the clusters—compare, for example, the blue-coloured branch corresponding to a zinc ion in Fig. 2 with the stretches for HEM or NRG with steeper slopes. The intercepts of the lines corresponding to the stretches reflect the density height inside the clusters. All clusters in the difference density, which do not correspond to bound ligands but rather to water molecules and other small features, form the rapidly decaying “background” region of the plot.

The density clusters for ligands eventually break into smaller fragments upon increase of the density threshold—see, for example, the HEM cluster in Fig. 2. At a low level, the density for the HEM entity is contiguous with that of several adjacent water molecules, especially those bound to its carboxylate groups. At ~ 1.4 sigma (point 1), the contiguous iso-surface fragments and the enclosed volume at this stage are those of the HEM and the sulphur atom of the cysteine residue (Cys186) that associates strongly with the iron atom of HEM. At higher threshold (approximately half-way between points 1 and 2 in the plot), the density iso-surface remains fully intact, and the reduction in volume of the cluster in this range corresponds to that predicted by Eq. (2). At point 2, the HEM cluster begins to fragment, initially with the separation of the carboxylate groups of HEM from the central pyrrole substructure. The other non-core carbon atoms—CBB and CBC— separate at point 3. The new, smaller density clusters enclosing these fragments are represented by the stretches of red points at lower-density volumes. At 5 sigma level (point 4), only the strong density of the HEM iron atom and the adjacent sulphur remains. The significant change in volume observed here results from the final separation of densities associated with both atoms.

**Shape matching—**An electron density map calculated from the ligand to be fit is compared to each potential density cluster based on their shapes. This is accomplished using seven shape features that provide a concise but thorough description of an object, described under Methods. The single highest-scoring match is taken forward for further ligand building.

**Deriving ligand stereochemistry—**Prior to construction of the ligand, its stereochemical description is required. Such information is not explicitly given in the input PDB file and must be obtained from the atomic coordinates. The protocols for stereochemical analysis are described later in the text. Ligand building subsequently proceeds via both label-swapping and Metropolis routines.

## Building the ligand

**Label swapping—**The label-swapping routine was first introduced in 2004,[9] but it is now applied in an improved manner, with the analysis of the output also advanced as evidenced in the following discussion. Within the routine, a search ligand could generally be well matched to one or more density clusters, as occurs in the construction of FAD depicted in Fig. 3.

We note that not all graph searches yield a complete model; generally, only a few do so (Fig. 4). For example, the assignment of the pivot atom shown in Fig. 3a to the middle of the sparse cluster (around the location of the phosphates of FAD) in Fig. 3b would not lead to the complete model. The best results were obtained following selection of up to 27 best-scored models and the use of two different sparse representations of each density cluster. The quality of each candidate solution is evaluated through a scoring function that uses inter-atomic distances, van der Waals repulsions, chiral centres and density height at atomic positions.

This results in up to 54 best-scored ligand-to-node assignments; one of them is shown in Fig. 3c. The assignments are the models of the search ligand with distorted stereochemistry, which are tidied up at a later stage.

**Metropolis-based ligand modelling—**Another 54 models are output by the Metropolis routine, and the results are merged with those from label swapping—thus, 108 models are taken forward for further processing. The time taken for model building via the Metropolis algorithm depends on the number of rotatable bonds in the ligands, with the dependency being approximately quadratic in our tests. A ligand with 10 rotatable bonds should be fitted in approximately 8 s, while a ligand with 50 rotatable bonds will require approximately 100 s.

### Real-space refinement

Real-space refinement is applied to each of the 108 models prepared. During the refinement, the density shape of each atom is described by a spherical Gaussian function, where its centre (*xyz*) and width are optimised. In addition, the two parameters scaling the observed electron density to the one for the modelled ligand are refined. Each of the models is optimised to fit density and the stereochemical targets (bond distances, angle-bonded distances and planes). The contributions from the density and the stereochemistry residuals are dynamically weighted to each other. Ribose rings are tested in conformations corresponding to the two puckers—the one providing the best fit to the density is selected. Other ring systems are currently only modelled in the input conformation, and this may be addressed in the future. In these instances, the user may input two different models—for example, one in a “boat” conformation and one in a “chair” conformation for non-aromatic six-membered rings—and use the map correlation coefficient of the built models in order to select the correct conformation.

### Selection of the output ligand

Three sequences of rankings are generated from the refined models: (1) the sum of the density values at atomic coordinates, (2) the r.m.s. shift of the model during the refinement and (3) the goodness of fit of the electron density calculated from the fitted model and the density cluster.

This goodness of fit also indirectly characterises the fit to the stereochemical targets. These three rankings are combined with weights of 0.68, 0.08 and 0.24, respectively. The highest weight corresponds to the density values at ligand atomic positions, which is the most characteristic feature for the quality of the fit. In contrast, the shift of the model during the refinement contributes little and thus has a small weight. These weights were obtained from training on a small test set of ligand ensembles, with the optimisation aiming to place those ligands built with smaller r.m.s.d. higher in the ranking list. Normally, the best single model is taken as the one with the best total rank. However, if there is another model amongst the top 20 that is sufficiently similar to the single best result, both are averaged and the “merged” model is refined again. Such ensemble averaging is repeated iteratively until convergence.

### Validation of the software

**The test set—**The number of unique ligands was 3462 (Fig. 5). Ligands of sizes 5 and 6 were most abundant and alone comprised one-third of the cases. These were mostly phosphates and sulfates.

Approximately 20% of the ligand-building cases concerned structures of low molecular mass (90–200 Da), thus testing the potential of the method to place fragment-like molecules.

The majority of the cases involved larger lead and drug-type ligands. The largest molecule was vancomycin containing 101 non-hydrogen atoms. In only 20 cases had the ligand more than 90 atoms. The resolution of the data ranged from 0.75 to 5.0 Å. However, there were only 10 cases with resolution in the range from 4.0 to 5.0 Å. In more than 50% of the cases, the ligand had a local map correlation of 80% or higher; only 33 cases had map correlations lower than 10%.

**Accuracy of the built ligands—**Overall results for building the largest ligand molecule (if more than one was present) are presented in Fig. 6. Building the ligand that is not the largest (e.g., attempting to build the *N*-omega-nitro-L-arginine before constructing protoporphyrin in the oxyreductase structure 1ed5) has shown only marginally lower performance. For ligands consisting of seven or more non-hydrogen atoms, the identification of their binding site was successful in over 80% of the cases (Fig. 6a)—in the previously reported method[9,10] in which the binding site was simply taken as the largest cluster in difference density, the binding site was only found in 70% of cases for the same type of ligand—and only when the largest ligand was fitted. The higher discrimination provided by our fragmentation tree and shape matching approach is apparently more accurate and rigorous than that used previously. Furthermore, the new approach should permit ligands to be modelled into a map in any order and not only from largest to smallest. When the binding site was determined, accurate construction of the ligand was possible in majority of the cases. For ligands of a size typical of drug molecules—20 to 40 non-hydrogen atoms—that were well seen in the density map, models with an accuracy of better than 1.0 Å were obtained in 75% of the cases, with the binding site being found in 96% of cases (Fig. 6b). When using the earlier method that utilised only a label-swapping routine, only 45% of similarly “good” ligands could be built. The significant improvement achieved is a good measure of the advantages in combining the method with a Metropolis optimisation.

Accurate building of the ligand entity was possible for ligands of diverse sizes. Figure 7 illustrates that, in cases of clear difference electron density, a small sulphur ion is built as accurately as a larger atorvastatin molecule.

A stringent nearest-neighbour r.m.s.d. limit of 1.0 Å was not always the best measure for evaluating success. For example, in the built model of a plant lumazine synthase inhibitor at 3.1 Å (Fig. 8a), an apparently inappropriate modelling of the inhibitor's benzene ring resulted in the incorrect placement of the hydroxylamine group and an apparent r.m.s.d. of 1.9 Å from the deposited ligand. Although we classify such a model as incorrect, we expect that it may be considered sufficiently accurate in many cases. Similarly, building S-adenosyl methionine into a disordered density leads to the observed r.m.s.d. of 2.5 Å. These two cases exemplify the directions in which future developments can be attempted.

## Conclusions

The obtained results convincingly demonstrate the efficiency of the combined approaches, where different methods complement each other. Evidently, ligands can be automatically and successfully built at various levels of crystallographic data quality and ligand complexity. The use of both label-swapping and Metropolis methods is superior to using either alone as the most appropriate algorithm can vary depending on the particular scenario. While the label-swapping method would be expected to be more sensitive to errors in phases as it works on a per-atom basis, the Metropolis search should compensate for this as it operates at the level of whole rigid groups of atoms. At the same time, the Metropolis search does not guarantee a convergence to the global minimum, but the label-swapping method building the model “from the seeds” has inherently high convergence properties. Amongst a subset of 3000 examples for which the deposited ligand was well pronounced in density and

for which the binding site was correctly located using the described approach, 821 were, on average, initially built within 1.0 Å only via the Metropolis method while 220 were accurately modelled only via the label-swapping routine. Amongst the former set, the average data resolution was 2.1 Å compared to 1.8 Å for the latter, indicating the required higher data quality for the label-swapping method as anticipated. Since both methods find their results in different ways, the combined model ensemble has a higher likelihood of containing the correct solution compared to either method alone.

The use of the fragmentation-tree approach in map analysis provides a very efficient means for identification of the binding site. Due to the fact that ligand molecules consist of bonded atoms, whose density shapes overlap (particularly at resolutions typical for MX structure determination), ligand electron density has properties distinct from those of solvent molecules and background noise. Since noise peaks are a result of random constructive interference of the Fourier terms, their spatial correlation is low; hence, their cluster volumes are high but only present at low-density levels. The density iso-surfaces of noise shrink faster as the contour level is increased. The fragmentation-tree-based approach is used to identify the ligand-binding site in cases where it is unknown or where it can be employed for validation purposes. Indeed, the automatic identification of a particular ligand at a particular location in the map may be taken as an indication that it is more likely bound there than at any other site.

We hope to extend our use of the fragmentation tree in the future, as it would appear to offer much information regarding the molecular structures described. One example might be its use to aid the actual identification of ligands from a cocktail of candidates. Most complex organic molecules with 10 or more atoms typically consist of a number of large cores or rigid groups that are connected by bonds around which conformational rotations may be possible. Due to their mobility, these connections may have lower electron density and serve as breakpoints of the density clusters into smaller blobs as the contour level is increased. In a fragmentation tree, this leads to a characteristic breakup pattern that gives hints as to the stereochemistry of a ligand molecule. In cases of ligands of unknown identity, the sizes and chemical content of rigid groups can, to an extent, be estimated from these breakup patterns.

Since the fragmentation tree is based on the overlapping density between bonded atoms, it is less powerful at very high resolutions, around 1.5 Å or better, where individual atoms are well resolved in the density. As a practical measure, for these cases, the difference electron density maps were computed from data truncated to 1.5 Å resolution. Conversely, at a resolution of 3.0 Å or worse, density clusters for ligands start merging with those for unmodelled solvent and noise in the map, which also complicates correct identification and building of ligand model.

The presented algorithms should be exceptionally useful in aiding the convenient and automated placement of ligands into density in crystallographic research, particularly in the area of knowledge-based drug design. In a typical real-life structure determination exercise based on crystallographic data, the researcher will initially build the protein model as fully as possible. At this stage, difference density maps can be prepared that should be of good quality provided that the protein building has been successful. Thus, in most cases, a ligand-building procedure of such broad applicability as presented should be successful in virtually all typical structure analyses. Solvent molecules can be modelled subsequent to the placement of the appropriate ligands in order to complete the structure.

There are a number of topics worthy of further investigation. As any other ordered molecule bound to the protein, ordered water molecules leave their imprint in the difference density map. In the fragmentation tree, the corresponding density blobs appear in the lower volume

regions. Indeed, the best-ordered water molecules define a “water horizon” below which it becomes impossible to identify the origin of the different blobs. Remarkably, the “water horizon” is always present, and its physical nature is known. Thus, it could possibly be used as a calibrating tool to put electron density maps on an absolute scale (electrons per cubic angstrom), in addition to the standard deviation (sigma) currently used. The fitting of partially ordered ligands is a real challenge even when done manually using an experienced researcher as a tool. It is desirable that automated procedures could advise on which groups of a bound molecule are giving rise to the observed density. Another under-explored area with high potential is the use of other shape descriptors (such as the abovementioned Zernike moments) that would be tailored to a particular family or stereochemical group of ligand candidates. Finally, there is a wealth of chemical information provided by the protein environment around the binding site. Based on our test set, we estimate that approximately 3% of ligands in the PDB are small fragment-type compounds that are shape symmetric and therefore have an ambiguity in atom placement that cannot be resolved solely using shape. These cases may require user intervention. The matching of ligands to hydrogen bond donors/acceptors and hydrophobic pockets in the protein, as widely used in docking approaches and evaluated using typical scoring functions,[21,22] should also aid the task of ligand modelling in MX, especially in these specific instances.

## Methods

### Selection of cluster points in the fragmentation tree

We note that one does not know in advance the contour level at which the similarity of the density cluster at the correct location to the search ligand is maximised. Therefore, we inspect the density at different contour levels $t$ ranging from 1.0 to 6.0 sigma above the mean in steps of 0.05 sigma. All clusters at every contour level are treated independently. The signal-to-noise ratio is very low at this stage, typically around 0.001, that is, for one correct ligand-binding density cluster, there are ~1000 other clusters.

At each density threshold, we consider the 11 clusters with the highest volume and select all branches of the fragmentation tree in which these clusters are located (examples of such branches are the sequences of the blue-, green- or red-coloured points in Fig. 2). We chose to select 11 clusters as testing showed that, in this case, there is an average 95% probability of including the correct cluster therein; further increases in selection size did not improve the results significantly. The selected branches are filtered so that, at any density threshold, they lie within the expected volume limits from $N$ to $10N$ Å$^3$ where $N$ is the number of non-hydrogen atoms in the search ligand. The described branch filtering provides an approximately 100-fold reduction in the number of cluster candidates.

### Creating an electron density from the input ligand

The $xyz$ coordinates of all atoms of the search ligand are used to generate an electron density blob trimmed to the resolution of the X-ray data. During this density generation, series terminations that result from truncation of electron density data are modelled by convoluting the density with a Gaussian kernel—this is equivalent to the application of an excess $B$-factor that introduces a required resolution-dependent smearing factor. A scaling factor is also used to ensure that the integral of the density distribution is appropriate. The electron density grid spacing is typically 0.5 Å; cell angles are those of the input map.

### Sparse density clusters

A sparse density cluster, created as now described, is used to model density clusters, both for calculating shape matches and for ligand building via label swapping. A density cluster corresponding to a ligand usually contains many more grid points than there are atoms in

that ligand. For example, at 1.8 Å resolution, the cluster for a 15-atom NRG molecule (Fig. 2) in the density contoured at 2.0 sigma has a volume of 58 $Å^3$, which corresponds to 268 grid points of the map at 0.6 Å grid spacing. Rather than processing all cluster grid points, we transform them into a pseudo-skeleton containing a smaller number of points. This transformation is similar to the reduction of the problem's complexity using the free atoms' concept of ARP/wARP[18], where free atoms with no particular chemical identity are placed into density at approximately inter-atomic distances. To build a pseudo-skeleton, we select a high-density pivot grid point, and we remove all neighbouring grid points within a radius of 1.1 Å. The next pivot is selected as the highest-density point within the distance range 1.1–1.6 Å to the previous pivot(s). This is iterated, resulting in a set of points that capture the spatial distribution of the density cluster; we denote such a representation a *sparse density cluster*. The number of points in the sparse cluster is set to be always higher than the number of ligand atoms. Since the distance range 1.1– 1.6 Å covers all bond lengths typically occurring in ligands, such a sparse cluster can be seen as a pseudo-molecule of interconnected atoms (Fig. 9), which can be directly compared to the structure of the search ligand.

### Automatic detection of ligand stereochemistry

A ligand connectivity tree that describes rigid groups and overall stereochemistry is generated automatically as follows. Pairs of atoms located at less than 2.5 Å distance are considered as potential bonding partners. Should the angle between two bonds involving the same atom be smaller than 80°, the longer bond is removed from the connectivity table. A graph search is applied to the connectivity table to identify closed polygons (*rings*). The hybridisation state of bonded atoms is inspected using a number of criteria. Inter-atomic distances are compared against tabulated distances for single, double and triple bonds between the most common elements present in carbon-containing compounds. Bonding angles close to 109°, 120° or 180° are taken as indications of $sp^3$, $sp^2$ or *sp* hybridisation. The local planarity is evaluated through a least-squares plane fitted to the coordinates of an atom together with its three neighbours. If none of the atoms deviates from the plane by more than 0.1 Å, the group of such connected atoms is considered locally planar. While such checks have been shown to be extremely robust in our tests, the input coordinates should be sensible—otherwise, evaluation of stereochemistry may be adversely affected.

### Shape features for choosing the appropriate binding site

The following seven shape features are used to compare the density modelled from the input ligand to the density clusters found using the fragmentation tree:

1. The ratio of surface points to the total number of points in a cluster. Surface points are those with at least one adjacent grid point having a density value below the threshold, with grid spacing always being set as close to 0.5 Å as possible;
2. The dimensions of the smallest rectangular box that fully encloses the cluster;
3. The eigenvalues of the moment of inertia tensor of the cluster about its centre of mass (computed from density-weighted *xyz* coordinates; the density values are always positive due to their thresholding at 1.0 sigma above the mean or higher);
4. The sum of the nonoverlapping volumes of the ligand and the cluster after their rigid-body superposition;
5. The eigenvalues of the sample covariance matrix of the mean-centred *xyz* coordinates of the points constituting the sparse cluster; similarly eigenvalues are computed for the search ligand model;

6. The difference vector between the two inter-atomic distance histograms of ligand and sparse cluster—treating all sparse cluster points as pseudo-atoms; both histograms are binned identically in steps of 0.4 Å up to a maximum distance of 40 Å; and
7. The same difference vector [as in (6)], calculated using geodesic distances. The geodesic distance is defined as the sum over bonded distances between atoms (or adjacent cluster points for a sparse density cluster) along a connectivity path through the molecule that links any two atoms; the shortest such path is selected.

Features 4, 5, 6 and 7 use the ligand molecule in the given conformation. For each feature except feature 4, the individual scores $S_i$ are computed as follows:

$$S_i = \exp\left(-k_i \cdot \sum_j (f_{ij}^{(o)} - f_{ij}^{(c)})^2\right) \quad (3)$$

where $k_i$ are empirical weights for each feature *i, j* denotes the summation index of the dimensions of each feature vector (1 for feature 1; 3 for features 2, 3 and 5; 200 for features 6 and 7) and $f^{(o)}$ and $f^{(c)}$ are the feature values for the observed and calculated clusters of the ligand density, respectively. These seven different features are at different scales, but each has a characteristic range of values for correct and incorrect cluster–ligand matches: the value of $S_i$ is between 0 and 1, approaching the latter for a perfect match. The individual scores $S_i$ are combined to yield the total score for each examined cluster:

$$S_{\text{ranking}} = q_0 + q_1 S_1 + \ldots + q_I S_I + q_{I+1} S_1^2 + q_{I+2} S_1 S_2 + \ldots + q_{\frac{(2+I)!}{2I!}-1} S_I^2 \quad (4)$$

The total score is a quadratic classifier and is expressed as a product of the weight vector $q$ and the feature vector $S$. The values of the vector $q$ were trained using a set of 1000 protein–ligand complexes with data resolution ranging from 1.3 to 3.1 Å and ligand sizes from 5 to 60 atoms. Only structures containing fully occupied ligands and those having local map correlation of higher than 75% were selected for the training set. The optimisation was done through a random 5000-step walk in the weight space; at each step, 20 random updates of the weight vector were generated, and the best update was accepted if it yielded a score (the sum of $S_{\text{ranking}}$ values for the 1000 training cases) higher than the one at the previous step. Thus, the optimisation targeted a maximisation of the total score across the range of all complexes in the training set.

## Label swapping

This method was introduced by Zwart *et al.*,[9] and only a brief description of the most crucial aspects of the procedure is given here. The task is to find the subset of a sparse cluster, a subgraph, which best matches the ligand. This may be seen as “swapping” the identities of the ligand atoms when they are mapped to a subgraph.

Such a procedure starts with a selection of the pivot ligand atom (Fig. 3a), which is successively assigned to each node of the sparse cluster (Fig. 3b). The search problem is thus split into as many smaller subtasks as there are nodes in the sparse cluster. From all trial models generated within each subtask at every step, only a subset with the highest scores need typically be kept for further extensions.

### Metropolis search

A Metropolis type of optimisation is used in various implementations of crystallographic ligand building.[5–7] It makes use of the conformational freedom of the ligand molecule around its rotatable bonds in order to optimally match it to the density cluster. In our implementation, the ligand in the conformation input by the user is placed into the density cluster to match its centre of mass and three principal axes (one out of four possibilities), and an initial score is calculated. This score reflects a density map correlation between the ligand and the sparse cluster.

The initial model is then rotated to sample all orientations in steps of 60°, with each orientation being subject to 100 steps of Metropolis optimisation of the score. The 12 best solutions are taken to the next round and are submitted to a further Metropolis optimisation at three different temperatures. The initial temperature is proportional to the initial score; after 4000 steps of optimisation at this initial temperature, 5000 steps are carried out at half the initial temperature and finally 200 steps at one-tenth of the initial temperature. The Metropolis optimiser works with an ensemble of randomly created initial models. From this ensemble, the best 54 models are output and merged with the result of the label-swapping algorithm.

### Data and software used

Model refinement and map calculation were done using REFMAC,[23] FFT and MAPMASK software from the CCP4 package.[24]

The developed methods were evaluated on a large set of ligand structures from the PDB using version 7.2 of the ARP/wARP software. The diffraction data were taken from the EDS (*E*lectron *D*ensity *S*erver)[25] and ligand coordinates from the heterocompound information centre in Uppsala, HIC-Up.[26] The EDS entries were filtered by the following criteria:

- **a.** the structure contains a protein and at least one ligand with five or more non-hydrogen atoms;
- **b.** the structure does not contain DNA/RNA chains; and
- **c.** at least one of the ligands in the structure must match the HIC-Up database with the compound name, the number of atoms and their stereochemical description.

Overall, 13,985 PDB entries containing 20,568 ligands were selected.

In order to eliminate model bias and closely mimic the real-life situation occurring in crystal structure determination, we removed all HETATM atoms including solvent from the PDB files and subjected the remainder to one cycle of restrained refinement with REFMAC. The difference electron density maps were then calculated for further analysis and ligand building.

During evaluation of software performance, a density cluster was interpreted as having been found correctly if it had at least one density grid point within 1 Å distance from any of the ligand atoms from the deposited model. After final model building, the rebuilt ligand models were compared to their PDB deposited structures, and the nearest-neighbour r.m.s.d. was computed. The PDB models were considered as absolutely correct reference structures, and an r.m.s.d. to them lower than 1.0 Å was interpreted as a successfully built ligand.

## Acknowledgments

This work was supported in part by the US National Institutes of Health grant R01 GM62612 to G.L. and G.E., and by the German Federal Ministry of Education and Research grant 05K10YEA to C.C. C.C. would also like to thank the European Molecular Biology Laboratory for an interdisciplinary postdoctoral fellowship (EIPOD).

## References

1. Mattos C, Ringe D. Locating and characterizing binding sites on proteins. Nat Biotechnol. 1996; 14:595–599. [PubMed: 9630949]
2. Hartshorn MJ, Murray CW, Cleasby A, Frederickson M, Tickle IJ, Jhoti H. Fragment-based lead discovery using X-ray crystallography. J Med Chem. 2005; 48:403–413. [PubMed: 15658854]
3. Murray CW, Callaghan O, Chessari G, Cleasby A, Congreve M, Frederickson M, et al. Application of fragment screening by X-ray crystallography to β-secretase. J Med Chem. 2007; 50:1116–1123. [PubMed: 17315856]
4. Bosch J, Robien MA, Mehlin C, Boni E, Riechers A, Buckner FS, et al. Using fragment cocktail crystallography to assist inhibitor design of *Trypanosoma brucei* nucleoside 2-deoxyribosyltransferase. J Med Chem. 2006; 49:5939–5946. [PubMed: 17004709]
5. Wlodek S, Skillman AG, Nicholls A. Automated ligand placement and refinement with a combined force field and shape potential. Acta Crystallogr, Sect D: Biol Crystallogr. 2006; 62:741–749. [PubMed: 16790930]
6. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr, Sect D: Biol Crystallogr. 2010; 66:486–501. [PubMed: 20383002]
7. Oldfield TJ. X-LIGAND: an application for the automated addition of flexible ligands into electron density. Acta Crystallogr, Sect D: Biol Crystallogr. 2001; 57:696–705. [PubMed: 11320310]
8. Terwilliger TC, Klei H, Adams PD, Moriarty NW, Cohn JD. Automated ligand fitting by core-fragment fitting and extension into density. Acta Crystallogr, Sect D: Biol Crystallogr. 2006; 62:915–922. [PubMed: 16855309]
9. Zwart PH, Langer GG, Lamzin VS. Modelling bound ligands in protein crystal structures. Acta Crystallogr, Sect D: Biol Crystallogr. 2004; 60:2230–2239. [PubMed: 15572776]
10. Evrard GX, Langer GG, Perrakis A, Lamzin VS. Assessment of automatic ligand building in ARP/wARP. Acta Crystallogr, Sect D: Biol Crystallogr. 2007; 63:108–117. [PubMed: 17164533]
11. Aishima J, Russel DS, Guibas LJ, Adams PD, Brunger AT. Automated crystallographic ligand building using the medial axis transform of an electron-density isosurface. Acta Crystallogr, Sect D: Biol Crystallogr. 2005; 61:1354–1363. [PubMed: 16204887]
12. Binkowski TA, Cuff M, Nocek B, Chang C, Joachimiak A. Assisted assignment of ligands corresponding to unknown electron density. J Struct Funct Genomics. 2010; 11:21–30. [PubMed: 20091237]
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
14. Cai W, Shao X, Maigret B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. J Mol Graphics Modell. 2002; 20:313–328.
15. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics. 2005; 21:2347–2355. [PubMed: 15728116]
16. Grandison S, Roberts C, Morris RJ. The application of 3D Zernike moments for the description of "model-free" molecular structure, functional motion, and structural reliability. J Comput Biol. 2009; 16:487–500. [PubMed: 19254186]
17. Venkatraman V, Chakravarthy PR, Kihara D. Application of 3D Zernike descriptors to shape-based ligand similarity searching. J Cheminf. 2009; 1:19.
18. Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat Protoc. 2008; 3:1171–1179. [PubMed: 18600222]

19. James RW. False detail in three-dimensional Fourier representations of crystal structures. Acta Crystallogr. 1948; 1:132–134.

20. Stenkamp RE, Jensen LH. Resolution revisited: limit of detail in electron density maps. Acta Crystallogr. 1984; 40:251–254.

21. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins. 2002; 47:409–443. [PubMed: 12001221]

22. Guo J, Hurley MM, Wright JB, Lushington GH. A docking score function for estimating ligand–protein interactions: application to acetylcho-linesterase inhibition. J Med Chem. 2004; 7:5492–5500. [PubMed: 15481986]

23. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr, Sect D: Biol Crystallogr. 1997; 53:240–255. [PubMed: 15299926]

24. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. Acta Crystallogr, Sect D: Biol Crystallogr. 1994; 50:760–763. [PubMed: 15299374]

25. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. The Uppsala Electron-Density Server. Acta Crystallogr, Sect D: Biol Crystallogr. 2004; 60:2240–2249. [PubMed: 15572777]

26. Kleywegt GJ, Jones TA. Databases in protein crystallography. Acta Crystallogr, Sect D: Biol Crystallogr. 1998; 54:1119–1131. [PubMed: 10089488]

## Abbreviations used

| | |
|---|---|
| **MX** | macromolecular crystallography |
| **PDB** | Protein Data Bank |

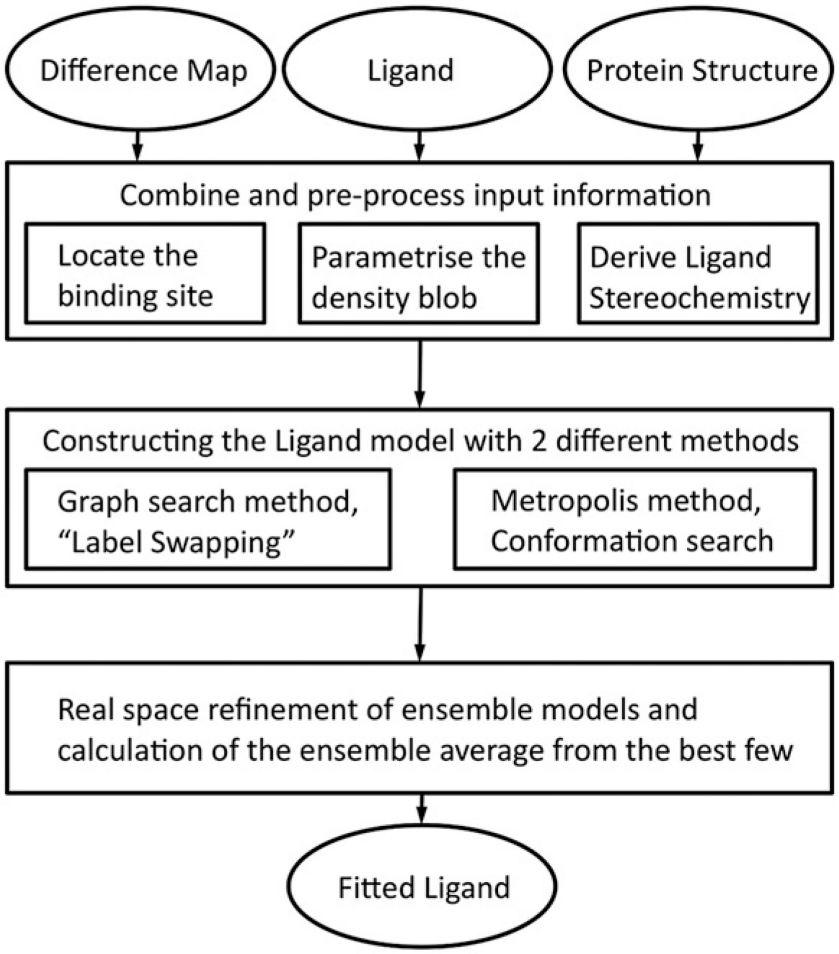


**Fig. 1.**
Schematic representation of the three-step procedure of crystallographic ligand building in ARP/wARP.

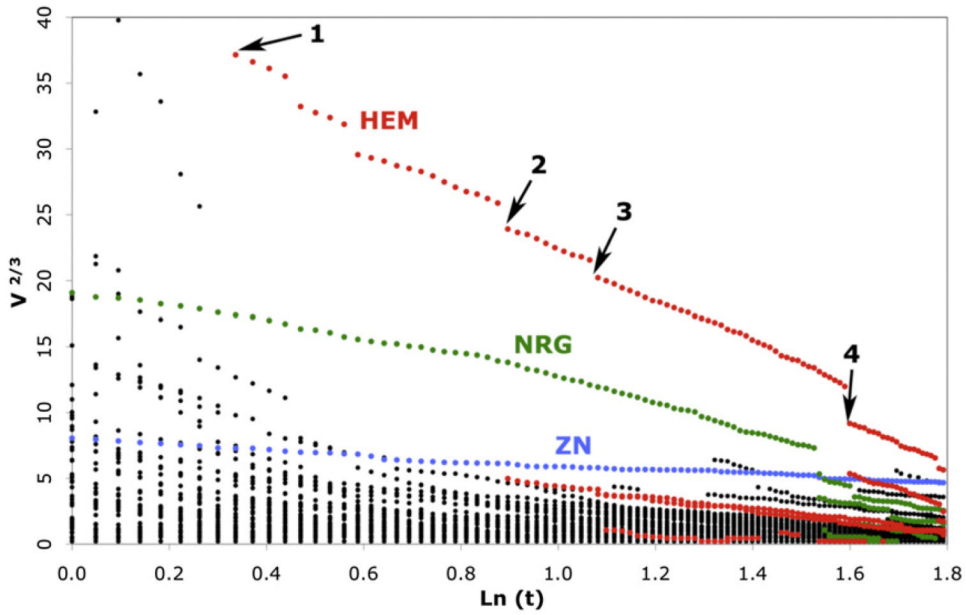


**Fig. 2.**

Fragmentation tree of the difference electron density of the oxyreductase structure 1ed5. The clusters for a protoporphyrin IX ring (HEM), *N*-omega-nitro-L-argi-nine (NRG) and a zinc ion (ZN) ion are shown in red, green and blue, respectively. Other features in the density map, including water molecules, are coloured black.

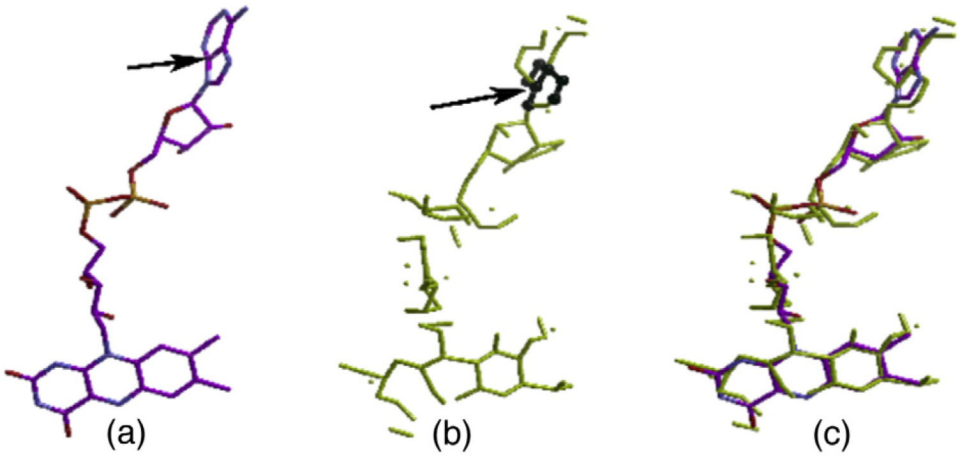


**Fig. 3.**

Construction of the FAD molecule (PDB code 2gf3) using the label-swapping algorithm: (a) the search ligand, (b) its sparse density cluster and (c) one high-scored subgraph-matching solution. The arrow in (a) points to the pivot ligand atom, which is assigned to each node. The arrow in (b) points to the group of nodes (marked as balls) from which the expansion leads to a complete model of the ligand (see also Fig. 4).

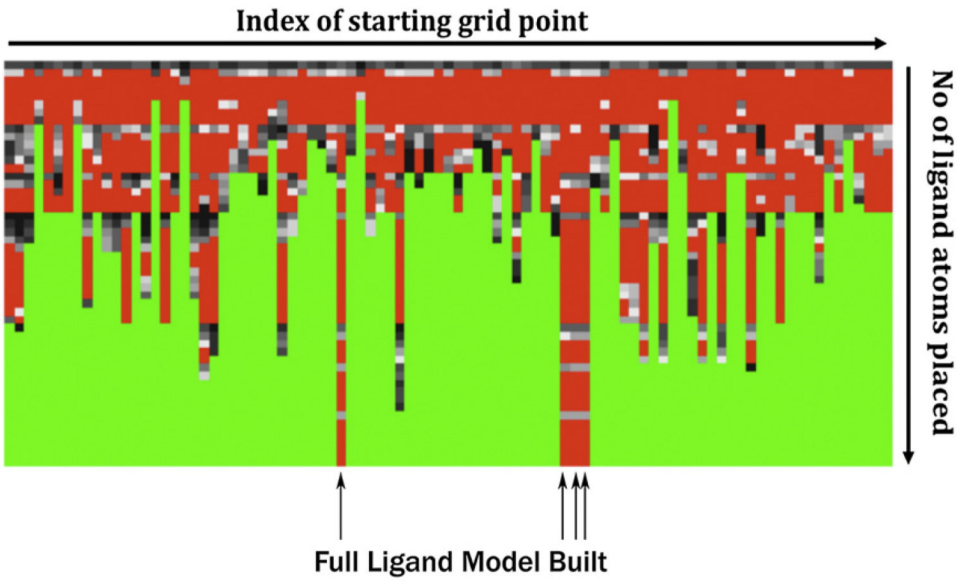


**Fig. 4.**

Matching of the FAD molecule (Fig. 3a) to the sparse grid shown in Fig. 3b. The evolution of the model building process for each starting node (of 98 nodes in total) is plotted on the horizontal axis. The number of ligand atoms assigned to a sub-cluster is on the vertical axis. The number of candidate models is delineated as follows: green stands for 0 models in a "stack" and red indicates that only the top 100 models are kept for further expansion; grey scale colours indicate intermediates between both extremes. Only the few labelled starting grid nodes allow for complete assignment of all ligand atoms.

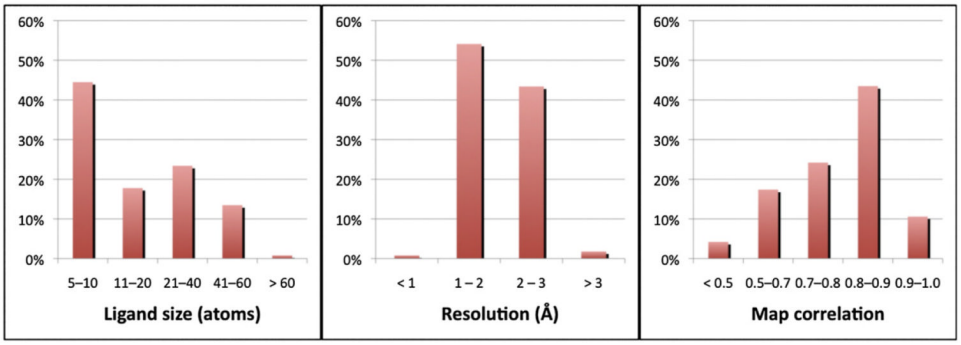


**Fig. 5.**
Characteristics of the ligand-building test set.

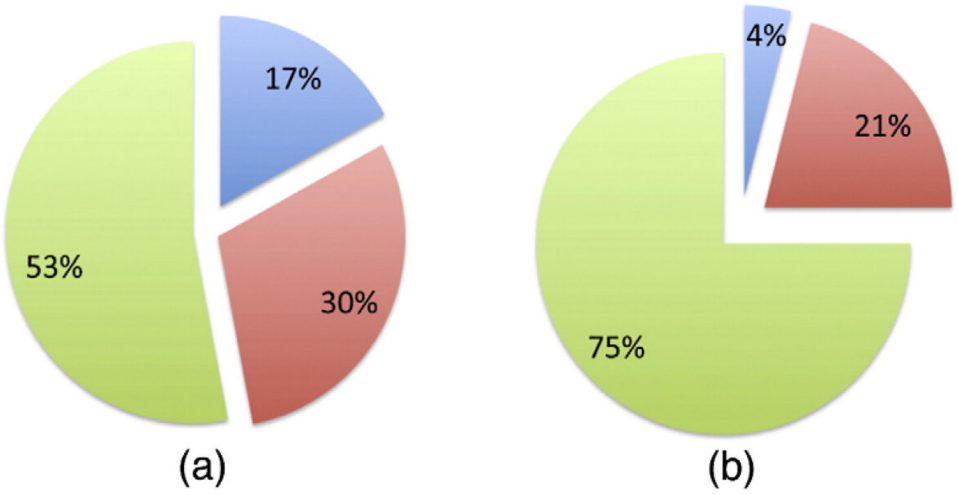


**Fig. 6.**

The overall performance of the ligand-building procedure. Green areas denote successful building with an r.m.s.d. from the PDB model of less than 1.0 Å, red indicates building at a correctly identified binding site but with an r.m.s.d. higher than 1.0 Å, blue areas correspond to ligand models built in the wrong place, (a) 9389 cases with seven or more non-hydrogen atoms and (b) 2773 cases with ligand well-pronounced in the density (real-space map correlation of 0.8 or higher) and sizes from 20 to 40 atoms.

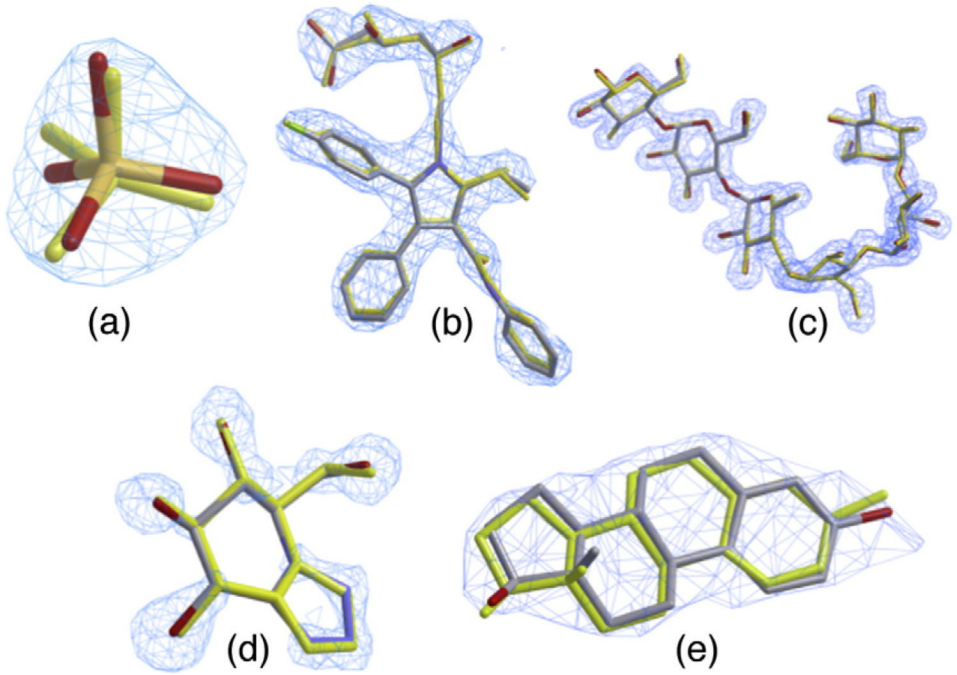


**Fig. 7.**

Examples of ligands of diverse sizes built in maps at various resolutions—deposited ligands are shown in atom colour, built ligands in yellow; the maps are contoured at a level of 1.5 sigma above the mean. (a) A sulphate ion in bovine pancreatic ribonuclease A built at 1.6 Å resolution with an r.m.s.d. of 0.45 Å to the reference structure (PDB code 1a5p); (b) the anti-cholesterol agent atorvastatin, bound to its biological target, HMG coenzyme A reductase (1hwk), at 2.2 Å with an r.m.s.d. of 0.23 Å; (c) a hexasaccharide ligand with 65 non-hydrogen atoms bound to a bacterial α-amylase (1qho) rebuilt with an r.m.s.d. of 0.31 Å at 1.7 Å resolution; (d) a transition-state analogue of a plant enzyme, myrosinase, (1e6q) built with a coordinate accuracy of 0.22 Å in a map at 1.35 Å; (e) 17-–-estradiol built with an r.m.s.d. of 0.31 Å in the map derived from a complex with the human estrogen receptor at 3.1 Å resolution (1ere).

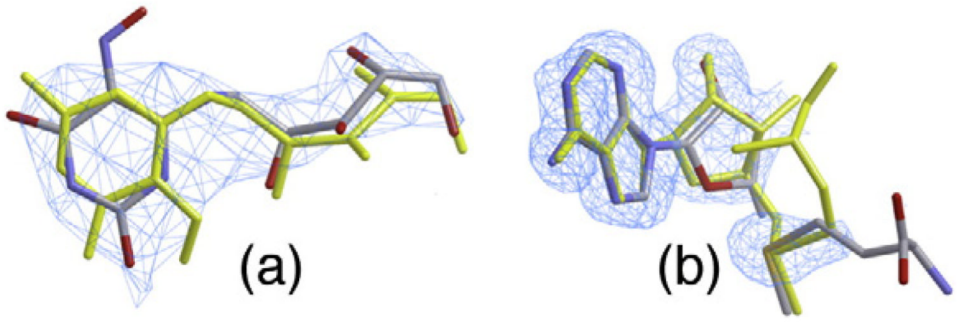


**Fig. 8.**

(a) A plant lumazine synthase inhibitor built into a 3. 1-Å protein structure (PDB 1c41), to an r.m.s.d. of 1.9 Å from the deposited ligand. (b) *S*-Adenosyl methionine modelled in a tRNA methyltransferase enzyme (PDB 1v2x); a long flexible aliphatic chain was apparently disordered, leading to little density to guide its placement. The variation in atom placement (the deposited model is shown in grey, and the built model is shown in yellow) results and explains the r.m.s.d. of 2.5 Å observed.

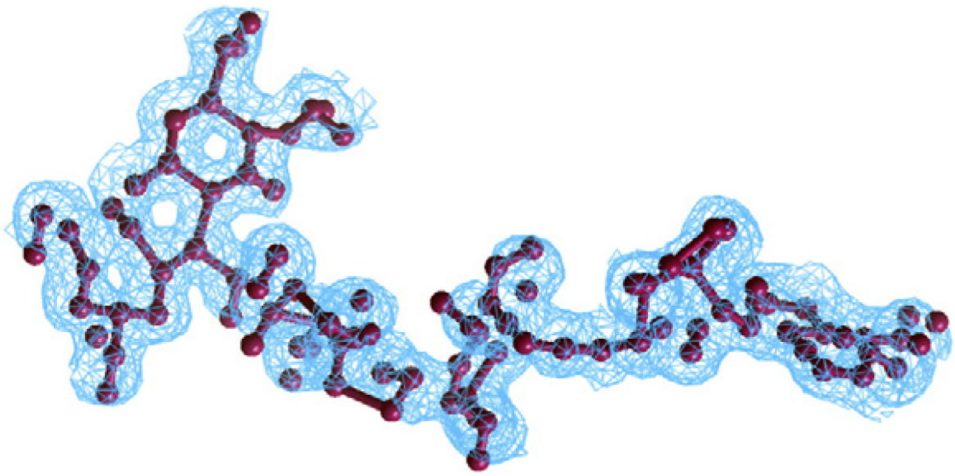

**Fig. 9.**

Sparse representation (magenta balls and sticks) of an electron density cluster (blue wire) for the FAD ligand in the structure 2gf3 of sarcosine oxidase at 1.3 Å resolution.