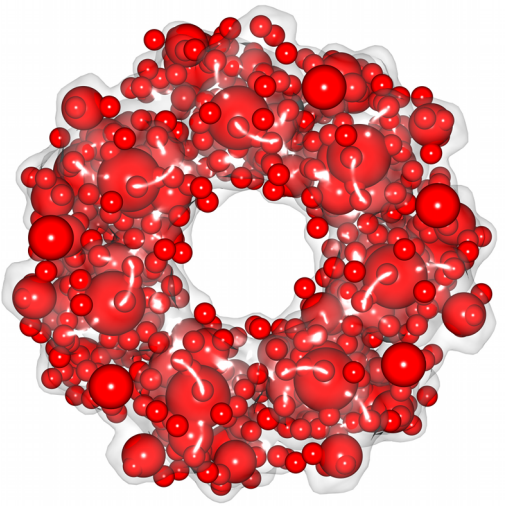
Similarity Searching in Macromolecular Electron Density Maps



Dissertation
zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik der Universität Hamburg

vorgelegt von

Axel Griewel

geboren in Hamm-Heessen.

Hamburg, 2011

Genehmigt von der MIN-Fakultät Fachbereich Informatik der Universität Hamburg auf Antrag von

Prof. Dr. Matthias Rarey (Erstgutachter) und

Prof. Dr. H. Siegfried Stiehl (Zweitgutachter)

Hamburg, 21. Dezember 2011

The title shows an isosurface of a synthetic electron density map of the bacterial chaperonin GroEL [34] as transparent surface. The red spheres of different sizes indicate keypoints detected at different scales. (© A. Griewel)

Abstract

From the very beginning of biology, the interpretation of images has been a major driving force of discovery. Especially the method of comparing images and identifying similarities and differences is widely used in various fields for the study of macroscopic objects. This work applies the generic approach of similarity searching to images of sub-nanometer scale objects, namely macromolecular electron density maps. Among others, these maps can be acquired by X-ray crystallography and single particle cryo-electron microscopy, and up to now more than 67 000 three-dimensional atomic structures of biomolecules have been elucidated using these methods. The comparison of atomic structures gives insight into various questions such as evolutionary relations between organisms and it can help in the process of drug development. Comparing biomolecules by registering the experimental electron density maps is a complex task since high resolution maps are exceedingly intricate. Furthermore, manual inspection of all available maps is impossible due to the sheer amount of maps. Therefore, methods for an automated, efficient, and accurate comparison of electron density maps are required. This work addresses the mentioned problem and introduces a method that is implemented in a software system coined *siseek*, which is geared to solving the problem of similarity searching in macromolecular electron density maps.

siseek is based on the scale-invariant feature transform and locates keypoints — image features — in salient, spherical regions of a given map. Each keypoint is assigned discrete orientations, which are determined based on the gradient in the keypoint's neighborhood. Orientations, in turn, are used for the computation of local neighborhood descriptors, which enable the identification of similar local neighborhoods in maps. Based on this information, map registration — the superposition of similar parts of two maps — is facilitated. Furthermore, an approach for molecule recognition based on feature vector similarity is described. *siseek* is parameterized in several large scale studies using a set of synthetically generated maps. The performance of *siseek* is first assessed by docking molecular subunits to distorted, synthetic maps of their corresponding assemblies. These experiments show that *siseek* is able to successfully locate atomic structures in intermediate and high resolution electron density maps requiring less time than other approaches. This finding is confirmed by exemplifying that *siseek* correctly registers atomic structures to various experimental maps acquired by single particle cryo-electron microscopy and X-ray crystallography. Additionally, pairs of X-ray crystallography maps are successfully registered by *siseek*. Furthermore, *siseek* is used in a proof of concept to identify molecules depicted in electron density maps. The experiments demonstrate that *siseek* facilitates similarity searching in macromolecular electron density maps and show that the method can be used to aid the process of interpreting these maps.

Zusammenfassung

Seit den Anfängen der Biologie liefert die Interpretation von Bilddaten wertvolle Erkenntnisse. Insbesondere der Vergleich von Bildern wird häufig für die Analyse makroskopischer Strukturen eingesetzt und ermöglicht es Ähnlichkeiten und Gemeinsamkeiten der dargestellten Objekte festzustellen. In dieser Arbeit wird die generische Methode der Ähnlichkeitssuche auf Objekte aus dem sub-nanometer Bereich angewendet, die in makromolekularen Elektronendichtekarten abgebildet sind. Diese Karten können z. B. aus Röntgenstrukturanalyse- und Kryo-Elektronenmikroskopie-Experimenten gewonnen werden und haben bislang die Bestimmung der dreidimensionalen, atomaren Struktur von mehr als 67 000 Biomolekülen ermöglicht. Aufgrund des hohen Detailgrades hochaufgelöster Elektronendichtekarten ist der Vergleich der Karten nicht trivial. Des weiteren liegt bei der großen Anzahl der Karten eine manuelle Analyse ohnehin nicht im Bereich des Praktikablen. Aus diesen Gründen sind Methoden zur automatisierten, effizienten und genauen Analyse von Elektronendichtekarten notwendig. In der vorliegenden Arbeit wird diese Fragestellung behandelt und eine Methode vorgestellt, die in einem Softwaresystem mit dem Namen *siseek* implementiert wurde und der automatisierten Ähnlichkeitssuche in makromolekularen Elektronendichtekarten dient.
siseek basiert auf der „scale-invariant feature transform“ und identifiziert Schlüsselpunkte („keypoints“) in hervorstehenden, kugelförmigen Bereichen einer Karte. Jedem Schlüsselpunkt werden Orientierungen auf Basis des umliegenden Gradientenfeldes zugewiesen, die wiederum dazu verwendet werden lokale Nachbarschaftsdeskriptoren zu berechnen. Jeder Deskriptor wird als Merkmalsvektor betrachtet und zur Berechnung der Ähnlichkeit zwischen den zugrundeliegenden Nachbarschaften genutzt. Basierend auf dieser Ähnlichkeit wird die Registrierung, also die Überlagerung ähnlicher Teile, von zwei Karten ermöglicht. Zusätzlich wird auf Basis von *siseek* ein Verfahren zur automatisierten Erkennung eines in einer Elektronendichtekarte dargestellten Moleküls vorgestellt.
siseek wurde in mehreren groß angelegten Studien mit Hilfe von synthetischen Karten parametrisiert und in Docking Experimenten mit synthetischen und experimentellen Karten analysiert. Diese Experimente zeigen, dass *siseek* zur Ähnlichkeitssuche in Karten mit mittlerer und hoher Auflösung eingesetzt werden kann und dabei weniger Zeit als andere Ansätze benötigt. Dieser Befund wird durch Registrierungsexperimente mit experimentell gewonnenen Karten aus Röntgenstrukturanalyse und Kryo-Elektronenmikroskopie untermauert. Außerdem wird eine Machbarkeitsstudie zur automatisierten Molekülerkennung in Elektronendichtekarten vorgestellt. Die Experimente zeigen, dass *siseek* zur Registrierung von Elektronendichtekarten eingesetzt werden und somit deren Interpretation unterstützen kann.

Contents

List of Abbreviations

Å Ångstrom — Unit of length equal to 10^{-10} m

°C Degree Celsius — Unit of temperature

C_α The α-carbon of an amino acid

2D Two-dimensional

3D Three-dimensional

B Byte — Standardized unit of digital information

CATH Class, Architecture, Topology, Homologous superfamily — A protein structure classification [254]

CLI Command Line Interface

CoA Coenzyme A

CPU Central Processing Unit

cryo-EM Cryo-Electron Microscopy

CTF Contrast Transfer Function

Da Dalton — Unit of mass equal to approximately $1.66 \cdot 10^{-24}$ g, also equal to the unified atomic mass unit (u)

DNA Desoxyribonucleic acid

DoG Difference of Gaussians

EMD EMDataBank — A public resource providing access to cryo-electron microscopy maps [196]

FPR False Positive Rate

FSC Fourier Shell Correlation Coefficient

FSSP Families of Structurally Similar Proteins — A protein structure classification [146]

G Giga — Unit prefix meaning multiplication by 10^9

g Gram — Standardized unit of mass

Gi Gibi — Unit prefix meaning multiplication by 2^{30}

h Hour — Unit of time equal to 60 minutes and 3 600 seconds

Hz Hertz — Standardized unit of frequency

ID Identifier

k Kilo — Unit prefix meaning multiplication by 10^3

Ki Kibi — Unit prefix meaning multiplication by 2^{10}

LoG Laplacian of Gaussian

M Mega — Unit prefix meaning multiplication by 10^6

m Metre — Standardized unit of length

m Milli — Unit prefix meaning multiplication by 10^{-3}

min Minute — Unit of time equal to 60 seconds

mol Mole — Standardized unit of the amount of a substance

n Nano — Unit prefix meaning multiplication by 10^{-9}

NMR Nuclear Magnetic Resonance

p Pico — Unit prefix meaning multiplication by 10^{-12}

pH Measure for the acidity of an aqueous solution

PM-correlation coefficient Pearson product-moment correlation coefficient

PSF Point Spread Function

RMSD Root Mean Square Deviation

RNA Ribonucleic acid

s Second — Standardized unit of time

SCOP Structural Classification Of Proteins — A protein structure classification [247, 66, 6, 7]

SIFT Scale-Invariant Feature Transform

SNR Signal to Noise Ratio

Ti Tebi — Unit prefix meaning multiplication by 2^{40}

tRNA Transfer ribonucleic acid

u Unified atomic mass unit — equal to approximately $1.66 \cdot 10^{-24}$ g, also equal to the unit Dalton (Da)

wwPDB Worldwide Protein Data Bank — A public resource providing access to atomic structures [23]

1. Introduction

From the very beginning of biology, the interpretation of images has been a major driving force of discovery. The analysis of differences and similarities in image data facilitates the detection of common features and divergences between structures found in living organisms. With modern imaging technologies, it is possible to depict structures of organisms at various scales in all three spatial dimensions [10, 355]. At the largest scale, organs and tissues and their interplay in the living organism are analyzed [234]. At the next smaller scale, the functioning of the elementary unit of all life, the cell, is examined. The disciplines of molecular and structural biology are concerned with the analysis of data at the lowest scale of general biological interest. They analyze the molecular foundations that make life viable [292, 17, 99].

Molecular structures are elucidated using various experimental techniques. The currently most frequently utilized method for the analysis of biomolecules is X-ray crystallography, which is based on the interpretation of reflection patterns that are converted to electron density maps [273]. The maps depict the spatial distribution of the molecule's electrons and are the basis for the deduction of atomic models. Single particle cryo-electron microscopy (cryo-EM) is a complementary technique, which also yields electron density maps. In contrast to X-ray crystallography, cryo-EM is better suited for the study of larger molecular assemblies and depicts the Coulomb potential of the biomolecule [94]. Most of the available cryo-EM maps to date have non-atomic resolution. However, the computing power that is available today and tremendous improvements in the method of cryo-EM over the last few years enabled the computation of high resolution cryo-EM maps of macromolecular complexes that allow for discerning single residues. For both, X-ray crystallography and cryo-EM, the objective is the determination of atomic models from electron density maps, which can then be used to analyze the structural basis of life on the atomic level [94, 273]. The long-term goal here is to provide a molecular view of life [24]. This does not only yield new insights into the functioning of biological processes but also enables more purposeful interventions — e. g., by developing new drugs targeted to specific proteins.

High resolution macromolecular electron density maps appear crowded due to the amount of information depicted and are thus hard to interpret by visual in-

spection. Furthermore, the amount of data gathered in various projects of structural biology is enormous and renders manual inspections of electron density maps impossible. Thus, automated methods for analyzing the harvested data are needed. In essence, electron density maps are three-dimensional (3D), gray-level images that depict the spatial arrangement of the atoms of the molecule of interest [94, 273]. The automated analysis of images in general has been improved tremendously over the last few years. Besides numerous approaches for the analysis of two-dimensional (2D) images [119, 115], the *registration* of 3D images is meanwhile also in the focus of digital image analysis [134, 119]. The task of image registration is to determine an alignment of two given images so that similar regions in the images are superposed. Solutions to the registration problem are of major interest for the interpretation of medical images, where healthy and diseased states of organs and tissues are studied by analyzing differences and similarities of registered 3D images [134].

The task of identifying structures in 3D electron density maps has been addressed in the context of different projects. The high resolution maps obtained by X-ray crystallography are interpreted using various image analysis techniques [273]: For a depicted protein, the known polypeptide chain is fitted to the electron density map. The knowledge of the chemical structure and the limited flexibility of the protein allow for the determination of the conformation of the molecule based on the density. In regions that are less well resolved, rigid molecular building blocks like annulated ring systems can be fitted to the density. Since the size of these systems is larger than for short, aliphatic structures, the location of the object can be more easily identified. In cryo-EM, the resolution of the maps is generally lower and does not allow for the direct interpretation in atomic detail. The low resolution of cryo-EM maps is due to various reasons, the most obstructive one being conformational flexibility of the depicted molecule. While in X-ray crystallography the depicted molecule is generally forced into one conformation, cryo-EM allows for the depiction of molecules in different conformational states. Averaging maps of proteins in different conformations reduces the resolution of the electron density maps and hence prevents the direct determination of atomic structures. Thus, these maps are frequently interpreted by fitting larger, rigid parts of known biomolecules to the density map [94].

The interpretation of electron density maps has contributed to the large archive of known molecular structures, which is now curated as the Worldwide Protein Data Bank (wwPDB) [23]. This database is the primary resource for atomic detail information of biomolecules and comprises currently more than 76 000 structures [383], which in large parts have been elucidated by X-ray crys-

tallography but also by cryo-EM and other techniques. Atomic models of the first resolved structure of myoglobin, a fairly small protein, up to the engineered models of the ribosome, which is among the largest molecular machines found in living organisms, are accessible through this resource. Additionally, the Electron Density Server [178] provides currently access to more than 48 000 experimental X-ray crystallography electron density maps, which were the basis for the determination of atomic structures deposited in the wwPDB. Electron density maps acquired by cryo-EM can be accessed through the EMDataBank [196], which currently holds more than 1 100 entries depicting various molecular structures. The interpretation of this enormous pool of data and the integration of information from different experimental techniques are going to be major challenges to molecular biology and bioinformatics in the upcoming years.

The objective of this work is the development of a method for similarity searching in macromolecular electron density maps. This includes not only the docking of atomic structures to experimental cryo-EM maps, but also the registration of experimental electron density maps. For this purpose, a software is to be implemented and the applicability to similarity searching in experimental maps is to be demonstrated.

The presented method is called *siseek* (SImilarity SEarching in Electron density maps using Keypoints). It is based on state of the art findings and techniques from digital image analysis and relies on an abstract map representation, which is based on scale-space theory [213] and the scale-invariant feature transform (SIFT) [218]. The map representation consists of salient features of all sizes in the depicted objects, and enables the description of image data at different levels of detail. Each of the features is represented by keypoints, which in turn are assigned local neighborhood descriptors. In this way, the map is decomposed into its intrinsic structures and salient parts are extracted. Based on the descriptors, keypoints can be compared and matched. This comparison of descriptors and thereby keypoint neighborhoods facilitates the identification of similarities in images based on the comprised features rather than using an exhaustive search.

siseek can be applied for solving several problems: For one, it enables the docking of atomic structures to high and intermediate resolution electron density maps. For this purpose, a synthetic map is generated from the atomic structure and subsequently registered to the experimental map. This task is frequently carried out for the interpretation of maps acquired by cryo-EM [94]. It also allows for the registration of two experimental electron density maps. This task is not carried out regularly, but allows for new insights that are not provided by other methods. The approach does not rely on an atomic detail interpretation

of the density map and does not consider sequence similarity. Therefore, it is a complementary approach for the registration of protein structures, which up to now is mainly done by computing sequence alignments and registering atomic structures based on the computed alignment. Furthermore, the method draws the attention to the genuine experimental data, the electron density, and therefore it is closer to the measurements than atomic structures, which represent one valid interpretation of an electron density map. This enables a more thorough analysis of the maps and can reveal similarities as well as inconsistencies such as unexplained densities. Furthermore, the abstract map representation can be used for similarity searching in a larger pool of protein structures, as shown here in a proof of concept. In this application, the content of an unknown electron density map is recognized based on a set of defined reference structures.

1.1. Main Contributions

siseek is a software system designed for similarity searching in electron density maps. The method is based on state of the art techniques for image analysis — the theory of scale-space and the SIFT — and allows for the identification of similar regions in macromolecular electron density maps. It contributes to the state of the art of image analysis in the following ways:

- The SIFT is generally applied to 2D images. *siseek* is a robust extension of this method to 3D images. It regards all degrees of freedom that have to be addressed in three dimensions, which is especially important for the proper handling of 3D orientations. All computations in *siseek* are based on interpolated intensity values rather than on the genuine intensities at voxel positions. Thus, it does not rely on a predetermined grid and is therefore only dependent on the depicted density, not the sampling.

- The method and all of its components are thoroughly tested and their performance is validated through various experiments. This includes the repeatability of keypoint detection, orientation assignment, and descriptor calculation as well as studies for the distinctiveness of the computed descriptors. This thorough validation on a large test set of 3D images yields the basis for successfully employing *siseek* in various experimental scenarios.

- The concept of resolution is essential for the interpretation of electron density maps and is defined as a quality in the frequency domain. Due to the central importance of resolution, the robustness of keypoint detection

with respect to resolution lowering was assessed in detail. It was found that the repeatability rate of keypoint detection in the SIFT is dependent on the ratio of resolution and sampling interval of the map. Thus, an optimal sampling interval in terms of the specified resolution is determined.

– The properties of the described 3D SIFT descriptor are assessed in detail and yield insights into the robustness of the descriptor with respect to various distortions. Based on this analysis, findings on the structure of the descriptor and their influence on matching feature vectors are summarized. These give insights into the general properties of descriptors using various parameters and allow for the determination of optimal parameters for *siseek*.

siseek is geared to similarity searching in electron density maps and parameterized for an optimal functioning in this application area. It distinguishes itself from previous work by the following properties:

– *siseek* reliably identifies similarities in high and intermediate resolution electron density maps using keypoints. The method is thoroughly tested on larger sets of synthetic and experimental proteins, and allows for the efficient registration of electron density maps. *siseek* directly references the resolution of the map and incorporates this information in the scale-space representation that is built for the input map. The computer program is shown to use less CPU time than other programs that are geared to the same problem because more demanding computations are only carried out for salient structures of the image.

– No sequence information and no atomic detail interpretation is necessary for similarity searching with *siseek*. The method solely relies on the information contained in the electron density map and therefore allows for similarity searching independent of any interpretation in terms of atomic structures.

– The map description employed in *siseek* can be used for the identification of an electron density map's content. This application is shown in case studies, which exemplify the high discriminative power of the computed descriptors. The presented study, however, is a proof of concept and means for improving the applicability are listed.

– The presented method is not limited to the interpretation of cryo-EM maps but explicitly addresses similarity searching in X-ray crystallography electron density maps. Structures acquired by X-ray crystallography

are regularly compared using the atomic detail interpretation of the maps. However, if no atomic model is available, *siseek* is capable of finding similarities of the given map to other electron density maps. Furthermore, *siseek* shifts the focus to data that is closer to the genuinely observed measurements, the electron density maps.

– The map description decomposes the depicted object automatically into its salient features. These features identify regions in the map that are either more or less dense than their surrounding. Each feature is assigned a scale and a descriptor, which allow for the automated comparison of features and the matching of images. Using this approach, similar subvolumes in the maps can be identified because the descriptors rely solely on local information. Using an exhaustive image comparison, such alignments are generally not identifiable if both maps comprise more than the considered domain. This is due to the mode of comparison in exhaustive searching, which always employs the complete volume of both maps.

All core modules of *siseek* are genuinely created for this work. The software is implemented in C++ and uses the basic functionality provided by the C++ standard library and the boost library. Additionally, an external implementation of an R-tree is used for the efficient location of points in 3D space. Further details on the software design, the implementation, and a description of the user interface can be found in Appendix A.9 on page 225.

1.2. Structure of this Work

Besides the introductory Chapter 1, this work comprises five more chapters and an appendix. The content of each part of this work is summarized as follows:

Chapter 2 provides a survey of the state of the art in relevant disciplines. The chapter comprises an overview of feature detection and image registration techniques, and gives an introduction to relevant techniques in the field without referring to biological problems. Subsequently, methods of structural biology for the elucidation of electron density maps are presented. This includes, on the one hand, default modes for the depiction of biomolecules and, on the other hand, experimental methods for acquiring electron density maps. Eventually, published methods are introduced that seek to solve similar problems as this work. This includes methods for the interpretation of data gathered in X-ray crystallography experiments, methods for docking of atomic structures to electron density maps, and tools for identifying molecular objects in larger databases.

Chapter 3 details the theoretical foundations of *siseek* and introduces all objects that are used in the method for describing electron density maps: Keypoints mark salient structures in the maps and are assigned orientations, which are determined by orientation histograms. According to the orientations, descriptors are computed. Based on this abstract map description, methods for the registration of electron density maps and for molecule recognition are proposed.

Chapter 4 describes the validation and parameterization of the proposed method. For this purpose, the repeatability of keypoint detection, orientation assignment, and descriptor computation is assessed based on synthetic maps of different resolutions and signal-to-noise ratios. Several experiments are performed using different parameters and the results are analyzed with respect to specified objectives.

Chapter 5 reports on the application of *siseek* in experimental scenarios with the parameters determined in Chapter 4. First, a validation of the complete software system using synthetic maps is performed. Subsequently, atomic structures are docked to experimental maps acquired by X-ray crystallography and cryo-EM. In the last outlined application scenario, *siseek* is used to identify the content depicted in an electron density map based on a database of reference structures.

Chapter 6 comprises a summary of the content of this work and outlines perspectives for future research in the field of image analysis in molecular biology.

The Appendix contains information that supplements the findings presented in this work. First, mathematical foundations, molecular identifiers, and properties of the utilized concepts are outlined. Then, plots and tables supplementing the presented findings are listed. Furthermore, depictions of proteins utilized in the test set are given. Eventually, programs used in this work as well as the software architecture of *siseek* are outlined.

Each chapter begins with an introduction and concludes with a summary of the main findings of the chapter. Each section commences with an introductory paragraph giving an overview of the content that is found in the section. All abbreviations and acronyms are explained in the list of abbreviations on page V. Throughout the work, identifiers for molecular structures and electron density maps are used as explained in Appendix A.2 on page 199. The index contains all central terms and is located on page 271. The definition of the mathematical notation can be found in Appendix A.1.

2. State of the Art

Techniques for similarity searching in images and methods for the elucidation of biomolecular structures have been improved tremendously in recent years. Their combined use has given new insights into the spatial composition of biomolecular structures and has allowed for the detailed analysis of various biochemical processes. In this work, a method based on image analysis techniques is presented and applied to structural biology data. This chapter introduces the topics relevant to this work, ranging from applicable image analysis techniques to a summary of pertinent methods of structural biology.

In Section 2.1, an abstract overview of methods for image registration is presented. Besides basic techniques for image analysis, especially methods for the detection of salient image features and their matching are introduced. Eventually, the concepts of scale-space and the scale-invariant feature transform (SIFT), which form the basis of this work, are summarized. This summary on image registration techniques is meant to be general and does not make explicit references to biochemical contexts. In Section 2.2, relevant methods that are employed in structural biology are introduced. First, the properties of atomic structures and electron density maps of biomolecules are described. Subsequently, the experimental techniques used for the elucidation of these structures — namely X-ray crystallography and cryo-electron microscopy — are explained. In the last Section 2.3, previously published methods for the docking of atomic structures to electron density maps and molecule recognition are summarized.

2.1. Image Registration

The process of aligning two or more images so that similar depicted structures are superposed is called *image registration* [41, 241, 119]. This task is generally trivial for the elaborate human cognitive system. However, the design of robust computer systems for image registration has proved challenging. This is due to differences in the circumstances under which images of an object are taken. Examples of these include changes in illumination or the presence of noise induced by the sensor. These have to be accounted for to enable a successful registration [89].

Image registration is of interest to almost all disciplines that make use of image data. Examples include the registration of satellite images [118], the identification of molecules in images depicting cells [99], and applications in medical science [134]. An example from the latter discipline is the registration of 3D image data to facilitate the detection of changes in tissue, which can be used for tumor monitoring [257] or for assessment in surgery [88]. For the latter purpose, images taken before and after a surgical intervention are registered, the differences of the images are identified and analyzed to affirm the outcome of the surgery. Registration also facilitates the combination of information from images taken with different sensors. An example application is the registration of images taken by magnetic resonance imaging and positron emission tomography, which allows for the combination of anatomic and metabolic information [306].

Methods for registration can be classified according to the following criteria [221, 92], which specify the method's capabilities and the application range:

Dimensionality The dimensionality of the images is generally 2D for planar pictures and 3D for volumetric images. However, higher dimensional image registration is also possible using, e. g., temporal volumetric image sequences.

Registration Basis The registration can be based either directly on image intensities or on image features. These include user-selected points or computationally determined attributes.

Transformation The result of the registration is a transformation that superposes the two given images. This transformation can either be rigid or elastic.

User Interaction Methods can be completely automatic, may require the user to supply an initialization, or solely support the user during registration.

Optimization procedure Refers to the method by which the images are brought into register and the employed target function. This can be based on an iterative optimization or can be ideally computed directly.

Modalities The means by which the images are acquired can differ. For medical imaging, standard techniques include X-ray computed tomography, magnet resonance imaging, and position emission tomography, while macromolecular electron density maps are generally acquired by cryo-electron microscopy and X-ray crystallography.

Methods for automatic image registration can be subdivided into the following stages [382, 119], which are going to be addressed in more detail in the next sections:

Image Processing Preparation of the raw image data for later stages. This can include the removal of noise or the detection of edges.

Image Feature Detection Salient, distinctive structures in the image are identified as image features. These can be points, corners, blobs, lines, curves, regions or image templates. Additionally, descriptors can be computed for characterizing the computed features.

Feature Matching Correspondences between detected features are identified using either computed descriptors or the distribution of the features. An identified correspondence yields a transformation of the images, which superposes common parts.

Using the matching, a transformation is determined and the images are superposed accordingly. It is, e. g., possible to resample one image using an appropriate interpolation technique and subsequently fuse the two images. Another application is the direct comparisons of image intensities for the detection of differences.

There exist various approaches to image registration, and the utilized terminology frequently differs depending on the author and the field of study. In the following, the first image to be registered is called *source image*. Other publications also utilize the term *reference image*, which explicates that the image is not changed during the registration. The second image is called *target image* or *sensed image*. The objective of image registration methods is to find the matching content of the target image in the source image. For this purpose, a *transformation* is determined, that maps the target image onto the source image using computationally detected *features*, which represent salient structures in the image.

In this work, an image is the two or three-dimensional depiction of an object using a square respectively cubic grid with a specified pixel or *voxel spacing*, which is also called *sampling interval*. If not stated otherwise, an image I — also called a map — has three dimensions and assigns an intensity value to each point on the cubic grid. The grid points of a 3D image are called volumetric picture elements — for short *voxels*. If no other definition is given, $\mathbf{x} = (x, y, z)^T$ identifies a voxel and $I(\mathbf{x})$ the corresponding intensity value in the image. The location $\mathbf{x}$ may also refer to points in the image that lie in between voxels. In this case, trilinear interpolation is used to calculate the intensity value at $\mathbf{x}$. To

analyze the signal underlying the image, partial derivatives are taken. The first-order partial derivatives are denoted by I_x, I_y, and I_z for the corresponding spatial direction while second-order derivatives are denoted as $I_{ab} \mid a, b \in \{x, y, z\}$. Furthermore, the partial derivative operator ∂_a is used. The *gradient* ∇ of the image function is defined as the vector field $\nabla I = (I_x, I_y, I_z)^T$. The *Laplacian* ∇^2 of an image yields again an image-like structure. It is denoted as the sum of all unmixed second-order partial derivatives $\nabla^2 I = I_{xx} + I_{yy} + I_{zz}$.

The following comprises an overview of image analysis and pattern recognition methods, which are related to this work. Relevant image processing techniques are presented first. Subsequently, methods for the detection and matching of image-features are introduced. The section concludes by giving an introduction to scale-space theory and the SIFT.

2.1.1. Image Processing

Raw image data is frequently preprocessed prior to feature detection. This can include image smoothing, which reduces the amount of noise in the image, but also blurs the image. This preprocessing is accomplished by convolving image $I(\mathbf{x})$ with a *filter* $F(\mathbf{x})$ resulting in an image $I'(\mathbf{x})$ [115] as shown in Equation 2.1 where $*$ denotes convolution.

$$I'(\mathbf{x}) = F(\mathbf{x}) * I(\mathbf{x}) \tag{2.1}$$

Filters F with the effect of smoothing an image are called low-pass filters since they attenuate high-frequency signal components. The *Gaussian function* is frequently employed for this purpose. This function has an infinite support, but can practically be truncated to compact support with minimal damage to the representation in both the spatial and the frequency domain. Furthermore, the Gaussian filter can be applied efficiently because it is a separable filter [115]. The normalized, isotropic, origin-centered Gaussian $G(\mathbf{x})$ in three dimensions is defined as in Equation 2.2, where $|\mathbf{x}|$ denotes the Euclidean norm of $\mathbf{x}$ and the standard deviation σ defines the width of the isotropic Gaussian function, and thereby the amount of smoothing. The larger the value of σ, the more high-frequency signal will be attenuated.

$$G(\mathbf{x}; \sigma) = \frac{1}{\left(\sqrt{2\pi} \cdot \sigma\right)^3} \cdot e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \tag{2.2}$$

Methods for the detection of edges — sharp changes in image intensities — have been vital in the development of image registration and various approaches

for their identification have been proposed. The Canny edge detector is frequently employed for identifying edges since it has good detection and location performance while giving a single response for a true edge [51]. It relies on first order derivative filters and a linking-procedure for the identification of edges. Another common method for edge detection is the Marr-Hildreth operator [225, 32] that relies on second-order derivatives. By applying the Laplacian to the Gaussian function, the *Laplacian of Gaussian (LoG)* filter is created. The 3D LoG is derived in detail in Appendix A.3 and its formula is displayed in Equation 2.3. Zero crossings of an LoG filtered image identify sharp changes in the image intensities and can be utilized for edge detection. However, for larger σ the localization error increases and thus the edge detection is less precise.

$$LoG(\mathbf{x};\sigma) = \nabla^2 G(\mathbf{x};\sigma) = \frac{1}{\left(\sqrt{2\pi}\right)^3 \sigma^5} \cdot \left[3 - \left(\frac{|x|}{\sigma}\right)^2\right] e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \tag{2.3}$$

The convolution with the Laplacian of Gaussian can be approximated using the *Difference of Gaussians (DoG)* approach. Here, the image is filtered with two Gaussians of different standard deviations and the difference of the two Gaussian filtered images yields a fast approximation of the convolution with the LoG.

Edge detectors based on the LoG are less accurate with respect to localization than the Canny edge detector. The principle of convolving images with derivatives of the Gaussian function to determine properties of the underlying signal is, however, essential for various image analysis methods.

2.1.2. Image Features

Features are salient structures in images that can be utilized for image registration. Different kinds of image features exist, the most prominent being corner points. Other image features include blobs, lines, regions, and image templates. Points mark salient elements of the image while blobs identify image segments that are either brighter or darker than their surrounding. Lines are frequently found in two-dimensional images depicting man-made objects and characterized by a straight, elongated edge in the image. Regions are computationally determined, contiguous, homogeneous patches of the image. Image templates are defined by the user and refer to small image regions that comprise structures of interest.

Points are the most frequently used feature of images [119]. They are referred to as feature points, interest points, point landmarks, control points or keypoints. In this section, the term *feature point* is used as a generic term allowing for different detection methodologies. In the methods section, however, the

term *keypoint* is used, which is specific to the SIFT. In the following, different approaches for the detection of feature points and regions are introduced, while the detection of blobs will be covered in Section 2.1.4.1. Line detection can be accomplished using the Hough transform [147, 151] or least squares fitting. The selection of image templates is generally a manual task and matching image templates is related to local neighborhood matching. Thus, it is described in Section 2.1.3.

Feature points are detected in image regions with high information content, e. g., at a local, sharp intensity change. The detection of feature points must be robust, stable and well-defined, so that in two different images of the same object the same feature points are detected. Many feature point detectors rely on differential methods and identify feature points in regions that have a high degree of variation in all directions. These regions are called image corners and are identified using a *cornerness* measure. Such measures were first defined for two-dimensional images [333] and later on generalized to three dimensions [277]. Most of the cornerness measures are based on differential operators and rely on the calculation of the structure tensor C or the Hessian matrix H.

The *structure tensor* is the dyadic product of the gradient of the image

$$\mathbf{C} = \begin{bmatrix} \widetilde{I}_x^2(\mathbf{x}) & \widetilde{I}_x(\mathbf{x})\widetilde{I}_y(\mathbf{x}) & \widetilde{I}_x(\mathbf{x})\widetilde{I}_z(\mathbf{x}) \\ \widetilde{I}_y(\mathbf{x})\widetilde{I}_x(\mathbf{x}) & \widetilde{I}_y^2(\mathbf{x}) & \widetilde{I}_y(\mathbf{x})\widetilde{I}_z(\mathbf{x}) \\ \widetilde{I}_z(\mathbf{x})\widetilde{I}_x(\mathbf{x}) & \widetilde{I}_z(\mathbf{x})\widetilde{I}_y(\mathbf{x}) & \widetilde{I}_z^2(\mathbf{x}) \end{bmatrix} \tag{2.4}$$

where $\widetilde{I}$ denotes a local averaging of the image. This averaging can, e. g., be performed using the arithmetic mean of the voxel intensities in a predetermined neighborhood, which may be spherical or cubic [276]. Furthermore, the image intensities can be weighted using, e. g., the Gaussian function [97]. This neighborhood is frequently called spherical window and — if a Gaussian weighting is applied — *Gaussian window*.

The *Hessian matrix* H is defined in Equation 2.5. It is determined in the same way as the structure tensor and comprises all second-order partial derivatives.

$$\mathbf{H} = \begin{bmatrix} \widetilde{I}_{xx}(\mathbf{x}) & \widetilde{I}_{xy}(\mathbf{x}) & \widetilde{I}_{xz}(\mathbf{x}) \\ \widetilde{I}_{yx}(\mathbf{x}) & \widetilde{I}_{yy}(\mathbf{x}) & \widetilde{I}_{yz}(\mathbf{x}) \\ \widetilde{I}_{zx}(\mathbf{x}) & \widetilde{I}_{zy}(\mathbf{x}) & \widetilde{I}_{zz}(\mathbf{x}) \end{bmatrix} \tag{2.5}$$

The first rotation-invariant feature point detector that makes use of the Hessian matrix is the Beaudet corner detector [16]. It determines the corner strength by analyzing the trace or the determinant of the matrix. The Kitchen and Rosenfeld [177] operator makes use of the first and second derivatives of the

image intensities and calculates the curvature of a level curve as the cornerness measure.

Prominent examples of measures utilizing the structure tensor include the Förstner [100, 101] and Rohr [274, 275] corner detectors. Furthermore, the Harris-Stevens [136] corner operator — also called Plessy operator — and the related Shi-Tomasi detector [300] make use of it. These approaches rely on an analysis of the eigen decomposition of the structure tensor. The determined eigenvalues of the tensor characterize the strength of the intensity variations in the local neighborhood and are used as cornerness measure. In the vicinity of a corner, no prominent direction in the local gradient field are to be found. Therefore, all eigenvalues are required to be of similar magnitude. This is analyzed using the determinant and the trace of the matrix, which respectively equal the product and the sum of the eigenvalues.

Besides these differential cornerness-measures, other approaches, which are based, e. g., on intensity variations in a local neighborhood [308, 283], have been proposed for two-dimensional images and are presented elsewhere [333].

The output of all the cornerness measures is an image comprising intensities that represent the corner strength at each image element. Using this information, feature points are computed by applying a non-extrema suppression and a threshold on the cornerness value. The remaining image elements are selected as feature points, and are used for the description of the relevant image content.

The performance of 3D corner detectors has been assessed on synthetic and medical images [137, 278, 138]. According to different criteria it was found that detectors using solely first order derivatives, which are the Förstner and the Rohr operator, show a better performance than the other examined operators.

Another class of detectors relies on the identification of spherical image regions, which are either brighter or darker than their surrounding. These regions are called blobs and they are identified using the Laplacian of Gaussian or the determinant of the Hessian operator in scale-space. These methods are described in more detail in Section 2.1.4.1.

Besides feature points and blobs, image regions are frequently used features [119]. Regions can be identified using *image segmentation*, which subdivides the image into meaningful partitions identifying single objects or building blocks of objects. These methods either rely on an intensity thresholding approach, make use of region growing, or segment contiguous patches in the image by the identification of boundaries. For single thresholding, a value is chosen interactively or automatically to partition the image. Subsequently, all image elements are assigned to two different classes depending on their intensity value being larger or smaller than the threshold. Region growing methods partition the image

based on seed-pixels, which are iteratively supplemented by neighboring image elements. Here, neighboring image elements are added to a region depending on a membership criterion.

2.1.3. Feature Matching

Correspondences between source and target image are identified by comparing the detected image features. Depending on the type of detected feature, different approaches can be utilized to calculate a matching. Methods using point-pattern matching can identify similarities in the distribution of feature points and thereby find similar regions in the images. Many image registration methods rely on a supplementary description of the local neighborhood of a feature point. Identifying and comparing image templates — small image regions — is therefore a related topic. Thus, point pattern matching algorithms are described first, followed by an overview of relevant template descriptors and their comparison.

Without supplementary information on the feature point neighborhood, all points are considered as compatible. The goodness of fit between two images is therefore only determined by the overlap of the two point sets. If a predetermined, non-collinear matching of at least three source to reference-image feature points is given, a rigid 3D transformation of the target image feature points can be computed [161, 162]. The quality of the superposition of the point sets is determined and yields an estimate of the similarity of the underlying images. The quality of the superposition can be assessed according to different measures. One measure is the Hausdorff distance [150], which is the maximum of the minimum distances between any two points of source and reference image that are closest to each other. If the two images correspond to each other, similar feature points have been detected and will be superposed. If exactly the same keypoints are found in both images, each keypoint from the target image finds a counterpart in the source image. However, usually the set of feature points in the two images is not exactly the same, since either some points have not been found, or superfluous points are detected due to noise. Therefore, it is important to account for outliers by, e. g., introducing a maximum threshold.

Algorithms for finding the superposition that maximizes the overlap of the point sets have been developed for cases, in which no association of feature points is known a priori. These algorithms compare triplets of non-collinear feature points in the two images and identify matchings using a voting scheme. For larger point sets, an exhaustive search is however not possible, since the number of matches that must be assessed lies in $O(n_s^3 n_r^3)$, with n_s and n_r being the number of feature points in the source and reference image. Therefore, algorithmic

strategies have been proposed to enable the identification of the correct match. The Generalized Hough Transform [12, 151] and Geometric Hashing [363] address this problem. In these schemes, each matched set of feature points casts a vote for either a final placement or a basis in the source image. In a second step, the votes are evaluated and superpositions of the images are determined. Alternatively, the probabilistic random sample consensus (RANSAC) method can be employed [91]. RANSAC generates a random, initial assignment and iteratively refines the model yielding an assignment of corresponding points in the two point sets. Furthermore, it is possible to identify correspondences in point sets using clique searching in a distance-compatibility graph [188, 40]. This, however, is only applicable to small point sets due to the complexity of the clique search algorithm.

The number of potential matches between points can be reduced if a descriptor capturing the local neighborhood is assigned to each point. Based on such a descriptor, matchings between feature points are excluded that have a substantially different local neighborhood. Therefore, the number of possible mappings from source to target feature points is reduced and thus the number of matchings that are assessed is also smaller.

One method for describing and comparing local neighborhoods is based on image template matching. For two matched feature points, equally sized neighborhoods N_1 and N_2 are determined from the local neighborhood in the underlying images. The intensities of the voxels can then be compared directly. A common similarity measure used for this purpose is the *Pearson product-moment correlation coefficient (PM-correlation coefficient)* . This coefficient is defined in Equation 2.6 where $\overline{N}$ denotes the mean intensity of the image. The determined value is bound to the interval $[-1;1]$ and yields a measure of the similarity of the intensity distributions in the two neighborhoods. Other measures that rely on a direct comparison of image intensities are mutual information or the sum of squared distances [119]. All of the above measures are not rotation-invariant and thus can only be used for comparing images that differ by a translational offset [79]. For determining the similarity between two images of unknown rotation, all relative orientation of the images must be assessed.

$$r = \frac{\sum\limits_{\mathbf{x}\in R} \left(N_1(\mathbf{x}) - \overline{N_1}\right)\left(N_2(\mathbf{x}) - \overline{N_2}\right)}{\sqrt{\sum\limits_{\mathbf{x}\in R} \left(N_1(\mathbf{x}) - \overline{N_1}\right)^2} \sqrt{\sum\limits_{\mathbf{x}\in R} \left(N_2(\mathbf{x}) - \overline{N_2}\right)^2}} \tag{2.6}$$

The PM-correlation coefficient can also be used to identify a target image of different size inside the source image by performing a complete translational search. It requires extensive computation if carried out in the spatial domain

since the PM-correlation coefficient must be computed for each image element. This process can be accelerated by computing the *cross-correlation* of the two images using the fast Fourier transform [9]. First, the complex conjugate of the source image's Fourier transform and the Fourier transform of the target image are computed. Subsequently, the image-element-wise product of the transformed images is calculated. The inverse Fourier transform of the resulting image contains at each voxel the non-normalized correlation that would be achieved for a corresponding offset. Maxima in this images identify the best translational match of the target image in the source image. This acceleration is frequently employed when searching for similarities in biomolecular images such as in protein-protein docking [164] and the fitting of atomic structures to electron density maps [53].

More abstract measures for local neighborhood similarity have been proposed that rely on a comparison of abstract descriptors [115, 119, 238, 79, 154]. Prominent examples include histogram based descriptors [8], steerable filters [98], invariant moments [103], spherical harmonics [294, 353] and, the SIFT descriptor [217]. The latter three descriptors have been applied in the context of similarity searching in electron densities, and therefore moments and spherical harmonics will be introduced here. An account of the SIFT descriptor is found in Section 2.1.5.

The intensity distribution in an image neighborhood can be described by the use of statistical moments. Three-dimensional *image moments* M_{pqr} of order pqr for an image region R are defined as

$$M_{pqr}(\mathbf{x}) = \sum_{\mathbf{x} \in R} x^p y^q z^r \cdot I(\mathbf{x}) \mid \mathbf{x} = (x, y, z) \tag{2.7}$$

Using a combination of moments of different orders, the underlying image texture can be described with the desired detail. It is also possible to compute moment invariants that do not change if the image is rotated or scaled [148, 288]. Therefore, moment invariants can be used for the rotation-invariant comparison of images. For the comparison of moment invariants, different measures such as the Euclidean distance or the PM-correlation coefficient between the values have been proposed. Shape descriptions based on moments have been employed in bioinformatics for various purposes. Related topics include the detection of planar objects in X-ray maps [141] or the superposition of objects in electron density maps [191].

Closely related to the use of moments are *spherical harmonics* descriptors. Spherical harmonics form an orthonormal set of functions on the unit sphere, and therefore can be used to describe any square integrable function on that sphere.

This is achieved by calculating expansion coefficients for the spherical harmonics as similarly done for the trigonometric functions in the Fourier transform. Here, the number of utilized spherical harmonics determines how close the representation resembles the genuine function on the unit sphere. Using this theory, it is possible to construct star-shaped objects from spherical harmonics. For this purpose, the distance from object surface to center is encoded as a function on the unit sphere and expansion coefficients are determined [353, 294]. Extensions to non-star shaped objects have been proposed [244] and a decomposition of the object into concentric shells has proved effective for object comparison [169]. For each shell, the expansion coefficients of the spherical harmonics are computed and a feature vector consisting of the coefficients for all shells is constructed. The comparison of the descriptors is then performed using the Euclidean distance between feature vectors. This approach has been used for the description of 3D objects and has also been employed for shape matching in bioinformatics. Applications range from small molecule comparison [204], to protein-ligand docking [49] and binding site comparison [243] but also include the docking of atomic structures to cryo-electron microscopy maps [107].

2.1.4. Multi-Scale Image Representations

For the complete description of an image, it is necessary to extract information on the depicted objects at all spatial sizes. For this purpose, scale-space and image pyramid representations have been developed, which are introduced in the following. Image data contains information at different scales. In imagery acquired by photography, a scaling of objects is introduced by the distance between camera and object. In three dimensions, the scale of the image and thereby the size of the comprised objects is generally known. In macromolecular electron density maps, this ranges from the depiction of single atoms [64] up to the cell level where only large macromolecular complexes can be identified [99]. Using other microscopic techniques and methods such as MRI [355], images of organisms can be recorded [10]. Using these sources, information from the molecular level can be combined and used for studying complete organisms [17].

2.1.4.1. Scale-Space

Scale-space theory was developed with the objective of extracting information from signals, which are present at different scales [213]. For this purpose, a one-parameter family of images is created, which is based on the genuine image. The *scale parameter* or, for short, *scale* $\sigma \geq 0$ is used to adjust the amount

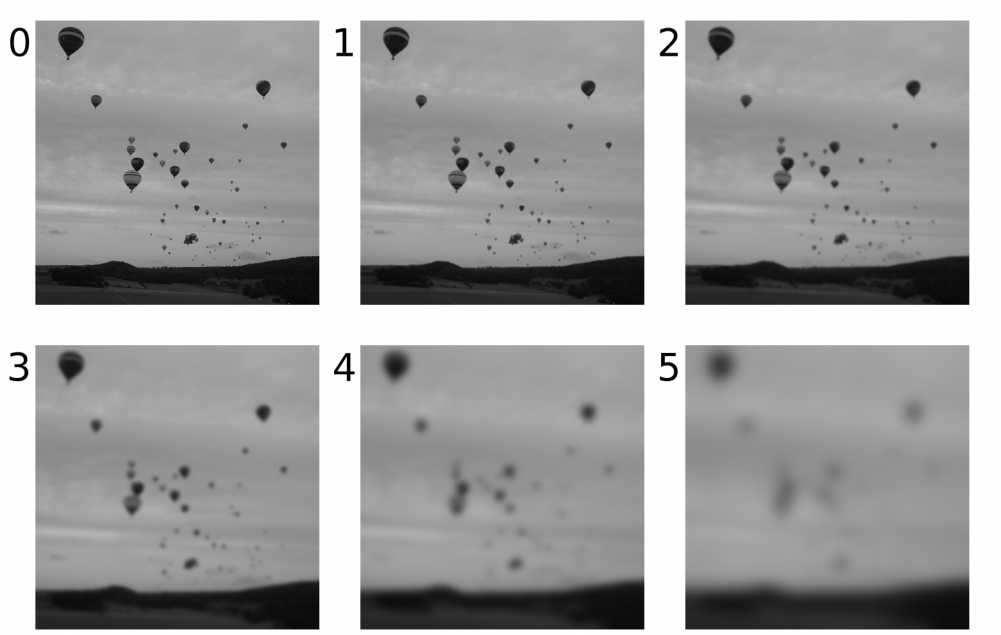


Figure 2.1 – Scale-space representation
A genuine image showing hot air balloons (0) and samples of the scale-space representation (1–5) are shown. The scale σ doubles from image to image. (© A. Griewel)

of structural detail present in the image.[1] For $\sigma = 0$, the genuine image is assumed, while larger values of σ suppress fine detail structures in the image. This is shown in Figure 2.1 for different values of σ.

Scale-space theory was first applied to one-dimensional signals [360]. It was shown that structural details can be attenuated by the convolution with a Gaussian kernel. The amount of structural detail contained in the resulting image can be controlled by specifying the standard deviation σ of the Gaussian function. More image detail will be attenuated if σ is larger, and therefore large scale structures will be more prominent. This allows for the interpretation of the signal at coarser scales [180].

Based on a genuine image $I(\mathbf{x})$, the *scale-space representation* L is computed by convolving I with a Gaussian kernel $G(\mathbf{x};\sigma)$ as shown in Equation 2.8. For $\sigma = 0$ the scale-space representation $L(\mathbf{x};0)$ is defined as the genuine image I

[1] The genuine formulation of scale-space theory uses a parameter $t = \sigma^2$. Here, the scale-space is parametrized in terms of σ, which makes the connection to the Gaussian function clearer.

while increasing the scale σ amplifies the smoothing effect of the Gaussian kernel. The filtering procedure has the approximate effect of suppressing structures with a diameter that is less than σ [208]. This, however, does not hold exactly and merely gives a coarse impression of the effect of the filtering.

$$\begin{aligned} L(\mathbf{x};\sigma) &= G(\mathbf{x};\sigma) * I(\mathbf{x}) \\ &= \left[\frac{1}{\left(\sqrt{2\pi}\cdot\sigma\right)^3} \cdot e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \right] * I(\mathbf{x}) \end{aligned} \tag{2.8}$$

The linear Gaussian scale-space is used frequently and defined using scale-space axioms [209]. It comprises the properties linearity and shift-invariance. The scale is required to be continuous, and a semi-group structure is imposed, which requires an infinitesimal generator. Another axiom is entitled non-creation of local extrema for increasing σ and requires that the number of local extrema decreases with larger σ. This is only true for the one-dimensional scale-space and does not hold for any higher dimensions. Therefore, it is not fulfilled for image data investigated in this work [205, 203]. Nevertheless, the non-enhancement of local extrema holds for all dimensions stating that local extrema are less pronounced at larger σ.

It was shown that the Gaussian kernel is the only kernel that fulfills all axioms [180]. The semi-group property of the Gaussian kernel with respect to convolution

$$G(\mathbf{x};\sigma_1) * G(\mathbf{x};\sigma_2) = G\left(\mathbf{x};\sqrt{\sigma_1^2+\sigma_2^2}\right) \tag{2.9}$$

can be used to create new scale-space samples iteratively. A new scale-space image with scale σ_2 can be created given an image with scale σ_1 according to the formula

$$L(\mathbf{x};\sigma_2) = G\left(\mathbf{x};\sqrt{\sigma_2^2-\sigma_1^2}\right) * L(\mathbf{x};\sigma_1) \tag{2.10}$$

Here, the targeted scale must be larger than that of the given image $\sigma_2 \geq \sigma_1$.

Using the scale-space representation, spatial derivatives of arbitrary order m, n, o can be defined at any level of scale $L_{x^m y^n z^o}(\mathbf{x};\sigma)$. Due to the commutative property of the derivative and the convolution operator, these can be computed by applying the derivative operator to the Gaussian kernel

$$L_{x^m y^n z^o}(\mathbf{x};\sigma) = G_{x^m y^n z^o}(\mathbf{x};\sigma) * I(\mathbf{x}) \tag{2.11}$$

The magnitude of the spatial derivatives $L_{x^m y^n z^o}(\mathbf{x}; \sigma)$ decreases with larger scale σ [207]. To facilitate the comparison of derivatives across scales, a normalization scheme is utilized [210]. The normalized derivative operator

$$\sigma \partial_a \mid a \in \{x, y, z\} \tag{2.12}$$

comprises a multiplication with a factor of σ. It accounts for changes introduced by adapting the scale and yields derivatives that are comparable across scales.[1]

Using normalized derivatives in scale-space, it is possible to implement algorithms for the identification of blobs — regions that are either brighter or darker than their surrounding [212]. The approach for detecting blobs is based on an extension of cornerness measures to the scale-space representation. This enables the *scale-invariant* detection of blobs, which can be used as image features.

Blobs may be detected using cornerness measures as the determinant of the Hessian [211] or the Harris Laplace approach [237]. The most commonly used blob detector, however, is the *scale-normalized Laplacian* [212, 213]

$$\nabla^2_{norm} L = \sigma^2 \cdot (L_{xx} + L_{yy} + L_{zz}) \tag{2.13}$$

Points in this normalized Laplacian scale-space representation, which are extremal with respect to both space and scale indicate the presence of a blob in the genuine image. The size of the blob corresponds to the current scale σ. Therefore, extrema in the normalized Laplacian scale-space are used to identify blob-like image features. These are qualified with a spatial location $\mathbf{x}$ and a scale σ.

2.1.4.2. Image Pyramids

An image pyramid is a multi-resolution[2] image representation [70, 47]. The scale-space approach presented in the previous chapter has the objective of smoothing the signal for further analysis while keeping the number of pixels constant in all sampled images. In contrast, image pyramids combine smoothing with downsampling and can thus be used for the reduction of information and therefore the compact representation of the image at different resolutions [206].

A pyramid is composed of different levels, which are created by iteratively downsampling the image. In most implementations, the image is downsampled

[1] The genuine definition of the normalized derivative operator $\sigma^\gamma \partial_{x_a}$ includes a parameter γ. This value is not utilized in the following and therefore omitted from discussion.

[2] The term resolution is generally used in image pyramid terminology, while scale-space theory relies on the term scale. However, the term resolution is also central to the analysis of macromolecular electron density maps. Thus, except for this section, the term *resolution* is used solely in the context of electron density maps.

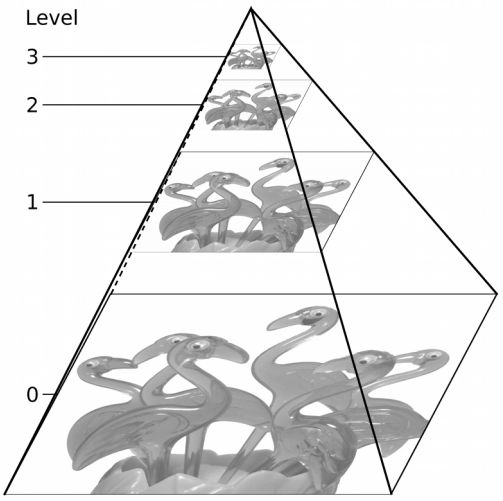


Figure 2.2 – Image pyramid representation
A pyramid representation of an image facilitates the analysis of the underlying signal at different resolutions. It is created by iteratively low-pass filtering and downsampling the genuine image. (© A. Griewel)

by picking every other voxel of the previous pyramid level [115]. Thus, a complete pyramid of an image of dimensions $N \times N \times N$ will consist of $\lfloor \log_2(N) \rfloor + 1$ levels where the number of voxels in the image decreases exponentially with the pyramid level. The metaphor *pyramid* becomes clear when considering 2D images as shown in Figure 2.2. Using the genuine image as base and stacking the downsampled images above it in logarithmic distance, the representation resembles a square pyramid.

High frequency signal must be removed before creating images at higher levels of the pyramid, since these images comprise less voxels and thus less information can be encoded in them [115]. This is achieved by applying a low-pass filter, which attenuates high-frequency signal, prior to downsampling — frequently, the Gaussian filter is used for this purpose. This enables a faithful representation of the low-frequency signal components in higher levels of the image pyramid.

A Laplacian pyramid can be created from an image pyramid by taking the difference of Gaussian filtered images between neighboring levels in the pyramid [46] following the DoG approach. From this representation, features such as edges and ridges, but also regions that are darker or lighter than their surrounding can be extracted [70, 71]. However, the quantization along the resolution direction is coarse in this method and it is not trivial to relate structures between levels of the pyramid [208].

2.1.5. Scale-Invariant Feature Transform

The *scale-invariant feature transform (SIFT)* is a method for the detection and description of features in two-dimensional images [217, 218]. These features

are called *keypoints* and have been utilized in various ways to facilitate the comparison of image data. Different applications have been implemented using the SIFT including, among others, object recognition [218], image stitching [43, 168], and gesture recognition [116]. This shows that the method is robust and applicable to a variety of imagery.

The method detects keypoints that represent blobs in the image. Such keypoints are detected using an elaborate sampling of scale-space, which comprises samples at scales that are separated by a multiplicative factor k. For each pair of neighboring images in the scale-space representation L, the element-wise difference is computed, which yields the DoG image D

$$D(\mathbf{x}; i; \sigma) = L(\mathbf{x}; k^{i+1}\sigma) - L(\mathbf{x}; k^{i}\sigma) \tag{2.14}$$

These images form the *DoG scale-space* and they are close approximations of the application of the scale-normalized Laplacian to the scale-space representation, if k is chosen reasonably small [208]. The exponential sampling along the scale-dimension yields the same normalization factor for all images D. Thus, DoG images approximate the application of the scale-normalized Laplacian operator to the genuine image.

SIFT combines the image pyramid representation with scale-space to analyze an image: First, an image pyramid is created where each level of the pyramid is called *octave*. For each octave o, a scale-space representation spanning scale interval $[\sigma_o; 2\sigma_o]$ is created. This representation is generated by sampling s maps from the scale space. Thus, s images are created at scales $\{2^{0/s}\sigma_o, 2^{1/s}\sigma_o, \ldots, 2^{s/s}\sigma_o\}$. In the SIFT the computation of the image pyramid and the scale-space representations are intertwined: The image with doubled scale 2σ serves as the base for the next octave. It is downsampled by a factor of two and used as the base for the scale-space of the next octave.

Keypoints are identified by analyzing each octave separately. Every extremum with respect to space and scale in the DoG scale-space is identified as a potential keypoint. Subsequently, the location of each keypoint is refined by interpolating the exact location of the extremum using a Taylor series [42]. To facilitate the interpolation, two additional DoG images are created for each octave. All potential keypoints are then subject to two screenings. The first screening discards potential keypoints with low absolute intensity in the DoG map . The second screening eliminates potential keypoints that are located on an edge and are therefore poorly defined. This assessment is performed by analyzing the Hessian matrix computed on the keypoint's associated DoG scale-space images.

To enable the comparison of keypoints, local neighborhood descriptors are computed. The SIFT descriptor is not rotation-invariant and therefore a local

coordinate system is computed, which serves as reference frame for descriptor computation. This coordinate system is computed using an orientation histogram, which accumulates information on the gradient in the local neighborhood of the keypoint. The gradient field is sampled for all pixels inside a circular window. All computed vectors are weighted according to a Gaussian centered on the keypoint and inserted into the histogram using interpolation to minimize discretization effects. A descriptor is then computed for all dominant orientations, which have been determined for this histogram.

The keypoint descriptor consists of a 4×4 array of square subregions. Each region contains an orientation histogram comprising eight bins. The histograms are populated using gradient vectors sampled at all voxels in a Gaussian window centered on the keypoint. Each gradient vector contributes not only to the subregion it is contained in, but its weighted magnitude is distributed to all neighboring subregions and corresponding bins using linear interpolation. This yields a vector comprising 128 entries, which is used for comparison according to the Euclidean distance. To allow for intensity changes, this feature vector is normalized to unit-length and large entries are truncated.

Keypoint detection and descriptor calculation rely on various parameters such as, e. g., the number of subregions used in the descriptor. These have been determined empirically in parameter studies [218].

The resulting image description by SIFT keypoints and descriptors has been used for object recognition. For this purpose, a database comprising descriptors from a set of reference images is generated. To identify an object in a given image, distinctive SIFT descriptors are determined from the database using an index. Distinctiveness is defined as the ratio of the similarity values of the best and second best descriptor match: If the Euclidean distance from the considered descriptor to the best match is significantly smaller than the distance to the second best match, the match is considered distinctive. These matches are used in a generalized Hough transform to discard false object identifications. Here, clusters of matches that agree on a relative orientation of source and target image are determined and yield the position of the identified object. Object recognition using the comparison of SIFT descriptors works reliably even in the presence of noise or if the object is partially occluded [218].

Recently, alternative 2D descriptors, which are related to the SIFT descriptor, have been proposed. Among these are the PCA-SIFT [170] and the GLOH [238] descriptor, which use principal component analysis to reduce the dimensionality of the feature vector. The RIFT descriptor [197] is a rotation-invariant descriptor of gradient vectors in a local neighborhood. The SURF method [15]

relies on integral images and utilizes the response of the Haar-wavelet in square neighborhoods of the keypoint.

The SIFT was also applied to three-dimensional image data. The properties of scale-space are directly transferable to higher dimensions [208]; the computation of descriptors, however, requires a more elaborate approach. Especially the tesselation of the sphere surface and the more profound properties of the 3D rotation group have to be addressed. While the isotropic discretization of the 2D circle into equal sized bins is trivial, it is impossible to find an even distributions for every set of points on the sphere[1]. Furthermore, three-dimensional rotations do not fulfill all properties that two-dimensional rotations do. On the one hand, the group formed by 3D rotations is non-abelian, in contrast to the two-dimensional rotation group. On the other hand, it is necessary to give three parameters to completely specify a 3D rotation, while in 2D one angle suffices. These three degrees of freedom are frequently specified using either a rotation matrix or quaternions [226]. It is also possible to specify a 3D rotation by an axis plus an angle, which specifies the rotation around the axis. These challenges have been addressed only partially in the approaches presented thus far.

The first SIFT descriptors in 3D have been used for the identification of keypoints in CT data [337]. However, the generated orientations were not rotation-invariant, but showed promising performance for equally oriented images. The first method addressing rotation-invariance has been applied to facilitate action recognition [297]. Orientations are assigned in this method using a binning of the sphere according to a geographic coordinate system. Gradient vectors are added to the bins and the most prominent bin is identified as the orientation of the keypoint. This, however, specifies only two of the three degrees of freedom of a 3D rotation. The rotation around the axis remains undefined.

A generalization of the SIFT for n dimensions was developed using a similar approach for the computation of orientations [57, 58, 253]. The method was applied to the registration of medical images and dynamic volumetric data. In the presented case studies, keypoints have been extracted and it was reported that 77 % of the keypoints are detected repeatedly if the images are rotated and scaled. The proposed descriptor for three dimensions comprises 2 048 entries — 4^3 cubes, 8 latitudinal and 4 longitudinal bins — and proved to be robust against different transformations and distortions of the image. Another implementation of a 3D SIFT was employed for the registration of ultrasound images [252, 251]. After a preprocessing of the genuine image, keypoints are detected as extrema in the Laplacian scale-space. Furthermore, the Rohr corner detector [274, 275] is applied to each generated image in the image pyramid. It was reported that

[1] See also Section 3.3.1.

the detection performance of keypoints is not invariant to rotation, but that it is possible to repeatedly detect 50 % of the keypoints for rotations of up to 10°.

A method achieving full rotation invariance uses a second, two-dimensional histogram [4]. In a first step, gradient vectors are accumulated in an orientation histogram, which is built using a geographic coordinate system. Then 2D histograms are assembled for prominent bins in the orientation histogram. For this purpose, all gradient vectors in the neighborhood of the keypoint are projected to the plane that is orthogonal to the direction of the histogram bin. Dominant orientations are eventually assigned using prominent bins of both histograms defining all three degrees of freedom of a 3D rotation. It was shown, that using the complete rotation invariance of descriptor calculation yields a significant improvement in descriptor comparison.

2.2. Structural Biology

The aim of structural biology is the elucidation of molecular structures that are found in matter of biological origin with special interest in the atomic structures of proteins and nucleic acids. In the last years, more and more detailed models of large molecules that carry out the most important functions in living cells have been determined using various experimental techniques. Representations of these structures are accessible via the internet and enable the elucidation of the molecular processes underlying life.

In this work, a method for similarity searching in electron density maps is presented. In the following, an introduction to the composition and depiction of relevant molecular structures is given to support the understanding of the analyzed data. A description of electron density maps and of methods for depicting these 3D images on 2D paper follows. Subsequently, the two major experimental techniques for acquiring electron density maps — namely X-ray crystallography and cryo-electron microscopy — are introduced. The section concludes with a summary of the presented information.

2.2.1. Atomic Structures

Molecules are entities that consist of two or more *atoms*, which are held together by *covalent bonds* [233]. The way, in which the atoms are connected by covalent bonds with certain bond orders, defines the *configuration* of the molecule. Thereby, the *chiral* properties of stereo-centers and the three-dimensional structures the molecule can adopt are defined. Covalent bonds are mediated by valence electrons, which repel each other. This induces typical arrangements of

bound atoms, which can be modeled using, e. g., the valence shell electron pair repulsion (VSEPR) theory [111, 186].

A specification of the three-dimensional arrangement of all atoms in space defines a *conformation* of the molecule. The molecular flexibility, which makes different conformations possible, is mainly introduced by rotations about single, non-ring bonds that are *rotatable*. By twisting the molecule at a rotatable bond, changes in the three-dimensional structure are induced and different conformations are adopted. This, however, does not change the configuration of the molecule because the relative arrangement of the bonds with respect to each atom remains the same. The concept of conformational flexibility is central to several processes in life such as protein folding [139] and protein-ligand binding [90, 182, 123].

Molecules are frequently described using either units of length or weight. A convenient measure of length on the molecular scale is the ångström (Å) [45]

$$1\,\text{Å} = 0.1\,\text{nm} = 10^{-10}\,\text{m} \tag{2.15}$$

The van der Waals radius of a carbon atom, for example, is 1.7 Å and the length of a carbon–hydrogen bond is smaller than 1.12 Å [358]. Molecular weight is specified using either the unified atomic mass unit (u) or the Dalton (Da). Both are defined as one twelfths of the weight of a carbon–12 atom in its ground state [242]

$$1\,\text{u} = 1\,\text{Da} = 1.66 \cdot 10^{-24}\,\text{g} \tag{2.16}$$

Biomolecules — organic molecules produced by living organisms — are generally composed of carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus atoms. Examples of biomolecules include desoxyribonucleic acid (DNA), saccharides, lipids, and proteins, which are of major interest to structural biology. Proteins are built in the ribosome using the genetic information. The building parts of proteins are the twenty proteinogenic *amino acids*, which are encoded in the genetic code and shown in Figure 2.3. Each amino acid comprises an amine and a carboxylic acid group, which are covalently bound to a central carbon atom — the C_α atom [152, 233]. Additionally, C_α is bonded to the side chain, which gives each amino acid specific properties. A protein is synthesized in the ribosome by linearly reading the genetic information and translating this information to the also linear amino acid code. The amino acids are attached to each other by a peptide bond between the amine and the carboxylic acid. Since the elements of water are removed during this process, the amino acids are also referred to as *residues*. The chain that is build by this process is also called *backbone* of the protein and comprises all protein atoms except those of the side chains. Furthermore, proteins are frequently referred to as *macromolecules* due

to their relatively high molecular weight. Large and highly complex protein structures are also called *biomolecular machines* or *macromolecular machines* since their components follow mechanical movements such as rotation, gliding, and lever motions.

Figure 2.4 shows the four levels of structure that are commonly discerned in proteins [179, 22]. The first level — the *primary structure* — describes the sequence of the protein and generally does not refer to the three-dimensional properties of the molecule. It does, however, include information on modifications as, e. g., the formation of disulfide bridges between cysteins. *Secondary structure* describes the conformational arrangement of the backbone of the protein excluding information on the side chain conformations or the relative orientation with respect to other segments. In proteins, α–helices and β–strands are the most frequent secondary structure elements [259] while other motives have also been described [163]. The *tertiary structure* of a protein specifies the complete conformation of the protein. In numerous proteins, the linear chain of amino acids folds up to globular shapes. The folding is induced by diverse chemical processes such as the hydrophobic effect or the formation of hydrogen bonds and salt bridges. The*quaternary structure* describes the spatial arrangement of two or more proteins that form one entity — a *complex*. The constituting chains of the complex are referred to as *monomers* or *subunits*. A complex is also known as *oligomer* if it comprises a small amount of subunits. An example of this type of complex is GroEL, which is shown in Figure 2.4. A complex is called *polymer* if it is composed of a theoretically unlimited number of subunits. An example of such a complex is a microfilament, which is formed by actin as shown in Figure 2.5. Furthermore, the attributes *homo* or *hetero* indicate whether a complex is built of subunits from one respectively various species.

There are various ways for the depiction of molecular entities — e. g., a 2D structure diagram [38] as shown in Figure 2.3. Proteins are frequently depicted using schematic drawings, which is exemplified in Figure 2.5. The figure comprises six panels showing an actin monomer and an image of a segment of a microfilament — a polymer formed by actin that is part of the cytoskeleton [81]. In panel A) of the figure, each atom is depicted as sphere. Carbon atoms are assigned white color, hydrogens are not shown and all other atoms are colored according to the Corey–Pauling–Koltun color convention [68, 181]. Panel B) shows the molecular surface [65], which covers the volume that no solvent molecule can occupy due to the atoms of the protein. From these depictions it is clear that the protein covers a dense, closed volume. C) shows a ball-and-stick model of the protein and D) uses solely sticks. These panels of the figure comprise a lot of information but they are also crowded and it is not easy to identify

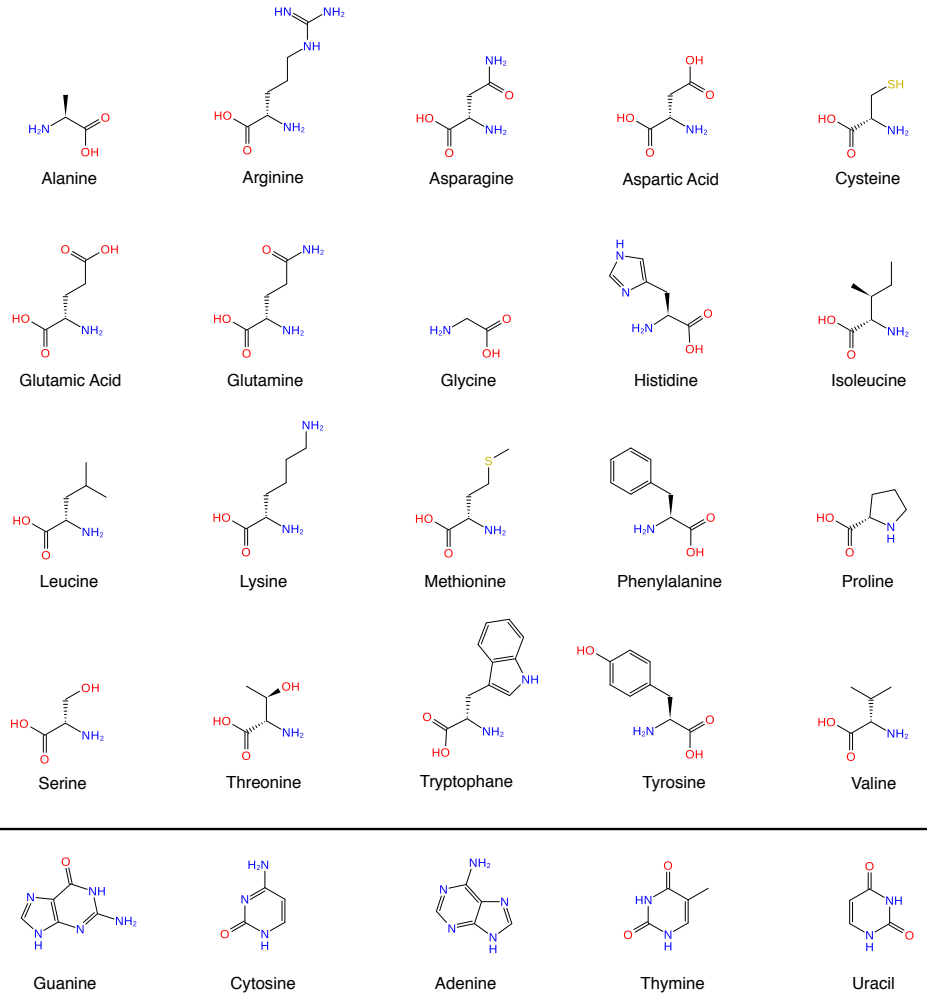


Figure 2.3 – Amino acid and nucleobase structure diagrams
The first twenty structure diagrams above the line depict the proteinogenic amino acids and are oriented so that the side chains point to the top. The five structure diagrams below the line represent the nucleobases. Guanine, Cytosine, and Adenine are found in both DNA and RNA. The fourth nucleobase of DNA is Thymine while RNA comprises Uracil. (© A. Griewel)

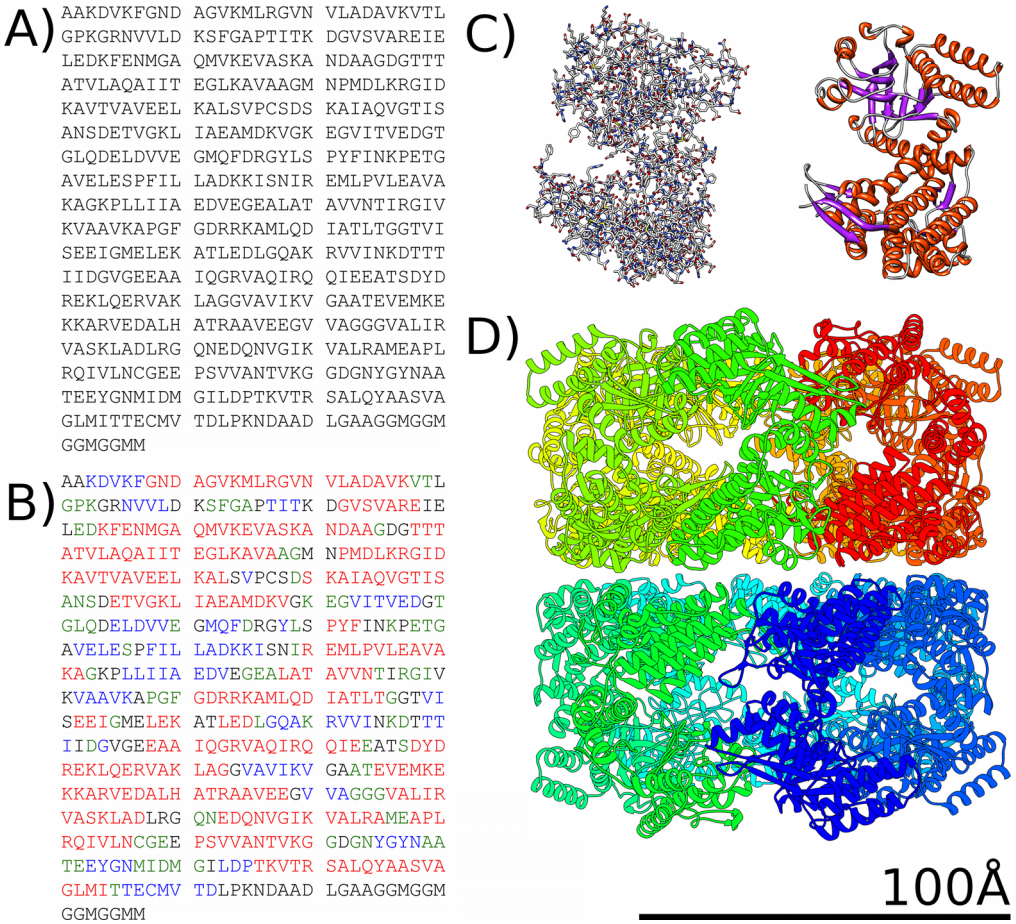
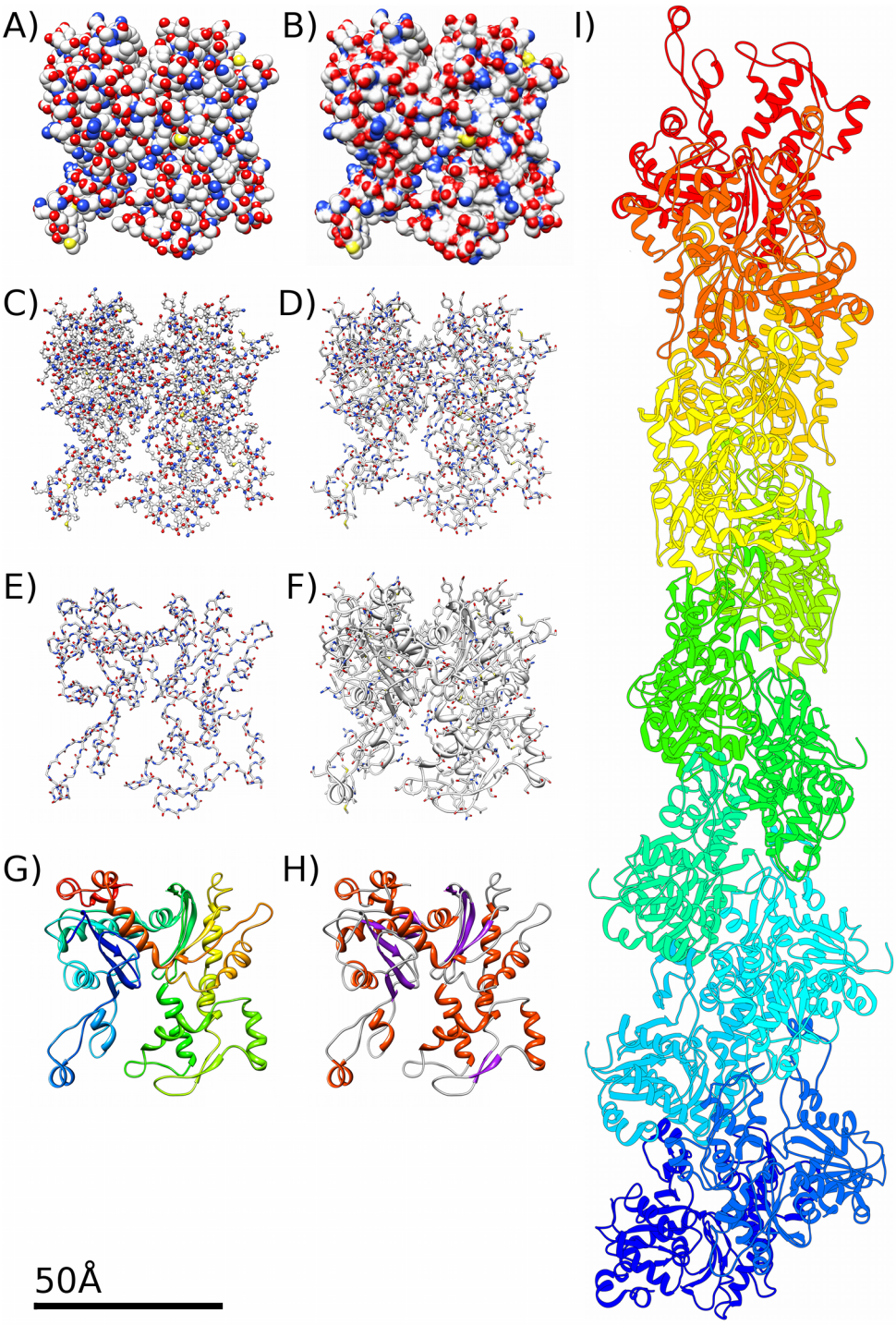


Figure 2.4 – The four levels of protein structure
A) The primary structure corresponds to the protein sequence. B) The secondary structure comprises a description of the conformational arrangement of the protein backbone without information on the global three-dimensional locations. The sequence of the protein is colored according to secondary structure: red for α–helices, blue for β–strands, and green for turns and bends according to DSSP (Define Secondary Structure of Proteins) [163]. C) The tertiary structure defines the location of all atoms and all bonds in the structure. A stick model of the protein is shown along a schematic drawing, which highlights the secondary structure elements in the tertiary structure. Here, α–helices are colored dark orange while β–strands are drawn in purple. D) The quaternary structure of the protein — the chaperonin GroEL [34] — consists of two back-to-back heptameric rings, which are colored here on a per-subunit basis. (© A. Griewel)

major building blocks of the protein. In panel E), only the backbone of actin is shown, which is also called backbone trace. The backbone is frequently used to represent proteins as a whole. For this purpose, a ribbon is drawn following the backbone as shown in F)-I). Panel F) of the figure shows the backbone and all side chains as stick models. Panel G) shows solely the backbone ribbon colored in rainbow colors while panel H) is colored according to secondary structure elements. Panel I) shows the quaternary structure of actin forming a microfilament. The structures were acquired through the wwPDB as explained below using ID `3G37` [245].

Similar to proteins, ribonucleic acid (RNA) and DNA are essential biomolecules, which are found in all forms of known life. These substances consist of long chains that are formed by linking nucleotides. These consist of a sugar, a phosphate, and a nucleobase — the latter are shown in Figure 2.3. RNA and DNA differ in the bound sugar unit — RNA employs ribose while DNA comprises desoxyribose — and the composition of the nucleobases. The three substances cytosine, guanine, and adenine are found in both RNA and DNA. The fourth nucleobase in DNA is thymine while in RNA uracil is used. DNA is commonly known for its function as carrier of the genetic information. RNA is also used to convey genetic information — e. g., in the form as messenger RNA. However, RNA is also assigned functional roles as it is used to carry out enzymatic reactions and as structural components of macromolecular machines. A prominent examples of the use of RNA in macromolecular machinery is transfer transfer RNA (tRNA), which supplies amino acids to the ribosome. This macromolecular machine is shown in Figure 2.8 and consists in its major parts also of RNA. The atomic structure of a phenylalanine tRNA as deposited in the protein data bank with ID `1EHZ` [299] is shown in Figure 2.6. Panel A) and B) comprise the already discussed sphere and stick models. Panel C) of the figure displays a schematic drawing of the tRNA using different colors for backbone and bases. In panel D), another schmeatic drawing of the tRNA is shown, which is rainbow

Figure 2.5 (following page) – Depiction modes for proteins
Panels A)–H) show an actin monomer. Atoms are depicted using A) spheres, B) the molecular surface, C) ball-and-sticks, and D) sticks only. Panel E) shows the atoms of the protein backbone as sticks while F) depicts the backbone as ribbon and the side chains as sticks. G) is colored in rainbow colors from blue at the N–terminus to red at the C–terminus while H) is colored according to secondary structure using dark orange for α–helices and purple for β–strands. I) shows a microfilament segment — a quaternary structure of actin. Each protein in the filament is colored according to a rainbow color palette and 3D effects are disabled for clearer view. (© A. Griewel)


A)
B)
I)
C)
D)
E)
F)
G)
H)
50Å

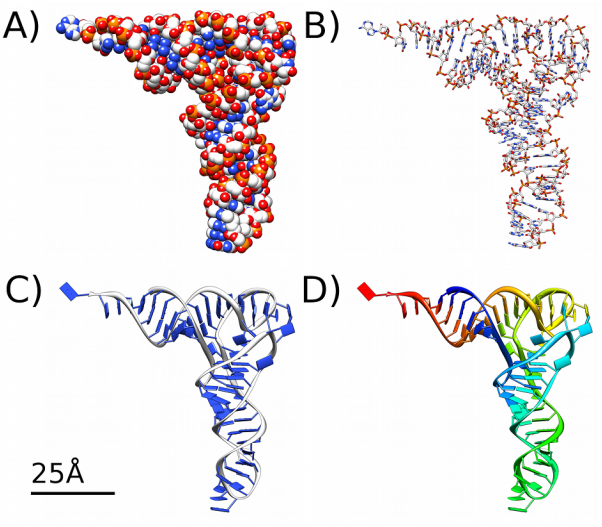


Figure 2.6 – Depiction modes for nucleic acids
Depictions of phenylalanine transfer RNA (tRNA) are shown using A) spheres and B) sticks for all atoms. C) shows the structure using a white backbone ribbon and blue tubes for the nucleobases. D) portrays the tRNA colored in rainbow colors from blue at the 5′ end to red at the 3′ end. (© A. Griewel)

colored from 5′ to 3′ end. Similar to proteins, RNA molecules have levels of structure, however, the elements of secondary structure differ and include, e. g., stems and loops in RNA.

Biomolecular structures can be determined using X-ray crystallography and cryo-electron microscopy, which are explained in Section 2.2.3 and Section 2.2.4. These methods are used to elucidate the tertiary structure of a biomolecule by assigning relative three-dimensional locations to its atoms. Besides these experimental methods, nuclear magnetic resonance spectroscopy [171] and small-angle scattering [319] yield insights into the structural composition of proteins.

Atomic models of biomolecular structures and associated experimental data can be deposited in the public wwPDB [23]. According to the website "The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community.". Atomic structures can be deposited and retrieved using any of the wwPDB member's interfaces namely those of the RCSB PDB[1] [25], the PDBe[2] [345], the PDBj[3] [175], or the BMRC[4] [335]. The archive was initiated at the Brookhaven National Laboratory, NY, USA, in 1971 [27] and initially

[1] http://www.pdb.org — Research Collaboratory for Structural Bioinformatics (RCSB), United States of America.

[2] http://www.pdbe.org — European Molecular Biology Laboratory — European Bioinformatics Institute (EMBL–EBI), United Kingdom of Great Britain and Northern Ireland.

[3] http://www.pdbj.org — Japan Science and Technology Agency — Institute for Bioinformatics Research and Development (JST–BIRD), Japan.

[4] http://www.bmrb.wisc.edu — Biological Magnetic Resonance Data Bank (BMRB), Department of Biochemistry, University of Wisconsin — Madison, United States of America.

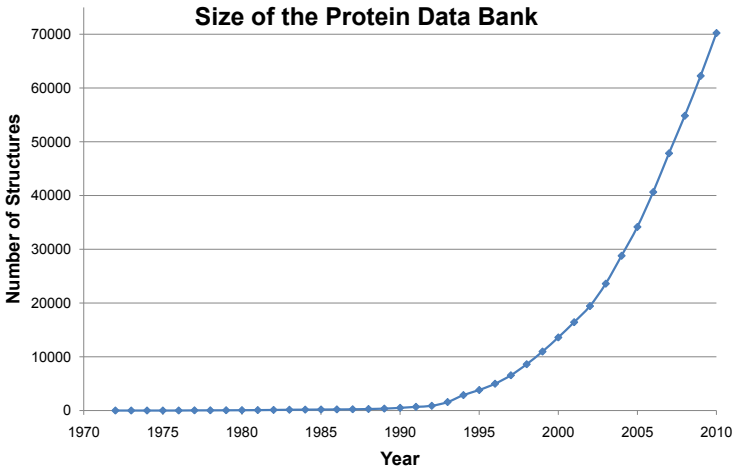


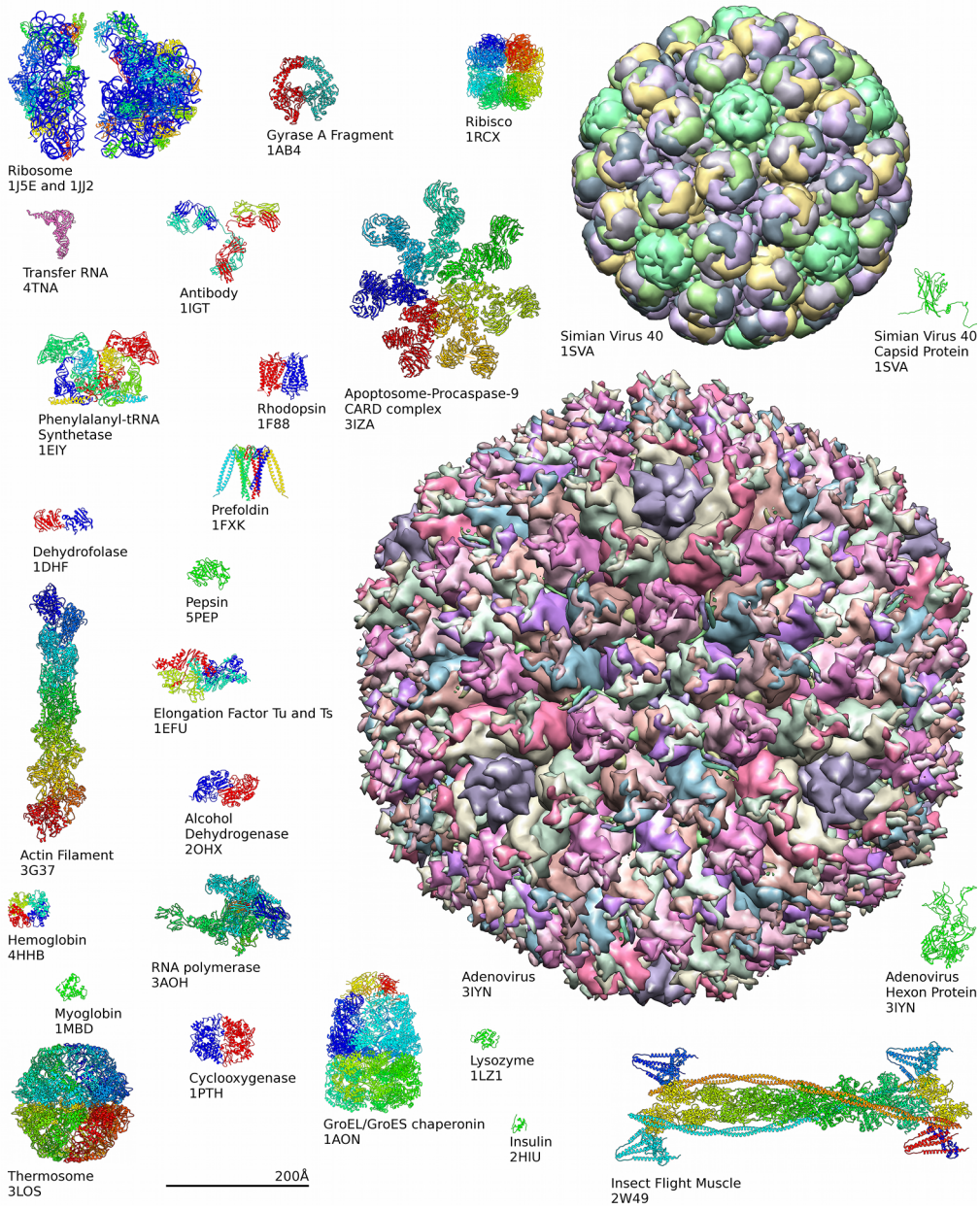
Figure 2.7 – Growth of the worldwide protein data bank in the last 40 years
The number of entries in the worldwide protein data bank has risen from seven structures in 1971 to more than 70 000 structures in 2010. (© A. Griewel)

contained seven structures. The number of available structures rose constantly over the last forty years as shown in Figure 2.7. In December 2010 it was announced, that "As the year 2010 draws to a close, the number of biomacromolecular structures available in the wwPDB archive now exceeds 70 000." [383] and currently there are more than 76 000 structures deposited in the wwPDB. A selection of structures available through the wwPDB demonstrating the variety of the molecular structures found in living organisms is shown in Figure 2.8 and also in Appendix A.7 on page 221.

Each wwPDB entry is assigned an ID, which consists of four characters — a number followed by three alphanumerical characters such as `1MBN`.[1] This identifier, `1MBN`, is the ID of the first elucidated three-dimensional protein structure of myoglobin [172]. Using this ID, a coordinate file can be downloaded, which specifies the location for each atom while bond information is only provided for non-standard entities in the structure such as ligands or modified amino acids [364]. All atoms are assigned to chains, which in turn are given a chain identifier. Furthermore, structures can be encapsulated in models, which allows

[1] The notation of identifiers in this work is explicated in Appendix A.2 on page 199.

Figure 2.8 (following page) – Examples of structures deposited in the worldwide protein data bank
Structures archived in the worldwide protein data bank are shown using schematic drawings of the biomolecules. For the two virus capsids only a surface is depicted. The main structural proteins constituting the capsid are shown as monomers on the lower right hand side of the capsids. (© A. Griewel)


Ribosome
1J5E and 1JJ2
Gyrase A Fragment
1AB4
Ribisco
1RCX
Simian Virus 40
1SVA
Simian Virus 40
Capsid Protein
1SVA
Transfer RNA
4TNA
Antibody
1IGT
Apoptosome-Procaspase-9
CARD complex
3IZA
Phenylalanyl-tRNA
Synthetase
1EIY
Rhodopsin
1F88
Prefoldin
1FXK
Dehydrofolase
1DHF
Pepsin
5PEP
Elongation Factor Tu and Ts
1EFU
Actin Filament
3G37
Alcohol
Dehydrogenase
2OHX
Hemoglobin
4HHB
RNA polymerase
3AOH
Adenovirus
3IYN
Adenovirus
Hexon Protein
3IYN
Myoglobin
1MBD
Cyclooxygenase
1PTH
Lysozyme
1LZ1
GroEL/GroES chaperonin
1AON
Insulin
2HIU
Thermosome
3LOS
200Å
Insect Flight Muscle
2W49

for the reuse of chain identifiers. Thus, a subunit of a molecular complex is unambiguously identified by specifying ID, chain, and model. Frequently, however, only chain or model identifiers are used, so that specifying either one is already sufficient. Some structures consist only of C_α atoms — so called backbone traces — showing merely the overall shape of the structure but not the positions of all atoms.

The wwPDB is the primary resource for structural biology. It yielded the basis for numerous investigations carried out not only in bioinformatics but also in chemoinformatics and medicine [24]. More than 25 new structures are deposited every day while more than 700 000 files are downloaded showing the central importance of this invaluable scientific resource [383]. As mentioned, the data stored in the wwPDB is used for various purposes, e. g., by parmaceutical companies to support the process of drug development using, e. g., virtual screening [301]. This is possible since data deposited in the wwPDB is "free of all copyright restrictions and made fully and freely available for both noncommercial and commercial use" [384].

2.2.2. Electron Density Maps

Atomic structures of biomolecules are most frequently computed using electron density maps acquired by X-ray crystallography while only a small amount of structures was determined using cryo-EM. Nuclear magnetic resonance spectroscopy (NMR) is another method for the elucidation of small molecular structures. It relies on spectra and distance geometry and therefore does not produce electron density maps. As of October 2011, the wwPDB holds more than 76 000 structures, which have been determined in 87 % by X-ray crystallography, in 12 % by solution nuclear magnetic resonance spectroscopy, and in 0.5 % by cryo-electron microscopy.

Electron density maps are three-dimensional images, which are sampled on a regular lattice. Maps acquired by cryo-EM depict the Coulomb potential of the investigated molecule and are generally sampled using a cubic lattice [94]. X-ray maps depict the spatial distribution of the electron clouds in the molecule. They are sampled on a lattice, which corresponds to the symmetry of the crystal [273]. Resampling of maps to other lattices and format conversions can be achieved using various tools [343, 365].

All electron density maps are assigned a resolution, which is determined during the experiment as explained below. For both X-ray crystallography and cryo-EM, resolution is defined as a quantity in the frequency domain and measured in ångström [341]. It specifies a limit on the maximally achievable resolvability of objects — i. e., it gives a lower bound on the size of objects that

can be discerned in the map. High resolution maps — i. e., maps, in which even small objects can be discerned — are assigned small resolution values that lie for X-ray crystallography frequently at approximately 2 Å. Low resolution maps are assigned higher values and are frequently found in cryo-EM.

In high resolution maps — generally acquired by X-ray crystallography but lately also by cryo-EM — side chains and even single atoms can be discerned. These maps are generated in an elaborate process that in some cases may introduce errors and result in flawed atomic structures [35, 160, 77, 376]. Furthermore, not all regions of the map are equally well resolved and show single atoms — especially the outer regions of the protein are frequently less well *resolved* [273]. This is especially cumbersome for the identification of ligand atom positions since these atoms are generally not covalently bonded to the protein and therefore no restrains on the atom locations are given [75]. Thus all structures deposited in the wwPDB must be seen as one valid interpretation of the measured data.

In lower resolution maps, less structural features can be identified and therefore less objects can be discerned [59, 158, 381, 214, 10]. Side chains become unrecognizable at resolutions lower than 4 Å. Also the backbone trace may break at this resolution at flexible segments of the molecule. However, detailed features of secondary structure elements such as the pitch of α–helices and the single strands of β–sheets remain identifiable for up to 5 Å resolution. At resolutions lower than 8 Å, also α–helices and β–sheets become unrecognizable. Maps at these low resolutions do not reveal detailed information on the molecule's internal structure but they provide information on the relative orientation of larger building blocks of the depicted molecules.

Maps are depicted by either showing an isosurface or a slice of a map. For an *isosurface*, a threshold is specified and a three-dimensional surface is calculated that cuts through the volume where the interpolated intensity values equal the threshold. This surface can be calculated using the marching cube algorithm [215]. Alternatively *slices* through the volume can be used for depicting the intensity values. These are planes that intersect the map and are colored according to the underlying intensity. On the one hand, this has the advantage of visualizing the density distribution in the two depicted dimensions in detail. On the other hand, the three-dimensional character of the information is lost. Frequently, also three orthogonal slices are used to give a better impression of the overall density. In Figure 2.9 and Figure 2.10 the atomic structures of two macromolecules and the associated electron density maps are shown using isosurfaces and slices. In Figure 2.9 a high resolution X-ray map is depicted while Figure 2.10 shows a cryo-EM map of lower resolution.

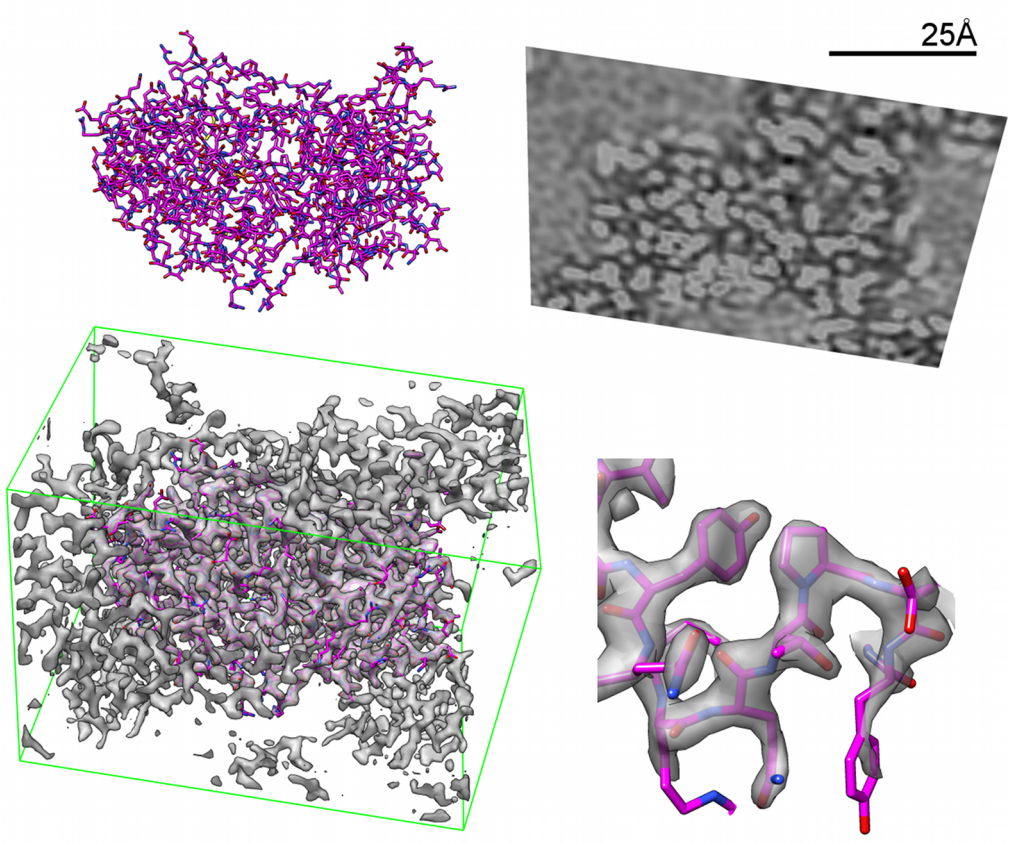


Figure 2.9 – Depiction modes of electron density maps (X-ray crystallography) The atomic model of a hydroxylase (1PBD [295]) is shown in stick mode. On the right hand side of the model a slice through the accompanying X-ray electron density map of 2.3 Å resolution is shown. Below the atomic structure, an isosurface depiction of the map superposed to the atomic model is located. In this depiction the green lines delineate the dimensions of the unit cell and through the transparent isosurface the match between density and atomic model is apparent. Furthermore, the unit cell also comprises density from proteins located in the neighboring unit cells. On the lower right, a close-up is located (not to scale). It shows that the resolvability of atoms in the electron density map is not equal in the whole map, which can be seen by comparing the densities defined for the upper and the lower tyrosine. (© A. Griewel)

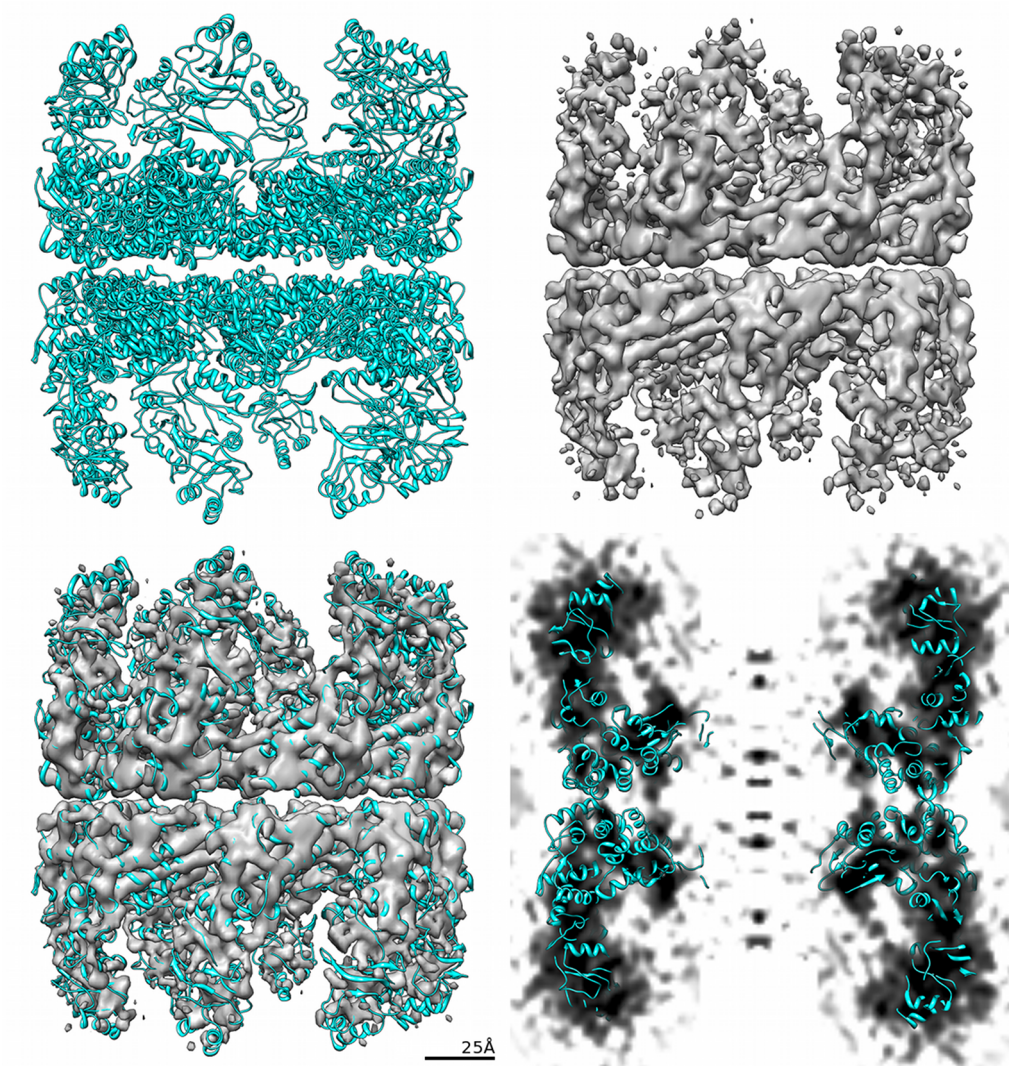


Figure 2.10 – Depiction modes of electron density maps (cryo-EM)
The images depict the open state of the Methanococcus maripaludis chaperonin (`EMD-5140` and wwPDB ID `3IYF` [377]) at 8 Å resolution. The upper left figure shows the atomic model in a schematic drawing, which was acquired by the cryo-electron microscopy map shown on the right. On the lower left, the superposition of map and structure is shown. The lower right picture shows a slice through the volume superposed to the ribbons, which cut the slice. (© A. Griewel)

2.2.3. X-ray Crystallography

In most cases, the conformation of a biomolecule is elucidated by X-ray crystallography. In this technique, a beam of X-rays strikes a crystal, which is formed by the molecule, generating a diffraction pattern. This pattern comprises the directions and intensities of the X-rays emitted from the crystal and is created by the interaction of the X-rays with the electron density distribution in the crystal. It is stored in a machine-readable format and a computer is used to simulate the effect of a lens focusing the X-rays. In this way, an electron density map is acquired that can be used to determine the relative locations of the atoms of the biomolecule. The following introductory overview of the method is based on the explanations and definitions in a standard textbook [273], which comprises further information and references.

The first and often most difficult step in a X-ray crystallography experiment is the growing of a high quality *crystal*. A crystal is a solid whose constituent atoms are arranged in an orderly, repeating pattern. Materials without long range order are called amorphous or vitreous. Protein crystals consist in large parts of the buffer the protein was solved in and are held together mainly by hydrogen bonds. Therefore, they are less stable than commonly encountered crystals such as diamonds. A crystal can be though of as being made of small regular volumes — the *unit cells*. All unit cells in a crystal are equal and their shape belongs to one of the seven lattice systems cubic, tetragonal, orthorhombic, rhombohedral, hexagonal, monoclinic, or triclinic. However, a unit cell does not necessarily comprise exactly one copy of the studied molecule but may include parts of proteins located in neighboring unit cells or even more than one protein of the crystal.

For an X-ray crystallography experiment to be successful, the grown crystal must be must be larger than 0.1 mm in all dimensions, without imperfections, pure in composition, and of regular structure. Several factors are decisive for the orderly crystallization of a protein in an aqueous solution. These include the purity of the protein, the pH of the solution, the concentration, the temperature, and the presence of precipitants. The latter are substances, which facilitate the formation of crystals such as the frequently used polyethylene glycol.

Protein crystallization is frequently achieved using vapor diffusion: A drop containing buffer, protein, and precipitant is deposited in a sealed system that comprises also a larger reservoir with buffer and precipitant in higher concentration. This causes the evaporation of buffer and precipitant from the droplet and therefor increases the concentration of the purified protein. If the concentration of the proteins is high enough and all other conditions are optimal, protein crystals are formed.

The crystal is transferred to a capillary along with parts of the remaining solution — the mother liquor — since dried crystals do not produce clear reflection patterns [26]. The capillary is mounted in a goniometer — a device that allows for the precise orientation of an object in 3D space using rotations. Then, the crystal is placed in an intense beam of X–radiation and the diffraction pattern of the crystal is recorded as it is rotated. X-rays are electromagnetic waves with length in the range of 1 Å to 10 Å. They can be created using an X-ray tube, in which high velocity electrons collide with a metal target such as copper and thereby emit X-rays as bremsstrahlung. However, for high resolution structures usually the beamlines of a synchrotron are used. The diffraction pattern of the crystal is recorded on the detector, which can be film, an imaging plate, or a charge-coupled device sensor.

X-rays are diffracted by the electron clouds of the crystallized molecule. The reflections for each orientation are recorded along their intensity and a three-dimensional index. Theoretically it is also possible to utilize a single molecule for diffraction. However, the amplitude of the reflections grows linearly with the number of scatterers and therefore a crystal concentrates the diffraction signal and makes it distinguishable from noise.

The index of each reflection is determined according to its location on the detector and the orientation of the probe, which in turn is determined by the goniometer. The indices are frequently called Miller indices and are commonly denoted by the three letters *hkl*. The total of all reflections recorded for the different orientations is then incorporated in the three-dimensional *reciprocal lattice*. This lattice yields information on the Fourier transform of the electron density map and regularly comprises 1 000 to 1 000 000 reflections.

The reflections yield information on the *structure factors* of the electron density in the unit cell. A structure factor consists of information on the *direction* of the ray, its *amplitude* and its *phase*. The direction of a reflection is specified by its Miller index *hkl* in the reciprocal lattice and specifies its frequency. The reflection pattern yields information on the size and shape of the unit cell as well as the symmetry of the crystal. The amplitude of a structure factor is proportional to the square root of the intensity of the corresponding reflection. According to Bragg's law, each reflection is caused by waves of the same phase. This phase, however, is not measurable during the experiment and determined computationally.

There are four standard techniques to solve the *phase problem*. Ab initio methods are regularly employed for small molecule crystallography, but are not suitable for the elucidation of macromolecular structures due to the large amount of structure factors. In the second approach — called multiple isomor-

phous replacement — heavy atoms are introduced in the crystal by soaking or co-crystallization. The changes in the diffraction pattern can then be used to determine initial phases. The third technique — called anomalous X-ray scattering — also relies on the introduction of heavy atoms in the crystal. From this crystal diffraction patterns are recorded at wavelengths far below, far above, and in the middle of the absorption edge of the introduced heavy atom. Again, the changes in the diffraction pattern can be used for determining initial phases. In protein crystallography the amino acid seleno-methionine is frequently employed for this purpose. In the fourth method — called molecular replacement — the phase problem is solved by taking initial phases from a similar molecular entity. Therefore, this technique is restricted to cases, in which either a structure of a homologous protein is known or if the structures differ only in small parts — e. g., in a co-crystallized ligand.

The diffraction pattern together with the solution of the phase problem yields the initial set of structure factors each comprising a location in the reciprocal lattice, an amplitude as well as a phase. An electron density map of the unit cell can be determined as Fourier sum of the structure factors. In the best case it is possible to determine the location of atoms from this maps. Generally, however, the first estimates of the structure factors are not yet exact. Therefore, refinement procedures are employed, which, e. g., sharpen the information contained in the map or bring the content of the initial map in line with the general knowledge about protein crystals.

Eventually, a molecular model is acquired from the electron density map using, e. g., skeletonization or the fitting of molecular fragments using the least-squares method. This also gives the possibility of incorporating prior knowledge such as stereochemistry and typical bond length and angles. Furthermore, the Ramachandran plot [268] can be used to calibrate the torsion angles of the bonds in the protein backbone. Refinement is an iterative procedure, which can be carried out in real space by improving the atomic model and in the frequency domain by adapting the employed phases.

Using the molecular model and the electron density map, the objective of the refinement is to bring the measured experimental reflections $\mathbf{F}_{\text{obs}}$ in accordance with the structure factors calculated from the computed atomic structure $\mathbf{F}_{\text{calc}}$. This agreement of the structure factors is often expressed using the *residual index* also known as *R-factor*, which is defined in Equation 2.17 and measures how well the computed model predicts the experimental data.

$$R = \frac{\sum_{\text{all reflecions}} \left| \, |\mathbf{F}_{\text{obs}}| - |\mathbf{F}_{\text{calc}}| \, \right|}{\sum_{\text{all reflections}} |\mathbf{F}_{\text{obs}}|} \tag{2.17}$$

An R-factor of 0 corresponds to perfect agreement of the observed and computed structure factors while a value of 0.6 indicates a random matching of reflections. The R-factor of the initial map frequently lies close to 0.3–0.4. A typical R-factor for a refined map depicting a macromolecule lies at 0.2 while electron density maps of small molecule crystals regularly achieve R-factors below 0.15.[1]. A statistically more meaningful criterion is the *free R-factor* R_{free}, which avoids the circular dependency that can be found in the R-factor. It is calculated by refining the map without using a small, random set of reflections — the test set. After refinement, a cross-validation is carried out by assessing the agreement between the predicted and observed reflections in the test set. It is reported that in most cases R_{free} correlates well with the R-factor especially in successfully refined maps. As of October 2011 the R_{free} was reported for 61 300 structures stored in the wwPDB with an arithmetic mean of 0.24 and a standard deviation 0.04.

The *resolution* of a map acquired by X-ray crystallography is the second central quality measure of the depicted electron density. It is determined as the frequency of the highest recorded reflection in the experiment. A map that has a resolution of 2 Å includes, e. g., reflections of up to $\frac{1}{2\,\text{Å}}$ in the reciprocal lattice. The resolution gives a limit on the potential *resolvability* of objects in the electron density map. In a map of 2 Å resolution, e. g., it is not necessarily possible to discern all atoms. However, most maps resolution can be interpreted using the knowledge that proteins consist of a chain of amino acids of generally known sequence. As of October 2011, the 67 156 X-ray crystallography structures in the wwPDB with annotated resolution have an average resolution of 2.18 Å with a standard deviation of 1.18 Å.[2]

If the crystal was perfect and all atoms were motionless, the resolution would be solely limited by the wavelength of the X-rays. However, in nature this is not the case and various kinds of disorder prevent the achievement of electron density maps at the theoretical resolution limit. On the one hand, macroscopical errors in the crystal such as point defects but also general packing disorder and mosaicity impair the diffraction pattern. On the other hand, there is also motion within the unit cell. For once, the conformations of the molecules can be different in the unit cells. Certain segments of the protein can be completely flexible especially if they are exposed to the solvent in the crystal. These atoms are frequently unresolved and not included in the model. Furthermore, side chains

[1] More than 95 % of the Cambridge Structural Database [5], which collects structures of small molecules have an R-factor lower than 0.15.

[2] The average and standard deviation is calculated using all resolution entries in the wwPDB — including electron microscopy maps. Due to the small number of maps acquired by electron microscopy, these have only a negligible influence.

of the amino acids can adopt a certain, fixed number of different conformations. In these cases, an occupancy factor is calculated per atom, which indicates the ratio in which the atom occupied the specified position. Additionally, atoms vibrate about their mean position, which is often referred to as *thermal motion*. This effect is measured by the B–factor, which is an indicator of the isotropic motion of the atom about its mean position. However, the B–factor is only insightful if no other errors are present.

Since February 2008, the wwPDB requires the publishing of the recorded structure factors along with the atomic model. Using this information, the Electron Density Server[1] [178] allows for the download of electron density maps. The software computes maps in an automated way using the coordinate and structure factor files from the wwPDB and makes them available if the calculated R-factor lies within 5 % of the published value. Currently more than 48 000 X-ray electron density maps available from this resource.

2.2.4. Cryo-Electron Microscopy

Cryo-EM is a method, which can be used for the elucidation of electron microscopy maps of large biomolecules. For this purpose, the single particle reconstruction technique is frequently used but also tomography methods are employed, which are not discussed here. In single particle reconstruction, the specimen is depicted in aqueous solution using a transmission electron microscope. This yields two dimensional projections of the specimen that are classified by the view they are showing and averaged for increasing the signal-to-noise ratio. By determining the orientation of the projections, a three-dimensional reconstruction of the electron density is accomplished. In the following, an introductory overview of the employed techniques is given, which is based on the explanations and definitions in the standard textbook of cryo-EM [94]. More detailed descriptions of the involved steps, related experimental techniques, and further references can be found in this book but also elsewhere [112, 307, 95, 155, 156].

A transmission electron microscope can be used to record projections of a specimen. For this purpose, the specimen must be enclosed in a thin material that is partially transparent to an electron beam. The electron beam is emitted from a cathode, accelerated by an anode and passed through an elaborate set of lenses, which focus the rays — all under high vacuum conditions since molecules in the gas phase would also scatter the electron beam. Eventually, the electron beam is transmitted through the specimen making it carry information on the specimen's Coulomb potential distribution. Afterwards, the beam is focused in

[1] http://eds.bmc.uu.se — Uppsala Universitet, Kingdom of Sweden.

the objective lens and the resulting image can be visualized using, e. g., a fluorescent screen, photographic film, or a charge-coupled device sensor connected to a computer.

Biomolecules are delicate objects, which are susceptible to radiation damage. Therefore, *micrographs* — images of the specimen — are taken under low-dose conditions, which equals an exposure of less than ten electrons per square ångström. Biomolecules in aqueous solution generally do not produce high contrast in the recorded images. Therefore, a large specimen with molecular weight larger than approximately 200 kDa — depending on the particle shape — is required for a successful reconstruction. Various preparation techniques such as negative staining have been used, which all have different advantages and disadvantages. Negatively stained specimen allow, e. g., only for the deduction on information on the outer shape of the molecule. From all techniques, the preparation of the specimen in vitrified — i. e., non-crystalline — ice was identified as being best suited for revealing all features of biomolecules.

Frozen hydrated specimen are prepared by first applying a few microliters of an aqueous solution containing the specimen to an electron microscopy grid. Subsequently, the grid is blotted to remove excess buffer assuring a thin layer of solution with a thickness of less than approximately 1 000 Å. The prepared grid with the specimen is plunged into a cryogen — usually liquid ethane cooled by a surrounding bath of liquid nitrogen — achieving cooling rates of 10^5°C per second, which prevents the water from forming crystals. Eventually, the grid is transferred to a liquid nitrogen bath, inserted in a cryo-holder, transferred to the microscope and images are taken — this all is done at temperatures typically below −160°C.

The complex image formation process in the electron microscope can be modeled using the point spread function (PSF), which describes the response of an imaging system to a single point of the depicted object. Thus, the image returned by an electron microscope is the convolution of the PSF with the imaged object's electron density. The Fourier transform of the PSF is called contrast transfer function (CTF). Equal to the convolution of the real object with the PSF, the imaging process can be modeled by taking the product of the CTF and the Fourier transform of the object's electron density. Thus, the CTF describes the transfer characteristics of the optical system with respect to different frequencies. Micrographs of biomolecules are taken in underfocus since biomolecules are not easily discerned from the background in focused images. The typical CTF in underfocus has a similar appearance as a band-pass filter and its general profile looks as follows: It starts with a relatively low value at frequency zero. Then, the CTF grows and remains at a relatively high level,

forming a plateau for a certain range of low frequencies. After this interval, rapid oscillations follow at higher frequencies, which act as virtual band limit.

For the interpretation of the recorded images, a CTF correction is employed, which includes estimating the parameters of the CTF and subsequently correcting for its effects. Furthermore, the CTF may vary depending on the image region, which has to be addressed additionally. This procedure yields CTF-corrected micrographs of the specimen taken under low-dose conditions. The next step in the method is to locate all depicted particles, to classify them, and to assign orientations to the views. This process is essentially a combined object recognition and image registration task, in which similar shapes are recognized, all particles are assigned to one shape, and eventually orientations are determined.

In a first step, the particles are located in the micrograph, which is referred to as *particle picking* or *boxing*. This process yields a collection of cutout images from the micrograph — the boxed particles — which depict projections of the specimen in various orientations. Only if the projections sample the orientational space well, a high resolution map can be constructed. This, however, is not always the case since many particles exhibit orientational preferences. An example of this can be found in micrographs of GroEL, which is shown in Figure 2.4 on page 31 D. The shape of the complex resembles a cylinder and frequently lies on the top or the side on the grid. However, a tilted orientation is observed less frequently, which limits the resolution as will become clear later on.

The boxed particle images have a very low signal-to-noise ratio. To amplify the signal, similar views of the particle are identified and the projections are averaged to decrease the influence of noise. Here, the amount of noise ideally decreases with the square root of the number of averaged projections. The first step in this process is the grouping of boxed particle images into classes, which depict the specimen in the same orientation — i. e., showing the particle from the same face neglecting rotation and translation in the image plane. This is done in a process called classification, in which the boxed particles are registered so that projections that have the same shape are placed in one group. Mainly intensity based registration methods are employed for this registration task due to the low signal-to-noise ratio.

By averaging the images contained in one group, an image with a higher signal-to-noise ratio is created — a *class average*. For the calculation of a three-dimensional model, the orientation of the projection of the class average needs to be determined. This orientation can be specified using, e. g., a fixed reference frame and three angles. The determination of the relative orientations of the

class averages to each other can be performed using the Fourier transform of the 2D images: The Fourier transform of a projection of a 3D image represents a slice of the Fourier transform of the 3D image. Here, all class averages depict projections of the specimen in different angles. Therefore, the Fourier transforms of two class averages must intersect and thus share a common line of intensities. This line can be used for determining the relative orientation in reciprocal- and eventually real space. Another method is the recording of a tilt series from the specimen, which yields images with known relative orientation. However, the tilted images have lower quality since they suffer from radiation damage and certain orientations cannot be recorded because of limitations on the tilting angle. In addition to these methods, there are special techniques for icosahedral and helical reconstruction.

After assigning orientations to the class averages, the three-dimensional reconstruction of the electron density map is acquired using weighted back projection. Similar to an inverse Radon transform, the two dimensional projections of the specimen are stacked in 3D space according to the determined orientations. The reconstruction is then performed by inversely projecting or "smearing out" each class average according to its determined orientation. This process can be carried out in the frequency domain to accelerate the computation. As mentioned earlier, some orientations may be missing or inaccessible. In these cases the resolution will be limited in the corresponding orientation.

As in X-ray crystallography, a final iterative *refinement* step is employed to achieve a well resolved map. An initial electron density map of the specimen was determined in the previous steps and is used as reference. Projections are calculated from the initial map and are used as references for refining the assigned orientations of the particle images. This enables the registration of the particle images with the calculated projections of the map and yields better resolved class averages. From these class averages, again a three-dimensional reconstruction is calculated, which shows the electron density map in more detail. This process is repeated until the generated electron density map does not improve during the iterations.

The resolution of electron density maps acquired by cryo-EM is most frequently determined using the Fourier shell correlation coefficient (FSC) [341, 344, 260]. Here, the set of particle images is separated randomly in two sets of images and two independent reconstructions are performed. Subsequently, Fourier-transformed three-dimensional images are created and the normalized cross-correlation coefficient between shells in Fourier space are calculated. The shell in which this value drops below a certain threshold determines a frequency

that is reported as the resolution of the map. In most published cryo-EM maps, a threshold of 0.5 is used for the computation of the resolution.

Cryo-EM has the advantage of depicting the specimen in its genuine surrounding, which promises more insightful electron density maps. However, the resolution of cryo-EM maps frequently remains limited, since various sources of noise impede the process of reconstruction. There is, on the one hand, the noise that is due to image variability and the low-dose conditions, under which the micrographs are taken. Furthermore, conformational variability of the specimen prevents the refinement to high resolutions, since averaging of different objects cannot yield a clear image. On the other hand, this source of noise also has its benefits as it allows for the elucidation of the specimen's conformational changes.

Electron density maps acquired by cryo-EM are available through the EMDataBank[1] [195, 196], which currently provides access to more than 1 100 maps [385].

2.2.5. Summary

Two experimental methods for the elucidation of electron density maps — X-ray crystallography and cryo-electron microscopy — have been introduced. Based on these maps, atomic models can be computed. While both methods are essential for the advancement of structural biology, they both have advantages and drawbacks. X-ray crystallography is capable of producing high resolution electron density maps from crystallized proteins. The proteins are generally of small size and only in rare cases, such as the ribosome, the elucidation of macromolecular machines was successful. The creation of the crystal is one of the most challenging tasks in X-ray crystallography. It requires more than 500 pmol of the protein — a large amount considering the size of biomolecules — whereas cryo-EM requires only 0.25 pmol for the recording of micrographs. Cryo-EM is limited to large particles with more than 200 kDa total mass. Since the specimen is depicted in aqueous solution, generally no unnatural interactions are encountered as can be found in the crystal where proteins form a regular lattice. Biomolecules are flexible objects, as they generally have multiple conformational states. While crystallized proteins are all forced into a similar conformational state due to the formation of the crystal lattice, cryo-EM frequently depicts different conformational states of the molecule. This allows for the analysis of the dynamics of the biomolecule, but also prevented the refinement of high resolution electron density maps [96, 74].

[1] http://www.emdatabank.org — PDBe, RCSB and National Center for Macromolecular Imaging, Houston, Texas, United States of America.

X-ray crystallography generally produces high resolution maps, which can be used to compute atomic models. Cryo-EM is theoretically equally capable of producing high resolution electron density maps. However, different sources of noise such as conformational variability of the specimen have prevented the elucidation of high resolution maps until recently. Examples of these high resolution maps include chaperonins [219, 377] and viral particles [379, 362], since in these structures the internal symmetry of the objects can be used to boost the resolution. From these maps, backbone traces and also complete atomic models have been determined.

The registration of electron density maps with atomic models and other maps yields information on the conformational states of biomolecules [282, 2, 313, 214, 11, 332]. When considering further available experimental techniques, it becomes clear that the integration of information from various experimental resources is going to be a major source for the elucidation of biological processes: By integrating information on protein-protein interactions it becomes possible to identify connections between these molecules [285, 99]. Combining these techniques with methods for the localization of proteins in cells enables the revelation of the whereabouts of biological process [292, 313, 17, 99]. This will eventually allow for both the disentanglement of the elaborate processes underlying living organisms [234, 10, 355] and a molecular view of biology [24].

2.3. Existing Methods for the Alignment of Macromolecular Structures

Different approaches for the comparison of molecular representations have been developed [85, 11]. On the one hand, there are methods for the registration of atomic structures to each other. These compute a matching of the residues and subsequently a superposition of the atomic models. On the other hand, docking methods for the fitting of atomic structures to electron density maps have been developed, which facilitate the interpretation of electron density maps. Furthermore, there are methods for the identification of the location of protein complexes in whole-cell tomograms. To document the current state of the art and to distinguish this work, a list of computer programs, which address similar problems as *siseek*, has been complied. In the following, a broader overview of the utilization of fitting techniques for the interpretation of macromolecular structures is given. Subsequently, available computer programs for fitting larger fragments of atomic models to electron density maps are presented in more detail. Then, approaches for identifying the content of an electron density map using a database of atomic reference structures are outlined.

The resolution of X-ray crystallography electron density maps is generally sufficient to directly perform an atomic detail interpretation of the map [273]. Cryo-EM maps, on the other hand, have been of limited resolution up to only recently [95, 219, 377, 362]. The interpretation of these maps is performed by fitting macromolecular entities to the density maps. Frequently, the atomic structures of subunits of a depicted complex have been elucidated by X-ray crystallography or NMR spectroscopy and have been used to interpret the cryo-EM map. This procedure has facilitated the interpretation of maps depicting, e. g., the ribosome [102], parts of the spliceosome [312], bacterial pili [199], or insect flight muscle [371] to name only four.

The historically first dockings were achieved using computer aided methods for the interactive placement of structures in maps. The assessment of the fit of the structure to the map was subject to the expert performing the docking and no objective criterion was employed. Thus, methods for semi-automated and fully-automated docking of atomic structures to electron density maps were developed. These employ objective criteria for the measurement of the goodness of fit between the atomic structure and the map.

The docking problem — i. e., the identification of the position of molecules in electron density maps — is also relevant to X-ray crystallography. The molecular replacement technique [84] uses the structure factors of a highly similar protein to solve the phase-problem. For this purpose, a registration of the similar proteins to the protein under investigation in the unit cell needs to be determined. This is accomplished by maximizing the similarity of the structure factors determined from the similar protein structure and the experiment. One approach for solving this problem is the computation of a registration of the structure factors. [84]

Methods for the rigid docking of atomic structures to electron density maps employ various measures to determine the similarity between maps, to determine trial placements, and to score placements. These approaches are outlined below. Methods for the flexible fitting of proteins to cryo-EM maps [367, 322, 74, 321, 240, 318, 296, 159, 287, 324, 329, 331, 380] are not discussed here since they generally rely on similar concepts as rigid docking and additionally employ techniques for handling the flexibility of the proteins. Equally, methods for the computation of assemblies from rigid docking placements [167, 194, 29, 286, 378] are not discussed. These approaches either rely on the results of rigid dockings or employ optimization methods for the simultaneous fitting of all components.

In the following, methods for the rigid docking of atomic structures to electron density maps and for molecular replacement are introduced chronologically. The list was assembled for this work and comprises available computer programs

as well as published approaches that have been developed for solving similar problems as those addressed in this work. Thereby, it facilitates the comparison of the presented work to the state of the art.

MOLREP [338, 339, 340] is a method for molecular replacement and was programmed for solving the phase problem in X-ray crystallography. Using the implemented concepts, it is also possible to compare electron density maps. The method employs two major steps, an orientational and a translational search. The relative orientation of the atomic model with respect to the measured structure factor is computed using the Patterson map [258], in which peaks correspond to inter-atomic distances of the depicted molecule. Thus, a map of one single protein is independent of the location of the molecule in the unit cell, however, it is covariant with the rotation of the molecule. The rotational matching is accomplished by using solely peaks in a spherical surrounding of the origin — those which most probably are not inter-molecule distances — and is accelerated using spherical harmonics [72]. The translational search is accelerated using the fast Fourier transform. This information together with a packing function, which avoids superpositions of the molecules in the unit cell, yields the 3D placement of the molecule in the unit cell. This program is not regularly employed for docking, but for molecular replacement and can be used to determine the highest cross-correlation coefficient that can be achieved for matching two maps.

CoAn – CoFi [350, 351, 349] (Correlative Analysis – Correlation based Fitting) uses a statistical approach for fitting atomic models to electron density maps of lower resolution. In a first step, initial placements are identified by assessing different orientations of the atomic structure on a coarse grid. The best placements are identified and refined in a second step using a finer grid in the surrounding of the initial placement. For both searches, the PM-correlation coefficient of the intensity values of a synthetic map generated from the atomic structure and the electron density map is employed. The surrounding of the final placements is statistically analyzed and a significance value for each placement is calculated.

Situs [369, 370, 367, 368, 53, 365] is a software package for the integration of biophysical data on molecular structures. It comprises CoLoRes, which is described below, and a docking tool that uses feature points. Besides these docking tools, various computer programs and libraries are included in the package facilitating, e. g., format conversions. Furthermore, it is tightly linked to Sculptor (see below).

The feature point docking program of Situs [369] uses topology representing neural networks [370] to place a user-defined number of codebook vectors in the molecular entities. Placements are determined by exhaustively combining all sets of codebook vectors from structure and map. Scores are computed according to the root mean square deviation of the locations of the matched codebook vectors. A Powell optimization [266] using the feature point description can be performed for improving the placements and the final placements are then returned as solutions.

DockEM [279] employs an exhaustive, real space search and uses a *local* PM-correlation coefficient as scoring function. For this purpose, a synthetic map is created that depicts the atomic structure at the specified resolution. This map is shifted through the electron density map with a constant stride and at each location the local PM-correlation coefficient is calculated. This correlation coefficient includes solely voxels for which the synthetic map has values differing from zero. The procedure is repeated for different orientations of the molecule and, eventually, placements with high local PM-correlation coefficients are identified.

EMFit [280, 281] is a program for placing molecular fragments in electron density maps. In a first step, it performs a coarse exhaustive search and thereby selects high scoring placements. In a second step, these placements are refined using an optimization procedure. The computer program offers various scoring methods including the sum of interpolated densities at atomic sites or the absence of atomic clashes between symmetry-related positions of the atomic structure.

Foldhunter [157] uses a cross-correlation search to localize a molecular structure in an electron density map. The atomic structure is rotated to different orientations and for each orientation the cross-correlation map of a synthetic map and the provided electron density map is calculated and thereby placements are identified.

Situs – CoLoRes [53] (Correlation based Low Resolution docking) uses a cross-correlation approach to determine the best placement of the atomic structure in the map. For this purpose, the structure is rotated to different orientations and synthetic maps are created for each orientation. The computation of the cross-correlation is accelerated by making use of the convolution theorem as well as the fast Fourier transform. The best scoring solutions can be subject to a Powell optimization [266] improving the placements to sub-grid accuracy.

CoLoRes employs an initial filter stage using the Laplacian operator and thereby creates maps that depict contour information for both objects. It was found that the contrast between scores of correct placements and false placements increases when using Laplace filtered maps. Thus, final placements are selected from cross-correlation searches of Laplace filtered maps from both the atomic structure and the electron density map. CoLoRes is widely used for docking and it is reported to reliably dock atomic structures to maps of resolutions as low as 30 Å.

Sculptor [30, 144, 29] is a software for the interactive visualization and docking of atomic structures and electron density maps. It provides access to the functionality comprised in the Situs package and further methods through a graphical user interface. The resulting docking placements can be explored interactively and the best placements can be identified using, e. g., a force-feedback device. For this purpose, the scores of different placements of the structure in the map are precomputed and the user is steered towards high scoring solutions using the force-feedback.

UCSF Chimera [262, 113] is a computer program for the interactive visualization and analysis of molecular structures. When provided with an atomic structure and an electron density map, it is capable of refining the placement of the atomic structure using an optimization procedure. In this procedure, placements are scored according the sum of the densities at the atom positions or using the PM-correlation coefficient and synthetic maps. The local optimization involves a small number of steps and is therefore generally computed within seconds.

3SOM [52] maximizes the overlap of isosurfaces determined from a synthetic map of the atomic structure and the experimental electron density map. The isosurfaces are specified by thresholding the map at a user-defined value. Using these surfaces, trial placements are determined by superposing surface voxels and aligning the maps according to the surface normals. Then, a rotational scan around the axis defined by the normal is carried out and all placements are scored according to the number of overlapped surface voxels. The placements with highest surface overlap can be reranked according to different scoring functions. As a consequence of the method, the atomic structure must have an exposed surface and sufficient detail of this surface must be depicted correctly to be able to identify correct placements. The method is reported to be very fast and proved effective in case studies. However, it was also reported that it is difficult to identify correct solutions in the abundant amount of generated placements since

these are not always assigned the highest score or because the root mean square deviation (RMSD)[1] of the placements is high [107].

EMatch [192, 82, 193] uses information on the spatial arrangement of secondary structure elements in the depicted proteins. In a first step, α–helices are identified in the electron density map using image-template matching and thresholding. The determined map is analyzed and the spatial arrangements of α–helices is computed. This information is used to query a database holding information on the secondary structure arrangements of reference proteins.

Mod-EM [328] uses real space PM-correlation coefficients to identify the best placement of the atomic structure in the map. The method creates a synthetic map from the atomic structure and uses different search strategies to identify trial placements. If there is only one molecule depicted in the map, then the center of mass of the two maps is superposed and a rotational scan is performed. Otherwise, optimization procedures based on the Monte Carlo optimization method can be employed.

ADP_EM [107, 183, 184] uses spherical harmonics to accelerate the fitting of atomic structures to electron density maps. In a first step, all voxels are identified at which the atomic structure can potentially be placed. For this purpose, the map is thresholded using a user specified density threshold and the resulting map is eroded by the minimum radius of the atomic structure. For each of these voxels, concentric spherical layers are described by means of spherical harmonics. By expressing the correlation function in terms of the calculated spherical harmonics, it is possible to accelerate the scan in the three rotational degrees of freedom. The remaining spatial degrees of freedom are scanned separately. ADP_EM makes use of an initial Laplacian filtering to increase its capabilities of docking into low resolution electron density maps.

ADP_EM is evaluated on a large test set and proved to be as accurate as CoLoRes for resolutions as low as 30 Å. The computer program is very fast and requires on average 34 s for the tested maps while the average run time of CoLoRes without optimization was 25 min.[2] Therefore, ADP_EM is used as benchmark in the results chapter.

Phaser [231, 271, 315, 232] is a software that provides methods for the solution of the phase problem in X-ray crystallography using either molecular

[1] See Equation 3.20 on page 91.

[2] The experiments were performed on a single machine with a 2.8 GHz processor.

replacement or experimental phasing methods. The provided molecular replacement employs a random walk and assesses various placements of the model structure in the unit cell. For each placement, structure factors are calculated and during an optimization process the similarity of these factors with the measured data is maximized. Multivariate statistics are applied to measure the quality of fit based on the similarity of the calculated and measured data: The log-likelihood gain indicates how much better the data can be predicted from the placement than from a random-atom model. This value can be used to compare the quality of different models against the same data. Furthermore, a Z-score is calculated, which indicates the clearness of the solution. It is defined as the multitude of root mean square deviation that the placements log-likelihood gain lies above the mean log-likelihood gain.

UROX [302, 250, 249] is an interactive tool for the fitting of atomic structures to electron density maps, which is inspired by the molecular replacement technique from X-ray crystallography. The user places the atomic structure inside the map and for each placement the cross-correlation of the synthetic map generated from the atomic structure and the electron density map is calculated. This is facilitated by accelerating the computation of the correlation in the frequency domain taking into account the symmetry of the map.

Segger [264, 263] uses a combination of the watershed method and scale-space analysis to segment electron density maps. The segments can be used for identifying fittings of atomic structures to the map. For this purpose, the center of mass of the structure is superposed with the segment's center of mass. The orientation is assigned by either superposing the principal axes of the objects or performing a coarse initial search. Subsequently, an optimization procedure is employed for the refinement of the position using the PM-correlation coefficient of the underlying voxel densities as score.

MOTIF-EM [290] is a computer program that uses the SIFT descriptor for comparing regions of electron density maps. In this way, it allows for the identification of similarities between atomic structures and maps, but also for the direct comparison of maps. The method assigns one orientation to each voxel of the map using the covariance matrix of the surrounding density. According to this reference frame, a fixed-size SIFT descriptor is computed and saved as 208-dimensional feature vector. Similarities in the maps are identified using a clique searching strategy to identify distance

compatible voxels with similar descriptors. No keypoints are calculated and therefore also no scale-invariance is achieved. Thus, the method compares fixed size sub-volumes of the input-maps.

MOTIF-EM was used for the direct comparison of electron density maps. In case studies it proved capable of identifying similar sub-volumes of simulated and experimental input maps, which can be used for the detection of conformational changes in macromolecules. Furthermore, it was used to locate atomic structures in electron density maps using synthetically generated maps. The required computations of the method are quite demanding since a descriptor is calculated for every voxel and not just for salient keypoints. This results in accumulated run times of 22 days for a single map comparison.[1]

The above methods can be utilized to identify the unknown content of a given electron density map using a database of reference molecules. Recently, first methods addressing this task have been published. These methods compare the scores measured when registering the reference structures to the map and thereby identify the content of the map. Thus, the goal of these methods is to find all similar proteins in the reference database.

SPI-EM [346] (Superfamily Probability In Electron Microscopy) aims to identify the CATH homologous superfamiliy of a protein depicted in an electron density map, i. e., it searches for specific fold motives. For single domain maps, the program FRM [183], which is similar to ADP_EM, is used to identify the best placement in the map and a local correlation coefficient for the best placement of the reference domain is determined. Subsequently, the statistical significance of the matches is assessed. This setup was tested using 28 synthetic protein domain maps with 8 Å resolution and was successful in 80 % of the cases. In a second step, multi-domain maps are analyzed using the program CoLoRes to identify the best placements of the domains. From this placement, the local correlation coefficient is computed and used for scoring. The method succeeded on synthetic maps depicting GroEL and DNA polymerase I. However, it did not succeed on experimental maps.

FREDS [173] (Fold Recognition Electron Density Search) uses the cross-correlation coefficient determined by MOLREP to identify a correct reference structure for a given query map. FREDS requires a segmented electron

[1] The experiments were performed on a 512 node compute cluster with 2.33 GHz processors.

density to be able to identify correct reference structures. For each reference structure, the best cross-correlation coefficient of atomic structure and map is determined using MOLREP. Subsequently, the correlation coefficients are normalized according to the domain size of the reference structure. The reference structure set is assembled by first filtering the wwPDB using a sequence identity criterion of 30 % and disassembling the remaining structures in 16 087 domains. The method was successfully tested on a set of nine manually segmented experimental cryo-EM maps with resolution of 6–8 Å and used 200–900 h of accumulated computing time.[1]

WS-MR [314] (Wide Search Molecular Replacement) uses a combination of the log-likelihood score and the Z-score determined by Phaser to score the correspondence between reference structures and the query map. The references comprise 95 000 domains found in the SCOP database and facilitate the search for molecular replacement models. Using three case studies, it was shown that the method is capable of identifying molecules that make up only a small part of the query map and proteins with low sequence identity. The calculations were performed on a computer cluster using an accumulated processing time of 800 days (19 200 h) per run.[2]

A Fingerprint Based Method [375] uses 3D Zernike moments to identify a molecule depicted in an electron density map. For this purpose, twenty Zernike moments are calculated, which allow for the description of 3D objects similar to geometric moments and spherical harmonics. The center for calculating the Zernike moments is the center of mass of the considered object. No keypoints are calculated in this method. The comparison of the moments is performed using the cosine distance. The electron density needs to be segmented prior to searching so that the electron density corresponds to the molecule that is about to be identified. In this way it can be assured that the center of mass of the query map and the reference structures coincide. All entries of the wwPDB are utilized as reference proteins. The search is very efficient, since a pre-filter on the molecular weight of the molecule is employed and in total one 20-dimensional feature vector is determined for each wwPDB structure. This allows for the screening of 800 000 structures per second in the performed test.[3] The method was successfully tested on two case studies using segmented densities from a

[1] Calculations performed on multi-core computers with 2.66 GHz processors.

[2] The computations were carried out using the Open Science Grid [265] relying on non-homogeneous hardware.

[3] Computations performed on a single machine with 3 GHz processor.

6 Å resolution GroEL map and a 5.5 Å resolution bovine metarhodopsin I map.

The presented list of related methods comprises twenty-one methods that are related to the problem of image registration and four methods for identifying the unknown content of a given electron density map. In Table 2.1 a summary of the methods and their approaches is found. The table lists the approaches for computing placements, scoring, and for post optimizing. Furthermore, methods are marked with a bullet in the placements column if two conditions are fulfilled. First, the method must not scan placements located on a regular grid. Second, it may not rely on any more input but the map and its properties. A bullet in the scoring column indicates that non-trivial descriptors are employed for determining the goodness of fit.

The table shows that most of the available computer programs rely on an exhaustive search on a regular grid. Only 3SOM, EMatch, Segger, and Situs do not perform a brute-fore scan of a set of placements on a regular grid while also not relying on an initial placement defined by the user.[1] This shows that only four out of the listed seventeen different methods identify placements based on image features rather than employing an exhaustive scan. A similar finding holds for the employment of image descriptors. Here, the six methods ADP_EM, 3SOM, the fingerprint based method [375], EMatch, MOTIF-EM, and Situs employ comparisons that are not solely based on the correlation of the superposed image intensities.

As shown in the previous sections, many state of the art image matching techniques are feature based. This facilitates, on the one hand, an abstract image representation and allows, on the other hand, for efficient image comparisons. Taking into account the growing number of available electron density maps and

[1] 3SOM requires the specification of a density threshold and Situs the specification of the number of feature points.

Table 2.1 (following page) – Existing Methods for the Alignment of Macromolecular Structures

The table lists the presented computer programs and summarizes the employed methods for generating placements, scoring, and for post-processing. Automated methods that do not rely on a grid scan and fit maps without further user-input are marked by a • in the Placements column. Those methods, which utilize an abstract descriptor, are marked by a • in the scoring and post-optimization column. FREDS, Sculptor, and WS-MR are listed along the methods they are based on. (PO: Post-optimization; CC: Correlation Coefficient; SIDA: sum of interpolated densities at atom positions)

Name	Placements		Scoring and Post-optimization	
ADP_EM [107, 183, 184]		Grid in a map sub-volume	•	Spherical harmonics
3SOM [52]	(•)	Surface alignment (Requires the specification of a density threshold.)	•	Surface overlap and correlation coefficient
Fingerprint Based Method [375]		Manual segmentation, Center of mass	•	Zernike moments
CoAn – CoFi [350, 351, 349]		Regular grid, Real space		CC
DockEM [279]		Regular grid, Real space		CC
EMatch [192, 82, 193]	•	Helix detection, Helix alignment	•	Helix matching; PO: CC, Least squares
EMFit [280, 281]		Manual placement, Discrete orientation scan		CC, SIDA, Number of atoms not covered by density; PO: Any scoring function, Least squares
Foldhunter [157]		Cross-correlation		CC
Mod-EM [328]		Superposition of center of mass; Orientation by Monte Carlo optimization		CC
MOLREP [338, 339, 340] (FREDS [173])		Regular grid, Reciprocal space		CC
MOTIF-EM [290]		Regular grid, Real space	•	SIFT descriptor
Phaser [231, 271, 315, 232] (WS-MR [314])		Random walk		CC
Segger [264, 263]	•	Segmentation, Center of mass superposition, Orientation by grid search or principal axes		CC; PO: UCSF Chimera
Situs – CoLoRes [53] (Sculptor [30, 144, 29], SPI-EM [346])		Cross-correlation		CC; PO: CC, Powell
Situs [369, 370, 367, 368, 53, 365] (Sculptor [30, 144, 29])	(•)	Vector quantization feature points, Point distribution (Requires the specification of the number of feature points.)	•	Feature point distribution; PO: Feature points, Powell
UCSF Chimera [262, 113]		Manual placement, Steepest ascent optimization		CC, SIDA
UROX [302, 250, 249]		Manual placement, Stochastic optimization		CC

the advances made in the field of image analysis, there is a clear demand for assessing the applicability of modern techniques to the problems of structural biology. The remainder of this work is geared to this topic and presents a method coined *siseek* for electron density map registration and molecule recognition, which is based on the SIFT.

2.4. Summary

The integration of image data of various levels of scale yields novel insights into the details of biochemical processes. Imaging methods for the elucidation of microscopic structures in complex organisms are readily available [234, 10, 355] and in recent studies, the location of macromolecular machines in whole-cell tomograms was identified, which eventually makes the assignment of processes to cell compartments possible [292, 313, 17, 99]. The integration of these findings — gained from these heterogeneous experimental resources — will continue to yield information on biological structures at different scales.

The state of the art in both image analysis and structural biology has been improved majorly in recent years. It is now with ease possible to register two-dimensional images using freely available computer programs. Also, methods for the registration of three-dimensional image data are employed regularly for the registration of medical image data [119]. Additionally, emerging technologies in structural biology have allowed for new insights into the composition of cells and organisms. While X-ray crystallography remains the major source of information on atomic structures of biomolecules [273], further methods have supplemented this technique in the last years. Cryo-electron microscopy is one of these methods and a major source of information on the spatial composition of large biomolecules [94].

The combination of high resolution data gained from X-ray crystallography and low-resolution information from cryo-EM has facilitated the model building of various biomolecular complexes [282, 2, 313, 214, 11, 332]. Examples include — among several others — models of the ribosome [102, 240], parts of the spliceosome [312], the type III secretion injectisome [224, 223], molecular motors [352, 109], bacterial flagella [246, 239], bacterial pili [199], and insect flight muscle [371].

The summary of related publications in Section 2.3 shows that the computer aided interpretation of macromolecular electron density maps has been in the focus of research already before bioinformatics was established as distinguishable branch of science and still is an active field of research. The compiled review demonstrates that most of the presented approaches rely on an exhaus-

tive scan of placements for computing a registration of two maps. Only four of the seventeen distinguished methods rely on non-trivial, feature based strategies for identifying candidate placements, while only six perform the scoring not directly based on a correlation of the corresponding density values. Given that many successful image analysis methods rely on feature based representations, research is needed that analyzes the applicability of contemporary, feature based techniques to the problem of interpreting macromolecular electron density maps.

The objective of this work is to employ state of the art image analysis techniques for the analysis of macromolecular electron density maps. For this purpose, a suitable theoretic foundation, the SIFT, was selected, extended, and implemented as detailed in Chapter 3. The resulting software system *siseek* was validated and tested thoroughly. It proved to be productively applicable for the registration of intermediate and high resolution macromolecular electron density maps, as detailed in Chapter 5.

3. Methods

The objective of this work is the development and validation of a method for similarity searching in electron density maps. To achieve this goal, keypoints in three-dimensional (3D) images are detected and descriptors for these keypoints are generated. The method is based on the scale-invariant feature transform (SIFT) [217, 218] and this chapter explicates the theoretical foundations, algorithmic concepts, and the parameters of the method. The determination of optimal parameter settings is discussed in the following chapter.

In the first four sections, methods for deriving an abstract map representation are introduced, as shown in an overview in Figure 3.1 A. First, an adaption of the sampling rate according to the map's resolution is discussed in Section 3.1. Keypoint detection is explained in Section 3.2. It relies on the theory of scale-space in a pyramidal setting and is equal to the method employed in the 2D SIFT [217, 218]. Orientation assignment to points in 3D images is introduced in Section 3.3, and a method for local neighborhood descriptor computation is outlined in Section 3.4. Both, the 2D SIFT orientation assignment and descriptor computation, are not directly transferable to 3D space. Thus, extensions to 3D space have been developed as described in the according sections.

Using this map representation, similarity searching in electron density maps is facilitated. An algorithm for the rigid registration of maps based on the determined keypoints, orientations, and descriptors is introduced in Section 3.5. Here, the probability that two keypoints originate from the same local neighborhood is assessed using a descriptor similarity measure. This information is used as basis for the computation of relative orientations of the 3D images that superpose similar regions as depicted in Figure 3.1 B. Furthermore, a method for the recognition of molecules that are depicted in an electron density map is described in Section 3.6. Here, map descriptions for reference protein structures are stored in a database and the map description of a given query map is compared against all references. The resulting list ranks the reference protein structures according to their similarity to the query map as illustrated in Figure 3.1 C.

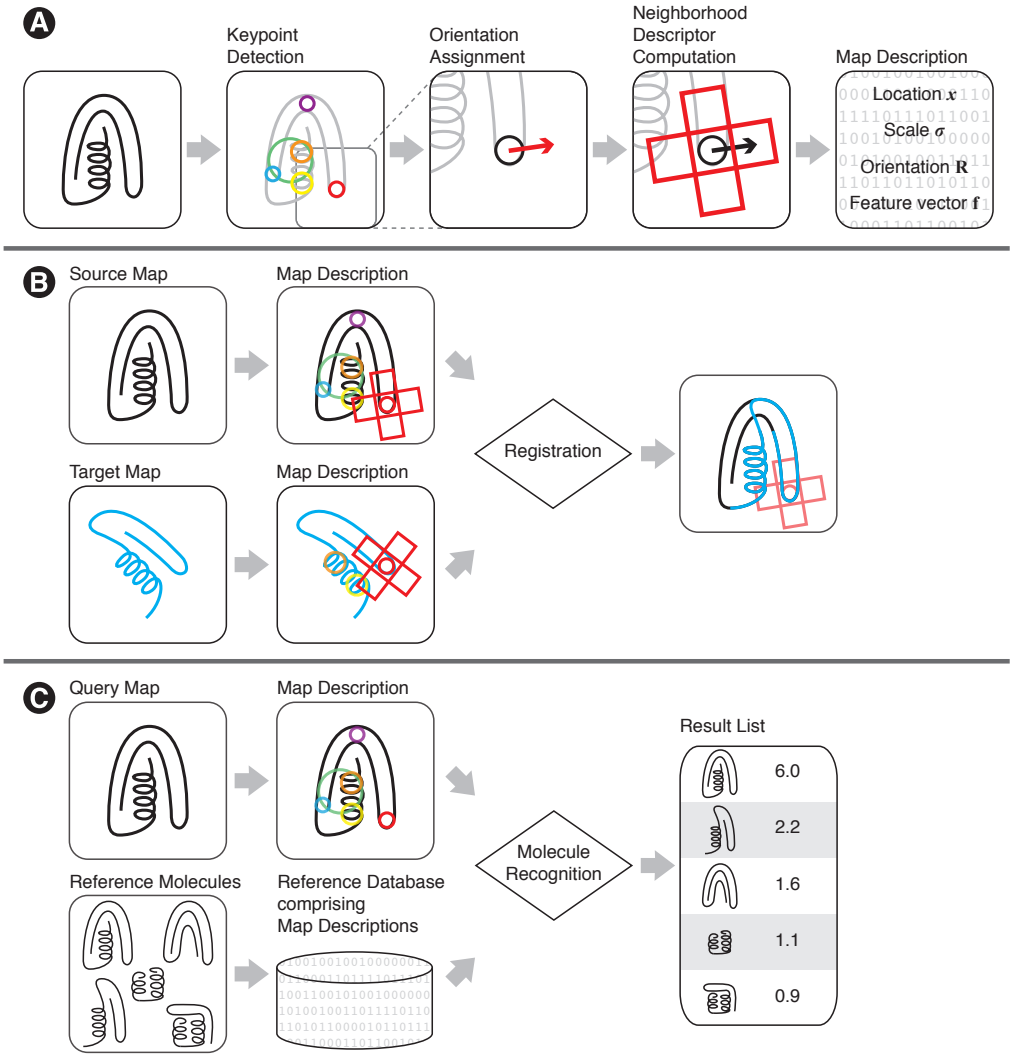


Figure 3.1 – General framework
Panel A illustrates the process of computing a map description, which consists of the stages keypoint detection, orientation assignment, and neighborhood descriptor computation. Panel B show a sketch of the approach for registration: By identifying similar descriptors, alignments for source and target map are computed and used for superposing the electron density maps. In Panel C an overview of the method for molecule recognition is shown. For a given query map, a map description is derived and compared to a database consisting of map descriptions of reference protein structures. The result list ranks the reference protein structures according to their similarity to the query map. (© A. Griewel)

3.1. Resolution Model

The definition of *scale* in scale-space theory is related to the concept of *resolution* in both X-ray crystallography and cryo-EM.[1] It can be interpreted as the level of resolution and therefore the amount of structural detail present in the image. However, there are different measures of resolution defined for cryo-EM and X-ray[2]: In X-ray crystallography the resolution is based on the highest frequency structure factor used for the computation of the map [273]. For cryo-EM, measures that correlate to the definition in X-ray crystallography are sought [94] and most frequently the Fourier shell correlation coefficient (FSC) [341, 344, 260] is used. In this work, the resolution of electron density maps is approximated using a unified model. Each map is assigned a Gaussian point spread function (PSF) [78, 94, 368] with a standard deviation that depends on the resolution of the map. This modeling of resolution is frequently used for generating synthetic maps [350, 279, 367, 320, 328, 159, 329, 264, 366] and simulates an isotropic, thermal motion for each atom. It does not account for flexible protein domains or concerted motions of larger segments of macromolecules since there is no consistent way of predicting these motions.

Maps are provided with an estimate of the resolution and are sampled on a cubic, isotropic grid. The standard deviation σ of the Gaussian PSF is determined according to the specified resolution R of the map. Different relations between σ and R have been used [350, 279, 367, 320, 328, 159, 329, 264, 366]. In this work, the normalization factor used in the Situs package [366] is employed

$$\sigma = \frac{R}{2\sqrt{3}} \tag{3.1}$$

This relation was empirically determined by the authors of Situs and verified on various maps. It proved effective in experiments and is thus chosen for modeling resolution in this work.

Given the point spread function, *synthetic* electron density maps at different resolutions can be created from atomic models as shown in Figure 3.2. This in turn facilitates the docking of atomic structures to experimental electron density maps. An exact simulation of the electron density of a single atom would require elaborate quantum mechanics. Small scale effects are, however, negligible at expected resolutions. Thus, each non-hydrogen atom is represented by a Gaussian function with a standard deviation corresponding to the defined resolution [350, 279, 367, 320, 328, 159, 329, 264, 366]. For atom i, the Gaussian

[1] Here, resolution does not refer to an image-pyramid as explained in Section 2.1.4.2 on page 22.

[2] See page 44 and page 48.

function is centered on the corresponding atom position $\mathbf{a}_i$. Furthermore, a multiplicative weighting is used to represent the element identified by atomic number Z_i. The electron density D for a complete molecule comprising n atoms is defined as the sum of the contributions of all atoms a defined in Equation 3.2. This function is sampled on an isotropic, cubic grid, and the sampling rate can either be chosen to correspond to another map, or it can be specified by the user. However, in order to avoid aliasing effects, the sampling interval must be sufficiently small. The dimensions of the map are chosen so that the molecule plus a padding of 6σ fits inside the volume. Thus, the truncation error for the sampling of the Gaussian function is negligible.

$$D(\mathbf{x}) = \sum_{i=1}^{n} \frac{Z_i}{\left(\sqrt{2\pi} \cdot \sigma\right)^3} \cdot e^{-\frac{|\mathbf{a}_i - \mathbf{x}|^2}{2\sigma^2}} \tag{3.2}$$

3.2. SIFT Keypoint Detection

Keypoints are detected in a multi-stage procedure, which equals the method applied in the SIFT [217]: After adapting the sampling rate of the provided map, an image-pyramid is created, and each level of the pyramid is analyzed using a scale-space representation. From each scale-space representation, keypoints are determined and saved for further processing. The details of this procedure are explained below and are depicted in Figure 3.4.

The genuine, provided map is called *input map*. For the reliable identification of keypoints, it is necessary to adapt the sampling interval of the input map with respect to its resolution. The resampled map is called *base map* and is used for creating the image-pyramid. For creating the base map, the standard deviation σ of the input map's point spread function is analyzed with respect to the sampling interval d

$$\{\sigma\}_{\text{vox}} = \frac{\sigma}{d} = \frac{R}{2\sqrt{3} \cdot d} \tag{3.3}$$

where R is the user-specified resolution of the map. The result of this calculation $\{\sigma\}_{\text{vox}}$ is the standard deviation expressed in terms of sampling intervals rather than in an absolute spatial unit like the ångström.

The sampling interval of the input map d_{in} is adapted so that the standard deviation of its point spread function $\{\sigma_{\text{in}}\}_{\text{vox}}$ equals the parameter $\{\sigma_0\}_{\text{vox}}$,

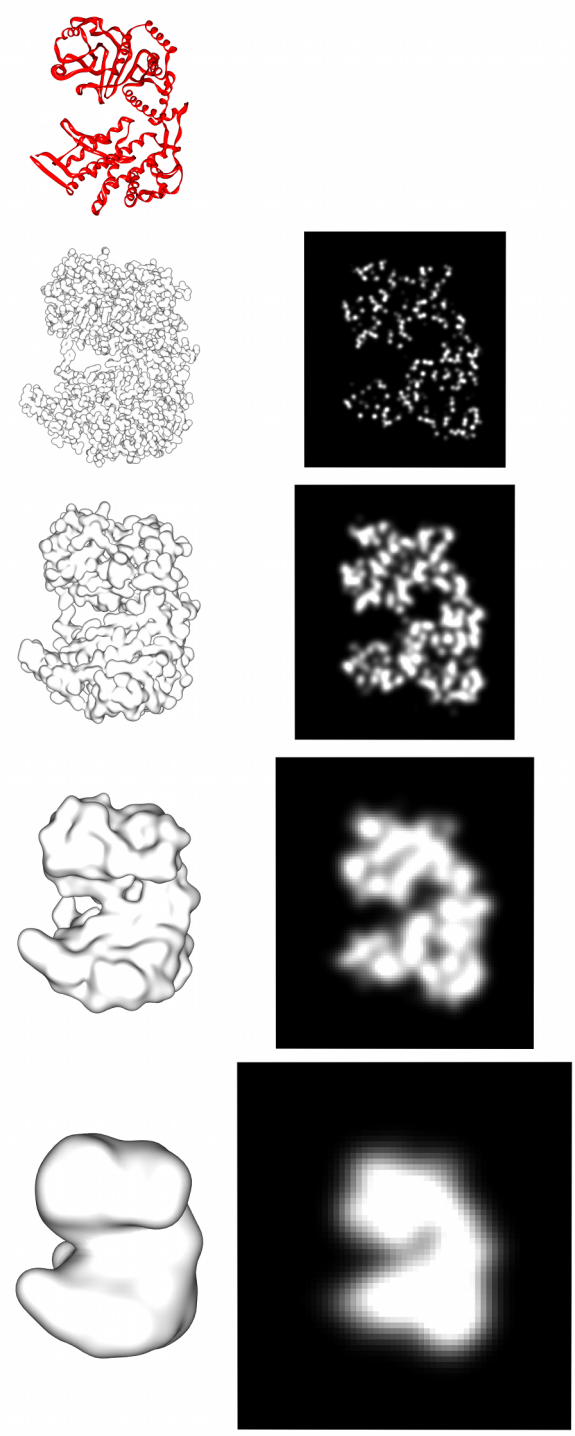


Figure 3.2 – Resolution lowering Maps of different resolutions generated with *siseek* for GroEL [34] are shown. The ribbon model of GroEL is shown on top. The following maps are blurred to resolution 2 Å, 4 Å, 8 Å, and 16 Å. The sampling interval is adapted to the resolution and equals 0.25 Å, 0.5 Å, 1 Å, and 2 Å respectively. On the left hand side, an isosurface of the respective map is shown, while the images on the right hand side display one slice of the map. The volume of the maps increases with larger resolution since the width of the point spread function is widened and therefore more volume must be covered to accommodate the signal. The number of voxels in the map, however, does not increase due to the larger sampling interval. (© A. Griewel)

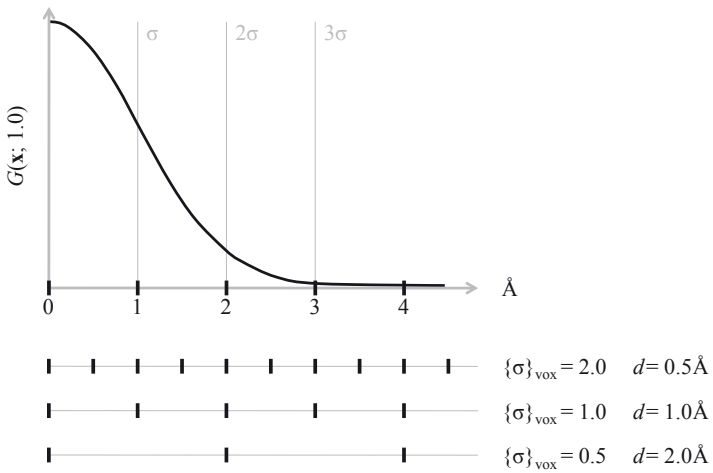


Figure 3.3 – Sampling of a Gaussian function
The plot shows the density profile of a Gaussian point spread function with standard deviation $\sigma = 1\,\text{Å}$. Below the plot, three samplings of the function are shown with different values for $\{\sigma\}_{\text{vox}} = \frac{\sigma}{d}$ and the resulting sampling interval d. (© A. Griewel)

which is empirically determined in Section 4.2. The sampling interval d_0 of the base map is determined as

$$d_0 = \frac{\{\sigma_{\text{in}}\}_{\text{vox}} \cdot d_{\text{in}}}{\{\sigma_0\}_{\text{vox}}} = \frac{\sigma_{\text{in}}}{\{\sigma_0\}_{\text{vox}}} \tag{3.4}$$

The effect of different values of $\{\sigma\}_{\text{vox}}$ on the sampling interval is illustrated in Figure 3.3 using a standard Gaussian function with $\sigma = 1.0$. The input map is resampled to a voxel spacing of d_0 using trilinear interpolation and the intensity values of the resampled map are normalized to the interval $[0; 1]$. The resulting map is used for further processing and called *base map* P_0.

The map P_0 serves as base for an image pyramid with a downsampling factor of 2. The resolution of P_0 is specified by the standard deviation σ_0 of the associated Gaussian point spread function. For each level $i+1$ in the pyramid, a map P_{i+1} is created from a low-pass filtered version of the map in the preceding level P_i. For this purpose, map P_i with resolution σ_i is low-pass filtered to a resolution of $2\sigma_i$. This is achieved by using the semi-group structure of the linear Gaussian scale-space

$$\begin{aligned} P_{i+1}(\mathbf{x}) &= G\left(\mathbf{x}; \sqrt{2\sigma_i^2 - \sigma_i^2}\right) * P_i(\mathbf{x}) \\ &= G\left(\mathbf{x}; \sigma_i\right) * P_i(\mathbf{x}) \end{aligned} \tag{3.5}$$

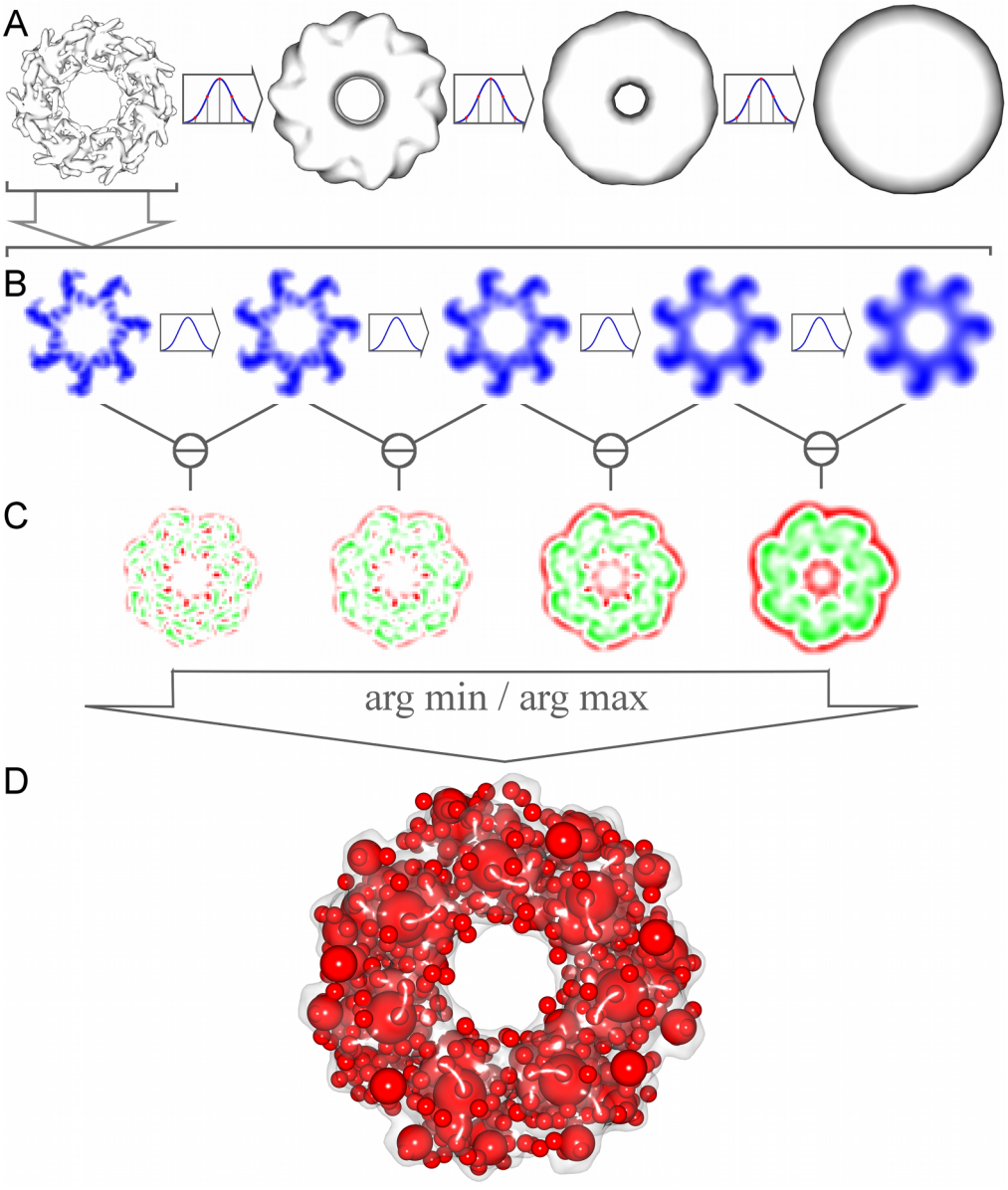
where $*$ denotes convolution. After downsampling, the resolution of P_{i+1} is $2\sigma_i$. With respect to the voxel spacing, however, the resolution is the same as in the base map, because the voxel spacing has doubled as well. This process is iterated, until the downsampled map has less than eight voxels in any dimension as shown in Figure 3.4 A.

Each map in the pyramid representation is analyzed using a linear Gaussian scale-space representation and called *octave*. Scale-space maps are sampled with different scales as shown in Figure 3.4 B. The process of creating a scale-space representation is equal for all levels of the pyramid. Hence, the process is described for octave i using P_i as the base of the scale-space $L(\mathbf{x}; \varsigma_0) = P_i(\mathbf{x})$. The scale of the first map in the scale-space is specified by $\varsigma_0 = \sigma_i$, where σ_i is the resolution of the image in the pyramid. The generated scale-space representation will be utilized for the detection of blobs at scales in the interval $[\varsigma_0; 2\varsigma_0]$. This is facilitated by sampling s *Difference of Gaussians maps* (DoG maps) plus one additional map, which is going to be used for interpolation in a later stage. Thus, in total $s + 2$ *Gaussian maps* are created in the scale-space by filtering. This yields $s + 1$ DoG maps after subtraction.

To incorporate scale-normalization [210], the Gaussian maps are separated by a constant, multiplicative factor k. Given the interval $[\varsigma_0; 2\varsigma_0]$ and the number of samples s along the scale dimension, the scale ς_j for the Gaussian scale-space map $j \in \{0, \ldots, s + 2\}$ is determined by

$$\varsigma_j = 2^{j/s} \cdot \varsigma_0 \tag{3.6}$$

Figure 3.4 (following page) – Keypoint detection
Keypoints are detected by sampling scale-spaces in a pyramidal setup, which is exemplified here using a map of GroEL [34]. (A) First, an image pyramid is built. From the left to the right the maps are low-pass filtered and downsampled by a factor of two. (B) For each level of the pyramid, a scale-space representation is created. Here, the scale-space representation of the first map on the upper left corner is shown by displaying one slice of the map. Each scale-space representation encompasses a scale-interval $[\varsigma; 2\varsigma]$ where ς is the scale of the first map in the octave. (C) Difference of Gaussians maps are created by subtracting neighboring maps in the scale-space representation. Red signifies positive and green negative intensity values. White areas have zero intensity. (D) Keypoints are identified as extrema in the Difference of Gaussians maps with respect to the spatial as well as the scale domains. They are depicted as red spheres where the size of the sphere corresponds to the keypoint's scale (larger spheres have been omitted for clearer view). (© A. Griewel)


A
B
C
arg min / arg max
D

Here again, the semi-group property of scale-space allows for the computation of maps for all sampled scales ς_j using the base map as input

$$L(\mathbf{x};\varsigma_j) = G\left(\mathbf{x};\sqrt{\varsigma_j^2-\varsigma_0^2}\right) * L(\mathbf{x};\varsigma_0) \tag{3.7}$$

This way of building the scale-space is referred to as *base formation*, since all scale-space maps are computed from the first map. Alternatively, the scale-space can be created using *incremental formation* by applying a Gaussian filter to the preceding map in scale-space

$$L(\mathbf{x};\varsigma_j) = G\left(\mathbf{x};\sqrt{\varsigma_j^2-\varsigma_{j-1}^2}\right) * L(\mathbf{x};\varsigma_{j-1}) \tag{3.8}$$

The base formation method employs larger filters due to the larger differences in scale. Therefore, the resulting maps are more exact, because the Gaussian filter is sampled with a higher frequency. The incremental formation, on the other hand, uses smaller filters and is therefore faster.

Difference of Gaussians maps D are calculated by subtracting neighboring Gaussian maps

$$D(\mathbf{x};\sigma) = L(\mathbf{x};\varsigma_{j+1}) - L(\mathbf{x};\varsigma_j) \tag{3.9}$$

as shown in Figure 3.4 C. This yields the DoG scale-space representations, which cover the range of relevant scales and facilitate the detection of blobs [218].

Keypoints are identified for each octave separately as shown in Figure 3.4 D. A keypoint is a local extremum in the four-dimensional scale-space hypervolume generated by the DoG maps of each octave. This hypervolume consists of three spatial and one scale dimension. Voxels with extremal intensities are detected using a sequential scan of all voxels. First, each scale-space map is scanned separately and voxels that are extremal with respect to their 6-neighborhood are identified. This neighborhood comprises all voxels that are within a distance of one sampling interval of the considered voxel. Subsequently, it is tested if the voxel is also a local intensity extremum with respect to the scale-dimension. For this purpose, the intensity value of the local extremum in the map is compared to all voxels that lie in the 6-neighborhood in the two neighboring scale-space maps. If the voxel intensity is an extremum with respect to the defined neighborhood, it is saved for further assessment as a potential keypoint.

After the initial detection of extremal voxels, the exact position of each local extremum is interpolated using a Taylor expansion up to the second order [42]. This yields off-grid keypoints and is especially helpful for higher octaves, where the sampling interval of the map is large. To determine the exact location $\widehat{\mathbf{x}}$ of the on-grid extremum $\mathbf{x}$, the Jacobian J and the Hessian H are calculated at $\mathbf{x}$

in the DoG map that corresponds to the scale of the extremum. These derivatives are approximated using finite differences between intensities of neighboring voxels. The exact position of the extremum $\widehat{\mathbf{x}}$ is then calculated as

$$\widehat{\mathbf{x}} = \mathbf{x} - H^{-1}J \tag{3.10}$$

and the DoG map intensity value at the interpolated position is given by

$$D(\widehat{\mathbf{x}};\sigma) = D(\mathbf{x};\sigma) + \frac{1}{2}J^T\widehat{\mathbf{x}} \tag{3.11}$$

If the calculated offset is larger than half the voxel spacing, the extremal voxel is not the closest voxel to the interpolated extremum. In this case, the interpolation is carried out once more using the voxel that is closest to the interpolated position.

The resulting list of interpolated extrema is screened and keypoints with a small intensity value are discarded using a thresholding approach [42]. In a first step, all keypoints with an intensity value smaller than a predetermined parameter t_{contrast} are discarded. Subsequently, the cornerness of the local neighborhood of the keypoint is assessed using the determinant of the Hessian matrix. The entries of the matrix are computed using differences of intensities that are interpolated in a sampling interval distance from the location of the keypoint on the corresponding DoG map. The eigenvalues of the Hessian are proportional to the principal curvature of the intensities in the local neighborhood. If the Hessian is indefinite, i. e., if the eigenvalues have different signs, the keypoint is discarded because it is located at a saddle point. For the remaining keypoints, the ratio of the largest to smallest eigenvalue is required to be smaller than a predetermined parameter $t_{\text{cornerness}}$. This ensures, that the principal curvature is not dominated in one direction and therefore ensures that the location of the keypoint is well defined.

Eventually, this process yields keypoints $(\mathbf{x}, \sigma, e)$ that are qualified with a spatial location $\mathbf{x}$ and a scale σ. Furthermore, a binary value e is saved, which indicates whether the keypoint is a maximum or a minimum in the DoG hypervolume.

3.3. Orientation assignment

Each keypoint is assigned orientations according to its local neighborhood in the image. For this purpose, discrete orientations are determined from the gradient in the keypoint's local neighborhood. These orientations are used to align descriptors and allow for computing a registration of two maps. To determine

orientations, the gradient in a local neighborhood is sampled. Based on the set of gradient vectors, a histogram is created, which in turn allows for computing dominant orientations.

In the following three sections, a 3D orientation histogram and a geodesic index are introduced. The *orientation histogram* tesselates the sphere surface into equally sized bins. The bins are populated by inserting vectors into the histogram, which is a computationally demanding task. Therefore, a *geodesic index* is used to accelerate the insertion. Using the assembled orientation histogram, *orientations* are computed by identifying peaks in the histogram. These peaks are used for the calculation of a rotation, which represents the orientation.

The presented method for determining dominant orientations in the local gradient field follows the procedures used in the SIFT, which determines 2D orientations using a circle-based histogram [218] as described in Section 2.1.5 on page 23. In the 2D case, the circle is discretized uniformly by splitting it into equally sized sectors. However, determining orientations in 3D space is more intricate than handling 2D orientations. It is, e. g., not possible to spread an arbitrary number of points uniformly on a sphere surface [289]. To address this problem, means for uniformly distributing points on the sphere surface have been identified and implemented for the orientation histogram. Furthermore, a geodesic index tailored to the orientation histogram's properties is developed in this work to speed up the process of inserting vectors into the histogram.

3.3.1. Orientation Histogram

An orientation histogram gathers information on a set of 3D vectors, which are computed from a local neighborhood in the map. The determination of orientations must be rotation covariant, because it is not known a priori in which orientation in 3D space the object is depicted. Thus, a uniform sampling of the sphere surface is necessary to avoid a bias in the calculation of dominant orientations. Representing each orientation bin by a point on the sphere surface, this requires a uniform distribution of points on the 3D sphere.

Uniform distributions of points on the sphere are found in various disciplines in science. Directly related to this work is single particle reconstruction in cryo-EM where a uniform distribution of points on the sphere is essential for a successful classification of particles. Besides that, various questions in several scientific disciplines such as climate modeling in meteorology [291], molecular structure in chemistry [111, 186], or viral morphology in biology [330] are related to the problem.

An ideal distribution of points on the sphere would be highly symmetrical. This, however, is only possible for 4, 6, 8, 12, and 20 points. The arrangements

of points for those numbers are described by the five Platonic solids, which are regular, convex polyhedra [143]. Each corner in one of these solids has the same number of neighbors at the same distance and therefore equal space around itself. Distributing any other number of points uniformly on the sphere has no exact solution and must therefore be approximated [289].

A tessellation of the sphere using polar coordinates creates the frequently used geographic coordinate system. However, this tesselation does not yield an equally distributed set of points on the sphere, since areas close to the poles are sampled finer than those at the equator. Even when configuring the sampling for bins with equal area, the distribution of points is not uniform, because the sampling between latitudes differs. Other approaches for distributing an arbitrary number of points uniformly on the sphere employ optimization procedures to maximize or minimize a quality criterion such as the smallest distance between points. An uniform distribution of points on the sphere can also be generated using a geodesic grid. These grids are computed by triangulating a Platonic solid, which yields an approximately uniform tesselation of the sphere surface [291]. A drawback of this approach is, that only certain numbers of points can be distributed by this method.

The icosahedron belongs to the set of Platonic solids and is shown in Figure 3.5. All 20 triangular faces are equivalent and the 12 corners of the icosahedron are distributed isotropically on the sphere each having the same distance to its neighbors. All 30 edges of the icosahedron are of equal length and subtend an angle of 63.4° yielding a coarse sampling of orientations. The resolution can easily be increased by quadruplicate subdivision of the triangular faces as also shown in Figure 3.5. The granularity of a geodesic grid is indicated by the *subdivision level*, which is 0 for the genuine icosahedron and increments by 1 for each subdivision.

During subdivision, each face of the geodesic grid is triangulated by introducing new corners at the midpoint of each edge. This splits each face into four new faces as shown in Figure 3.6. While the corners of the icosahedron have five neighbors, all newly introduced corners are connected to six corners. The coordinates of the newly introduced corners are adjusted, so that they lie on the circumsphere of the icosahedron, which yields a finer tesselation of the sphere surface. The number of corners rises exponentially with the level of subdivision as shown in Figure 3.7 and analyzed in Appendix A.4.

For subdivided icosahedra, the distribution of points on the sphere is not uniform. This is shown by the red bars in Figure 3.7 that correspond to the minimum and maximum subtended angle of an edge in the geodesic grid. The diagram shows that the absolute deviation of the edge length is small. Further-

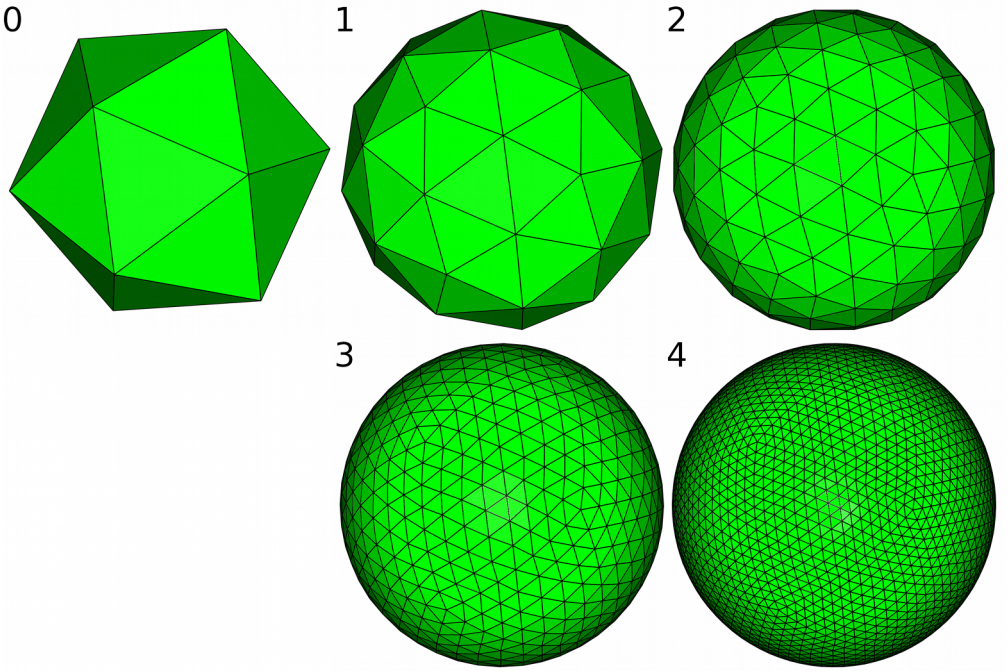


Figure 3.5 – Geodesic grids at different subdivision levels
Geodesic grids of subdivision levels 0–4 are shown. The grid is based on the icosahedron (subdivision level 0) and iteratively subdivided using triangulation. All corners of the genuine icosahedron are connected to five neighbors, while corners that have been introduced at a higher level have six neighbors. (© A. Griewel)

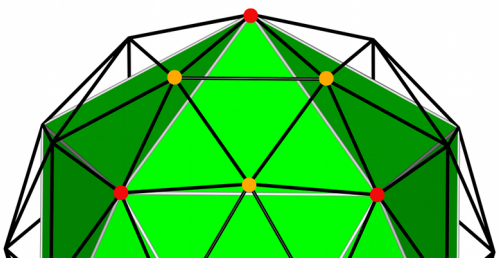


Figure 3.6 – Geodesic grid subdivision
For the subdivision of an icosahedron face (red dots), four new faces are introduced. The distance from the icosahedron center to new corners (orange dots) is adapted so that the corners lie on the circumsphere of the icosahedron. (© A. Griewel)

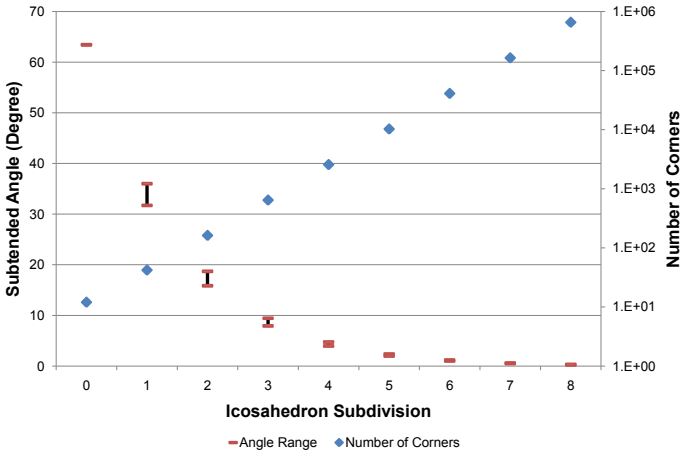


Figure 3.7 – Geodesic grid properties
The plot shows the properties of the geodesic grid based on the icosahedron with increasing subdivision level. The number of corners in the grid is scaled according to the right, logarithmic ordinate. The angle range describes the minimum and maximum subtended angle by an edge in the geodesic grid and is scaled according to the left ordinate. See also Appendix A.4 on page 201. (© A. Griewel)

more, the relative deviation — i. e., the deviation with respect to the absolute length of the edges of the geodesic grid — is small. Therefore, the distribution of points on the sphere using geodesic grids based on a triangulated icosahedron is suitable for sampling orientations.

Each vector from the center to a corner of the geodesic grid represents a bin and is called *bin vector*. Bins are populated by inserting vectors into the histogram. For the reasons mentioned previously, it is important to assure a uniform sampling of the sphere surface. Therefore the magnitude of each inserted vector $\mathbf{v}$ is not added to one single bin, but distributed among the bins that are closest to its direction. Given the structure of the geodesic grid, which consists solely of triangular faces, the three closest bins are identified for $\mathbf{v}$. The contribution w_a to one of the three closest bins $b_a \in \{1; 2; 3\}$ is determined using inverse distance weighting [298]. The angle between $\mathbf{v}$ and the corresponding bin vectors $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ is utilized as distance measure. The total weight of the contribution equals the length of the vector $|\mathbf{v}|$

$$w_a(\mathbf{v}, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3) = \frac{(\sphericalangle(\mathbf{v}, \mathbf{c}_a))^{-1}}{\sum_{i=1}^{3} (\sphericalangle(\mathbf{v}, \mathbf{c}_i))^{-1}} \cdot |\mathbf{v}| \tag{3.12}$$

The defined orientation histogram is based on the concepts of the SIFT. Differing from other 3D SIFT approaches, the 3D version of the orientation histogram developed for this work addresses the additional degrees of freedom in 3D space. It faithfully represents information on the orientations of a set of vectors. This is facilitated by a uniform distribution of points on the sphere surface and an insertion function that interpolates the contributions of each inserted vector among its closest bins.

3.3.2. Geodesic Index

The efficient identification of the closest bins to an inserted vector is critical for the overall performance of the algorithm. Using an exhaustive search — i. e., identifying the closest bins by calculating the angles between all bin vectors and the inserted vector — results in asymptotically exponential run time with respect to the subdivision level of the geodesic grid. To speed up the identification of the closest bins, a geodesic index was developed for this work, which facilitates the lookup of the closest bins in asymptotically constant time.

All bin vectors are assigned to tiles. These are precomputed surface patches of the sphere and allow for an efficient access to its contents. The indexing is based on polar coordinates, since these can be computed efficiently for a given vector. The sphere is subdivided into tiles by latitudes and longitudes as shown in Figure 3.8 using an arbitrary reference coordinate frame specifying a fixed pole $\mathbf{N}$ and a fixed orthogonal 0° longitude $\mathbf{G}$.

Each tile must contain at least one corner of the geodesic grid. This is accomplished by first subdividing the sphere into L equally sized spherical sectors centered on $\mathbf{N}$. The width w_{lat} of a sector is measured in radians on a great circle on the unit sphere that touches $\mathbf{N}$ and $\mathbf{G}$. It is chosen to be larger than the maximal geodesic length w_{geo} of a geodesic grid edge

$$L = \left\lfloor \frac{\pi}{w_{\text{geo}}} \right\rfloor \qquad\qquad w_{\text{lat}} = \frac{\pi}{L} \tag{3.13}$$

This guarantees that for each geodesic grid face touching a sector at least one corner of the face comes to lie on that sector. Each sector is assigned an index $\{0, \ldots, L-1\}$ emanating from the pole and arriving at the pole's antipode.

Sectors are subdivided along longitudes into spherical quadrilaterals, which are called *tiles*. The width of each tile $w_{\text{long}}(i)$ is measured in radians around the pole and depends on the associated sector i. To determine $w_{\text{long}}(i)$, the

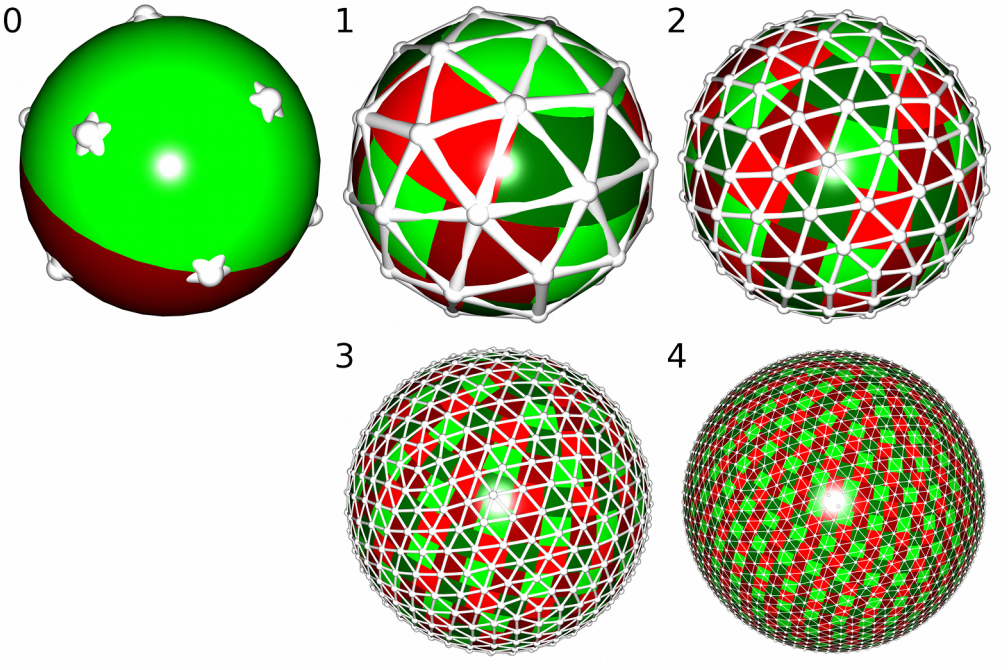


Figure 3.8 – Geodesic index
The first five levels of subdivision of the geodesic grid (white) are shown. For the efficient identification of neighboring bins of an inserted vector, geodesic grid corners are indexed according to their polar coordinates (red and green spherical quadrilaterals). (© A. Griewel)

minimum of the circumferences $m_{\text{lat}}(i)$ of the two small circles delimiting sector i is determined on the unit-sphere

$$m_{\text{lat}}(i) = \min \left\{ \begin{array}{l} |2\pi \sin(i \cdot w_{\text{lat}})|, \\ |2\pi \sin((i+1) \cdot w_{\text{lat}})| \end{array} \right\} \tag{3.14}$$

Sector i is then subdivided into M tiles each of width $w_{\text{long}}(i)$

$$M = \left\lfloor \frac{m_{\text{lat}}(i)}{w_{\text{geo}}} \right\rfloor \qquad\qquad w_{\text{long}}(i) = \frac{m_{\text{lat}}(i)}{M} \tag{3.15}$$

This yields a tesselation, in which the sides of all tiles are longer than the longest edge in the geodesic grid. Therefore, it is guaranteed, that at least one corner of the geodesic grid lies on each tile.

Each geodesic grid corner $\mathbf{c}$ is assigned to the tile it lies on. This tile is identified by first determining the index i of the sector

$$i = \frac{\sphericalangle(\mathbf{c}, \mathbf{N})}{w_{\text{lat}}} \tag{3.16}$$

Subsequently, the index j of the tile on the sector is identified using the angle between $\mathbf{c}$ and $\mathbf{G}$ about the pole $\mathbf{N}$. Thus, each tile is assigned the geodesic grid corners, which lie on the tile and thus the corners are indexed by their polar coordinates.

The index is used to identify the three closest bin vectors to a given vector $\mathbf{v}$. In other words, the index is used to identify the triangular face F of the geodesic grid that $\mathbf{v}$ points to. The algorithm is illustrated in Figure 3.9. It commences by first determining the set of bin vectors that lie on the same tile as $\mathbf{v}$ using the same procedure as for assigning geodesic grid corners to tiles. In this set, the bin vector $b_\circ$ with smallest angular distance to $\mathbf{v}$ is identified. Due to subdivision in tiles, $b_\circ$ is not necessarily a bin vector that touches the geodesic grid face F. Thus, a bin vector $b_\triangle$ touching F is identified by examining $N(b_\circ)$, the set of all corners that are connected to $b_\circ$ by a geodesic grid edge. The three bins that span F and thus are closest to $\mathbf{v}$ are identified subsequently by examining $N(b_\triangle)$. The three bin vectors in $N(b_\triangle)$ with smallest angular distance to $\mathbf{v}$ are returned for further processing in the orientation histogram.

Using the geodesic index, the closest bins in the geodesic grid are determined in asymptotically constant time for arbitrary subdivision levels of the geodesic grid. The polar coordinates of the inserted vector can be determined in constant time. Subsequently, the three closest bins are determined by analyzing the neighborhood of two geodesic grid corners. Each grid corner holds a list of

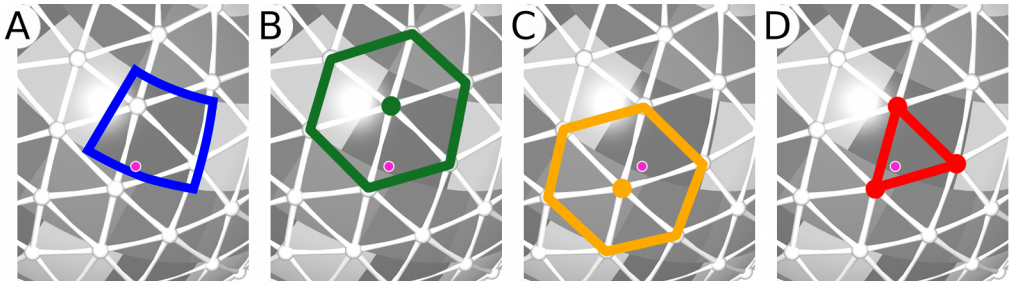


Figure 3.9 – Geodesic grid algorithm
The three closest bins to an inserted vector $\mathbf{v}$ (pink) are determined by first (A) identifying the tile $\mathbf{v}$ is pointing to. (B) Subsequently, the closest bin vector $b_\circ$ (green dot) to $\mathbf{v}$ on the tile is determined. (C) A bin vector $b_\triangle$ (golden dot) lying on the triangular geodesic grid face that $\mathbf{v}$ is pointing to is identified by analyzing the neighborhood of $b_\circ$ (green hexagon in B). (D) Eventually, the three bin vectors closest to $\mathbf{v}$ (red) are identified by analyzing the neighborhood of $b_\triangle$ (golden hexagon in C). (© A. Griewel)

neighbors allowing for the identification of the relevant grid corners in constant time. The angle between the inserted vector and up to seven geodesic grid corners is also computed in asymptotically constant time. Therefore, the run time of one insertion lies in $O(1)$.

The performance of the geodesic grid in an application scenario was affirmed by measuring the run time of inserting 1 000 000 randomly oriented vectors into an orientation histogram. The required time is compared to an exhaustive search and an R*-tree, [18] — a generic spatial index. For the R*-tree each bin is represented as the geodesic grid corner and a nearest neighbor search is executed when inserting a vector. The resulting timings are shown in Figure 3.10 and demonstrate that the geodesic index requires a constant amount of time independent of the level of subdivision of the geodesic grid.

From Figure 3.10 it is clear that the geodesic index does not accelerate the insertion into orientation histograms that consist of a genuine icosahedron with respect to a brute force search. This is comprehensible since the number of points in the icosahedron is small and therefore an exhaustive search is performed quickly. Furthermore, the tesselation of the sphere in tiles is very coarse at this level as shown in Figure 3.8. However, for all higher levels of subdivision the performance advantage of the geodesic index is evident.

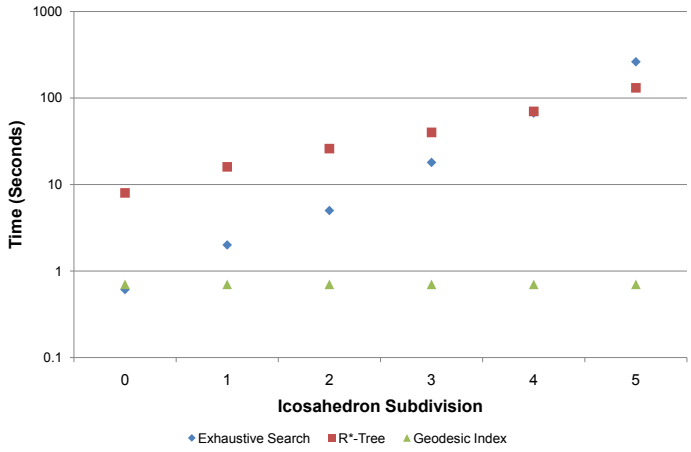


Figure 3.10 – Geodesic index run time
Run time for inserting 1 000 000 randomly oriented vectors into an orientation histogram of various levels of subdivision (logarithmic ordinate). (© A. Griewel)

3.3.3. Dominant Orientations

The objective of orientation assignment is to determine a discrete set of orientations for each keypoint, which is used for descriptor alignment and registration. Orientations are determined as illustrated in Figure 3.11 and summarized in the following: Gradient vectors are sampled in a Gaussian window that depends on the keypoint's location and scale — the latter determining both, the size of the window and the scale-space map, which is used for calculating the gradient. All sampled gradient vectors are accumulated in an orientation histogram, and dominant orientations are subsequently identified in a two step procedure: First, prominent peaks in the histogram yield a first axis of rotation. Subsequently, a 2D histogram is created, in which again prominent peaks are identified. The axis identified in the 3D histogram together with the prominent bin in the 2D histogram completely specify a 3D rotation, which is saved as orientation for the keypoint.

Each keypoint is detected as extremum in one specific scale-space map in one specific octave. This map is used for sampling gradient vectors and thus for determining dominant orientations. Since the sampling interval differs from octave to octave, the sampling of gradient vectors is not performed according to the voxel spacing but rather dependent on the scale of the keypoint. Thus, gradient vectors are sampled on a cubic, isotropic lattice centered on the keypoint as shown in Figure 3.11. The sampling interval of the lattice is set to $w_{\text{samp}}\sigma$, where σ is the scale of the keypoint and w_{samp} is a parameter. For each lattice point lying inside a circular truncation window, a gradient vector is computed using finite difference approximation. Since the density values needed for cal-

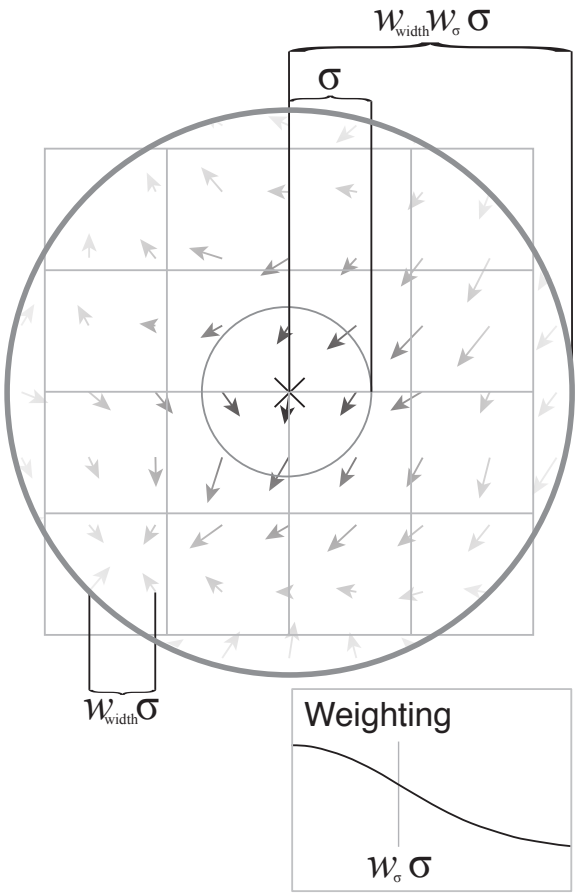


Figure 3.11 – Sketch of the gradient sampling for orientation assignment
For each keypoint with scale σ, dominant orientations are determined. These are computed by sampling the gradient on an isotropic, cubic grid with sampling interval $w_{\text{samp}}\sigma$ inside a Gaussian window (intensities of arrows) centered on the keypoint (cross). In the sketch, the grid depicts the voxel spacing while the arrows depict sampled gradient vectors illustrating that the sampling of the gradient is independent of the sampling interval of the map. The keypoint's scale σ and truncation-radius $w_{\text{width}} w_\sigma \sigma$ are shown on top of the figure while the Gaussian weighting function with standard deviation $w_\sigma \sigma$ and the sampling interval $w_{\text{samp}}\sigma$ is shown at the bottom. (© A. Griewel)

culating the differences are not at voxel positions, these values are determined using trilinear interpolation.

For orientation assignment, the gradient is considered only in a Gaussian window. Therefore, the magnitude of each gradient vector is multiplied by a Gaussian weighting function, which depends on the distance between sampled gradient and keypoint. The standard deviation of the Gaussian weighting function is determined as $w_\sigma \sigma$, where σ is the keypoint's scale and w_σ is a parameter. For speeding up the calculation, the Gaussian window is truncated at $w_{\text{width}} w_\sigma \sigma$, where w_{width} is a parameter.

The weighted gradient vectors are accumulated in an orientation histogram, which captures the gradient in the neighborhood of the keypoint. The subdivision level of this histogram is determined by the parameter h_{g}^{3D}. Calculating the 3D orientation assignment in this way makes it scale-invariant, since the com-

Figure 3.12 – Orientation histogram
Gradient vectors (gold) are sampled in a local neighborhood (sphere) of the keypoint. Each gradient vector is weighted according to its distance to the keypoint. This can be seen by the length of the gradient vectors, which is smaller in outer regions of the sphere. All gradient vectors are added to an orientation histogram (red), which is used for orientation assignment. (© A. Griewel)

putation solely depends on the scale of the keypoint σ and not on the sampling interval of the map.

Dominant orientations are computed by first identifying *dominant bins* in the orientation histogram and subsequently fixing a rotation around the bin vector. This algorithm is described below and illustrated in Figure 3.13. Let m^{3D} be the maximum entry in the histogram. Each bin that has a value larger than $h_t^{3D} m^{3D}$ is considered as dominant bin, where h_t^{3D} is a parameter. For each dominant bin, a 2D orientation histogram is assembled using the entries of the 3D orientation histogram. Each bin in the 3D histogram is interpreted as bin vector with length according to the entry. These vectors are projected to the plane that is orthogonal to the dominant bin and assembled in a 2D histogram consisting of h_g^{2D} bins. This histogram gathers 2D vectors similar to its 3D counterpart using inverse distance weighting for each inserted vector and its two adjacent bins. The same thresholding strategy as before is applied. Using the maximum entry m^{2D} in the 2D histogram and a threshold h_t^{2D}, all bins that are larger than a threshold $h_t^{2D} m^{2D}$ are selected for the computation of dominant orientations.

Eventually, discrete orientations $\mathbf{R}$ are computed using the peaks in the orientation histogram and the dominant 2D bins. Orientations are saved as rotations relative to a given coordinate frame: A dominant peak of the orientation histogram specifies the 3D rotation axis for an orientation. The dominant 2D

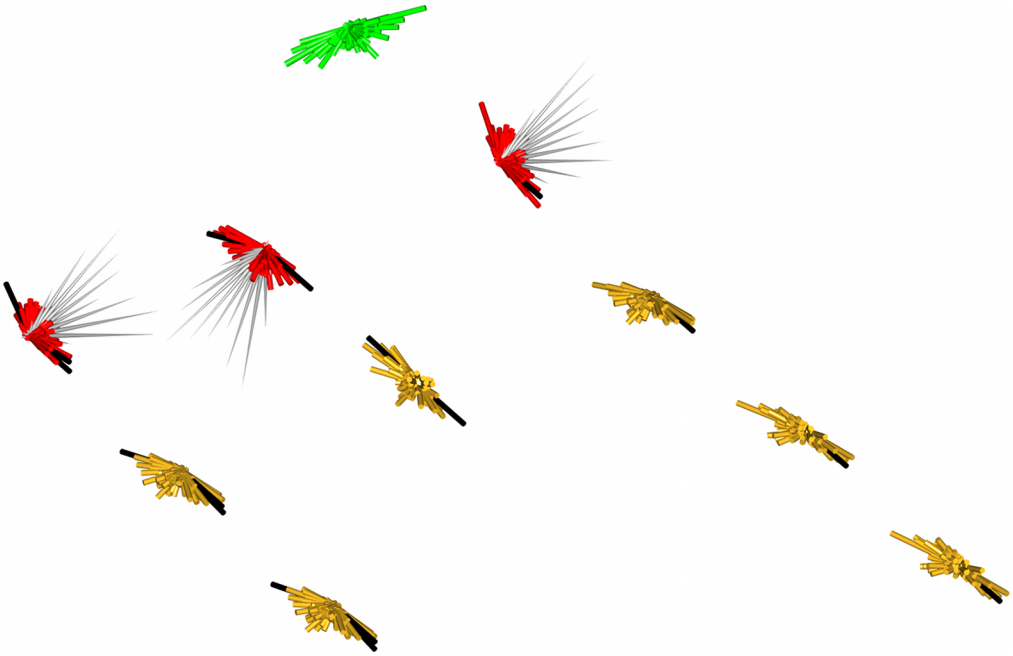


Figure 3.13 – Dominant orientations
Dominant orientations are computed from an orientation histogram (green) in a two stage procedure. Each red orientation histograms corresponds to one dominant bin in the 3D histogram and the dominant bin is colored black. From top right to bottom left, new dominant orientation are analyzed (previously processed dominant bins remain black). The red and golden orientation histograms are aligned so that the considered dominant bin points to the lower right corner. For each dominant bin — i. e., red orientation histogram — a 2D histogram (gray) is assembled. Dominant bins in this 2D histogram specify a rotation angle around the axis defined by the dominant bin. The complete orientation is then calculated as rotation around the dominant bin by the angle determined in the 2D histogram. In the displayed case, six dominant orientations are computed for the genuine orientation histogram. These are displayed as golden orientation histograms, which are rotated accordingly. (© A. Griewel)

bins define the rotation around this axis. The combination of axis and rotation angle around the axis completely specifies the 3D rotation relative to the given coordinate frame and thus determines the orientation. Thus, the computation addresses all degrees of freedom of a 3D rotation and therefore rotation invariance for orientation assignment is achieved. The resulting dominant orientations $\mathbf{R}$ serve as reference coordinate frame for calculating descriptors and for computing registrations.

3.4. Neighborhood Descriptor Computation

For each orientation of every keypoint, a neighborhood descriptor is created. The descriptor consists of orientation histograms that represent the gradient in spatial neighborhoods of the keypoint. Using all entries in the orientation histograms, a feature vector is generated from the descriptor, which is employed for matching and similarity searching. The process of calculating descriptors and the associated feature vectors is described in the following and summarized in Figure 3.14.

The descriptor is calculated by first setting up a cubic lattice, which is centered on the keypoint and aligned with the considered orientation. The points of the lattice are separated by a constant distance $\Delta = \delta\sigma$, which depends on a parameter δ and the scale of the keypoint σ. Each lattice point is assigned the cubic volume according to the lattice. The number of cubes that are comprised in the descriptor is specified by the parameter r. For all points that lie inside or on the surface of a sphere with radius $r\Delta$, an orientation histogram is created. Thus, the spatial extent of the descriptor solely depends on the scale of the keypoint. Therefore, the descriptor is scale-invariant since it is not dependent on the sampling interval of the map. A sketch of a two-dimensional, central slice of a descriptor exemplifying the effects of r is shown in Figure 3.15. Three-dimensional images of descriptors for various values of r are shown in Figure 3.16.

For each cube, an orientation histogram is calculated. All orientation histograms of the descriptor are of the same level of subdivision g. The bins of the histograms are populated by gradient vectors interpolated inside the corresponding cube, as similarly done for the computation of the orientation histogram. These are calculated at positions that are specified by a second, cubic lattice, which spans the inside of the cube and comprises p points per dimension. Therefore, p^3 gradient vectors are calculated and added to each orientation histogram. The sampled gradient vectors are weighted using a keypoint-centered Gaussian function of standard deviation $\sigma_{\mathrm{d}} r\Delta$ — similar to the orientation assignment. This gives larger weight to gradients that are closer to the keypoint.

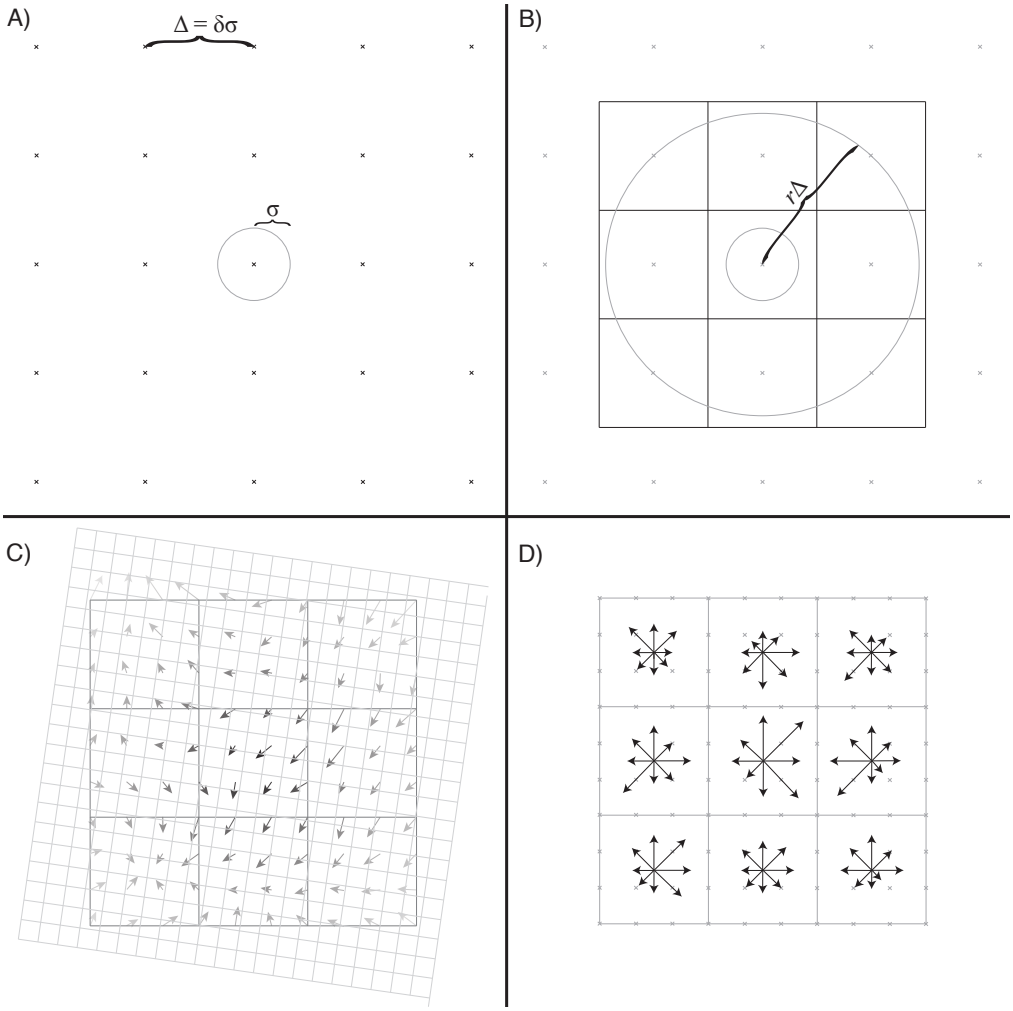


Figure 3.14 – Sketch of the construction of a local neighborhood descriptor
A keypoint is shown in the center of A) and its scale is indicated by a circle of radius σ. The descriptor is calculated by setting up a keypoint centered, cubic lattice with spacing $\Delta = \delta\sigma$, which is aligned to the given orientation. B) All lattice points inside a sphere of radius $r\Delta$ are included in the descriptor. C) For each cube surrounding these lattice points, p^3 gradient vectors (arrows) are sampled on a lattice relative to the cube — the sampling of the map is depicted as the underlying grid. Each of the gradient vectors is weighting according to a Gaussian function with standard deviation $\sigma_d r\Delta$, which is illustrated using the intensities of the arrows. D) For each cube, the weighted gradient vectors are inserted into the orientation histogram. (© A. Griewel)

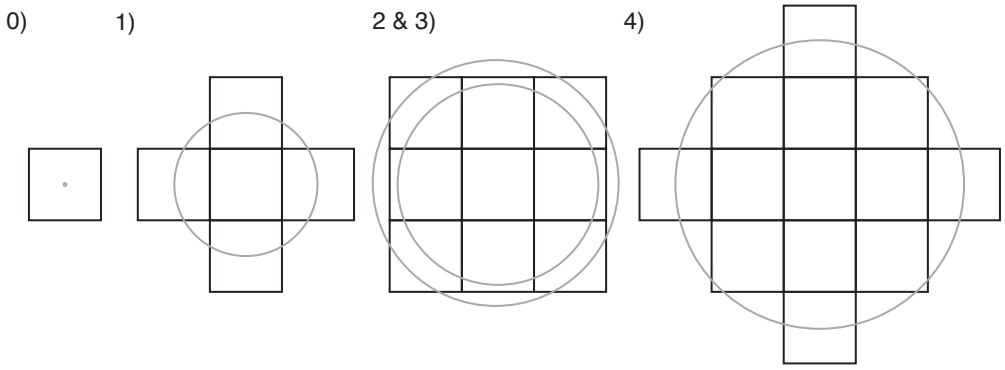


Figure 3.15 – 2D sketches of local neighborhood descriptors
A sketch of the central slice of descriptors with different numbers of cubes is shown. The number of cubes is specified by the parameter r, which determines the radius of a sphere in terms of cube edge length. All cubes that lie inside the sphere and on a cubic lattice centered on the keypoint are included in the descriptor. Due to the spherical restriction and the properties of the Euclidean distance, the number of cubes changes only if r equals the square root of the sum of three squared integral values. Here, examples are shown for $\sqrt{0}$, $\sqrt{1}$, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{4}$. There is no difference in the two-dimensional central slice for $\sqrt{2}$ and $\sqrt{3}$. However, the three-dimensional shapes of the descriptors differs as shown in Figure 3.16. (© A. Griewel)

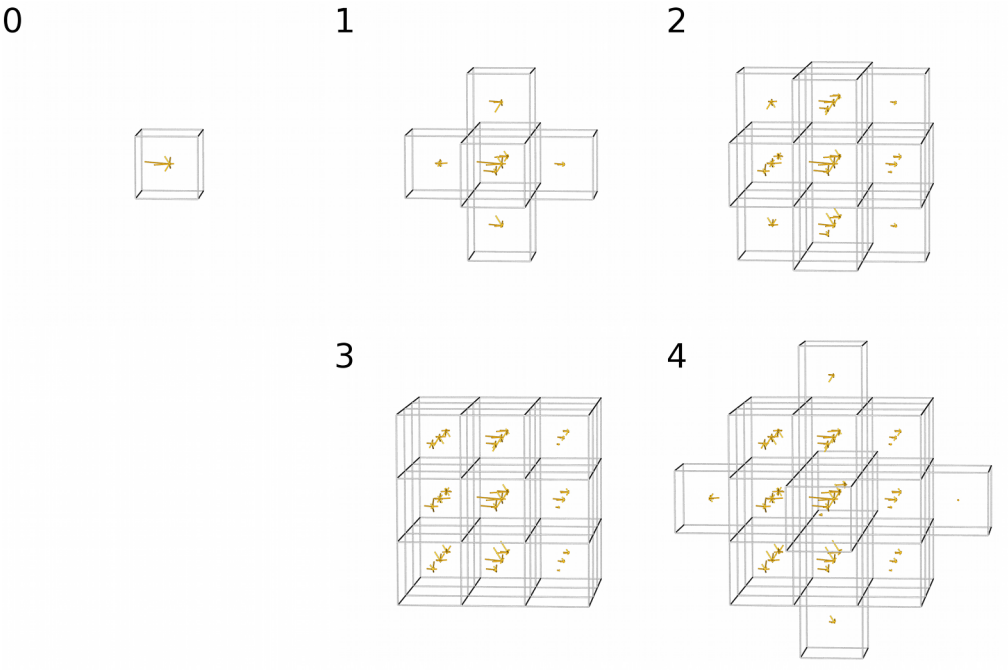


Figure 3.16 – 3D examples of local neighborhood descriptors
A keypoint descriptor is created by subdividing the local neighborhood into cubic cells and calculating orientation histograms inside each cube. All cubes with a center that is closer to the keypoint than r are considered where r is specified relative to the cube edge length Δ. Here, descriptors are shown for $r \in \{\sqrt{0}; \sqrt{1}; \sqrt{2}; \sqrt{3}; \sqrt{4}\}$. (© A. Griewel)

		$\sqrt{0}$	$\sqrt{1}$	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$	r
g	$B(g)$	1	7	19	27	33	$C(r)$
0	12	12	84	228	324	396	
1	42	42	294	798	1 134	1 386	
2	162	162	1 134	3 079	4 374	5 346	

Table 3.1 – Feature vector dimensionality
The feature vector dimensionality is determined by the total number of orientation histogram bins in the descriptor. This again depends on the number of cubes in the descriptor C, which depends on the parameter r. The number of orientation histogram bins per cube B is given by the subdivision level g. The total dimensionality of the feature vector is determined as $C(r) \cdot B(g)$.

After creating all orientation histograms, a feature vector $\mathbf{f} = (x_1, \ldots, x_d)$ is assembled by concatenating the histogram entries in a canonical order. The total dimensionality d of the feature vector depends on the number of cubes and the subdivision level of the orientation histograms in the descriptor. A list of the resulting dimensionality of the feature vector for a selection of parameter combinations r and g is shown in Table 3.1.

The distance D_{descr} of two feature vectors $\mathbf{f}^a = (x_1^a, \ldots, x_d^a)$ and $\mathbf{f}^b = (x_1^b, \ldots, x_d^b)$ is defined as the Euclidean distance

$$D_{\text{descr}}(\mathbf{f}^a, \mathbf{f}^b) = \sqrt{\sum_{i=1}^{d} (\mathbf{f}_i^a - \mathbf{f}_i^b)^2} \tag{3.17}$$

Furthermore, each feature vector is normalized to unit length using the Euclidean norm. This allows for the identification of similarity in volumes that have a similar structure, but have different overall intensity.

The final descriptor data structure gives access to all information that was gathered during its computation. This includes the spatial location of the keypoint $\mathbf{x}$, the keypoint's scale σ, the orientation $\mathbf{R}$ used for calculating the descriptor, and the feature vector $\mathbf{f}$. The set of all keypoints, orientations, and descriptors that have been computed for one map is called the *map description*.

3.5. Map Registration

The matching of descriptors is based on the Euclidean distance of the feature vectors. Distinct descriptor matches yield correspondences between keypoints,

which are used to compute a registration of the given maps. Here, each keypoint creates one placement of the target map onto the source map. Eventually, final placements are selected by assessing all created placements for consistency.

Matching is performed on a per-keypoint basis. For each source keypoint, the best matching, *compatible* keypoint from the target map is identified. Two keypoints are compatible if the following two conditions are fulfilled. First, the keypoints must both correspond to either a maximum or a minimum in DoG scale-space. Second, the scales σ_1 and σ_2 of the keypoints must lie within half an octave distance

$$\frac{1}{\sqrt{2}} \leq \frac{\sigma_1}{\sigma_2} \leq \sqrt{2} \tag{3.18}$$

The registration is based on a set of *keypoint matches*, i. e., associations of source- and target keypoints. Each source keypoint is matched to at most one target keypoint based on the associated descriptors. For this, two requirements must be fulfilled. First, the keypoints must be compatible. Second, there must be a pair of associated descriptors, which has a distinctively low feature vector distance as defined below. The distance of a given pair of source and target keypoint is then defined as the minimum distance between any pair of associated descriptors.

For the registration, only relevant matches are to be considered. Therefore, an *absolute* threshold τ_{abs} can be enforced on descriptor similarity discarding all matches with larger distances. For certain combinations of parameters, an absolute threshold does not perform well. In these cases a threshold on the distinctiveness of the best match can be utilized (cf. [218]). The *distinctiveness* is analyzed using the ratio between the distance values computed for the closest and second closest match. If this value lies below a predetermined value τ_{dist}, the matches are considered distinct and are saved for the calculation of a transformation of the target map.

A registration of two given maps is calculated for each identified source-target keypoint match. For this purpose, the spatial information contained in the matched descriptors comprising location and orientation for both the source $(\mathbf{x}_\text{s}, \mathbf{R}_\text{s})$ and the target $(\mathbf{x}_\text{t}, \mathbf{R}_\text{t})$ keypoint is utilized. Based on this information, a rigid transformation $A(\mathbf{c})$ superposing the matched descriptors can be calculated. This transformation superposes the matched keypoints of the target map to the matched keypoint in the source map in the corresponding orientations. Assuming that the orientations $\mathbf{R}_\text{s}$ and $\mathbf{R}_\text{t}$ are specified as rotation matrices, the transformation can be calculated as shown in Equation 3.19. For each of the determined keypoint matches, the target map is transformed according to A and the resulting coordinate frame is saved as one *placement*.

$$A(\mathbf{c}) = \mathbf{R}' \cdot \mathbf{c} + (\mathbf{x}_\mathrm{s} - \mathbf{R}' \cdot \mathbf{x}_\mathrm{t}) \mid \mathbf{R}' = \mathbf{R}_\mathrm{s} \cdot {\mathbf{R}_\mathrm{t}}^T \tag{3.19}$$

An example of the outcome of a registration is shown in Figure 3.17. There are five clusters of placements identified for the displayed registration. The cluster in the middle of the figure is created by the protein that is contained in the center of the unit cell. The four clusters that are found on the periphery of the figure are induced from neighboring proteins, which protrude into the unit cell of this experimental X-ray map.

The figure shows that many placements lie closely together and therefore hold redundant information. Thus, relevant placements are identified by analyzing the distribution of the placements by means of single linkage clustering. The distance measure for clustering is the *root mean square deviation (RMSD)* metric. For two placements of the same molecule $A = \{\mathbf{a}_i\}$ and $B = \{\mathbf{b}_i\}$ — where $i \in \{1, \ldots, n\}$ identifies the atom and $\mathbf{a}_i$ as well as $\mathbf{b}_i$ the atom location in the two placements — the RMSD is calculated using the Euclidean distance between the position of corresponding atoms. This calculation is carried out in asymptotically constant time using an efficient formulation of the RMSD, which is applicable to rigid objects only [270]. The distance threshold is defined as four times the voxel spacing of the base map. If the target map does not comprise an atomic model, the locations of those voxels that are larger than an intensity threshold are utilized for the calculation of the RMSD. This intensity threshold defaults to the mean plus the standard deviation of the intensity values in the map and can be altered interactively. Each cluster is represented by only one placement, which has the highest score determined according to one of the scoring functions detailed below.

$$\mathrm{RMSD}(A, B) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} |\mathbf{a}_i - \mathbf{b_i}|^2} \tag{3.20}$$

Each placement is assigned a score as a measure of the goodness of fit between the superposition of source and target map. These measures can either be based on the calculated keypoints, but may also utilize the map intensities or the atomic model, if provided. The following scoring functions are available:

Keypoint Matches A keypoint match is encountered, if a compatible source- and a target keypoint are spatially closer than an error margin. This error margin is defined as the minimum of the scales of the two keypoints. Furthermore, it is required that the keypoints are compatible. Keypoints that lie spatially close to each other are determined by executing a nearest

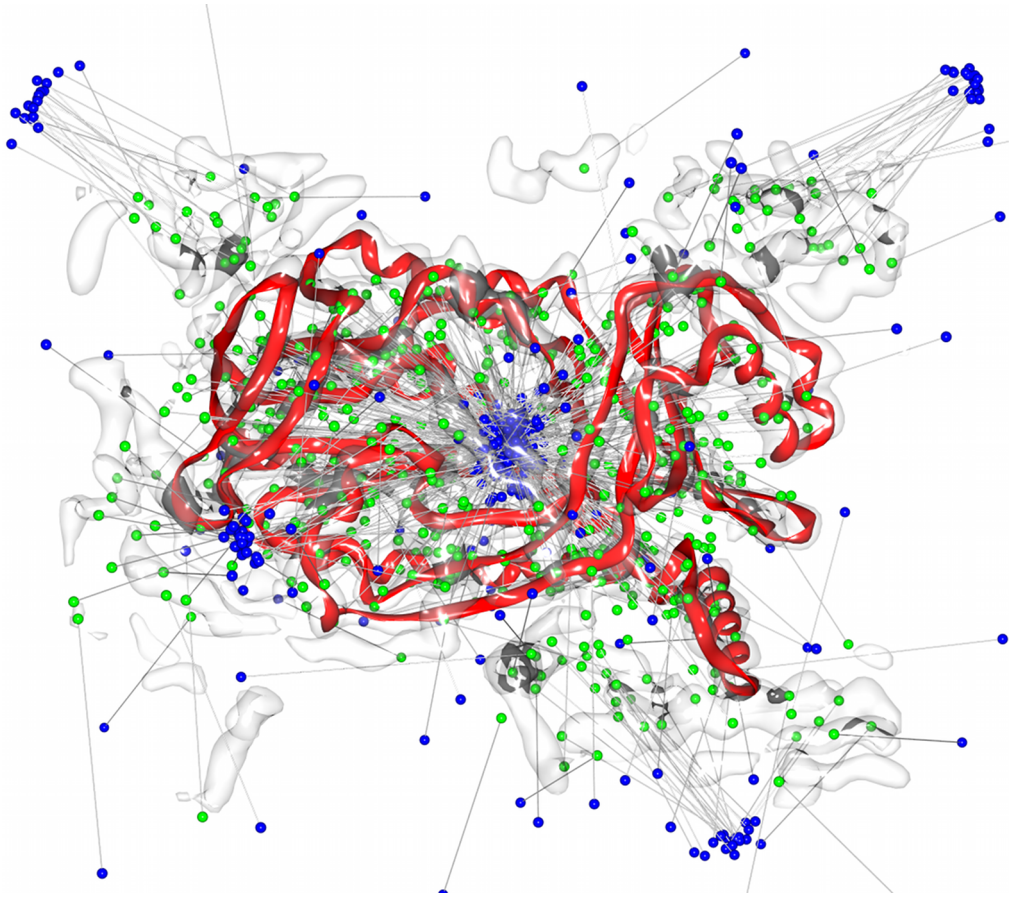


Figure 3.17 – Map registration
The target is a synthetic map of a hydroxylase (red ribbon; wwPDB ID `1PBD` [295]), which is registered to its experimental electron density map (transparent surface, low pass filtered map for clearer view). All source keypoints that have been matched are displayed as green spheres. Each matched source keypoint generates one placement, which is determined using its match partner. The placements are shown as blue spheres that are connected to their corresponding source keypoint using a line. (© A. Griewel)

neighbor query on an R*-tree [130, 18, 132]. This operation is highly efficient, and therefore the run time of this scoring function is low.

Weighted Keypoint Matches As in the previous scoring function, keypoint matches are determined. These, however, are weighted by their spatial distance giving less weight to inexact matches. The total score of a matching s is determined using the distance d between the matched keypoints and the scale of the source keypoint σ

$$s = 1 - \frac{d}{\sigma} \tag{3.21}$$

This yields a value that lies in the interval $[0; 1]$ for each match. The total score comprises the sum of all keypoint match scores.

Correlation This score is implemented for a pair of map and atomic model. It determines the PM-correlation coefficient between the source map and a synthetic map generated from the atomic model [368]. Only points that have an intensity value larger than zero in the synthetic map are considered. This is the computationally most expensive scoring function.

Atom Interpolation This scoring function can be applied for evaluating the placement of an atomic model in an electron density map. It interpolates the intensities at each position of an atom and reports the sum of the interpolated values as a score [281]. The function is quickly evaluated and well suited for docking larger atomic models into high resolution electron density maps.

Enclosed Atoms The score function requires an atomic model and a thresholded electron density map. Voxels with an intensity higher than the threshold are set active and the remaining voxels are inactive. The score is defined as the number of atoms that lie on an active voxel of the segmented map.

Eventually, the resulting placements can be optimized with respect to the specified scoring function. A simulated annealing [176] method is available, which iteratively processes all determined placements. The purpose of this optimization is to localize a close local minimum of the scoring function and thus only 1 000 steps, a cooling factor of 0.9, and an initial temperature of 10^{-3} of the initial score are used. In each step of the stochastic optimization, the placement is randomly altered. The translational displacement is calculated using an uniform distribution of points in a spherical surrounding of 2 Å around the previous placement. The rotation is determined by choosing a random point on the sphere [67] and rotating around the axis formed by sphere center and the

random point. The rotation angle is determined by drawing from an uniform distribution in the interval $[-2°; 2°]$. All parameters have been determined in parameter studies using synthetic maps of various proteins.

3.6. Molecule Recognition

The computed descriptors yield an abstract representation of the macromolecule depicted in an electron density map. This representation can not only be used for registration, but also for the identification of the content of a map. No atomic interpretation of the map is necessary for this task, since the method solely relies on the information given by the electron density. Thus, the identification of a molecule in a map — molecule recognition[1] — can be performed using the map description without relying on modeled atomic structures.

For molecule recognition, a set of predetermined *reference molecules* is used, which comprises all structures that are to be identified. The proposed method is generically applicable to any set of reference molecules, which can contain either complete structures or building blocks of the molecules. The references can be supplied as atomic structures or as density maps giving the possibility to search in non-interpreted electron densities. Furthermore, the set can be focused for a specific set of molecules, or it can contain a broad amount of molecular structures.

In Section 5.3, molecule recognition is used to identify arbitrary proteins in electron density maps. To identify all known proteins successfully, the reference set must comprise all known building blocks of proteins. Different classifications of protein structures have been assembled, which can be used as reference sets [179]. These rely on evolutionary or structural criteria or a combination of both and are either automatically generated or manually curated. The most prominent examples of such classifications are FSSP [146], SCOP [247], and CATH [254]. To enable successful molecule recognition, the reference set must comprise an instance of the molecule or a part of the depicted molecule in a similar conformation. Equally sufficient is a close homolog or part of that homolog that must also have a corresponding conformation. If no structure in the reference set corresponds to the protein depicted in the query map in both sequence and overall conformation, the map descriptions are going to differ largely and therefore no similarities on the image level will be identified. Thus, the reference protein set must cover a large amount of the known proteins. However, the selection of references is not part of the method, since the method is generic and

[1] Not to be confused with molecular recognition, which is a term used in chemistry and biology [108].

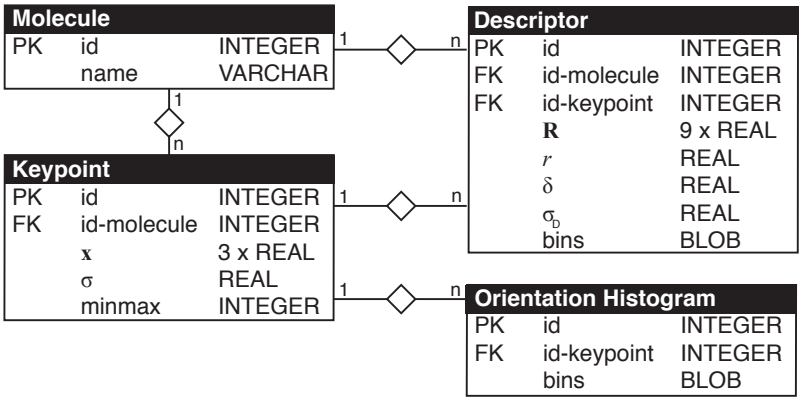


Figure 3.18 – Database entity relationship diagram
Entity relationship diagram of the utilized data model. Primary keys (PK) and foreign keys (FK) are indicated in the first row, the second row describes the content as explained in the text, and the third row defines the type of the attributes. (BLOB is an acronym standing for *binary large object* and is not to be confused with blobs that are detected in images.) (© A. Griewel)

applicable to any set of references. Thus, the selection of a suited database is discussed in the specific test of molecule recognition in Section 5.3 on page 172.

All keypoints that have been detected in the reference structures are stored in a relational database system [145] using the entity-relationship model shown in Figure 3.18. The database scheme is not normalized, so that it is possible to query for descriptors and to directly identify the corresponding reference models using foreign keys. For each *query descriptor*, the closest reference descriptors are identified using the feature vector distance defined in Equation 3.17 on page 89 as similarity measure. This approach is similar to the k-nearest neighbor problem, in which the k nearest neighbors to a query object are to be identified according to a given distance measure. No indices or parallel computing are currently used to accelerate the processing of the high dimensional feature vectors. However, there are indices that support nearest neighbor queries in higher dimension [21, 19]. Furthermore, algorithms tailored to modern parallel hardware have been proposed for accelerating nearest neighbor queries [105, 106, 201, 202, 165].

The content of a *query* electron density map is identified using a voting scheme. To evaluate the voting, each query keypoint is assigning a *match list* of best matching reference structures. The list is created by comparing each descriptor of the query keypoint to all compatible[1] reference descriptors in the

[1] See Section 3.5 on page 90.

database. The result is a list of descriptor matches. Each match is assigned a score, which corresponds to the smallest distance value that is encountered for a pair of compatible reference- and query descriptor.[1] The list potentially comprises reference structures more than once, since more than one reference descriptor may match the query descriptor. For the voting procedure, duplicates are removed from the match list. Thus, the resulting list for each keypoint contains unique reference structures assigned with the smallest distance encountered during comparison.

The comparison of feature vectors using the Euclidean distance is done in high dimensional feature spaces, which introduces problems that are frequently referred to as the *curse of dimensionality* [20, 28, 140]. Among other findings, it was shown that distance measures lose their discriminative power in higher dimensions. This is induced by the exponential increase in volume, which is caused by adding dimensions to the feature space. To address this problem, appropriate *scoring schemes* are required to allow for a robust recognition method. Here, two scoring schemes are devised, which are used for weighting the matches to reference structures that are saved for each keypoint in its match list. The first scoring scheme, termed 1-NN′, is based on a method developed for the SIFT [218]. It is a modification of a 1-nearest neighbor search for each keypoint. The second scoring scheme, LR, uses local regression [61, 62, 140] and determines the content of the electron density map depending on a set of feature vectors that are close to the query feature vector. These *weight functions* are explained below. Both scoring methods yield a *result list*, which is ordered according to the score. The first entry of this list identifies the depicted molecule. It is also possible, that the first entries identify the content of the electron density map. This is the case if the reference structures make up only parts of the map or if the depicted molecule matches a class of reference structures rather than one exact reference.

In 1-NN′ scoring each query keypoint has one vote, which it may cast for one reference structure. To decide whether the vote is valid, the two closest compatible reference proteins with feature vector distance of η_0 and η_1 are extracted from the match list. The vote of a keypoint is only cast, if the distance of the closest match η_0 is significantly smaller than the distance of the second closest η_1. Similar to the procedure used during matching, this requirement is assessed using a fixed threshold τ_{NN}, which determines the maximal allowed ratio of the distance of closest and second closest match

$$\frac{\eta_0}{\eta_1} < \tau_{\mathrm{NN}} \tag{3.22}$$

[1] Cf. Section 3.5 on page 90.

If this requirement is fulfilled, a vote of weight one is cast for the reference structure that belongs to the matching reference descriptor. The results of evaluating the match lists of the query keypoints are summed up per reference structure in the result list.

In LR scoring, each matched descriptor is assigned a weight according to a weight function. In local regression, the tri-cube function is frequently employed for this purpose but other functions as the Gaussian or the Epanechnikov kernel are equally applicable [140]. For the weighting, a function is desired, which is capable of assigning high values to true matches and low values to false matches. Here, a weight function based on the tri-cube function is employed. The tri-cube function t assigns a reference descriptor with a feature vector distance η to the query descriptor a weight in the interval of [0;1]. The non-normalized formula of the tri-cube function is

$$t(\eta) = \begin{cases} (1 - |\eta|^3)^3 & \text{if } |\eta| \leq 1 \\ 0 & \text{else} \end{cases} \tag{3.23}$$

For the presented application, normalization is not required and the range of the function is adapted by a variable transform to the interval $[0; \tau_{\mathrm{LR}}]$, where τ_{LR} is a parameter. Furthermore, the slope of the function is of interest, which can be controlled by the parameter λ_{LR}. This yields the final weight function

$$w(\eta) = \begin{cases} \left(1 - \left(\frac{\eta}{\tau_{\mathrm{LR}}}\right)^{\lambda_{\mathrm{LR}}}\right)^{\lambda_{\mathrm{LR}}} & \text{if } |\eta| \leq \tau_{\mathrm{LR}} \\ 0 & \text{else} \end{cases} \tag{3.24}$$

As before, all scores of all matches are gathered in the result list. For LR scoring, the calculated, single scores are summed up per reference structure and yield the result list.

3.7. Summary

An approach for identifying similarities in 3D macromolecular electron density maps is described in this chapter. The method relies on scale-space theory [213] and the scale-invariant feature transform (SIFT) [217, 218]. It uses an abstract description of the map that consists of keypoints, orientations, and descriptors. All parameters for descriptor calculation depend on the scale of the keypoint. Thus, the map description in its entirety is scale-invariant since it does not depend on the sampling properties of the map, but rather on the properties of the depicted molecule.

While scale-space theory is equally applicable to 3D space, this is not true for all parts of the SIFT. The keypoint detection method of this work is equal to the SIFT. For orientation assignment and descriptor computation the data structures were extended to address the additional degrees of freedom in 3D space. For this purpose, an orientation histogram has been developed, which is accelerated by a geodesic index and comprises uniformly distributed bins. Both, the orientation assignment and descriptor computation rely on various parameters that allow for detailed parameter-studies and therefore enable the identification of suitable settings.

The compact map description can be used for various purposes. Two applications are explicated here: map registration and molecule recognition. For both applications, descriptors are converted to feature vectors. These are compared using the Euclidean distance, which allows for analyzing the similarity of local neighborhoods of the maps. Based on this comparison, placements are proposed in the map registration application. These placements are subsequently validated using detailed scoring functions and clustering methods. In molecule recognition, the similarity value of two feature vectors is used to identify a protein depicted in an electron density map. Here, again, feature vectors are compared and similar reference structures are identified in a voting procedure. Compared to other approaches for the docking of molecular structures into electron density maps, this method allows for new means of comparing the content of maps. Its abstract description is constrained to salient features in the map, and there is no need for brute-force comparisons of two maps.

All tasks — i. e., calculating the map description, computing a registration, and performing molecule recognition — rely on various parameters, which have been introduced in this chapter. A summary of all parameters is listed in Table 3.2 along with a short description and the page number the parameter is defined on. In the following Chapter 4, the robustness of each component of the method is assessed in detail and optimal values for each parameter are determined.

Keypoint Detection		
Parameter	Description	Page
$\{\sigma_0\}_{\text{vox}}$	Initial PSF standard deviation	66
s	Number of scale-space samples per octave	69
t_{contrast}	Contrast threshold	72
$t_{\text{cornerness}}$	Cornerness threshold	72

Orientation Assignment		
Parameter	Description	Page
w_{samp}	Sampling interval of gradient vectors	81
w_σ	Standard deviation of the Gaussian weight function	82
w_{width}	Width of the truncation window	82
h_{g}^{3D}	Subdivision of the 3D geodesic grid	82
h_{t}^{3D}	Threshold for the 3D orientation histogram	83
h_{g}^{2D}	Number of bins in the 2D histogram	83
h_{t}^{2D}	Threshold for the 2D histogram	83

Neighborhood Descriptor		
Parameter	Description	Page
r	Radius of the descriptor, i. e., number of cubes in the descriptor	85
δ	Cube width	85
g	Subdivision level of the orientation histograms	85
σ_{d}	Standard deviation of the Gaussian weight function	85
p	Number of sampled gradient vectors in one dimension per cube	85
τ_{abs}	Absolute similarity threshold	90
τ_{dist}	Distinctiveness similarity threshold	90

Table 3.2 – Parameters
The table lists the parameters of the method for the three stages keypoint detection, orientation assignment, and descriptor computation. For each parameter, a short description and the page number on which the parameter is introduced is given.

4. Validation and Parameterization

In the previous Chapter 3, a method for similarity searching in electron density maps based on the SIFT [218] was described. It relies on a map description, which is determined in three main stages, namely keypoint detection, orientation assignment, and descriptor computation. Furthermore, means for computing a registration of two maps and for molecule recognition have been described. For all these methods, parameters have been introduced. Different settings for these parameters are studied in this chapter with the goal of identifying a parameter set that enables map registration and molecule recognition. The resulting parameter set will then be used in Chapter 5 for similarity searching in experimentally acquired electron density maps.

Optimal parameters are determined in repeatability experiments that are performed on synthetically generated maps. Similar studies assess the repeatability of keypoint detection in 2D [237, 218, 238] and 3D images [137, 278, 138]. Here, the influence of discretization noise, resolution lowering, and additive white Gaussian noise is investigated. These analyzes are based on average repeatability rates, since there are no predetermined landmarks in electron density maps of proteins. The aim of the parameterization is to maximize the similarity of the generated map descriptions regardless of the introduced distortions. The presented studies assess this question for several reasonable combinations of parameters.

First, the training set and the experimental setup are introduced. Subsequently, the repeatability of keypoint detection is determined in Section 4.2. Here, different parameter combinations are assessed and an optimal set of parameters is selected. The same is done for orientation assignment in Section 4.3. In Section 4.4 the properties of the descriptor are analyzed in detail and two sets of suitable descriptor parameters are identified. The chapter concludes with a summary giving an overview of the determined parameter set.

4.1. Experimental Setup

A training set comprising fifteen diverse protein structures is used to assess the repeatability of keypoint detection, orientation assignment and descriptor

computation. The members of the training set are manually selected from the CATH database [254], which classifies protein domains by fold motives using a hierarchical system. This system consists of the four levels class, architecture, topology, and homologous superfamily. The lowest level — homologous superfamily — is clustered into sequence families using sequence identity thresholds. For the training set, a representative is selected for each CATH architecture that comprises more than 500 entries. This results in a set consisting of the fifteen protein domains listed in Table 4.1 and shown in Figure 4.1.[1]

The training set covers a wide range of molecular structures, which can be depicted in X-ray crystallography and cryo-EM experiments. In cryo-EM, however, larger molecules are required for a successful experiment, which are not found in the training set. This is, on the one hand, due to the complex computations performed for parametrization. On the other hand, these molecules also consist of domains that are represented in CATH. Therefore, only large scale keypoints that correspond to the assembly of domains are omitted from the parametrization. Considering the protein domain level, the training set comprises a representative set of the main CATH architectures. Since CATH architectures cluster similar protein domains, the training set consists of a diverse range of protein structures, which could equally be subject to registration or molecule recognition. Therefore, the training set comprises a representative selection of protein structures.

Three disturbances have been identified as having an influence on the repeatability of keypoint detection, orientation assignment, and descriptor computation. These include the discretization error, the signal-to-noise ratio (SNR), and the resolution of the map. Discretization error is introduced by the sampling of the underlying signal. Its influence on the repeatability is assessed using synthetic maps generated from a molecule in random orientations using different resolutions and sampling intervals. Since there is no ground truth for the location of keypoints in a protein map, a set of reference and test map descriptions is created. By comparing the descriptions computed from all test maps to the descriptions determined from all reference maps, the repeatability is determined as an average value.

Using a specified test molecule, n reference maps and n test maps are created with the specified resolution. Each map depicts the molecule in a different, random orientation. The orientations are generated by picking a point on the sphere-surface [67] that yields a rotation axis. Subsequently, the angle of rotation around the axis is specified by drawing a number from a uniformly distributed in the range $[0°; 360°[$. If specified, noise is added to the test maps, but not to

[1] The notation of identifiers in this work is explicated in Appendix A.2 on page 199.

Class / Architecture	Domain	Ref.	Description
Mainly Alpha			
Up–down Bundle	2LIS:A00	[185]	Sperm lysin
Orthogonal Bundle	1OAI:A00	[122]	Nuclear RNA export factor
Alpha Horseshoe	1IHG:A02	[325]	Cyclophilin 40
Mainly Beta			
Beta Barrel	1GVK:B02	[166]	Elastase 1ja
Sandwich	2HNU:A00	[200]	Oxytocin-neurophysin 1
Ribbon	1H8P:A02	[354]	Seminal plasma protein PDC-109
Distorted Sandwich	1M3Y:A01	[248]	Major capsid protein of PBCV-1
Roll	1NH2:D02	[31]	Transcription initiation factor IIA small chain
Mixed Alpha–Beta			
Alpha–Beta Complex	1JOP:A00	[255]	Cytochrome c3
Roll	3DLK:B01	[14]	p51 RT
Alpha–Beta Barrel	2EIY:B02	[120]	Branched-chain amino acid aminotransferase
2–Layer Sandwich	1C0P:A02	[336]	D-amino acid oxidase
3–Layer (aba) Sandwich	2HBA:A00	[60]	50S ribosomal protein L9
3–Layer (bba) Sandwich	2QJ2:A01	[327]	Hepatocyte growth factor
4–Layer Sandwich	1B25:A01	[149]	Formaldehyde ferrodoxin oxidoreductase

Table 4.1 – List of CATH domains contained in the training set

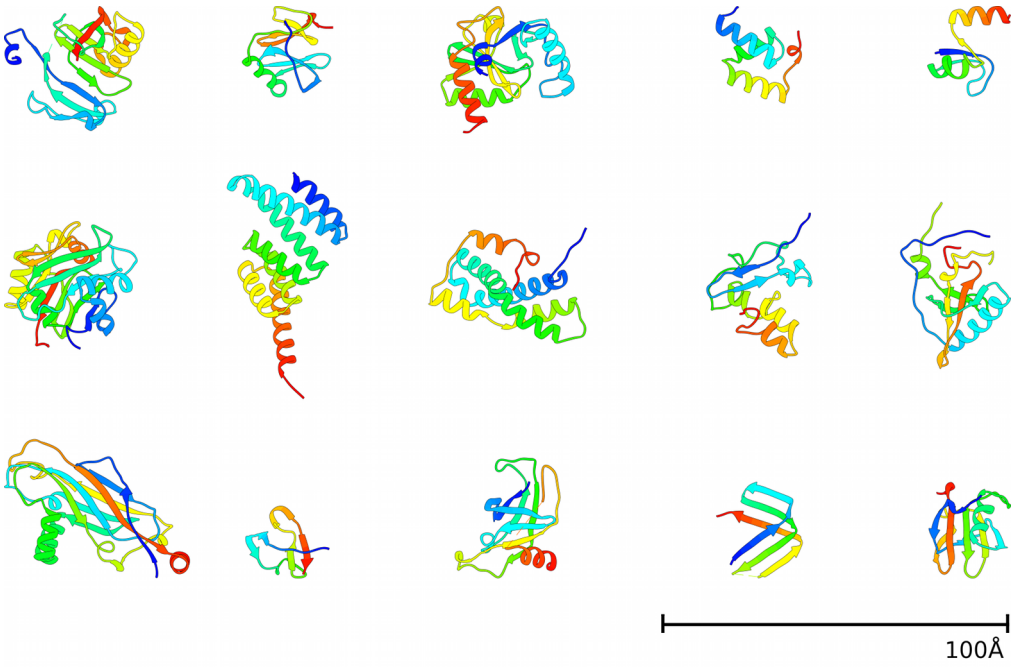


Figure 4.1 – Training set depiction
CATH domains contained in the training set. From left to right and top to bottom the domains 1C0P:A02, 2HNU:A00, 2EIY:B02, 1OAI:A00, 2HBA:A00, 1B25:A01, 1IHG:A02, 2LIS:A00, 1J0P:A00, 3DLK:B01, 1M3Y:A01, 1H8P:A02, 1GVK:B02, 1NH2:D02, and 2QJ2:A01 are shown. (© A. Griewel)

the reference maps. Eventually, keypoints and descriptors are determined for all $2n$ maps.

This process yields n^2 pairs of reference and test map descriptions. All maps are generated synthetically and depict the molecule in different orientations. The comparison of the map descriptions is performed by applying the inverse rotations to the keypoints and descriptors superposing the underlying molecules. The resulting, rotated keypoints and descriptors can be compared directly since their coordinate frames are the same after rotation. A pairing of reference and test map is called *trial pair*.

The SNR of the test maps is lowered using additive white Gaussian noise. All intensities in the maps are positive, therefore the SNR is defined as the quotient of the mean signal intensity μ_{signal} and the standard deviation of the noise σ_{noise} [115] as shown in Equation 4.1.

$$\text{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{noise}}} \tag{4.1}$$

Varying the SNR of a test map is accomplished by adding a random value to the intensity of each voxel of the map. This value is drawn from a Gaussian distribution [33] with a standard deviation of

$$\sigma_{\text{noise}} = \frac{\mu_{\text{signal}}}{\text{SNR}} \tag{4.2}$$

Examples for maps at different signal-to-noise ratios are shown in Figure 4.2.

The mean intensity μ_{signal} is determined independently for each map. It is solely based on voxels with an intensity larger than $10^{-3}m$, where m is the maximal intensity of the image. Thus, all voxels that have an intensity of less than one thousandths of the maximum intensity in the image are ignored when computing μ_{signal}. This means effectively that only voxels in an envelope around the protein are considered. Using this calculation, the influence of the orientation and shape of the depicted protein are minimized as can be seen in the following example: Given a rod-like protein depicted in two maps — once parallel to the map's grid, once diagonal to the grid while both maps share the same padding of one voxel. The second map would mainly consist of zero-intensity voxels. Thus, the mean μ_{signal} would be lower than for the first map when considering all voxels. This would result in a lower standard deviation σ_{noise} for the noise generating function and thus yield less distortion only due to a differing orientation. Using the above explained calculation of μ_{signal}, the resulting σ_{noise} depends on intensities inside an envelope around the protein volume only and is independent of the orientation of the depicted protein.

Using this setup, the repeatability of keypoint location and orientation assignment as well as the descriptor properties can be assessed. For one molecule, this

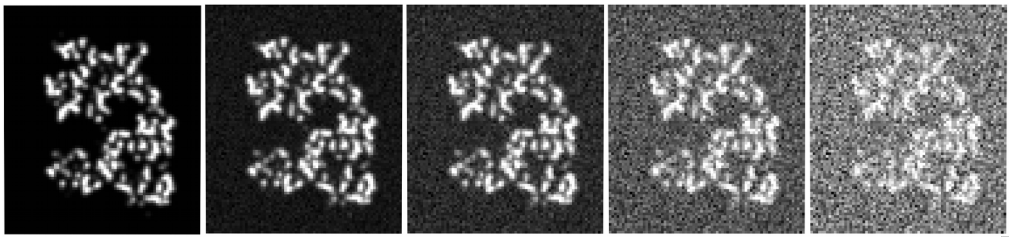


Figure 4.2 – Adaption of the signal-to-noise ratio
A slice through a synthetic map of the molecule GroEL at 1 Å sampling interval and 3.5 Å resolution is shown (left). In the following slices the signal-to-noise ratio is adapted to 10, 5, 2, and 1 from the left to the right using additive white Gaussian noise. (© A. Griewel)

is done by generating all trial pairs and determining the average value of the considered attribute. In the following keypoint detection study, $n = 5$ reference and test maps are used yielding in total 25 trial pairs. Larger values for n do not show significant changes in the results. The final, reported value is the average value of all fifteen molecules in the training set.

4.2. Keypoint Detection

A robust detection of keypoints is vital for the success of identifying similarities in electron density maps: Only if a keypoint has been identified in a neighborhood, descriptors will be generated, which can be used for comparison. An excess detection of keypoints, however, is not beneficial for solving the problem of similarity searching, since it will introduce errors and lower efficiency.

Following related studies [137, 278, 138], three objectives are specified that are required to be fulfilled for a high quality keypoint detection method

Sufficient Amount of Repeatable Keypoints A large number of distinctive, repeatable keypoints facilitates the description of the map in its entirety.

High Ratio of Repeatable Keypoints A stable description determines keypoints in the same location relative to the molecule independent of noise and orientation.

Small Number of Excess Keypoints Keypoints that are detected due to noise must be avoided since they introduce false matchings and decrease efficiency.

To quantify the performance of the method with respect to the objectives, three values are determined for each trial pair. These are the number of reference keypoints n_{ref}, the number of test keypoints n_{test}, and the number of repeatable keypoints n_{match}. The latter measure n_{match} is defined as the number of reference keypoints that have a test keypoint in their neighborhood that is compatible[1]. Here, the neighborhood of the reference keypoint is defined as a sphere using the keypoint's scale as radius.

Using these three values, which are computed for each trial pair, three quantities are defined for the repeatability test, which are shown in Equation 4.3 and allow for assessing the fulfillment of the objectives. The three quantities are the averages of the number of detected reference keypoints $P_{\#\text{ref}}$, the ratio of repeatedly detected keypoints P_{perc}, and the excess ratio P_{exc}, which quantifies the amount of test keypoints with respect to the number of reference keypoints

$$P_{\#\text{ref}} = \langle n_{\text{match}} \rangle \qquad P_{\text{perc}} = \left\langle \frac{n_{\text{match}}}{n_{\text{ref}}} \right\rangle \qquad P_{\text{exc}} = \left\langle \frac{n_{\text{test}}}{n_{\text{ref}}} \right\rangle \tag{4.3}$$

Here, the chevrons $\langle \cdot \rangle$ denote the arithmetic mean of the enclosed quantity, which is calculated per trial pair. The repeatability values presented in the following are the arithmetic mean of the above mentioned quantities for the complete training set.

The described test is performed under different *external conditions* and for different *internal parameter sets* for every protein domain in the training set. The external conditions that are assessed include the initial voxel spacing, resolution, and signal-to-noise ratio. The voxel spacing is set to 1 Å, 2 Å, and 3 Å while initial resolutions of 3.5 Å, 6.9 Å, and 10.4 Å are considered. These values corresponds to Gaussian point spread functions with standard deviations of 1 Å, 2 Å, and 3 Å. Furthermore, the signal-to-noise ratio is adapted to different levels. One noiseless setup — i. e., without added white Gaussian noise — is complemented by setups with SNR 5, 2, and 1. This yields $3 \cdot 3 \cdot 5 = 45$ different combinations of sampling intervals, resolutions, and SNRs that are termed *external conditions* in the following. The assessed resolutions and sampling rates are typical for maps accessible through the internet. A resolution of 3.5 Å is at the lower range of resolutions encountered in X-ray crystallography and therefore allows for a parameterization for high resolution maps. Maps at the intermediate resolution 10 Å are frequently acquired by cryo-EM maps and comprise structural features such as secondary structure motives. Maps of lower resolution are not included in the study, since the protein domains in the test set are not sufficiently large. However, the results have been validated for large molecules in low resolution maps in subsequent, selected case studies.

[1] See Section 3.5 on page 89.

The relevant *internal parameters* need to be set so that the repeatability is optimal with respect to the objectives specified on page 106. These parameters include the initial sampling interval, which is determined according to the parameter $\{\sigma_0\}_{\text{vox}} \in \{0.8; 0.9; \ldots 1.4\}$. This parameter specifies the standard deviation of the Gaussian point spread function associated with the base map in terms of voxels. Closely related to this parameter is the number of samples $s \in \{4; 5; \ldots 10\}$ that are created for each octave. Additionally, the thresholds on contrast t_{contrast} and cornerness $t_{\text{cornerness}}$ need to be specified. These latter two parameters are set to 0.02 and 10 in the following experiments and their determination is explained at the end of the section.

The findings on keypoint repeatability are summarized in this section using selected plots, which demonstrate the effects of the internal parameters under different external conditions. A complete list of plots showing the effects for all combinations of external and internal parameters is shown in Appendix A.5 on page 203. For all experiments, the scale-space representations are built using base formation, i. e., by convolving the base map of each octave with a Gaussian to achieve the desired resolution. The results for incremental formation are also discussed at the end of the section.

Figures 4.3, 4.4, 4.5 and 4.6 are split into nine panels. These correspond to the external conditions applied to the experiment, i. e., initial voxel spacing and resolution. For each SNR — the third external distortion — one set of plots is shown. Inside each panel, a heat map is located, which uses a rainbow color scheme and summarizes the outcome of the experiment when applying different values for s and $\{\sigma_0\}_{\text{vox}}$. Each panel in the heat map represents a *parameter set*. The three panels on the lower left are empty here. These fields correspond to experiments, in which the initial sampling rate is not able to represent the signal faithfully according to the sampling theorem: The width of the Gaussian point spread function is lower than 1.0 in terms of voxels, which introduces aliasing effects [115]. In practice, it is assumed that all maps are sampled in a way that the underlying signal is represented faithfully. Therefore, maps with insufficient sampling rate are neither considered in the repeatability tests nor for the determination of parameters. For completeness, the plots of these maps are listed in Appendix A.5. In the following, the main observations made during the analysis of the plots are noted first. Subsequently, conclusions are drawn from these observations, which yield an optimal parameter set.

The number of keypoints detected by *siseek* in reference maps $P_{\#\text{ref}}$ is shown in Figure 4.3 for different values of s and $\{\sigma_0\}_{\text{vox}}$. These keypoints are detected in noiseless maps. The value $P_{\#\text{ref}}$ depends on the ratio of the initial standard deviation to the number of samples in the octave: $\frac{\{\sigma_0\}_{\text{vox}}}{s}$. If the value of $\frac{\{\sigma_0\}_{\text{vox}}}{s}$

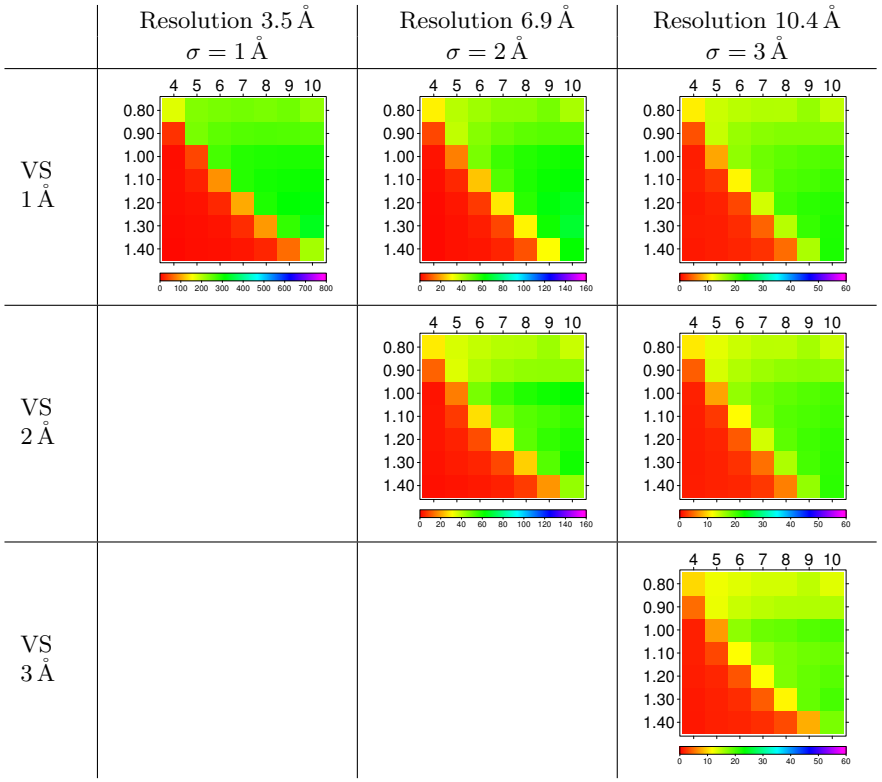


Figure 4.3 – Number of detected keypoints $P_{\#\text{ref}}$ in noiseless maps
The heat maps show values for different resolutions (table columns) and voxel spacings (VS, table rows). The initial sampling rate $\{\sigma_0\}_{\text{vox}}$ (heat map rows) and the number of samples per octave s (heat map columns) are varied for each combination of resolution and voxel spacing.

is high, few keypoints are detected (lower left corner of the heat map). For lower ratios — especially above the diagonal in the plot — the number of detected keypoints rises. Additionally, the number of detected keypoints rises with the number of samples s. With respect to the first objective, these plots suggest to choose a parameter combination located on or above the diagonal in the plot, which satisfies

$$10\,\{\sigma_0\}_{\text{vox}} - 4 \leq s$$

The ratio of repeatedly detected keypoints P_{perc} is initially analyzed using noiseless test maps and the resulting average values are shown in Figure 4.4. The repeatability for parameter combinations below the diagonal in the plots is relatively small and close to 40 %. Larger rates of 70 %–80 % are observed for

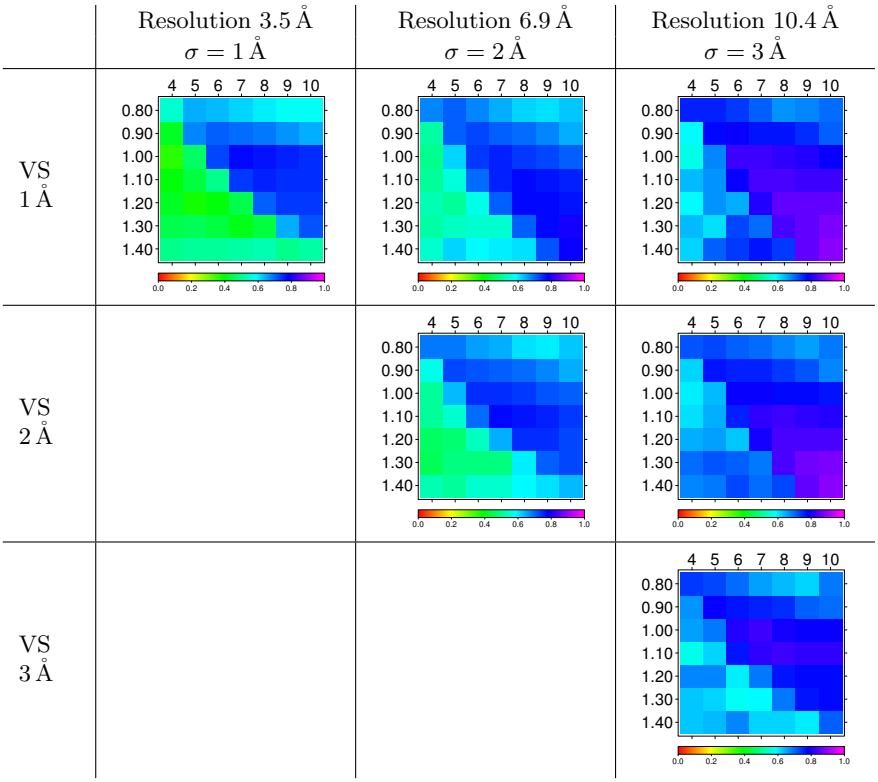


Figure 4.4 – Ratio of repeatedly detected keypoints P_{perc} in noiseless maps
The heat maps show values for different resolutions (table columns) and voxel spacings (VS, table rows). The initial sampling rate $\{\sigma_0\}_{\text{vox}}$ (heat map rows) and the number of samples per octave s (heat map columns) are varied for each combination of resolution and voxel spacing.

parameter sets above the diagonal. Here, the repeatability for parameter sets above the diagonal with $\{\sigma_0\}_{\text{vox}} \leq 1$ is lower than for larger values of $\{\sigma_0\}_{\text{vox}}$. With respect to the second objective, these plots also suggest a parameter combination located on or above the diagonal in the plot with $\{\sigma_0\}_{\text{vox}} > 1$.

Following the assessment of noiseless conditions, the influence of noise on the repeatability of keypoint detection is assessed. The results for P_{perc} at SNR 5 are shown in Figure 4.5. As expected, the repeatability rates deteriorate with larger amounts of noise but still resemble the findings for the noiseless scenario. For values that are located on the diagonal, repeatability rates of up to 65 % are reached. Parameter combinations located above the diagonal yield even higher rates for $\{\sigma_0\}_{\text{vox}} \geq 1$. In external conditions with larger amounts of noise — i. e.,

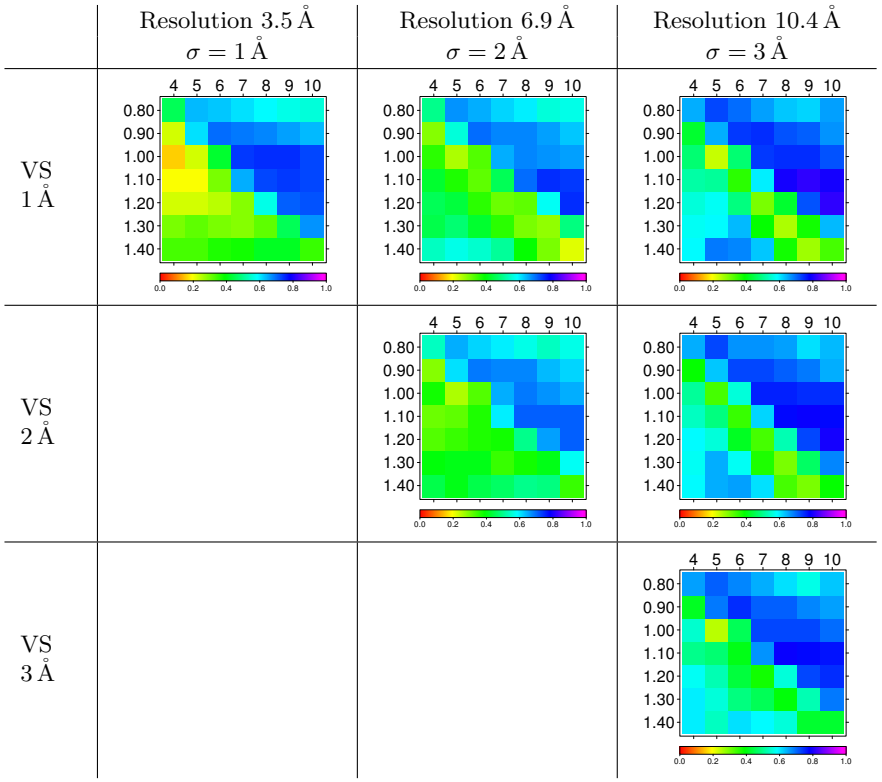


Figure 4.5 – Ratio of repeatedly detected keypoints P_{perc} in maps with SNR 5
The heat maps show values for different resolutions (table columns) and voxel spacings (VS, table rows). The initial sampling rate $\{\sigma_0\}_{\text{vox}}$ (heat map rows) and the number of samples per octave s (heat map columns) are varied for each combination of resolution and voxel spacing.

SNR 2 and SNR 1 — the observed value P_{perc} deteriorates to approximately 30 % for parameter combinations located on the diagonal of the plot.

The excess ratio P_{exc} lies close to one for all experiments in noiseless maps. This means that the number of detected keypoints is on average the same in reference and test map. With added noise, the number of detected keypoints in the test map deviates from the number of keypoints detected in the reference maps, which results in excess ratios P_{exc} differing from one. In Figure 4.6 the values of P_{exc} at SNR 5 are shown. For almost all parameter sets above the diagonal, the excess ratio is larger than 2. This means that twice as many keypoints are detected in the test map, as in the noiseless reference map.

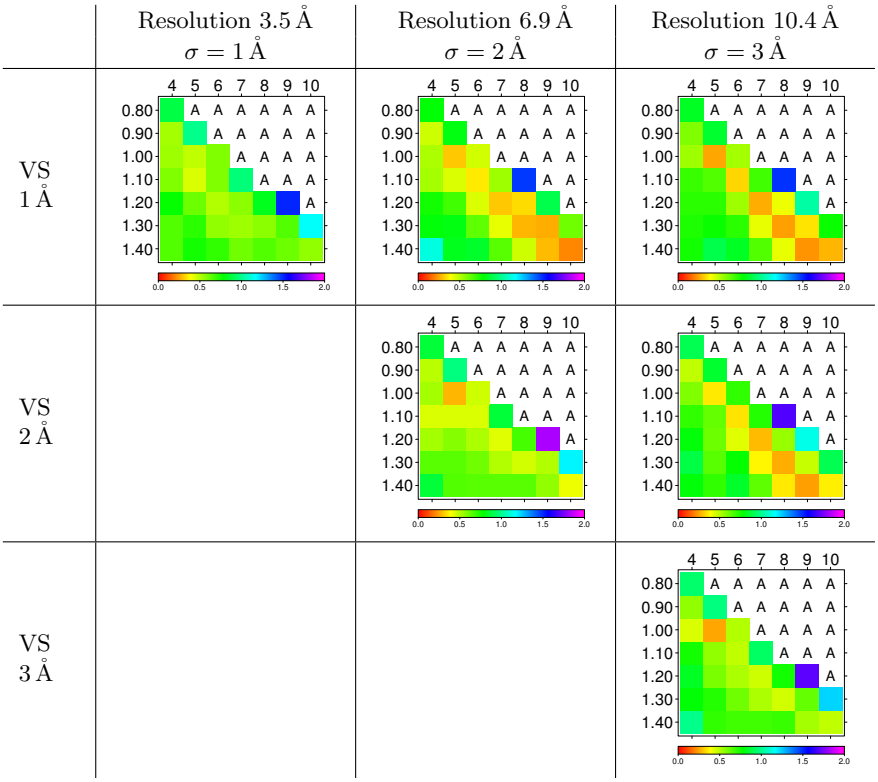


Figure 4.6 – Excess ratio P_{exc} in maps with SNR 5
The heat maps show values for different resolutions (table columns) and voxel spacings (VS, table rows). The initial sampling rate $\{\sigma_0\}_{\text{vox}}$ (heat map rows) and the number of samples per octave s (heat map columns) are varied for each combination of resolution and voxel spacing. An "A" in the heat maps indicates that the measured value lies above the displayed scale.

The interpretation of the results is based on the effects of the two studied parameters. While s specifies the number of samples in each scale-space representation, $\{\sigma_0\}_{\text{vox}}$ determines the sampling interval of the base map as shown in Figure 3.3 on page 68. For larger values of $\{\sigma_0\}_{\text{vox}}$, a smaller sampling interval is required. Therefore, maps consisting of more voxels are generated. Smaller values of $\{\sigma_0\}_{\text{vox}}$ have the contrary effect and are beneficial in terms of efficiency, since these require less voxels in the base map.

A main result from the examination of the heat maps is that the repeatability is only large for parameter combinations above the diagonal. This is analyzed by first considering only one $\{\sigma_0\}_{\text{vox}}$ — i. e., one row in the heat map — and subsequently the whole heat map: The small number of keypoints for parameter combinations with small $\{\sigma_0\}_{\text{vox}}$ on the left of the heat maps can be attributed to the method by which keypoints are detected. These are local extrema in the DoG scale-space that are identified by comparing intensities of neighboring voxels. Here, not only the spatial neighborhood, but also the neighborhood with respect to scale is considered. If the spacing of samples s with respect to scale is too large, it is not possible to identify the relevant extrema. A finer sampling of the scale-dimension allows for the identification of less pronounced local extrema. These corresponds to blobs that are only slightly lighter or darker than their surrounding. Thus, a larger number of samples results in an increased amount of detected extrema. If the sampling along the scale-domain is, however, too fine, the unstable extrema are detected that are frequently induced by noise. This is documented by the large excess ratios for parameter combinations above the diagonal under noisy external conditions.

Considering all combinations of s and $\{\sigma_0\}_{\text{vox}}$, it is apparent that the repeatability depends on the ratio of the number of samples s and the standard deviation of the initial Gaussian point spread function in terms of voxels $\{\sigma_0\}_{\text{vox}}$. The larger $\{\sigma_0\}_{\text{vox}}$, the more scale-space samples s are required for a repeatable keypoint detection. In other words: a finer sampling of the spatial domain requires also a finer sampling of the scale-domain. This can also be explained by the keypoint detection method: If the sampling frequency of a map rises, the intensity differences of neighboring voxels become generally smaller. The initial scale of an octave depends on the point spread function of the base map. If $\{\sigma_0\}_{\text{vox}}$ is larger, the width of the Gaussian function used for building the scale-space with respect to voxels is also larger. Therefore, extrema in the DoG maps become less pronounced in finer sampled maps because a wider filter is used. Increasing the number of samples s addresses this effect by increasing the resolution along the scale-domain and allowing for the detection of less pronounced local extrema.

The repeatability of keypoint detection is decreased, if the input map is resampled so that the initial sampling interval is smaller than the standard deviation of the Gaussian point spread function $\{\sigma_0\}_{\text{vox}} \leq 1$. In these cases noticeable aliasing effects occur because the signal is not represented faithfully as explained by the signaling theorem. This can be explained by the Fourier transform of the Gaussian function, which is again a Gaussian function. This function approaches, but never reaches zero. Therefore, a finite approximation will always introduce aliasing effects when sampling a Gaussian signal. These effects are insignificant if $\{\sigma_0\}_{\text{vox}}$ is sufficiently large and sufficient padding is used. For small values of $\{\sigma_0\}_{\text{vox}} \leq 1$, however, these effects become significant and impair the repeatable detection of keypoints.

Assembling the above findings yields that parameter combinations located on the diagonal of the plots are most suitable for repeatedly detecting keypoints. At this, the sampling rate must allow for a faithful representation of the signal. Therefore, the standard deviation of the Gaussian point spread function of the first map in the scale space map must be sufficiently large to avoid significant aliasing. Inspecting the presented plots and those in Appendix A.5, an optimal parameter combination with respect to the presented objectives is $\{\sigma_0\}_{\text{vox}} = 1.1$ and $s = 7$. This parameter combination lies on the diagonal of the plots and yields a large number of keypoints that are highly repeatable while showing small excess ratios.

All presented measurements are performed on scale-spaces that are created using base formation and a sampled Gaussian kernel. Using a discrete Gaussian kernel [205] did not improve the repeatability of keypoints. In a second experiment, the incremental formation is used for creating the scale-space representations. It was found, that only half the keypoints are detected using this setup while the repeatability ratios P_{perc} are comparable. However, this setup is more susceptible to noise, which can be explained by the small standard deviations employed in the Gaussian filters. These are regularly less than one voxel and result in significant aliasing.

The parameters contrast threshold t_{contrast} and cornerness threshold $t_{\text{cornerness}}$ remain to be determined. They are utilized to discard poorly defined keypoints that are either located in regions with little information content — i. e., only small changes in the signal — and keypoints that are located along edges. The test indicated that the repeatability will attain a maximum for $t_{\text{cornerness}} = 10$ and $t_{\text{contrast}} = 0.05$. These values have been validated with respect to experimental electron density maps and it was found that a contrast threshold of 0.05 discards valuable keypoints. This is caused by gradual differences in overall intensity, which can be observed in experimental data. Therefore, a lower

contrast threshold of $t_{\text{contrast}} = 0.02$ was determined by comparing experimental and synthetic maps, while keeping $t_{\text{cornerness}} = 10$.

4.3. Orientation Assignment

Seven parameters have been introduced that are required for orientation assignment. These are the width of the window w_{width}, the standard deviation of the Gaussian weight function w_σ, and the relative sampling interval for gradient vector calculation w_{samp}. These first three parameters specify spatial lengths with respect to the scale of the keypoint. Furthermore, there are the histogram parameters granularity h_{g}^{3D} and h_{g}^{2D} as well as the cutoffs h_{t}^{3D} and h_{t}^{2D}.

The parameters are determined using an experimental setup similar to the one used for determining the repeatability of keypoints. The same voxel spacings, resolutions and SNRs are assessed. However, the locations of the keypoints in this setup are constant with respect to the protein. This is accomplished by determining a fixed set of keypoints on the genuine protein. These are rotated according to the same random rotation that is applied to the protein. Therefore the orientation histogram is assembled in the same location with respect to the protein.

Besides the experimental setup, a distance measure for the comparison of two rotations is needed. A frequent choice for this task is the angle between the quaternions representing the rotations [187]. Here, however, a measure is needed that determines the minimum angle of rotation about any axis and does not rely on the distance between the rotation axes. The reason for this is found in the properties of the descriptors, which are robust in matching for rotations of up to 10° as shown in Section 4.4 on page 119.

In the described setup, the distance between a reference rotation $\mathbf{R}_{\text{ref}}$ and a test rotation $\mathbf{R}_{\text{test}}$ is measured. The similarity of the two rotations is assessed using the transforming rotation $\mathbf{R}_{\text{trans}}$, which is defined as the rotation that yields $\mathbf{R}_{\text{test}}$ when applied prior to $\mathbf{R}_{\text{ref}}$. In terms of rotation matrices, $\mathbf{R}_{\text{trans}}$ is defined as the rotation that yields $\mathbf{R}_{\text{test}}$ when right multiplied to $\mathbf{R}_{\text{ref}}$. The transforming rotation is then derived by solving this equation for $\mathbf{R}_{\text{trans}}$ using the properties of rotation matrices

$$\begin{aligned} & \mathbf{R}_{\text{ref}} \cdot \mathbf{R}_{\text{trans}} &&= \mathbf{R}_{\text{test}} \\ \Leftrightarrow \quad & {\mathbf{R}_{\text{ref}}}^T \cdot \mathbf{R}_{\text{ref}} \cdot \mathbf{R}_{\text{trans}} &&= {\mathbf{R}_{\text{ref}}}^T \cdot \mathbf{R}_{\text{test}} \\ \Leftrightarrow \quad & \mathbf{R}_{\text{trans}} &&= {\mathbf{R}_{\text{ref}}}^T \cdot \mathbf{R}_{\text{test}} \end{aligned} \tag{4.4}$$

The distance between two rotations is defined by interpreting the rotation $\mathbf{R}_{\text{trans}}$ as pair of a rotation axis and rotation angle $d_{\measuredangle}$. The angle $d_{\measuredangle}$ is used as

Window	Histogram
$w_\sigma \in \{1; 1.5; \mathbf{2}; 2.5; 3\}$	$h_\mathrm{g}^\mathrm{3D} \in \{2; \mathbf{3}; 4; 5\}$
$w_\mathrm{width} \in \{\mathbf{2}; 3\}$	$h_\mathrm{t}^\mathrm{3D} \in \{\mathbf{0.8}; 0.9\}$
$w_\mathrm{samp} \in \{\mathbf{0.5}; 0.75; 1\}$	$h_\mathrm{g}^\mathrm{2D} \in \{\mathbf{36}; 72\}$
	$h_\mathrm{t}^\mathrm{2D} \in \{0.8; \mathbf{0.9}\}$

Table 4.2 – Tested parameter sets for orientation assignment
The study for identifying optimal parameters for orientation assignment included all combinations of the above shown parameters. The parameters w_σ, w_width, and w_samp specify lengths with respect to the scale σ of the corresponding keypoint. The combination of parameters that yield optimal repeatability of orientation assignment are bold.

the distance measure between $\mathbf{R}_\mathrm{ref}$ and $\mathbf{R}_\mathrm{test}$. It describes the amount of rotation that is necessary to transform the reference rotation into the test rotation.

The following studies are based on protein domain `[1OAI:A00]`$_{\texttt{CATH}}$ from the previously introduced test set. This protein was chosen because it comprises different structural features and is of medium size. All combinations of parameters shown in Table 4.2 have been assessed to identify the most suitable setup. The parameters w_σ and w_samp specify parameters with respect to the scale of the considered keypoint. The width of the window w_width is specified relative to the standard deviation of the Gaussian weight function w_σ. Using this setup, the average number of orientations is determined and the repeatability of the orientation assignment is assessed. This is defined as the average ratio of detected keypoints that have at least one orientation with $d_\measuredangle < 10°$ in reference and test map.

Four objectives are identified for the selection of an optimal parameter set:

High Repeatability To determine comparable descriptors, it is essential that orientations are assigned within the defined error bound $d_\measuredangle < 10°$.

Small Number of Orientations per Keypoint A small amount of orientations per keypoint minimizes the amount of calculated descriptors and therefore increases efficiency.

Small Window The orientation of the keypoint is to be determined based on the information contained in the local neighborhood. Therefore, a small window width expressed in w_width and w_σ is desired.

Small Number of Samples The number of samples computed for the determination of the orientation histogram influences the run time. Thus, a small

number of samples per orientation histogram and therefore a large value for w_{samp} is beneficial.

In the analysis of the generated repeatability values, the parameters marked in Table 4.2 were identified as optimally fulfilling these objectives. For all assessed resolutions, the repeatability rate lies above 97 % for noiseless maps and is larger than 90 % for SNR 1. This is achieved by assigning on average 7.7 orientations to a keypoint.

The experiments show that the repeatability of orientation assignment rises when increasing the standard deviation w_σ up to 2σ. Larger values do not improve the repeatability. Here, a window cutoff w_{width} of $2w_\sigma$ suffices for a robust orientation assignment. A three-times subdivided geodesic grid $h_{\text{g}}^{\text{3D}} = 3$ for the orientation histogram in combination with a 2D histogram comprising $h_{\text{g}}^{\text{2D}} = 36$ bins proved as optimal for the chosen criterion $d_{\measuredangle} < 10°$. Finer granularities of the histograms do not improve the repeatability. The thresholds of the histograms h_{t}^{3D} and h_{t}^{2D} have a limiting effect on the number of assigned orientations. They are chosen to achieve maximal repeatability at a minimum amount of assigned orientations. The last parameter — subsampling w_{samp} — is essential for achieving a high repeatability rate. Using no subsampling, the repeatability drops to 87 %. Even for a subsampling of $w_{\text{samp}} = 0.75$ the repeatability is only at 92 %. Therefore, the smallest assessed subsampling of $w_{\text{samp}} = 0.5$ is utilized.

These parameters have been determined using keypoints that are located at the same position with respect to the protein. The keypoint location may, however, vary within the keypoint's scale. To analyze the repeatability of orientation assignment for genuinely detected keypoints, the test setup introduced in Section 4.1 is utilized. The ratio of matched keypoints that have been assigned the same orientation is determined with respect to the total number of matched keypoints. The results for SNR 5 are shown in Figure 4.7, while the remaining plots are located in Appendix A.5. The plots show that for the selected parameter combination of $s = 7$ and $\{\sigma_0\}_{\text{vox}} = 1.1$ a repeatability rate of more than 85 % is achieved. In the noiseless scenario — shown in the appendix — more than 90 % of the orientation assignments are repeatable. For the smaller SNRs 2 and 1 this value deteriorates to approximately 70 % and 60 % respectively.

These findings show that the orientation assignment is robust against the allowed changes in the location of the keypoint and also to large amounts of noise. Therefore, it is a robust basis for the computation of descriptors.

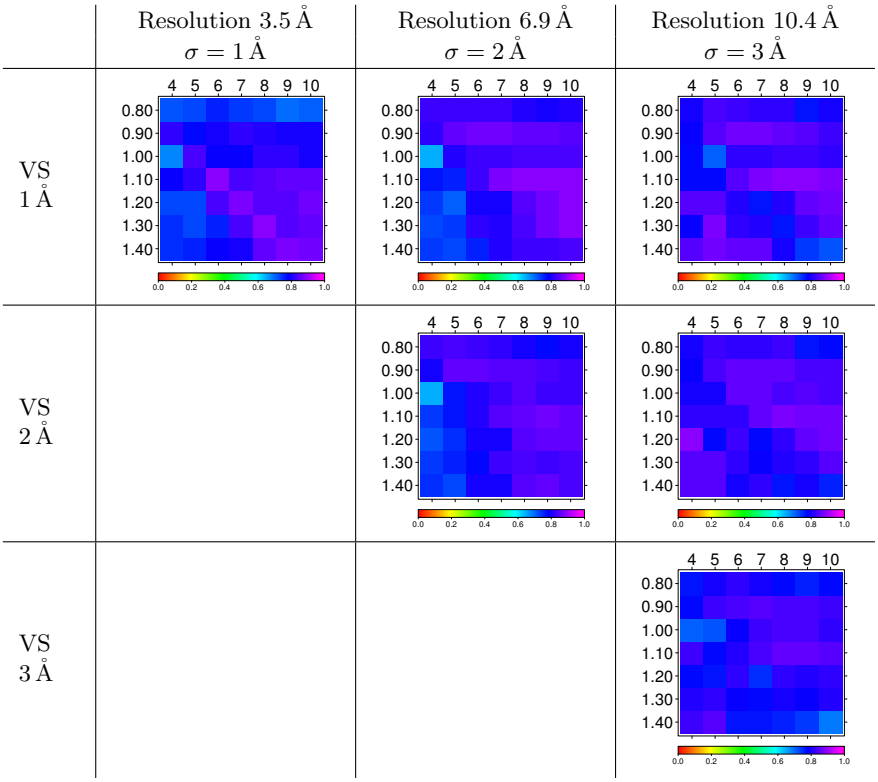


Figure 4.7 – Repeatability of orientation assignment in maps with SNR 5
The heat maps show values for different resolutions (table columns) and voxel spacings (VS, table rows). The initial sampling rate $\{\sigma_0\}_{\text{vox}}$ (heat map rows) and the number of samples per octave s (heat map columns) are varied for each combination of resolution and voxel spacing.

4.4. Neighborhood Descriptor Computation

The computation of robust and distinguishable descriptors allows for the comparison of keypoints based on their local neighborhood. The descriptor yields a feature vector, which facilitates the comparison of the surrounding of keypoints in feature space. Based on this comparison, keypoints are matched and irrelevant matches are discarded, which effectively reduces the number of possible keypoint assignments.

A descriptor captures the gradient field in the surrounding of the keypoint. The neighborhood of the keypoint is subdivided into cubic volumes, which in turn are represented using an orientation histogram as described in Section 3.4. Five parameters have been introduced that control the form of the descriptor. The edge length of the cubes δ is specified relative to the keypoint's scale σ. The number of considered cubes is specified by a radius r. For each cube, an orientation histogram of subdivision level g is assembled and the histogram is populated by inserting p^3 gradient vectors that are calculated inside the cube. Additionally, the gradient vectors are weighted according to a Gaussian function with standard deviation σ_{d}.

Four objectives are defined for the selection of suitable descriptor parameters:

High Discriminative Power The descriptors must be capable of discriminating between descriptors from similar and non-similar local neighborhoods.

Low Dimensionality High dimensional feature vectors require longer run times for the computation of the distance function. Furthermore, high dimensional feature vectors do not necessarily increase the discriminative power of descriptors when employing nearest neighbor searching using the Euclidean distance. This metric looses contrast in high dimensional spaces, which is also known as the curse of dimensionality [20, 28, 140]. For these two reasons, a small number of cubes and a low subdivision level of the orientation histogram are beneficial for efficiency.[1]

Locality Enlarging the volume comprised in the cubes in turn increases the discriminative power of the descriptor. However, enlarging the volume makes the descriptor comprise information on a more global and not a local neighborhood. Therefore, a small total volume of the descriptor is desired, which yields feature vectors based on the keypoint's close surrounding only.

Small Total Number of Samples For each cube in the descriptor, p^3 samples are calculated and inserted in the orientation histogram. For an efficient

[1] These two parameters determine the dimensionality of the feature vector as shown in Table 3.1 on page 89.

and effective calculation, the value of p must be chosen as low as possible while guaranteeing a stable computation of feature vectors.

4.4.1. Robustness to Distortions

A *robustness test* of the descriptor properties and the resulting feature vector is performed in a first step. This test elucidates the properties of the computed feature vectors and their distance in feature space with respect to the allowed spatial distortions. Thus, it enables the characterization of the distance function with respect to the utilized parameter set as explained in the following.

A synthetic, noiseless map of the training set protein domain $\texttt{[1OAI:A00]}_{\texttt{CATH}}$ at resolution 3.5 Å and sampling interval 1 Å is used for the robustness test. The protein domain was chosen since it belongs to a large architecture class in CATH, and the high resolution scenario because the largest amount of descriptors is determined in this scenario. For each descriptor that is detected in the synthetic map, the robustness with respect to a rigid transformation on that map is assessed. The transformation comprises a random dislocation in a spherical volume that has the radius of the corresponding keypoint's scale σ. The rotation is performed around a random axis that is determined by randomly picking a point on the sphere [67] while the rotation angle φ is a parameter.

The influence of the allowed random translational displacement for a keypoint on descriptors and their comparison is analyzed at $\varphi = 0°$. In combination with rotations $\varphi > 0°$ about a random axis, also the influence of orientation assignment is evaluated. Here, only positive values $\varphi \in [0°; 80°]$ are considered since negative φ yield the same results. For each keypoint in the map, 100 random rigid transformations are performed and for each displacement the average Euclidean distance to the genuine descriptor is determined. The distance $\rho(\varphi)$ for a given rotation φ is defined as the average of all computed feature vector distances.

The robustness test is performed for various combinations of parameters that are listed in Table 4.3. In the following, only the results for $p = 4$ yielding $4^3 = 64$ insertions to each orientation histogram are shown. Using solely $2^3 = 8$ insertions does not yield stable results and the calculation of $8^3 = 512$ gradient vectors shows only minor improvements over $p = 4$. Figure 4.8 shows the results of the robustness test for $\varphi \in [0°; 40°]$ and $\sigma_\mathrm{d} = 1.0$ using an orientation histogram of subdivision level 0 and 1, respectively. The results for the angular range $\varphi \in [0°; 80°]$ and $\sigma_\mathrm{d} \in \{0.5; 1.0; 1.5\}$ are shown in Appendix A.6.1 on page 212.

The influence of a random translational offset can be analyzed at the Y–intercept in Figure 4.8 where no random rotation is applied ($\varphi = 0°$). Four

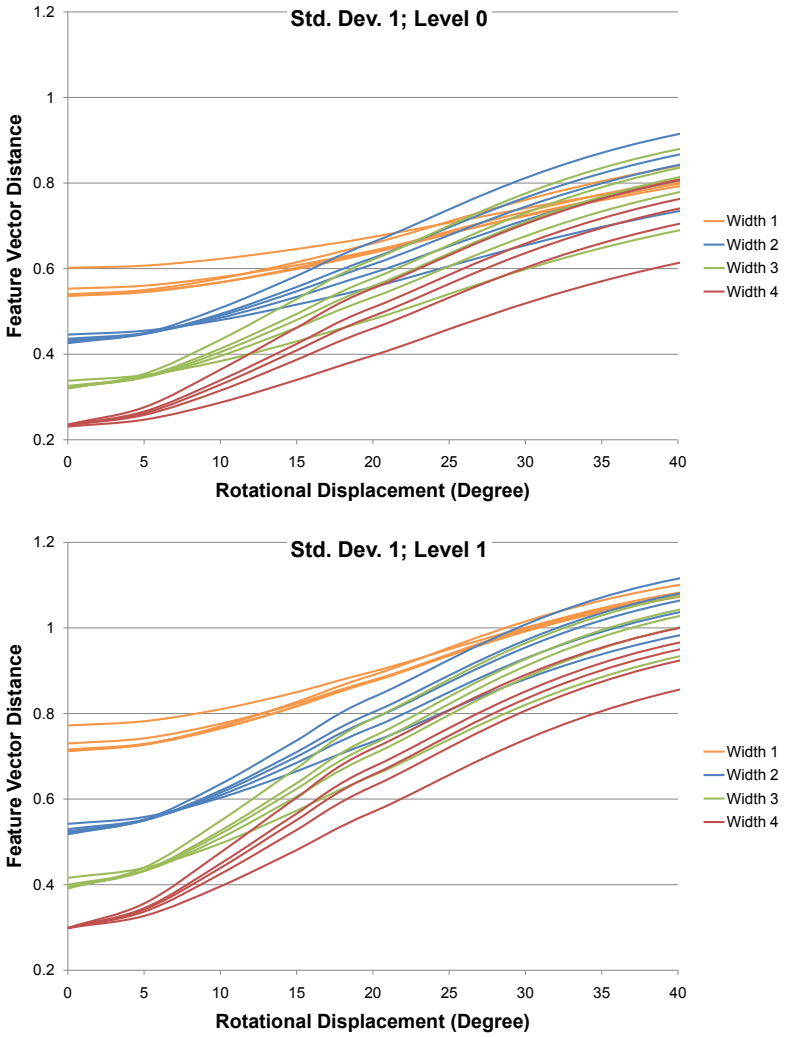


Figure 4.8 – Effect of random displacements on feature vector distance
The plots show the change in feature vector distance ρ (ordinate) when applying a random rigid transformation to the descriptor with specified rotation angle φ around an arbitrary axis (abscissa). For the upper plot, an orientation histogram of subdivision level 0 is used. The lower plot shows the result for subdivision level 1. Each line corresponds to a combination of δ and r. The parameter sets can be grouped into cohorts (colors) by cube width δ. Lines in each cohort are not distinguished by r for clearer view. The value of r can be determined by the slope of the curve: The steeper the curve, the larger r. (© A. Griewel)

$\delta \in \{1; 2; 3; 4\}$	$p \in \{2; 4; 8\}$
$r \in \{\sqrt{1}; \sqrt{2}; \sqrt{3}; \sqrt{4}; \sqrt{5}\}$	$g \in \{0; 1\}$
$\sigma_\mathrm{d} \in \{0.5; 1.0; 1.5\}$	

Table 4.3 – Tested descriptor parameter sets
The properties of the descriptor with respect to random displacements are analyzed using all combinations of the above shown parameters.

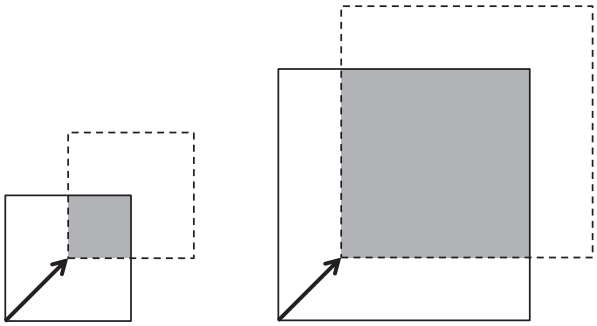


Figure 4.9 – Effect of random displacement on descriptors with respect to feature vector volume
Two squares with relative edge length 1 and 2 (solid line) are shown on the left and the right hand side. The squares are translated by the same vector (arrow) to a new location (dashed line). For the smaller square only 25 % of the same area is covered after dislocation. For the larger square this values is 56 %. (© A. Griewel)

cohorts can be identified in both diagrams, which correspond to the width δ of one descriptor cube. Small values for the cube width δ yield large differences in $\rho(0°)$ while larger values decrease the average distance. This is caused by the larger ratio of the volume in one cube that remains the same when the cube is translated. A sketch exemplifying this fact for a square is shown in Figure 4.9: If the displacement relative to the cube width is large, the part of the map underlying each cube is going to change considerably. If the displacement is relatively small, the underlying volume of the map is going to be equal in large parts. In the latter case, the orientation histogram of the cube is mainly populated by gradient vectors calculated in the same location as before the displacement and thus the feature vectors are more similar.

A significant increase in the average distance ρ is reached above approximately 10° for all sets of parameter combinations with $\delta > 1$. The tolerance to a rotational offset, which keeps the average distance ρ low for rotations of less than 10°, is introduced by the interpolation of the inserted gradient vectors to the

three neighboring bins. This 10° limit was used for the parametrization of the orientation assignment in Section 4.3. There, the distance measure $d_{\sphericalangle} < 10°$ determines the amount of allowed rotation around an arbitrary axis. This measure is equal to restraining the analyzed random rotation to $\varphi < 10°$. Thus, the allowed error in orientation assignment is parametrized so that the average distance ρ for two descriptors computed within these bound is significantly smaller than for descriptors with large rotational offsets — and for this reason also random descriptors.

The slope and the absolute levels of all curves in Figure 4.8 is smaller for orientation histograms of subdivision level 0 than for level 1. This is caused by the finer sampling of the sphere surface and the larger dimension of the feature vector when using orientation histogram of subdivision level 1. The slope of $\rho(\varphi)$ is smaller for cohorts of smaller δ for both values of g, which can be attributed to the large initial distance $\rho(0°)$. Furthermore, the slope of ρ differs within each cohort and corresponds to the number of cubes and thus the feature vector dimensionality.

In summary, the average distance ρ depends on the cube width δ and the dimensionality of the feature vector. Larger cubes make the descriptor more robust to a rigid transformation and decrease the initial average distance $\rho(0°)$. A larger dimensionality of the feature vector increases the slope of ρ. On the one hand, this increases the contrast in the distance function and identifies matches more clearly. On the other hand, this requires matches to be more exact in order to be identified as correct.

Different values for the weighting of gradient vectors σ_d yield similar results as shown in Appendix A.6.1 on page 212. A standard deviation of $\sigma_\mathrm{d} = 0.5$ diminishes the magnitude of the gradient vectors sampled in large distances from the keypoint and therefore the influence of these gradient vectors during comparison. However, it also decreases the discriminative power of the descriptor, which is indicated by the smaller slope in the average feature distance $\rho(\varphi)$. A comparison of the results using standard deviations of 1.0 and 1.5 shows only little differences. Therefore, $\sigma_\mathrm{d} = 1.0$ is used for weighting the gradient vectors that are used for descriptor calculation.

4.4.2. Parameter Preselection

Following the analysis of the descriptor's properties with respect to random rigid displacements, a *preliminary selection* of parameters is performed using a synthetic database. In this test, the capability of the computed feature vectors to identify correct matches in a larger database is assessed as explained below. This test serves to identify reasonable parameter settings that yield descriptors, which

$\delta \in \{2; 3; 4; 5\}$	$p \in \{4; 8\}$
$r \in \{\sqrt{0}; \sqrt{1}; \sqrt{2}; \sqrt{3}; \sqrt{4}\}$	$g \in \{0; 1\}$
$\sigma_\mathrm{d} \in \{0.5; 1.0; 1.5\}$	

Table 4.4 – Tested descriptor paramter sets
During the parameter preselection all combinations of the above shown parameters are assessed. This table deviates from Table 4.3 in δ and r.

are able to discriminate true from false matches in a larger pool of descriptors. The parameter settings determined in this test will be analyzed in more detail in the following section.

The database comprises map descriptions generated for protein domains contained in the training set — the *query proteins* — and supplementary *decoys*. The decoy set is assembled according to two objectives. On the one hand, no descriptor from the query protein domains is to be found in the decoy set. On the other hand, the protein domains need to share a certain degree of similarity to model a realistic classification scenario. Therefore, the decoy set comprises domains from the protein classification CATH that are taken from the same CATH architectures that were also used in the training set. The CATH IDs of the utilized protein domains in the decoy set are `1AOC:A00`, `1GK9:A01`, `1HQ0:A00`, `1LB3:A00`, `1N2M:C00`, `1RWH:A03`, `1TIG:A00`, `2AWK:A00`, `2B49:A00`, `2GTQ:A05`, `2R0X:A00`, `2V9L:A00`, `2VHK:A00`, `3D1G:A01`, and `7ODC:A02`. For all combinations of parameters shown in Table 4.4, a database consisting of query- and decoy descriptors is created. For this purpose, synthetic maps are created for all protein domains using a sampling interval of 1 Å and a resolution of 3.5 Å. This yields 4 315 and 28 806 *query* keypoints and descriptors as well as 6 743 and 45 554 *decoy* keypoints and descriptors. In total, 11 058 keypoints and 74 360 descriptors are contained in the database.

The created databases — one for each combination of parameters listed in Table 4.4 — are queried using descriptors computed from the query protein domains. The location and orientation of the query keypoints are determined on a noiseless synthetic map. To simulate the influence of the allowed translation and rotation, the descriptor coordinate frames are dislocated randomly as described in the previous test on page 120. Furthermore, white Gaussian noise is added to the map prior to descriptor computation yielding a SNR of 2. The database is queried for all descriptors found in the query proteins. The keypoint compatibility requirement is dropped in this test making the identification of the correct match more challenging because more descriptor matches are allowed. For each set of parameter combinations, the success rate, i. e., the rate of key-

points that are matched correctly, is calculated. A threshold on the absolute value of the feature vector distance $\tau_{\text{abs}} = 0.5$, the absolute criterion, is used in this test. This value was chosen by visually inspecting the plots shown in the robustness test and serves as a first estimate. For the distinctiveness criterion a threshold of $\tau_{\text{dist}} = 0.9$ is employed, which proved effective in preliminary test. Both thresholds are enforced separately, to allow for an assessment of their performance.

Figure 4.10 shows a scatter plot of the success rates for different combinations of parameters when using $\tau_{\text{abs}} = 0.5$ and $\tau_{\text{dist}} = 0.9$. The plot shows that descriptors with small cube volumes $\delta = 2$ and descriptors with only one cube $r^2 = 0$ have limited discriminative power. This is also true for $r^2 = 1$ in combination with several other parameters. Parameter setting that perform well with respect to both, the absolute and the distinctiveness criterion, have a relatively low dimensionality and cube widths of $\delta \geq 3$. For high dimensional descriptors, the distinctiveness criterion yields larger success rates than the absolute threshold criterion. The robustness test of the descriptors showed that the slope of the average distance ρ with respect to the average rotation φ increases with the dimensionality of the feature vector. Thus, high dimensional feature vectors have a larger distance when applying the same random dislocation and therefore lie more frequently above 0.5. This implies that the absolute threshold criterion is violated more frequently when using these combinations of parameters. Summarizing, the plot shows that two requirements must be met for computing descriptors that are capable of discriminating local neighborhoods: A large number of dimensions in the feature vector and a large volume. These properties are displayed as size and shape of the symbols in Figure 4.10.

The plots in Figure 4.11 and Figure 4.12 show magnifications of the area $[94\,\%; 100\,\%] \times [94\,\%; 100\,\%]$ of Figure 4.10. They allow for the discrimination of parameter sets that have high success rates with respect to both criteria. These parameter sets are also listed in Table 4.5, which additionally details parameter sets that perform well with respect to the distinctiveness criterion only. The plot in Figure 4.11 illustrates that also a low-dimensional feature vector $(r; \delta; g) = (\sqrt{1}; 5; 0)$ with 84 dimensions is capable of successfully identifying correct matches in more than 94 % of cases using either criterion. The plot in Figure 4.12 gives information on the total width of the descriptor and the number of samples necessary for calculating the descriptor using a fixed number of samples $p^3 = 64$. This number of samples is chosen since it proved as suitable in the robustness test and because larger values require more computing time. The plot also endorses the mentioned parameter set, which yields descriptors that require few gradient vectors to be calculated and have a small total width.

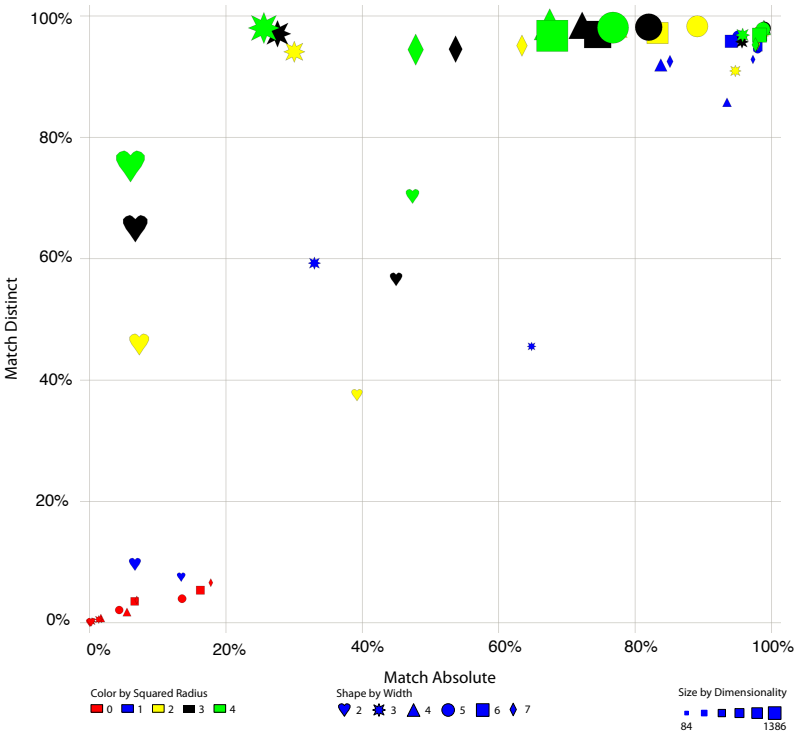


Figure 4.10 – Parameter preselection: Success rates
The plot shows the success rates determined in the parameter preselection test using the absolute threshold $\tau_{\text{abs}} = 0.5$ (abscissa) and the distinctiveness criterion $\tau_{\text{dist}} = 0.9$ (ordinate). The color of each symbol indicates the number of cubes as determined by r while the shape shows the width of each cube δ. The size of each symbol corresponds to the dimensionality of the feature vector. (© A. Griewel)

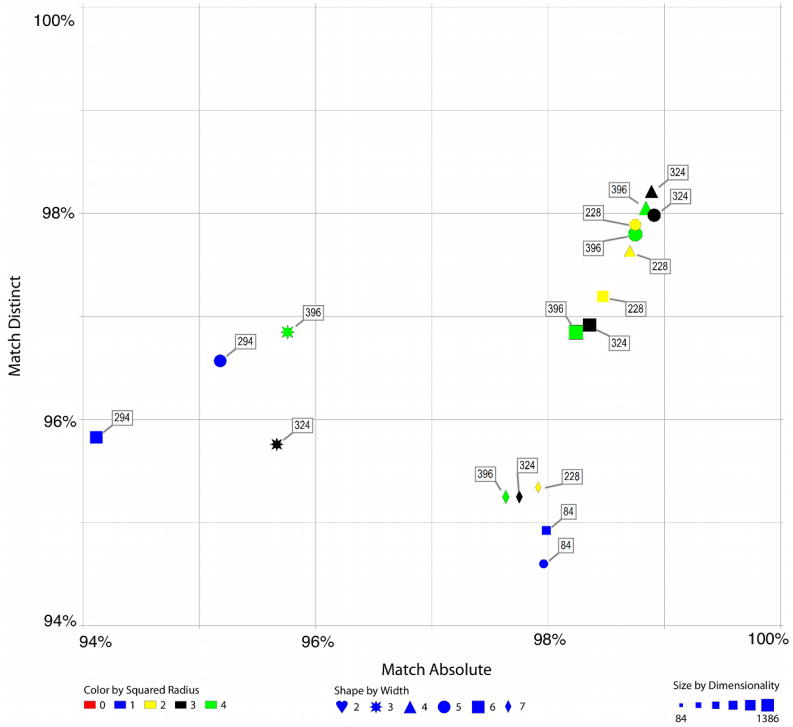


Figure 4.11 – Parameter preselection: High success rates (I)
The plot shows a magnifications of the area $[94\,\%; 100\,\%] \times [94\,\%; 100\,\%]$ of Figure 4.10. The color of each symbol indicates the number of cubes as determined by r while the shape shows the width of each cube δ. The symbol size corresponds to the dimensionality of the feature vector and the labels show the exact dimensionality. (© A. Griewel)

Considering these findings, the descriptor parametrization $(r^2; \delta; g) = (1; 5; 0)$ yields the best descriptor with respect to the specified objectives in this test, where the correct matching descriptor is to be identified in a pool of 74 360 descriptors.

Using a larger value for weighting the sampled gradient vectors, $\sigma_\mathrm{d} = 1.5$, deteriorates the success rates slightly while a smaller value $\sigma_\mathrm{d} = 0.5$ increases the success rate. As explained earlier, smaller values for σ_d make the descriptor less discriminative and therefore the increase with the feature vector distance becomes less pronounced. Using a larger number of sampled gradient vectors per cube $p = 8$ increases the success rates consistently by a small amount. Thus, a larger number of samples can be used to produce more exact results, which requires more computing time.

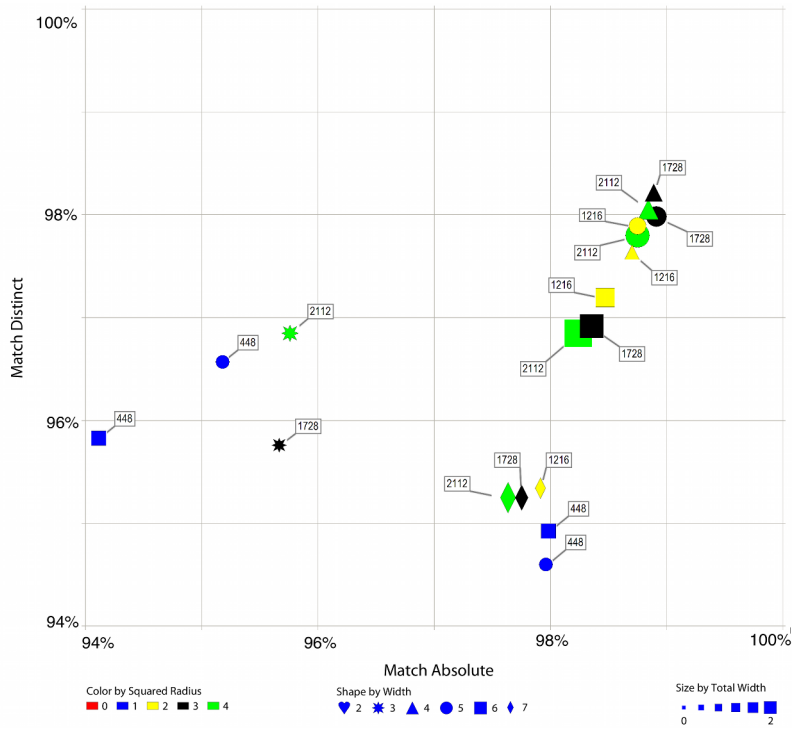


Figure 4.12 – Parameter preselection: High success rates (II)
The plot shows a magnification of the area $[94\,\%; 100\,\%] \times [94\,\%; 100\,\%]$ of Figure 4.10. The color of each symbol indicates the number of cubes as determined by r while the shape shows the width of each cube δ. The symbol size corresponds to the total width of the descriptor $r\delta$ while the labels indicate the number of sampled gradient vectors. (© A. Griewel)

g	r^2	δ	Dimensionality	Total Width	# Samples
0	1	5	84	5.0	448
0	1	6	84	6.0	448
0	2	4	228	5.66	1 216
0	2	5	228	7.07	1 216
0	2	6	228	8.49	1 216
0	2	7	228	9.9	1 216
0	3	3	324	5.2	1 728
0	3	4	324	6.93	1 728
0	3	5	324	8.66	1 728
0	3	6	324	10.39	1 728
0	3	7	324	12.12	1 728
0	4	3	396	6.0	2 112
0	4	4	396	8.0	2 112
0	4	5	396	10.0	2 112
0	4	6	396	12.0	2 112
0	4	7	396	14.0	2 112
1	1	5	294	5.0	448
1	1	6	294	6.0	448
1	2	4	798	5.66	1 216
1	2	5	798	7.07	1 216
1	2	6	798	8.49	1 216
1	3	4	1 134	6.93	1 216
1	3	5	1 134	8.66	1 216
1	3	6	1 134	10.39	1 216
1	4	4	1 386	8.0	2 112
1	4	5	1 386	10.0	2 112
1	4	6	1 386	12.0	2 112

Table 4.5 – Properties of parameter combinations with high success rates
The table shows the parameter sets, for which success rates of more than 94 % are achieved as shown in Figure 4.10. The three leftmost columns show the parameters while the three right most columns show the determined properties dimensionality, total width, and number of samples. While parameter sets above the line perform well with respect to both criteria $\tau_{\text{abs}} = 0.5$ and $\tau_{\text{dist}} = 0.9$, the settings below the line perform well only with respect to the distinctiveness criterion.

During parameter preselection, combinations for descriptor-parameters have been identified that are capable of discriminating true from false matches in a set of 74 360 descriptors. With respect to the defined objectives, the parameter set $(r; \delta; g; p; \sigma_\mathrm{d}) = (\sqrt{1}; 5; 0; 4; 1.0)$ was shown to be capable of successfully identifying descriptor matches. However, this test does not yield information on the properties of the calculated distances with respect to the chosen parametrization. This aspect is analyzed in more detail in the next section.

4.4.3. Classification Performance

The parameter sets chosen in the preselection are analyzed in more detail to determine exact values for τ_abs and τ_dist. For both, query- and decoy descriptors, the minimal distance to a descriptor in a database containing only query descriptors is determined. The resulting distributions are plotted and on this basis two parameter settings for descriptor computation are selected. The first parameter set yields a 84-dimensional descriptor, which is capable of identifying correct matches in a smaller set of descriptors. The second set yields a 1 134-dimensional descriptor and separates matches and decoys more clearly. For both parameter sets, optimal thresholds are selected based on the analysis of the diagrams.

For this analysis, a database containing keypoints and descriptors computed from the query proteins used in parameter preselection is created. This database is queried using descriptors that are detected in both query- and decoy proteins. The descriptors are computed in synthetic maps at SNR 2 in positions that are within the allowed range from the keypoints in the noiseless maps — i. e., within a 10° random rotation and a translation of maximal length σ. For each keypoint, the distance η_0 to the best matching descriptor in the database is determined. Furthermore, the smallest distance η_1 to a descriptor that originates from a different keypoint is stored. Based on these two values, the distribution of the values of τ_abs and τ_dist is analyzed using the 4 315 query keypoints with 28 806 descriptors as well as the 6 743 decoy keypoints with 45 554 descriptors. Optimal values for the thresholds τ_abs and τ_dist are determined from this sample distributions.

The distance values for all descriptors have been determined for all parameter sets listed in Table 4.5. For each parameter set, a kernel density plot [293] showing the distribution of the distance for query- and decoy descriptors is shown in Appendix A.6.2 on page 216. Additionally, the sample quantiles, means, and standard deviations for these distributions are tabulated in the appendix. First, the parameterization $(r^2; \delta; g) = (1; 5; 0)$, which was determined to be effective in the parameter preselection, is chosen for closer analysis. Second, a descrip-

tor that separates true from false matches more clearly is required for large databases, as, e. g., a database comprising all domains of CATH. Based on the computed density plots, the parameter set $(r^2; \delta; g) = (3; 6; 1)$ is chosen because it separates the distribution of the distances of query- and decoy descriptors more clearly. In the following, the former parameter set is going to be referred to as parameter set A while the latter set is called parameter set B.

The plots in Figure 4.13 show the distributions of the distances of query- and decoy descriptors for both selected parameter sets. For parameter set A, the sample mean and standard deviation of the distributions are 0.22 ± 0.04 for true matches and 0.38 ± 0.05 for decoys. These values are larger for parameter set B with 0.38 ± 0.06 and 0.69 ± 0.05. This is in accordance with the properties of the descriptor that have been determined in the robustness test in Section 4.4.1. It was found that, on the one hand, the total distance decreases with larger cube widths and, on the other hand, increases with higher dimensionality. Thus, the average distance values for parameter set B are larger since the dimensionality of this descriptor is 13.5 times the dimensionality of descriptors computed with parameter set A. For both parameter sets, the amount of overlap in the query- and decoy distance distributions is small. Therefore, an absolute threshold on the Euclidean feature vector distance yields a sensible criterion for discarding false matches. To allow for maximal detection of true positives, a value of $\tau_{\text{abs}} = 0.3$ is chosen for parameter set A and a value of $\tau_{\text{abs}} = 0.6$ for parameter set B.

The distinctiveness criterion is equally effective in separating true matches from decoy matches. The plots show distributions for the distinctiveness criterion on the right hand side. For parameter set A, the sample mean and standard deviation are 0.61 ± 0.13 and 0.96 ± 0.04 for query- and decoy matches respectively. For parameter set B, these values lie in a similar range at 0.56 ± 0.12 and 0.98 ± 0.02. From the plots, it can be deduced that decoy matches have a distinctiveness score of close to one. This means that the distance of the second best matching descriptor is of a similar magnitude as the distance of the best matching keypoint. True matches, on the other hand, have a smaller distinctiveness score. Here, the ratio of the distance of best to the second best match is large indicating that the best match is distinct. For both parameter sets, A and B, a threshold of $\tau_{\text{dist}} = 0.9$ separates the two classes well while discarding a minimal amount of true positives.

The values for the thresholds τ_{abs} and τ_{dist} have been chosen by visual inspection of the diagrams. The quality of the selected thresholds is verified by determining recall, precision, and false positive rate (FPR) for the two parameter sets [323]. These rates are defined according to the table of confusion, which

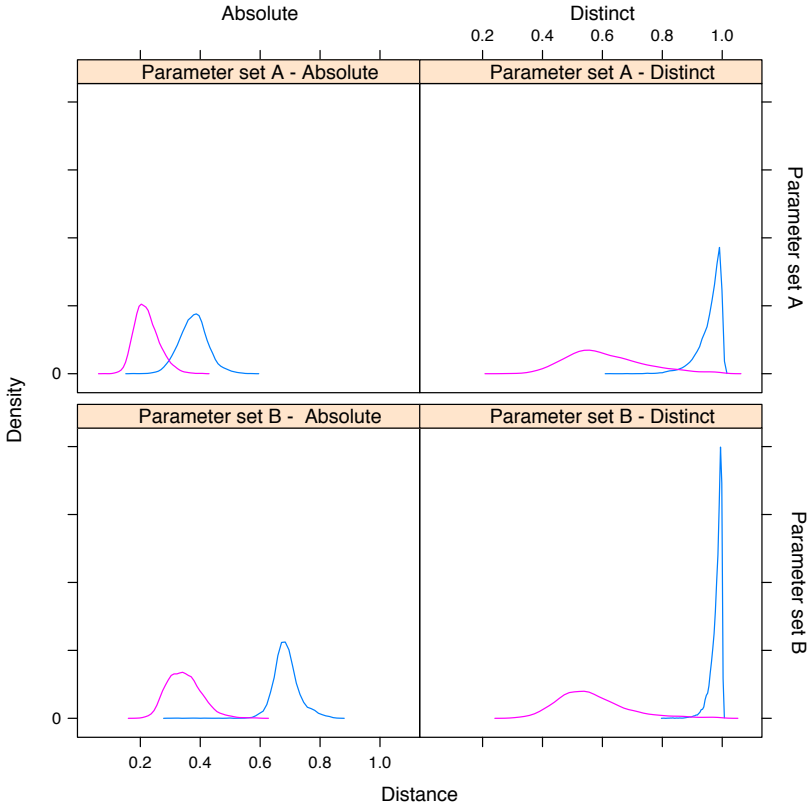


Figure 4.13 – Sample distributions of feature vector distances
The kernel density plots show the *relative* distribution of distances achieved by true hits (pink) and decoys (blue) when querying a test-database containing only the query descriptors. The left plots show the smallest Euclidean distance η_0, which is achieved by each descriptor. The right plots depict the distances calculated with the distinction criterion $\frac{\eta_0}{\eta_1}$. The utilized parameter set is indicated for each diagram with parameter set A $(r^2; \delta; g) = (1; 5; 0)$ and parameter set B $(r^2; \delta; g) = (3; 6; 1)$. (© A. Griewel)

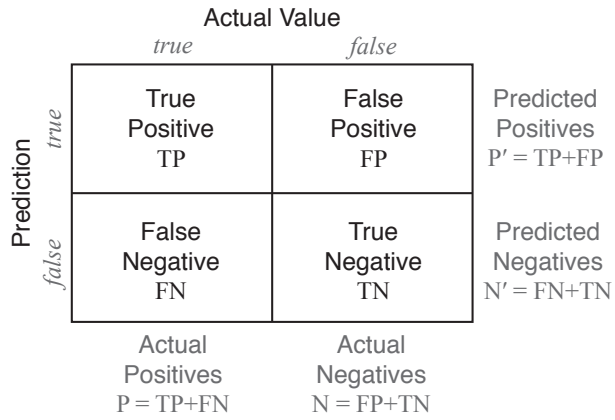


Figure 4.14 – Confusion matrix
The table of confusion classifies the outcome of a binary classification experiment. The input set is separated into actual positives P and actual negatives N. Furthermore, it is divided into predicted positives P′ and predicted negatives N′. This partitioning creates four classes: The correctly classified true positives (TP) and true negatives (TN) and the wrongly classified false negatives (FN) and false positives (FP). (© A. Griewel)

is shown in Figure 4.14. Recall — also known as true positive rate — specifies the rate of true matches that are identified successfully. Precision is the rate of correctly predicted true matches with respect to the total number of predicted matches. The FPR is determined as the ratio of false positives to the total number of negatives. According to the quantities introduced in Figure 4.14, these ratios are defined as

$$\text{Recall} := \frac{\text{TP}}{\text{P}} \qquad \text{Precision} := \frac{\text{TP}}{\text{P}'} \qquad \text{FPR} := \frac{\text{FP}}{\text{N}} \tag{4.5}$$

Precision, recall, and FPR are calculated using three different classifiers. The first classifier consists solely of the absolute criterion $\eta_0 < \tau_{\text{abs}}$ while the second classifiers uses only the distinctiveness criterion $\frac{\eta_0}{\eta_1} < \tau_{\text{dist}}$. The third classifier uses a combination of both criteria $\eta_0 < \tau_{\text{abs}} \wedge \frac{\eta_0}{\eta_1} < \tau_{\text{dist}}$. The results are listed in Table 4.6 and show that the descriptors are highly discriminative.

For parameter set A, recall and precision for the absolute criterion τ_{abs} lie above 95 % and 92 %. For the distinctiveness criterion, a recall of more than 97 % and a precision of more than 88 % is achieved. Combining the two criteria diminishes the recall to 94 % and raises the precision to 96 % while the FPR lies at 2.14 %. Thus, the combination of both criteria yields a highly discriminative scoring function. In cases where a high recall rate is most important and a high

Parameter set A $(r;\delta;g) = (1;5;0)$			
	$\tau_{\text{abs}} = 0.3$	$\tau_{\text{dist}} = 0.9$	$\tau_{\text{abs}} = 0.3 \wedge \tau_{\text{dist}} = 0.9$
Recall (%)	95.64	97.27	94.32
Precision (%)	92.39	88.23	96.54
FPR (%)	5.04	8.03	2.14

Parameter set B $(r;\delta;g) = (3;6;1)$			
	$\tau_{\text{abs}} = 0.6$	$\tau_{\text{dist}} = 0.9$	$\tau_{\text{abs}} = 0.6 \wedge \tau_{\text{dist}} = 0.9$
Recall (%)	100.00	98.49	98.49
Precision (%)	97.23	99.04	99.43
FPR (%)	1.82	0.61	0.36

Table 4.6 – Classification performance on the test set
Recall, precision and false positive rate (FPR) are listed in percent for the matching experiments. These measures were calculated for three different classifiers: the absolute criterion τ_{abs}, the distinctiveness criterion τ_{dist}, and the combination of both criteria.

precision is not the primary objective, the application of the distinctiveness criterion alone yields the best results. These are situations as, e. g., the registration to electron density maps where only a few resulting matches are reported that can be postprocessed easily. Therefore, only the distinctiveness criterion with $\tau_{\text{dist}} = 0.9$ is used for the registration task.

As already seen in Figure 4.13, descriptors determined using parameter set B separate decoys more clearly from true matches. All recall rates lie above 98 % with precisions of more than 97 % and FPRs of less than 1.9 %. Using the combination of both criteria yields the highest precision of 99.43 % while achieving a recall of 98.49 %. This indicates that descriptors computed using parameter set B are well capable of separating true matches from decoys. Thus, this parameter set is used for database searching using the combination of both criteria with $\tau_{\text{abs}} = 0.6$ and $\tau_{\text{dist}} = 0.9$.

With higher dimensionality of the feature vector, the mean of the distribution of the Euclidean distance increases as shown in Appendix A.6.2 on page 216. This explains the poor performance of high dimensional descriptors during the parameter preselection. Here, a large quantity of the correctly matched distances lies above the chosen threshold $\tau_{\text{abs}} = 0.5$ and is therefore discarded.

In a nutshell, these findings yield two parameter sets that perform optimal with respect to the specified objectives. On the one hand, the parameter set $(r; \delta; g; p; \sigma_\mathrm{d}) = (\sqrt{1}; 5; 0; 4; 1.0)$ yields a compact descriptor with low dimensionality. It proved effective in all tests and is capable of separating true from false matches using the threshold $\tau_\mathrm{dist} = 0.9$. On the other hand, the parameter set $(r; \delta; g; p; \sigma_\mathrm{d}) = (\sqrt{3}; 6; 1; 8; 1.0)$ yields a clearer separation of true and false matches at the cost of more run time, larger total volume, and higher dimensionality. However, in applications where a multitude of descriptors is compared — as in database searching — this setting is more appropriate. Here, the two thresholds $\tau_\mathrm{abs} = 0.6$ and $\tau_\mathrm{dist} = 0.9$ have been determined for separating true from false matches.

4.5. Summary

The performance of keypoint detection, orientation assignment, and descriptor matching has been evaluated separately. For this purpose, a representative training set of protein domains has been assembled. Based on synthetic maps generated from these protein domains, controlled experiments have been performed to assess the performance of each of the mentioned tasks. The influence of discretization, resolution, and additive white Gaussian noise on the computation of keypoints, orientations, and descriptors has been assessed thoroughly. From the experimental results, the parameter set listed in Table 4.7 has been determined, which fulfills the defined objectives best and yields the most repeatable results under the test conditions.

Using this parameter set, a keypoint repeatability of 74 % is achieved in a noiseless scenario. The assignment of orientations is repeatable — i. e., within a rotation of 10° around an arbitrary axis — in 90 % of the repeatedly detected keypoints using on average seven orientations per keypoint. For the descriptor, two parameter sets are determined that are capable of separating true and false matches. One of the parameter sets yields a small descriptor, which can be computed quickly. Using this descriptor, a recall of more than 97 % is achieved at a precision of 88 %. Using the larger descriptor makes the matching more expensive in terms of time and memory, but makes the discrimination clearer with a recall of 96 % and a precision of more than 99 %.

These values have been achieved using the standard PSF for modeling the imaging system. Furthermore, a representative training set of protein domains was used for generating the synthetic maps. Therefore, the parameter set is considered optimal for the purpose of similarly searching in electron density

Keypoint Detection	
Initial PSF standard deviation	$\{\sigma_0\}_{\text{vox}} = 1.1$
Number of scale-space samples per octave	$s = 7$
Contrast threshold	$t_{\text{contrast}} = 0.02$
Cornerness threshold	$t_{\text{cornerness}} = 10$

Orientation Assignment	
Sampling interval of gradient vectors	$w_{\text{samp}} = 0.5$
Standard deviation of the Gaussian weight function	$w_\sigma = 2$
Width of the truncation window	$w_{\text{width}} = 4$
Subdivision of the 3D geodesic grid	$h_{\text{g}}^{\text{3D}} = 3$
Threshold for the 3D orientation histogram	$h_{\text{t}}^{\text{3D}} = 0.8$
Number of bins in the 2D histogram	$h_{\text{g}}^{\text{2D}} = 36$
Threshold for the 2D histogram	$h_{\text{t}}^{\text{2D}} = 0.9$

Neighborhood Descriptor			
		Param. set A	Param. set B
Radius of the descriptor	r	$\sqrt{1}$	$\sqrt{3}$
Cube width	δ	5	6
Subdivision level of the orientation histograms	g	0	1
Standard deviation of the Gaussian weight function	σ_{d}	1.0	1.0
Number of sampled gradient vectors in one dimension per cube	p	4	8
Absolute similarity threshold	τ_{abs}		0.6
Distinctiveness similarity threshold	τ_{dist}	0.9	0.9

Table 4.7 – Selected parameter set
The tables show the parameter set that is selected for keypoint detection, orientations, and descriptors from electron density maps. For neighborhood descriptors, two parameter sets are listed: Parameter set A is more efficient with respect to run time, while parameter set B separates true from false matches more clearly.

maps. The overall performance of *siseek* using this parameterization is assessed in the following chapter using both synthetic and experimental maps.

5. Results and Discussion

The proposed method, called *siseek*, was engineered for the purpose of similarity searching in electron density maps of macromolecules and macromolecular assemblies. The previous two chapters introduce the constituents of the method — keypoint detection, orientation assignment, descriptor computation, registration, and molecule recognition — and elaborate on the settings of their respective optimal parameters. In the following, the performance of *siseek* is analyzed with respect to effectiveness, i. e., "How close is the registration to ground truth data", and efficiency, i. e., "How large is the runtime". The goal of this chapter is to demonstrate that *siseek* can effectively be used for registration based on selected experiments involving typical, experimental electron density maps. Also, a proof of concept for molecule recognition is shown in this chapter.

In a first step, docking scenarios are created using synthetic maps, for which ground truth information is available. These are used for assessing the performance of *siseek* at various resolutions and signal-to-noise ratios. Subsequently, atomic models are docked into experimental electron density maps that have been determined by cryo-EM and X-ray crystallography. In the following tests two experimental X-ray crystallography maps are registered. These experiments are solely based on the electron density maps and do not require the atomic detail interpretation of the maps. Eventually, the map description is used for molecule recognition, i. e., to identify the content of an electron density map using a database of reference structures.

siseek efficiently and effectively registers intermediate and high resolution macromolecular electron density maps. This is first shown on synthetic maps and then demonstrated on selected experimental X-ray crystallography and cryo-EM maps. The experiments also show that the performance of *siseek* for registering low resolution maps is limited. In a proof of concept, it is shown that *siseek* is also applicable to the problem of molecule recognition. In most experiments, the correct molecules are identified, however, due to the large demands on computing power the number of test cases is limited and thus further research in this area is required for clearly determining the capabilities and limitations of *siseek* in molecule recognition.

5.1. Synthetic Data

A test set based on complexes downloaded from the Worldwide Protein Data Bank (wwPDB) is assembled to assess *siseek*'s performance for registering macromolecular electron density maps. By creating synthetic maps from the atomic models, a known reference — also called gold standard or ground truth — is available, which can be used to analyze the quality of the placements. The synthetic maps enable a detailed inspection of the method's effectiveness and efficiency with respect to resolution lowering and additive Gaussian noise. They also facilitate a comparison to other approaches. This is neither possible for experimental cryo-EM nor for experimental X-ray crystallography maps because the atomic models that are determined from these maps are already interpretations of the experimental data.

The test set consists of 23 atomic protein complexes, which are listed in Table 5.1 and shown in Appendix A.7 on page 221. The complexes are built from 35 distinct polypeptide chains and comprise 234 subunits in total. This selection is based on a previously published test set [107], which is altered by removing incomplete models and C_α traces since these do not correspond to biologically relevant structures. Furthermore, the test set is supplemented with complexes comprising smaller proteins. The subunits of the test set complexes have chain lengths ranging from 4 amino acids in a small polypeptide to 1 045 residues in a large protein.

The complexes are disassembled and all subunits are saved separately. One experiment in this test consists of docking a single subunit to a synthetic map that is created from the atomic model of the complete complex. All subunits are rotated and translated randomly to avoid bias. This is necessary since it was shown in Section 4.2 that the keypoint detection is susceptible to discretization noise. For each complex, five maps are created with resolutions of 2.5 Å, 5 Å, 7.5 Å, 10 Å, and 12.5 Å using a voxel spacing of one fifth of the resolution. This yields a well sampled Gaussian point spread function with a standard deviation of 1.44 voxels in a 6σ window. After creating a map at the desired resolution, the signal-to-noise ratio (SNR) is adapted as described in Section 4.1. For each resolution, one noiseless map as well as maps at SNRs of 10, 5, 2, and 1 are created.

In each experiment, the position of one subunit of a complex in a map depicting the whole complex is to be determined. The accuracy of the docking is measured using the root mean square deviation (RMSD)[1] metric. Here, the distance between the ground truth placement and the closest, reported placement

[1] See Equation 3.20 on page 91.

ID	Ref.	Chain	Length	#M	Description	#Total
1A6D	[80]	A	545	8	Thermosome α Subunit	**16**
		B	543	8	Thermosome β Subunit	
1AW5	[83]	A	340	8	5-Aminolevulinate Dehydratase	**8**
1E6V	[121]	AD	553	1	Methyl-Coenzyme M Reductase I α Subunit	**6**
		BE	443	1	M.-C. M R. I β Subunit	
		CF	258	1	M.-C. M R. I γ Subunit	
1FPY	[110]	A–L	468	1	Glutamine Synthetase	**12**
1G8G	[334]	AB	511	3	Sulfate Adenylyltransferase	**6**
1GD1	[305]	OPQR	334	1	Holo-D-Glyceraldehyde-3-Phosphate Dehydrogenase	**4**
1GK8	[326]	ACEG	475	2	Ribulose-1,5 Bisphosphate Carboxylase Large Chain	**16**
		IKMO	140	2	R. B. C. Small Chain 1	
1H2I	[304]	A–K	209	1	DNA Repair Protein RAD52 Homolog	**11**
1IJG	[303]	A–L	309	1	Bacteriophage Φ29 Upper Collar Protein	**12**
1J2P	[125]	A–G	246	1	Proteasome α Subunit	**7**
1K32	[37]	A–F	1 045	1	Tricorn Protease	**6**
1KF6	[153]	AM	602	1	Fumarate Reductase Flavoprotein	**8**
		BN	243	1	F. R. Iron-Sulfur Protein	
		CO	130	1	F. R. 15 kDa Hydrophobic Protein	
		DP	119	1	F. R. 13 kDa Hydrophobic Protein	
1L1F	[310]	A–F	505	1	Glutamate Dehydrogenase 1	**6**
1MFR	[131]	A–X	176	1	M Ferritin	**24**
1N6D	[174]	A–F	1 071	1	Tricorn Protease	**12**
		G–L	4	1	RVRK	
1NIC	[1]	A	340	3	Nitrite Reductase	**3**
1PMA	[220]	AC–O	233	1	Proteasome	**28**
		12BP–Z	211	1	Proteasome	
1Q5B	[142]	ABC	880	1	EP-cadherin	**3**
1RUZ	[104]	HJL	328	1	Hemagglutinin	**6**
		IMK	160	1	Hemagglutinin	
1SX4	[56]	A–G	524	1	GroEL	**21**
		H–N	524	1	GroEL	
		O–U	97	1	GroES	
1W3A	[222]	A	315	6	Hemolytic Lectin LSLA	**6**
1XMV	[372]	A	356	6	RecA	**6**
7AHL	[311]	A–G	293	1	α–Hemolysin	**7**
					Total	**234**

Table 5.1 – Test set protein structures
The table lists the wwPDB ID (ID), the corresponding reference (Ref.), the comprised chains (Chain), the chain length (Length), the number of models (#M), a description, and the total number of subunits (#Total).

according to the distance metric is reported as the result of the experiment. All experiments without placements or with a minimal RMSD larger than 5 Å are regarded as failure. The total *accuracy* for one combination of SNR and resolution is computed as the average RMSD of all non-failed experiments. The *failure rate* is determined as the ratio of all experiments without placements or with an RMSD larger than 5 Å with respect to the total number of experiments, which is 234.

siseek, with the parameters listed in Table 4.7, is used to determine the placements of the subunits in maps of their corresponding complexes. Each docking commences by reading the map of the complex and the atomic coordinates of the rotated atomic structure, for which a synthetic map is created. Subsequently, map descriptions are calculated for both maps. Then, the map descriptions are matched and placements are saved. All placements are clustered using a threshold of four times the sampling interval of the base map to remove duplicates. The position of each remaining representative is optimized using the atom interpolation scoring function.

The performance of the method is compared to the docking software ADP_EM[1] [107], which was developed for docking atomic structures to cryo-EM maps. This software is chosen since it is reported to implement one of the most efficient methods available. Furthermore, it was thoroughly evaluated on a larger test set, is fully automated, and reported to be the most reliable tool for this docking task [107]. In its default configuration, ADP_EM employs the Laplacian filter to boost the robustness against resolution lowering. ADP_EM creates spherical harmonics representations of concentric spherical layers surrounding every other voxel for both the atomic model and the map. This representation is used to determine placements by exhaustively comparing all computed representations. Two experiments — with and without the Laplacian pre-filtering — are performed using the default parameters of ADP_EM. This yields up to fifty placements for each experiment out of which the minimal RMSD to the ground truth placement is determined and reported as a result.

The accuracy and failure rates for both programs and all combinations of resolution and SNR are shown in Figure 5.1. Parameter combinations, in which more than 30 % of the docking experiments fail, are not shown in the accuracy plot.

Overall, *siseek* works well on high resolution maps, while the performance drops, as expected, when lowering the SNR or the resolution. The performance analysis of *siseek* at different levels of resolution shows:

[1] See Section 2.3 on page 55.

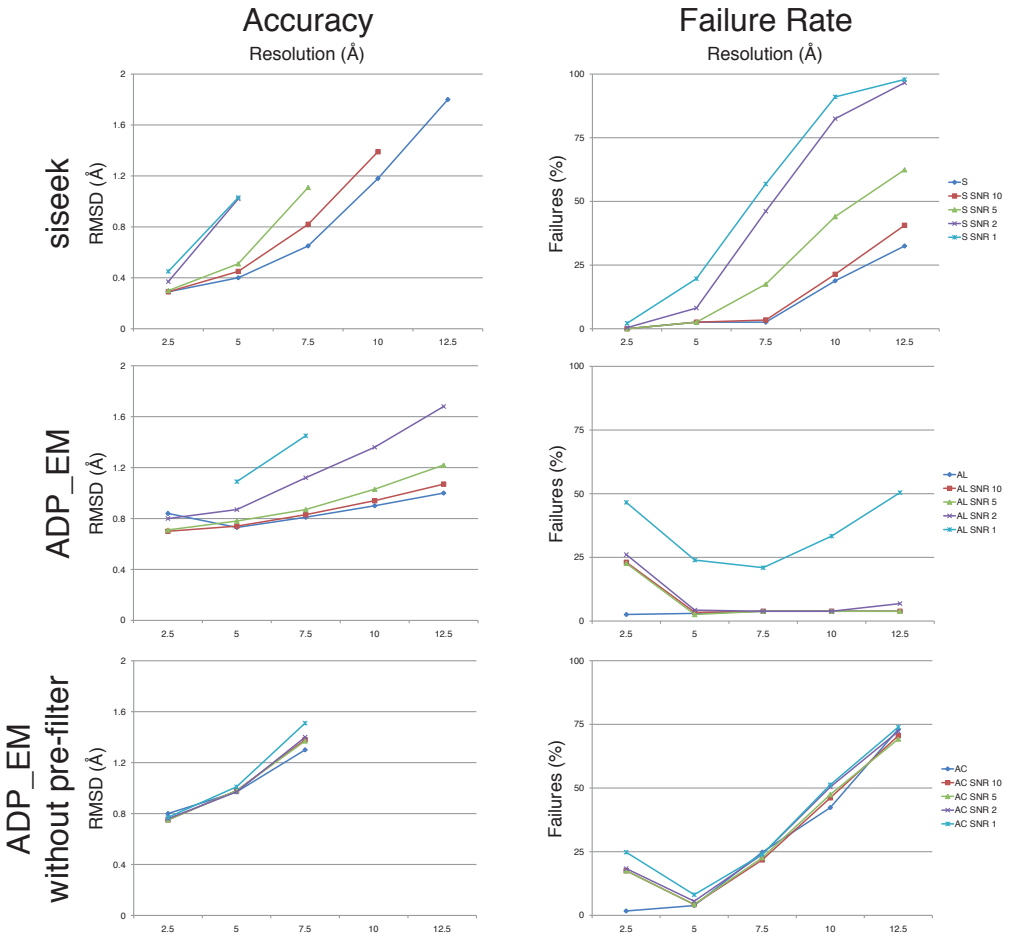


Figure 5.1 – **Test results**
The plots show the accuracy and failure rates for *siseek* (S), ADP_EM with Laplacian pre-filtering (AL), and ADP_EM without using the Laplacian pre-filter. Dockings with no placement or an RMSD of more than 5 Å are considered as failure and are reported as failure ratio in the plots on the right hand side. Dots in the accuracy plot are only shown for combinations of SNR and resolution with a failure rate of less than 30 %. (© A. Griewel)

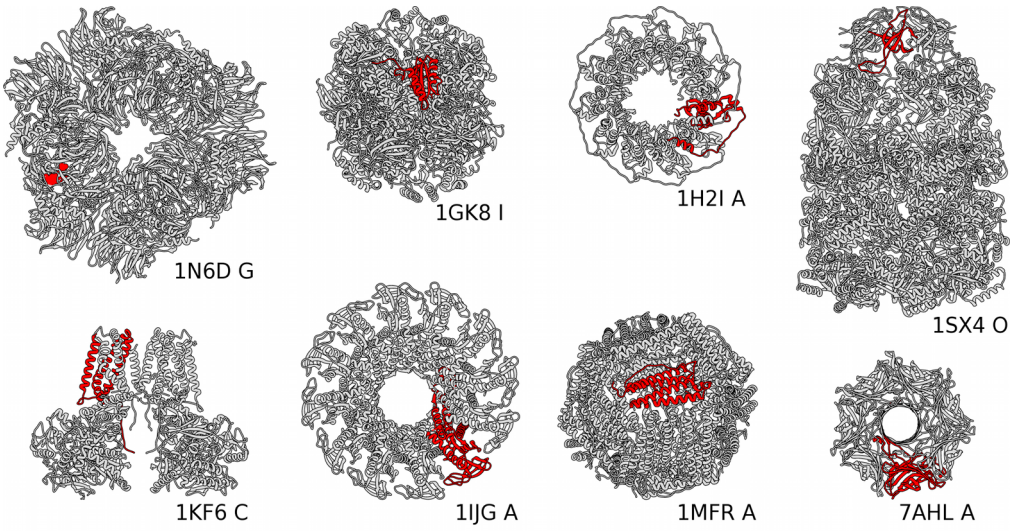


Figure 5.2 – Small or unstructured proteins
Examples of small or unstructured proteins in the test set include `[1N6D:G-L]`, `[1GK8:IKMO]`, `[1H2I:A-K]`, `[1SX4:O-U]`, `[1KF6:CO]`, `[1IJG:A-L]`, `[1MFR:A-X]`, and `[7AHL:A-G]`. One subunit of each protein is colored red in the otherwise gray-colored complexes. The polypeptide `[1N6D:G]` consists of four residues only and is shown here as sphere model for clearer view. (© A. Griewel)

2.5 Å The correct placement for all subunits is identified even at low SNRs. The average RMSD of the best placements is less than 0.5 Å and therefore smaller than the sampling interval of the map.

5.0 Å The accuracy of the placements remains high with an average RMSD of less than 0.5 Å for maps with SNR ≤ 5. Six failures are reported for these cases that correspond to the chains G–L from `1N6D`, which comprise solely four amino acids as shown in Figure 5.2. For SNR ≤ 2, the docking of subunits of `1H2I`, `[1GK8:IKMO]`, and `1MFR` fails. These are either small or non-globular subunits, which cannot easily be discerned from their neighboring subunits.

7.5 Å For SNR ≤ 10 the docking accuracy remains high and no further dockings fail. At SNR 5, the previously mentioned subunits are partially docked incorrectly. For lower SNRs, more than 38 % failures are reported. The failed experiments include subunits from `[1SX4:O-U]`, `[1KF6:CO]`, and `1IJG` besides the already mentioned subunits, which are all shown in Figure 5.2. These subunits have little internal structure, are very elongated or form

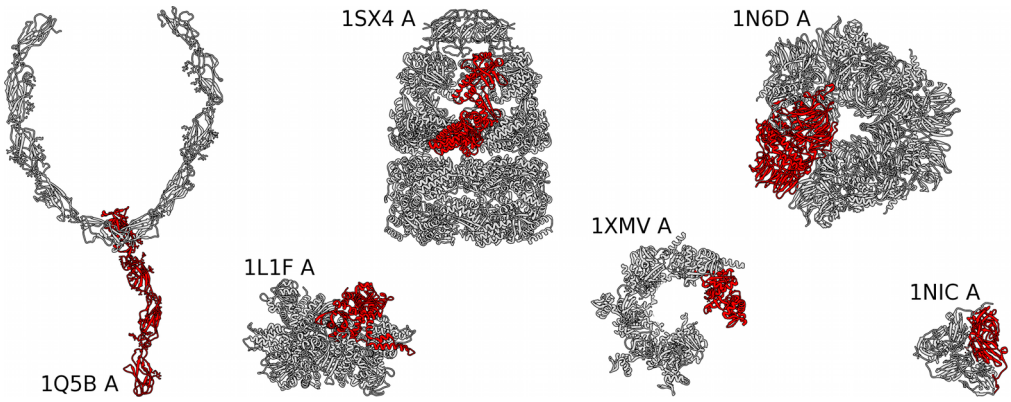


Figure 5.3 – Large or discernible proteins
Examples of large or discernible molecules in the test set include `[1Q5B:A-C]`, `[1L1F:A-F]`, `[1SX4:A-G]`, `[1XMV:A]`, `[1N6D:A-F]`, and `[1NIC:A]`. Instances of these proteins are shown in red while the remainder of the complexes is colored gray. (© A. Griewel)

close connections to their neighboring subunits. Thus, they are not easily discerned from their surrounding in the complex and therefore the docking fails.

10 Å For SNR ≤ 2, the docking fails in more than 80 % of the cases. Only for very pronounced subunits with high amounts of structural detail, placements are identified correctly. Examples of these subunits are shown in Figure 5.3. At SNR 5, 44 % of the docking experiments fail, which corresponds to the subunits mentioned previously plus subunits from `7AHL`, `1NIC`, `1AW5`, `1RUZ`, and `1PMA`. For SNR ≤ 10, the docking is successful except for single subunits of `1H2I`, `[1GK8:IKMO]`, `1MFR`, and `1IJG`.

12.5 Å More than 95 % of the docking experiments fail at SNRs 1 and 2. For SNR $= 5$ more than 60 % of the docking experiments fail. For the scenarios with high SNRs, failure rates of 32 % and 42 % are measured. These failures originate from the already mentioned subunits with little internal structure.

The observations demonstrate that the method is able to successfully identify the position of the subunits in high-quality maps depicting molecular complexes. The success of the method depends, on the one hand, on the resolution and SNR of the map, which must allow for the identification of the subunit. On the other hand, the size and shape of the subunit as well as its mode of interaction with neighboring subunits are decisive. While the small polypeptide `[1N6D:G-L]` is

only detected at a resolution of 2.5 Å, prolonged structures such as `1Q5B` are successfully docked even at 12.5 Å resolution. Structures that are less globular and comprise little secondary structure content are not identified at lower resolutions. In the test set, these structures are intertwined with their neighboring subunits. This causes a superposition of their densities at lower resolutions and therefore prevents the correct localization of keypoints.

The influence of the added white Gaussian noise on the docking results is stronger at lower resolutions. This can be explained by the sampling interval of the base map, which depends on the resolution. For high resolution maps, the number of created octaves is larger than for low resolution maps since the sampling interval in the base map is smaller. This allows for the attenuation of noise in higher octaves. If the resolution of the map is low, the effect of noise is larger because less octaves are created and therefore the noise is not attenuated as well.

As a comparison, the software ADP_EM [107] is used for the docking of atomic subunits to maps of their complexes. At first, the default configuration using a preprocessing of the input map by the Laplacian filter is analyzed. The resulting accuracies and failure rates are shown in Figure 5.1. ADP_EM with the Laplacian preprocessing yields correct placements for almost all scenarios. Even at 12.5 Å resolution with SNR 2 correct placements are identified for more than 90 % of the test cases. For noisy maps with 2.5 Å resolution, the docking fails in more than 20 % of the cases. These failures are partially caused by an abnormal termination of the program. However, they also originate from the sampling of rotational space: In high resolution maps the angular space needs to be sampled finely to identify matches. The default configuration of ADP_EM uses a coarse sampling of angular space, which does not suffice for identifying correct placements in all maps. Furthermore, the docking into maps at SNR 1 fails for all resolutions in more than 20 %, which can be attributed to the low SNR and the properties of the Laplacian, which amplifies noise.

Using ADP_EM without the Laplacian filter increases the failure rate for low resolutions significantly. The failure rate is only lower for SNR 1 at resolutions better than 7.5 Å. This can be explained by the amplification of noise, which is induced by the Laplacian. Besides this fact, the failure rate rises proportionally to the resolution of the map for all SNRs. This shows, that the Laplacian filter is essential for successfully docking subunits to low resolution maps when using correlative methods. Applying the Laplacian in *siseek*, however, does not improve the performance, but diminishes the repeatability rate of keypoint detection.

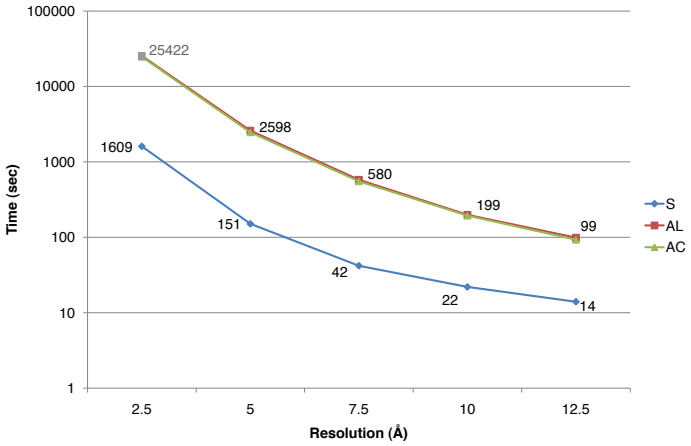


Figure 5.4 – Test set run time
The average run time of the programs is plotted against the resolution using a logarithmic ordinate. The timings for *siseek* (S) as well as ADP_EM with Laplacian pre-filtering (AL) and without pre-filtering (AC) are shown. The icons for ADP_EM at 2.5 Å are grayed out to indicate that the program aborted some of the experiments with an abnormal termination and that this number is based only on the data from the successful runs. (© A. Griewel)

A comparison of the average run times of ADP_EM and *siseek* is shown in Figure 5.4. The plot demonstrates that *siseek* requires less run time especially for high resolution maps and is on average an order of magnitude faster than ADP_EM for high resolution maps: For 2.5 Å resolution, *siseek* is 15.8 times faster than ADP_EM. However, ADP_EM failed in several experiments at 2.5 Å resolution due to an abnormal termination of the program. Considering all experimental scenarios, the ratio for 2.5 Å is probably higher since the ratio for 5 Å lies at 17.21. For lower resolutions, the ratio of the run times is falling to 13.81 at 7.5 Å, 9.05 at 10 Å, and 7.07 at 12.5 Å.

The run time for ADP_EM is higher because it creates and compares descriptors for every other voxel of the map. In *siseek*, the sampling interval of the map is adapted to the resolution and therefore a representation is chosen that is adequate for the content of the signal. Furthermore, the methods used in *siseek* depend on the depicted object and do not perform an exhaustive search for comparing the maps. Thus, the run time does not increase with the number of voxels, but rather with the information content of the map. Therefore, the run time of *siseek* is considerably lower than the run time of ADP_EM.

Considering all scenarios, ADP_EM is more robust to resolution lowering and noise than *siseek*. However, maps from X-ray crystallography generally have high resolutions and recently also maps from cryo-EM allowing for atomic detail interpretation have been acquired [219, 377, 362] . For very low resolution maps, as frequently acquired by cryo-EM, the docking accuracy of ADP_EM remains higher, which is facilitated by applying the Laplacian filter to the map prior to docking. This results in a bandpass filtered map, in which edges — i. e., transitions from the interior of the protein to the solvent — are marked. Using this preprocessing, the contrast [53] in the map is increased, which is beneficial for most docking scenarios. ADP_EM performs an exhaustive search using probes on a regular grid and therefore no keypoints are employed. This results in increased run times. However, the identification of keypoints is the major source of error in *siseek* and thus is one cause of its higher failure rates. This problem can be addressed by a closer investigation of the keypoint detection method, the assessment of alternative approaches, or by also computing descriptors for points on a regular grid inside the map.

Summarizing, these findings shows that *siseek* is able to successfully locate proteins of different sizes and shapes if the map comprises sufficient detail a priori. For intermediately sized proteins, placements are identified for resolutions as low as 7.5 Å, even if noise is present. *siseek* makes use of the given resolution of the map and is therefore not dependent on the sampling interval of the genuine map. Thus, it is possible to dock very small molecules to maps if the resolution and sampling of the signal are sufficient. Furthermore, the run time depends on the content of the signal rather than on the number of voxels in the map. If there is only little variation in the signal, few keypoints are identified and the run time of the program is small.

5.2. Map Registration

In this section, *siseek* is used to register experimental maps acquired by cryo-EM and X-ray crystallography to demonstrate the applicability of the method to experimental data. In the first section, atomic structures are registered to high resolution cryo-EM maps, which is a frequent task carried out for the interpretation of electron density maps. Here, the recently published structures of chaperonins [219, 377] and viral capsid proteins [379, 362] are used as docking targets since the resolution of these maps are sufficiently high. Subsequently, atomic structures are registered to their corresponding X-ray crystallography maps. These experiments show that *siseek* is able to identify even only partially depicted proteins in X-ray crystallography maps. In the remaining two

registration experiments, two experimental X-ray crystallography maps are registered. By superposing two experimental maps, similarities and differences of the depicted objects can be identified. This is the case if only certain parts of the depicted molecules are equal in configuration, or if the conformations of the proteins are only similar for certain parts of the molecule.

The utilized experimental maps are acquired through the internet. Cryo-EM maps are downloaded from the EMDataBank [196] and are identified by the letters EMD followed by a four digit number such as `EMD-5001`. Atomic structures are provided by the wwPDB [23] and are assigned a four letter identifier — a number followed by three alphanumerical characters. Using this identifier, it is possible to download X-ray crystallography electron density maps for many atomic structures from the Electron Density Server [178]. Maps that are not sampled on a cubic grid are converted to a cubic grid using Situs [365]. In the following text, the key properties of the maps are summarized. A list of all attributes of the utilized maps including experimental method, resolution, dimension, sampling interval, the number of detected keypoints, and the number of computed descriptors is found in Table 5.4 on page 171. All performance measurements are performed on one core of a computer equipped with a 2.67 GHz Intel Core i5 CPU with 8 192 KiB cache and 8 GiB main memory.

5.2.1. Cryo-Electron Microscopy

Recently, high resolution cryo-EM maps with atomic- and sub-nanometer resolution have been published [385]. The acquirement of these maps was possible due to the high symmetry of the depicted objects as well as the tremendous advances made in the field of cryo-EM, and it is expected that even more high resolution maps will be acquired in the future [342, 158, 95, 64]. All maps presented here were calculated using the single particle reconstruction technique and the specified resolutions have been determined using the Fourier shell correlation coefficient (FSC) with a cutoff at 0.5. The docking is performed using the parameters listed in Table 4.7. Subsequent to the matching procedure, placements are clustered and optimized using the atom interpolation score as described in Section 3.5 on page 93.

5.2.1.1. GroEL

While the structure of many synthesized proteins is encoded in their sequence, some proteins require assistance in folding. Chaperonins aid this process by enclosing newly synthesized or misfolded proteins in a cavity. Thereby, the proteins are shielded from other proteins in the solution that might interfere

with the process of folding. Upon release from the cavity into the surrounding solution, the protein is folded correctly or, if the fold is still incorrect, it might bind again to a chaperonin. During this process chaperonins undergo large conformational changes and consume energy. [139]

Chaperonins are found in all domains of life [361] and can be subdivided into two groups [129] based on the mechanism of encapsulation of the substrate: Group I chaperonins require a separate lid for sealing the cavity while group II chaperonins posses a built-in lid. The bacterial chaperonins — also found in the endosymbiotic organelles mitochondria and plastids — belong to group I and are well studied using the model system GroEL/GroES. The archeal chaperonin [80], also known as thermosome, and the eukaryotic chaperonin [63], also known as TRiC or CCT, belong to group II. All chaperonins consist of a multimeric double ring. Each ring is formed by 7–9 monomers, each having a molecular weight of approximately 60 kDa [190].

The bacterial chaperonin GroEL from E. coli aids protein folding together with its co-chaperonin GroES. Each of the fourteen subunits of GroEL has a molecular weight of approximately 53 kDa and consists of three domains. The equatorial domain is located at the center of the complex and links the two rings, which comprise seven subunits each. The intermediate domain connects the equatorial to the apical domain, which in turn interacts with GroES and undergoes considerable rearrangements upon substrate binding [87].

The double-ring complex of GroEL, consisting of fourteen subunits, was depicted at 4.2 Å using cryo-EM [219]. The acquired map is available through the EMDataBank with accession code `EMD-5001`. A C_α trace was calculated from the map and deposited as wwPDB structure `3CAU`. This trace is not suitable for docking using *siseek* since C_α traces merely represent the overall shape of a protein. Therefore, one subunit from an atomic model of GroEL acquired by X-ray crystallography and deposited as `1XCK` [13] in the wwPDB is docked to the experimental map.

The experimental cryo-EM map consists of 8×10^6 voxels at a sampling interval of 1.06 Å. For the atomic model of the GroEL monomer, a synthetic map consisting of 0.5×10^6 voxels is created automatically at the same voxel spacing using the resolution of the experimental map. *siseek* computes 997 keypoints and 7 012 descriptors for the synthetic map and 12 764 keypoints and 65 499 descriptors for the experimental map. The keypoint detection for both maps is accomplished in 8 min, the subsequent comparison of descriptors in 2 min, and the post-optimization in 17 s. The complete registration of the monomer to the experimental electron density map is therefore performed in 10 min. All 14 subunits in the experimental map are identified successfully and the resulting

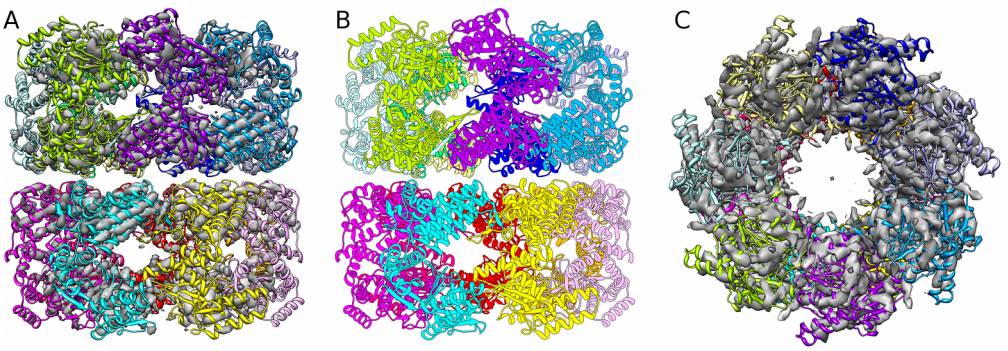


Figure 5.5 – Docking: GroEL
The assembly generated by docking subunits to an experimental cryo-EM map of GroEL at 4.2 Å resolution is shown. In A, a side view of the complex superposed to the electron density is depicted while C shows a top view. Panel B does not include the electron density and allows for a clearer view of the assembly. (© A. Griewel)

assembly is shown in Figure 5.5. The average all-atom RMSD[1] to an optimally fitted model of `1XCK` is 0.72 Å. Thus, the calculated placements resemble the reference structure almost perfectly with only minimal deviations.

5.2.1.2. Methanococcus Maripaludis Chaperonin

This group II chaperonin is found in the cytosol of the archaeon Methanococcus maripaludis (Mm–cpn) and is composed of sixteen identical subunits. Each subunit has an approximate molecular weight of 58 kDa and consists of an equatorial, intermediate, and apical domain. In group II chaperonins the apical domain also comprises a protruding lid-segment, which is used to seal the cavity that is formed to aid protein folding. [129]

The mechanism of chamber closure of Mm–cpn has been studied using a wild-type chaperonin and a "lidless" mutant [377]. The mutant is able to bind unfolded polypeptides and hydrolyze adenosine triphosphate, but it lacks the ability to aid folding of a stringent substrate. Two atomic models of these complexes have been determined by cryo-EM single particle reconstruction. These maps depict the wild-type complex in the closed conformation and the lidless Mm–cpn in the opened conformation.

A map depicting the closed state of wild-type Mm–cpn at 4.3 Å resolution is accessible under `EMD-5137`. An atomic model was determined from this map

[1] See Equation 3.20 on page 91.

and deposited as wwPDB structure 3LOS. Docking one subunit to the map and thereby assembling this complex takes 16 min. The experimental map consists of 7.1×10^6 voxels with a sampling interval of 1.33 Å. In this map, 20 963 keypoints with 166 226 descriptors are computed. The synthetic map, which is created for a monomer from 3LOS, comprises 0.3×10^6 voxels at the same sampling interval. For this map, 899 keypoints with 5 637 descriptors are computed. Keypoint detection and descriptor computation is accomplished in 13 min. The subsequent docking and post-optimization are accomplished in less than three minutes. The resulting placements have an average RMSD to the deposited atomic model of 2.79 Å and resemble the complex closely, as shown in Figure 5.6.

The opened state of lidless Mm–cpn is depicted in map EMD-5140 at 8 Å resolution and supplemented with an atomic model deposited in the wwPDB with ID 3IYF. The experimental map comprises 13.8×10^6 voxels while the synthetic map consists of 0.4×10^6 voxels. Both maps have a sampling interval of 1.33 Å. Since the resolution is considerably lower for this experimental map, less keypoints are detected. The experimental map is described by 2 439 keypoints and 16 987 descriptors, while for the synthetic map of one subunit of 3LOS only 149 keypoints with 797 descriptors are computed. The registration of the synthetic map of the monomer to the complex is performed in less than 4 min. The resulting placements are shown in Figure 5.7. They resemble the reference complex closely with an average RMSD of 0.72 Å so that the complex is clearly assembled correctly.

5.2.1.3. Rotavirus Particle 6

Rotavirus belongs to the family of Reoviridae and is the cause of severe diarrhea among infants and young children. Viral particles are up to 770 Å in diameter and contain — among other molecules — the viral genome, which consists of 11 molecules of double-stranded RNA. The virus capsid is non-enveloped and consists of three layers, which have icosahedral symmetry. The outer layer is lost during cell entry and a double-layered particle remains. The outer layer of this double-layered capsid is mainly formed by 780 copies of viral particle 6 (VP6), which is highly antigenic. [261]

The atomic structure of VP6 has been determined by X-ray crystallography at 1.95 Å resolution (1QHD) [227]. Three molecules of VP6 arrange in a tower-like trimer, which forms the capsid and can be identified in low-resolution single particle reconstructions of the virion. A monomer is approximately 95 Å long, has a molecular weight of 47 kDa and can be subdivided into two domains: The proximal B-domain, which comprises mainly α–helices, and the distal H-domain, which consists of a β-barrel [36].

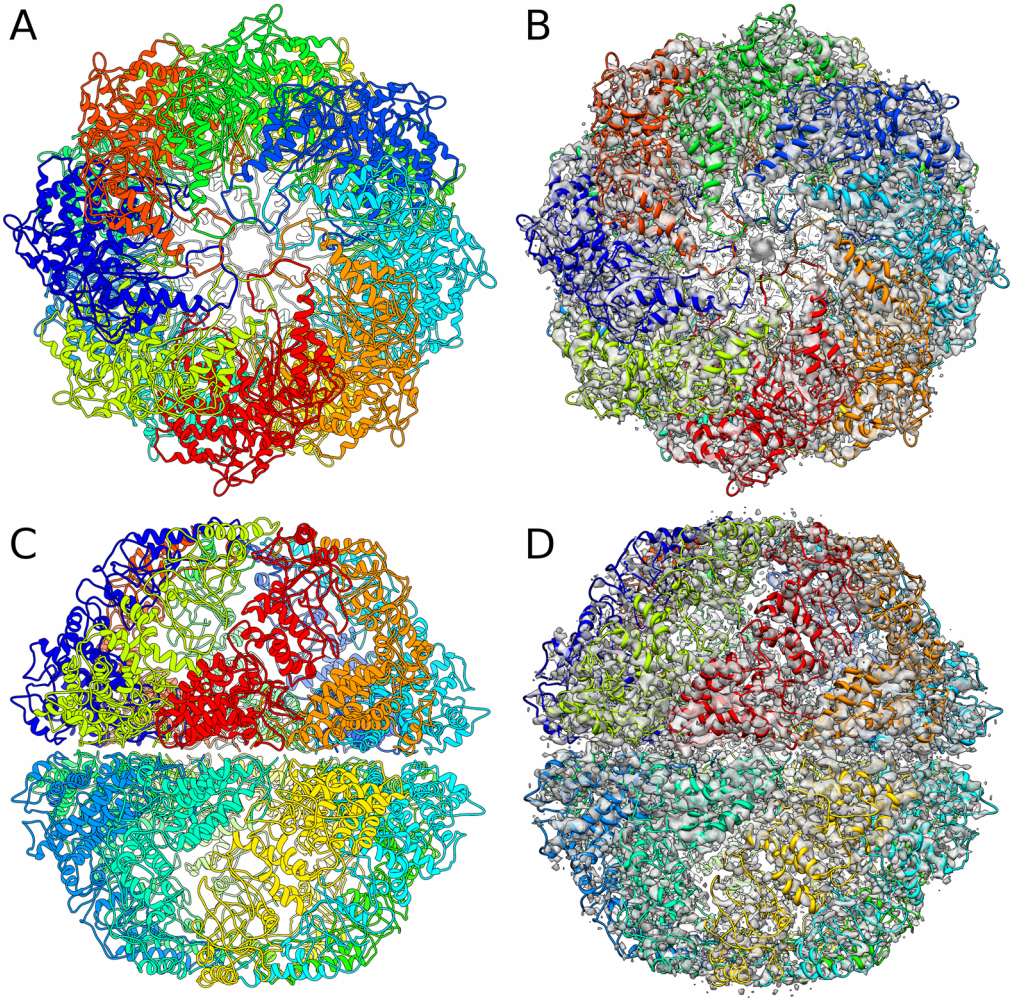


Figure 5.6 – Docking: Methanococcus maripaludis chaperonin closed conformation
An assembly of Methanococcus maripaludis chaperonin in the closed conformation is shown. The complex was generated by docking subunits to an experimental 4.3 Å resolution cryo-EM map. Panels A and B show a top view while panels C and D depict a side view of the chaperonin. In A and C no electron density is included to allow for a clearer view of the complex. Panels B and D also contain the electron density and show the high level of detail in the map and also the considerable amount of noise. (© A. Griewel)

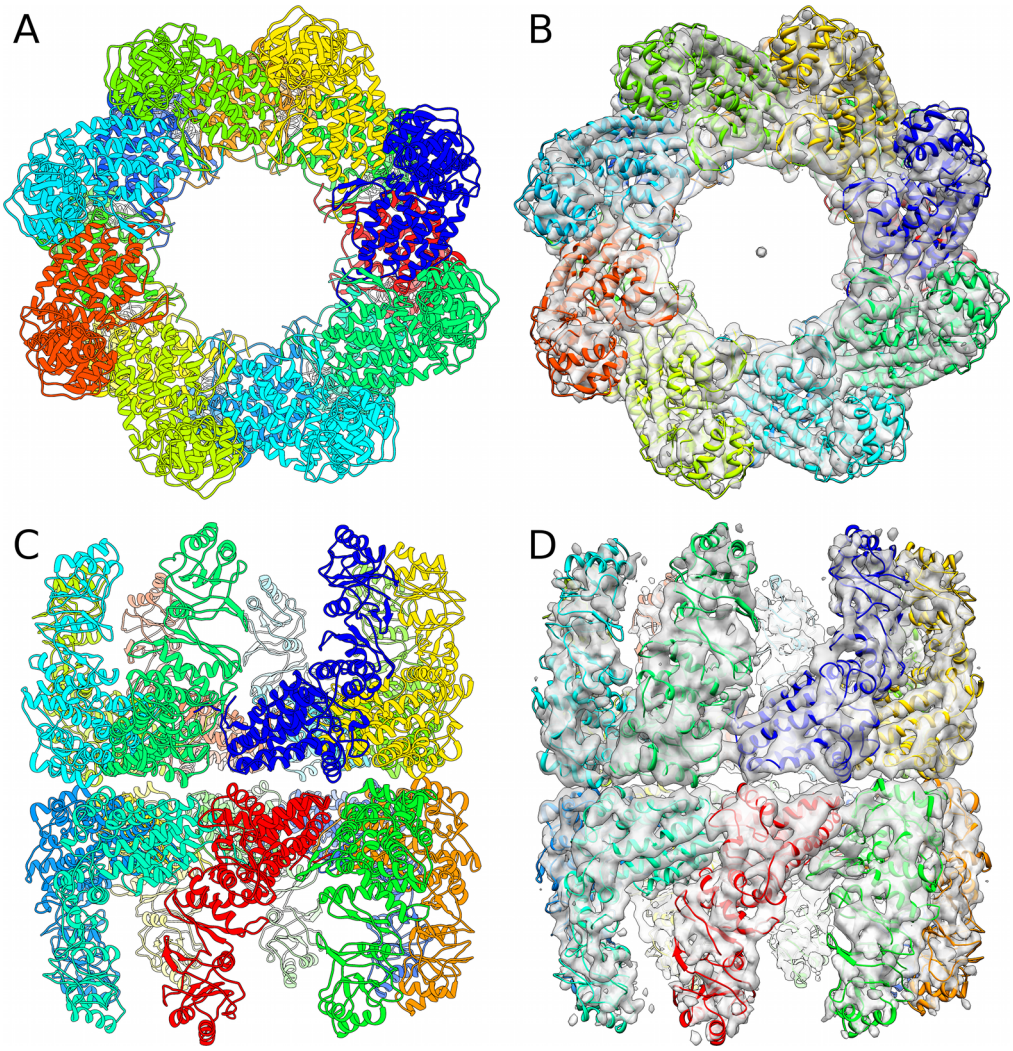


Figure 5.7 – Docking: Methanococcus maripaludis chaperonin opened conformation
An assembly of a Methanococcus maripaludis chaperonin mutant is shown, which was created by docking subunits to an experimental 8 Å resolution cryo-EM map. Panels A and B display top views of the complex while C and D depict a side view. The electron density depicted in B and D shows that α–helices are clearly discernible in the map. (© A. Griewel)

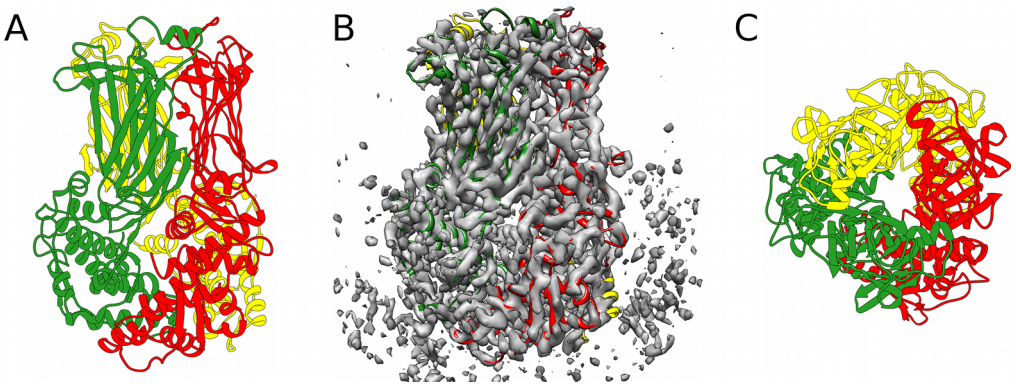


Figure 5.8 – Docking: Rotavirus particle 6
The assembly of a trimer of rotavirus particle 6 was generated by docking an atomic model of one monomer to the experimental cryo-EM map. Panel A shows a side view and C shows a top view of the trimer — both without superposed density. The electron density displayed in B demonstrates that the backbone of the map and the docked monomers are superposed well. (© A. Griewel)

A high resolution structure of VP6 was recently acquired by cryo-EM single particle reconstruction [379]. The experimental map `EMD-1461` consists of 1.7×10^6 voxels at a sampling interval of 1.23 Å with a resolution of 3.8 Å. For the docking, a synthetic map of a monomer from the atomic structure `1QHD` is created. This map has 0.3×10^6 voxels and uses the same resolution and voxel spacing. For the experimental map 10 291 keypoints and 70 751 descriptors and for the synthetic map 910 keypoints and 6 134 descriptors are computed. The docking is performed in less than 7.5 min using 5 min for computing keypoints and descriptors and the remaining 2.5 min for matching and optimization. The resulting placements are shown in Figure 5.8 with an average RMSD of 1.26 Å to an optimally fitted model of `1QHD`, which is larger than in the previous experiments. However, the placements fit well into the electron density as shown in Figure 5.8 B and perfectly resemble VP6 as shown in panels A and B of Figure 5.8.

5.2.1.4. Papillomavirus Structural Protein L1

Bovine papillomavirus type 1 causes tumors in cattle and is used as a model system for studying the properties of other members of the family of Papillomaviridae in molecular biology. The virions are non-enveloped and comprise an icosahedrally symmetric capsid, which is 550–600 Å in diameter and contains the circular, double-stranded DNA genome. [50]

The capsid of bovine papillomavirus type 1 is formed by the structural proteins L1 and L2, which are encoded in the viral DNA [44]. The capsid comprises 72 L1 pentamers, which have a molecular weight of approximately 250 kDa. The molecular structure of L1 in the capsid was studied using cryo-EM single particle reconstruction and yielded high resolution maps, in which the trace of the L1 chain is clearly depicted. The map shows the β-jelly roll fold [36], which forms the core of the protein L1 and revealed that the protein's N- and C-terminal segments mediate nearly all inter-pentamer contacts. These segments penetrate neighboring pentamers and intricately participate in their folded structure. This interaction is further stabilized by ionic interactions and disulfide bonds which collectively facilitate the structural integrity of the assembly of the viral capsid [362].

Two maps were acquired in a study of bovine papillomavirus type 1 [362]. In map `EMD-5155` icosahedral averaging was applied and the map depicts seven pentamers of the capsid as discussed below. For map `EMD-5156`, additionally non-icosahedral averaging of the pentamers was used, which allowed for the creation of a high resolution map depicting one pentamer. The map consists of 1.7×10^6 voxels with a sampling interval of 1.24 Å, has a resolution of 4.2 Å and shows features equal to an X-ray density map of about 3.5 Å resolution. An atomic model was determined from this map and deposited in the wwPDB with ID `3IYJ`. The synthetic map generated from one subunit of this structure comprises 0.4×10^6 voxels at the same sampling interval and resolution. For the experimental map 7 930 keypoints and 62 141 descriptors and for the synthetic map 1 132 keypoints with 7 744 descriptors are computed. The run time of the docking totaled less than 7 min, where 5 min are used for keypoint and descriptor computation and the remaining 2 min for matching and post-optimization. The resulting assembly is shown in Figure 5.9 and has an average RMSD of 0.28 Å to the deposited map. This shows that the computed placements have a minimum deviation to the reference structure and that the assembly was computed correctly.

The icosahedrally averaged map `EMD-5155` shows seven pentamers and has a resolution of 4.9 Å. The experimental map is very large and consists of more than 22.2×10^6 voxels at a sampling interval of 1.24 Å. For this map 93 735 keypoints and 732 659 descriptors are computed. For the atomic structure, a synthetic map with 0.6×10^6 voxels is generated, for which 909 keypoints and 6 481 descriptors are computed. The docking is performed in 88 min using 49 min for keypoint detection and descriptor computation, 20 min for matching, and 19 min for the final optimization. As shown in Figure 5.10 and Figure 5.11, the icosahedrally symmetric segment of the capsid with seven pentamers is correctly

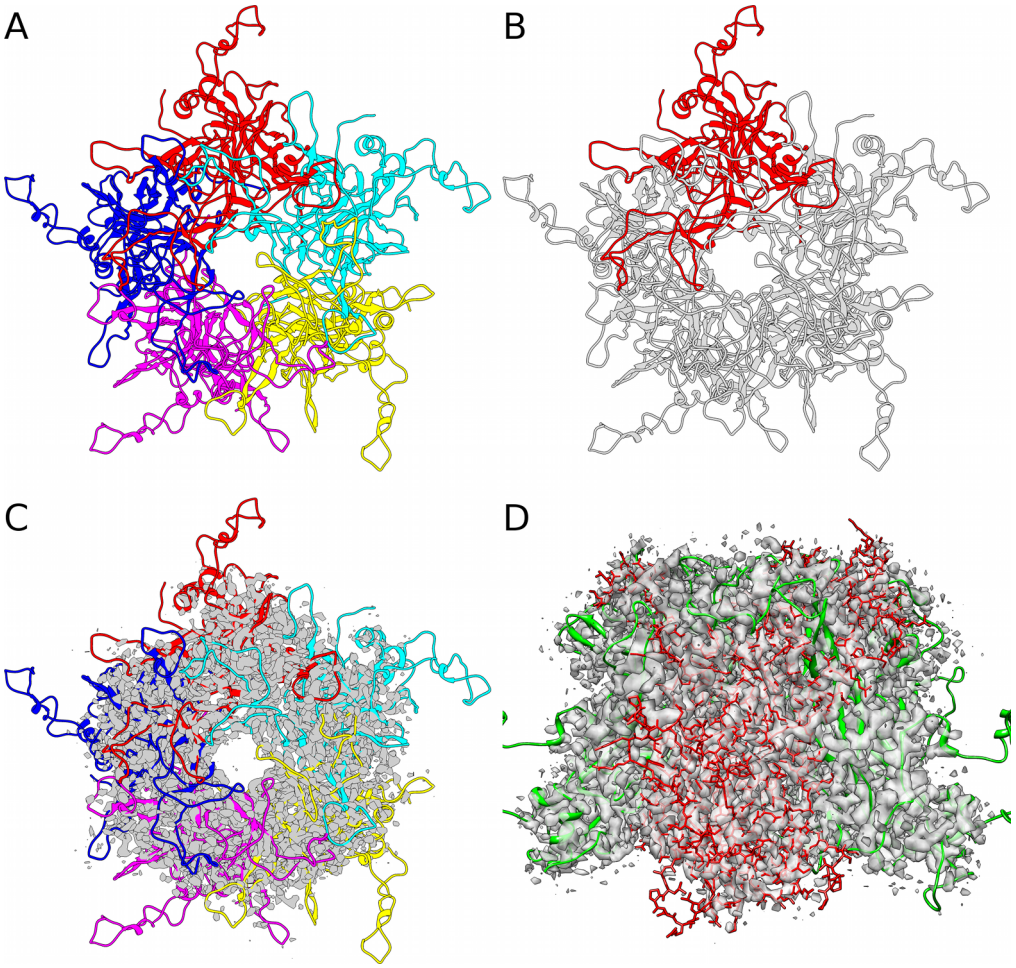


Figure 5.9 – Docking: Bovine papillomavirus pentamer
The assembly of the pentamer of bovine papillomavirus type 1 structural protein L1 was created by docking an atomic model of one monomer to an experimental cryo-EM map. Panel A of the figure shows the complex with each monomer colored differently while B demonstrates the intricate fold of the pentamer by coloring solely one monomer in red. In C, a top view of the superposition of the atomic structures and the electron density is shown. Panel D depicts a side view of the pentamer with one monomer drawn in stick mode. This view shows that large parts of the structure are clearly resolved while certain regions remain blurred. (© A. Griewel)

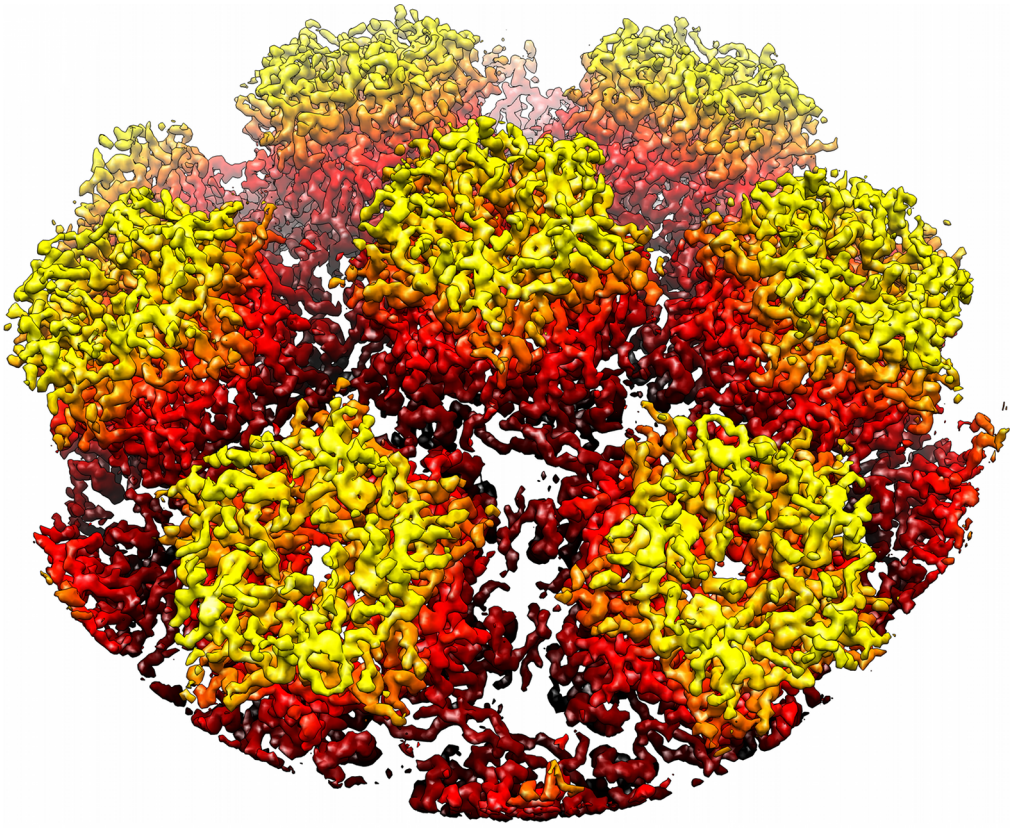


Figure 5.10 – Docking: Bovine papillomavirus icosahedral unit (I)
The cryo-EM map depicts the icosahedral unit of the bovine papillomavirus type 1 capsid comprising seven pentamers of the viral protein L1. The density is colored according to the distance to the center of the capsid, with lighter colors indicating larger distance. (© A. Griewel)

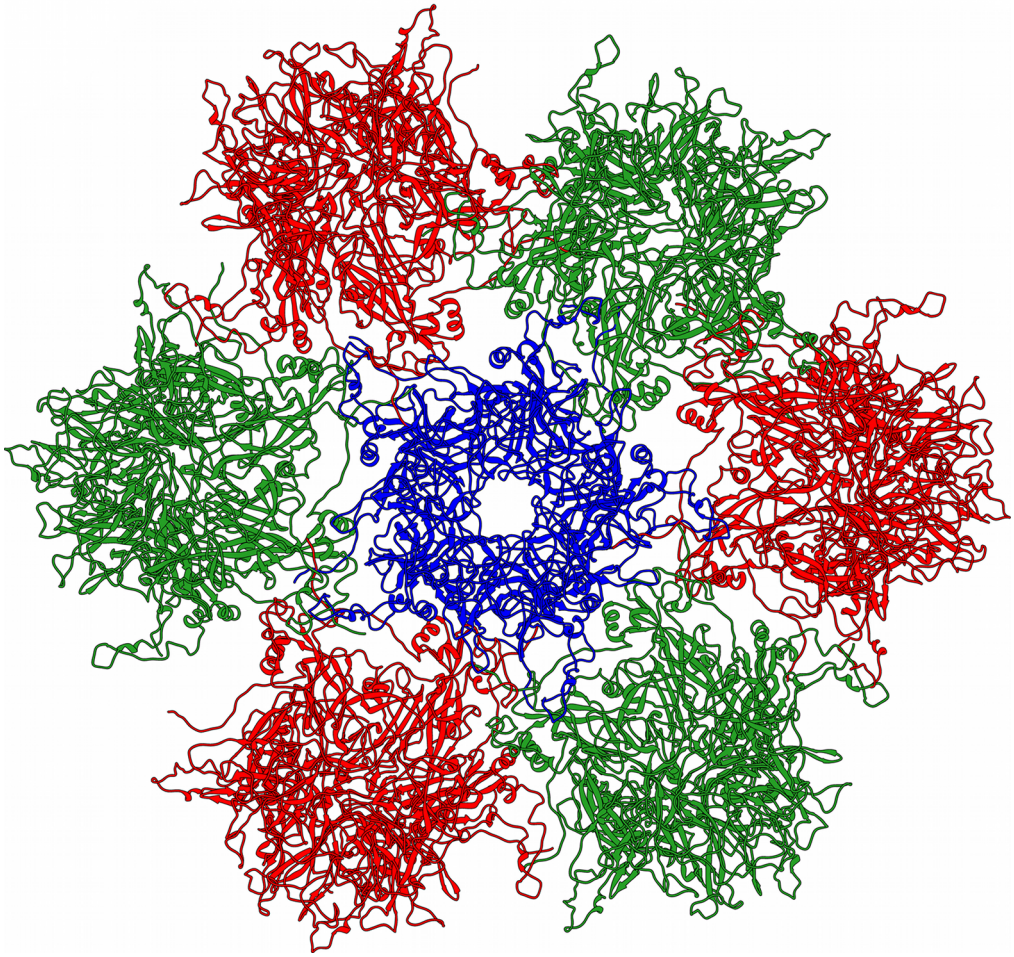


Figure 5.11 – Docking: Bovine papillomavirus icosahedral unit (II)
The assembly was generated by docking one monomer to the electron density shown in Figure 5.10 and shows that in this large map all placements of the monomer are identified correctly. (© A. Griewel)

assembled, which demonstrates that also very large maps can be analyzed using the presented method.

5.2.2. X-ray Crystallography

An abundant amount of atomic structures of biomolecules has been determined using X-ray crystallography. For all atomic models published in the wwPDB after February 2008 the deposition of the experimentally measured structure factors is required. Using these measurements and the deposited atomic structures, the Electron Density Server [178] provides access to experimental maps that are automatically generated using the experimental structure factors deposited in the wwPDB. Generally, X-ray crystallography maps are not sampled on cubic grids. Therefore, the downloaded maps are resampled using the Situs software package [365].

In the first two presented experiments, atomic structures of acetyl-coenzyme A synthetase and erythrocruorin are registered to their corresponding electron density maps. In these experiments, it is shown that also partially depicted molecules are successfully identified by *siseek* and that these proteins can also be found in large maps. The following two experiments align two experimental X-ray crystallography maps. In the first experiment, similar segments of DNA gyrase are successfully identified in two maps. In the second experiment, equine and human carboxyhemoglobin, which share less than 90 % sequence identity, are registered. These experimental setups were selected to demonstrate different application scenarios of *siseek*. The experimental data was chosen so that the electron density maps sample different resolutions typically found in the wwPDB and different molecular structures.

5.2.2.1. Acetyl-Coenzyme A Synthetase

The synthesis of acetyl-coenzyme A (acetyl-CoA) is catalyzed by acetyl-CoA synthetase, which has an approximate molecular weight of 72 kDa. The protein consists of a C- and an N-terminal domain that are connected by a hinge region. An electron density map of the protein complexed with adenosine-5′-propylphosphate and CoA is deposited as wwPDB ID `1PG4`. In this structure, the C-terminal domain is rotated so that the two domains of the protein form binding pockets, which accommodate the ligands. [128, 126]

In a first step, the atomic structure of `1PG4` is docked to the corresponding experimental electron density map. The map has a resolution of 1.75 Å and consists of 5.5×10^6 voxels with a sampling interval of 0.6 Å. The unit cell comprises two complete molecules of acetyl-CoA synthetase and several partially

depicted copies of the protein, which originate from neighboring cells. For this map, 37 002 keypoints and 231 871 descriptors are computed. For the synthetic map — generated for one monomer and consisting of 1.5×10^6 voxels — 4 380 keypoints with 32 631 descriptors are computed. The docking of the synthetic map is performed in 54 min using 21 min for creating the map description and the remaining 33 min for matching and optimization. The resulting placements are shown in Figure 5.12 and have an RMSD of 0.02 Å to the structure deposited in the wwPDB.

Figure 5.12 A shows that the calculated position of the protein fits well into the density map. In Figure 5.12 B a superposition of the deposited atomic structure `1PG4` and the predicted placements is displayed. Figure 5.12 C shows the positioning of the proteins in the complete unit cell and reveals that several copies of acetyl-CoA synthetase are partially contained in the unit cell. Figure 5.12 D shows a tilted view of the unit cell where all detected copies of the protein are shown without the electron density. The copies of the proteins are arranged in seven lines in the unit cell, which are depicted in different colors for clearer view. These proteins have been detected, even though they are depicted only partially in the unit cell.

5.2.2.2. Erythrocruorin

Erythrocruorins are respiratory complexes that are found in annelids and serve the same function as erythrocytes in other life forms: the transport of oxygen. The complex is not encapsulated in cells but contained freely in solution. It is composed of multiple copies of both oxygen-carrying globins and structural subunits and has a high molecular weight of more than 3.5 MDa. The X-ray crystallography map of `1X9F` [316], which is considered here, depicts a part of the structure of the erythrocruorin found in Lumbricus terrestris — the earth worm. The globin subunits, which form a dodecameric sub-complex in the erythrocruorin, have a molecular weight of approximately 16 kDa each. They form tetramers, which are aligned around a three-fold symmetry axis. Thus, the globin sub-complex is arranged as a trimer of tetramers. The asymmetric unit of the crystal includes one dodecameric complex — i. e., three copies of each type of globin. The sequence identity of the four globins ranges from 28 % to 47 %. [284, 316]

In this experiment, all four subunits are docked to the experimental X-ray map of `1X9F`, which has a resolution of 2.6 Å and a sampling interval of 0.9 Å. The number of voxels, keypoints and descriptors for both, the experimental and the synthetic maps, are listed in Table 5.2. The docking is performed in less than 13 min where 8 min are used for computing the map description and

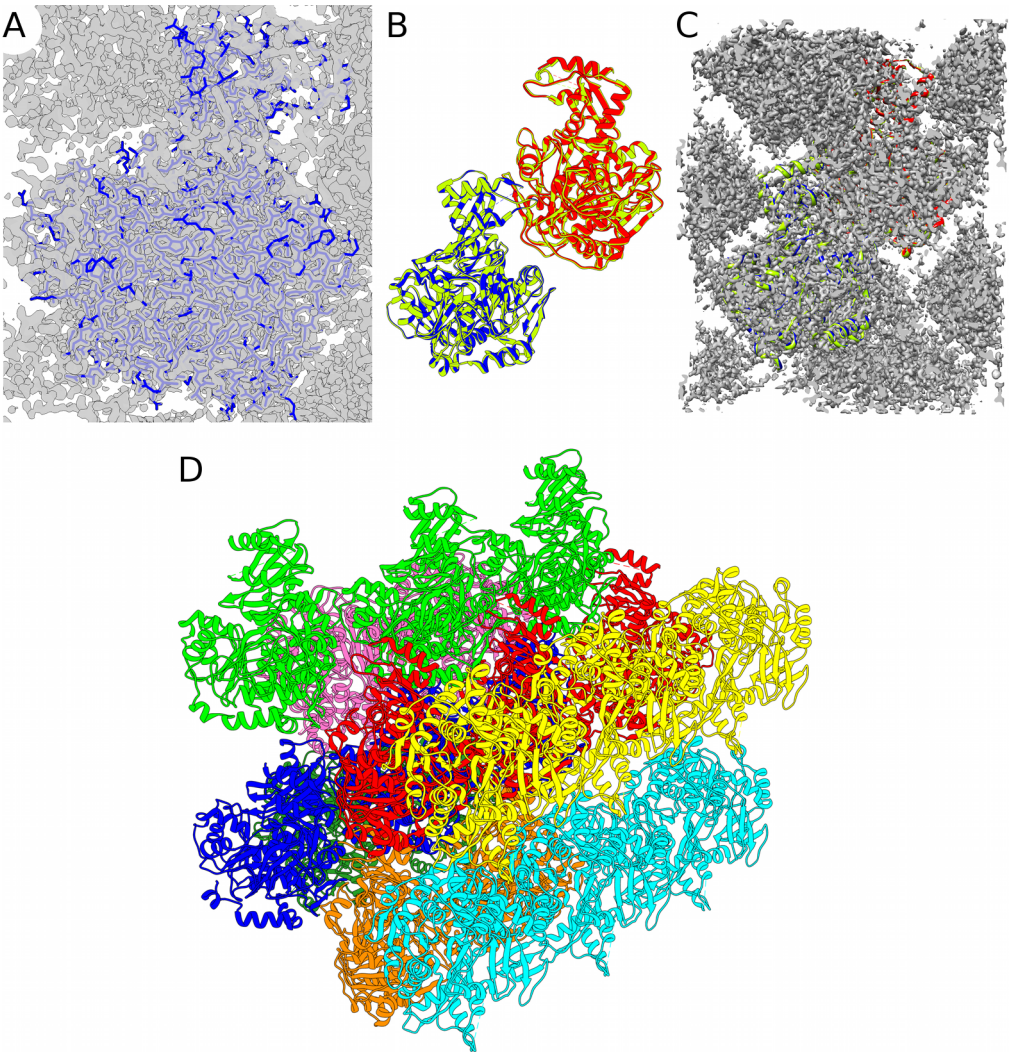


Figure 5.12 – Docking: Acetyl-coenzyme A
The figures were created by docking an atomic model of acetyl-coenzyme A synthetase to an experimental X-ray crystallography electron density map. Panel A shows a stick model of the computed placement superposed to the electron density while B depicts the superposition of the atomic model from the wwPDB and the computed placements. Panel C exemplifies that the two instances shown in B make up only a small portion of the unit cell and that additionally many partially depicted synthetases are contained in the unit cell. Panel D comprises all copies of partially depicted proteins that are identified by the docking. Thereby, an assembly of all proteins depicted in the unit cell is created. (© A. Griewel)

Description	Dimension	# Voxels	# Keypoints	# Descriptors
Experimental map	137×144×102	2 M	16 233	92 922
[1X9F:AEI]	54× 58 ×60	0.2 M	556	3 746
[1X9F:BFJ]	59× 66 ×57	0.2 M	590	3 484
[1X9F:CGK]	59× 64 ×59	0.2 M	558	3 488
[1X9F:DHL]	51× 64 ×57	0.2 M	621	3 634

Table 5.2 – Erythrocruorin properties
The table lists the properties of the maps that are used for assembling erythrocruorin.

5 min for matching and optimization. The calculated placements are shown in Figure 5.13. They have an average RMSD of 0.47 Å to the deposited structure showing that the placements correctly resemble the macromolecular assembly of erythrocruorin.

5.2.2.3. DNA Gyrase

DNA gyrase is a type II topoisomerase — these allow one DNA molecule to pass through another molecule of DNA. In prokaryotes, DNA gyrase relieves strain in the circular prokaryotic DNA when it is unwound by a helicase, which is especially important during DNA replication. This process is facilitated by binding DNA at two positions. The DNA held in position one is cleaved and the DNA held in the second position is passed through the opening. Afterwards, the cleaved strand is resealed. Since DNA gyrase is not found in humans, it is a frequent target for antibiotics. [229, 357, 117]

DNA gyrase is a dimer and each monomer consists of an A and B subunit. Subunit A can be subdivided further into a structural domain and a DNA binding domain, which are connected by α–helices to the dimerization domain (see Figure 5.14). Here, two experimental X-ray crystallography maps depicting parts of DNA gyrase subunit A are registered. The first map with wwPDB ID `1AB4` [48] shows a 50 kDa segment of a monomer of subunit A, which comprises all four domains. The map has a resolution of 2.8 Å, is sampled with an interval of 0.9 Å and consists of 0.7×10^6 voxels. The second map, `1X75` [73], depicts a complex of a subunit A dimer with a dimer of the toxin CcdB. The gyrase segments have a molecular weight of 26 kDa and comprise the dimerization domain plus the connecting α–helices. The toxic CcdB dimer binds to the gyrase and has a molecular weight of 20 kDa. The map consists of 0.5×10^6 voxels and has the same resolution and voxel spacing as `1AB4`.

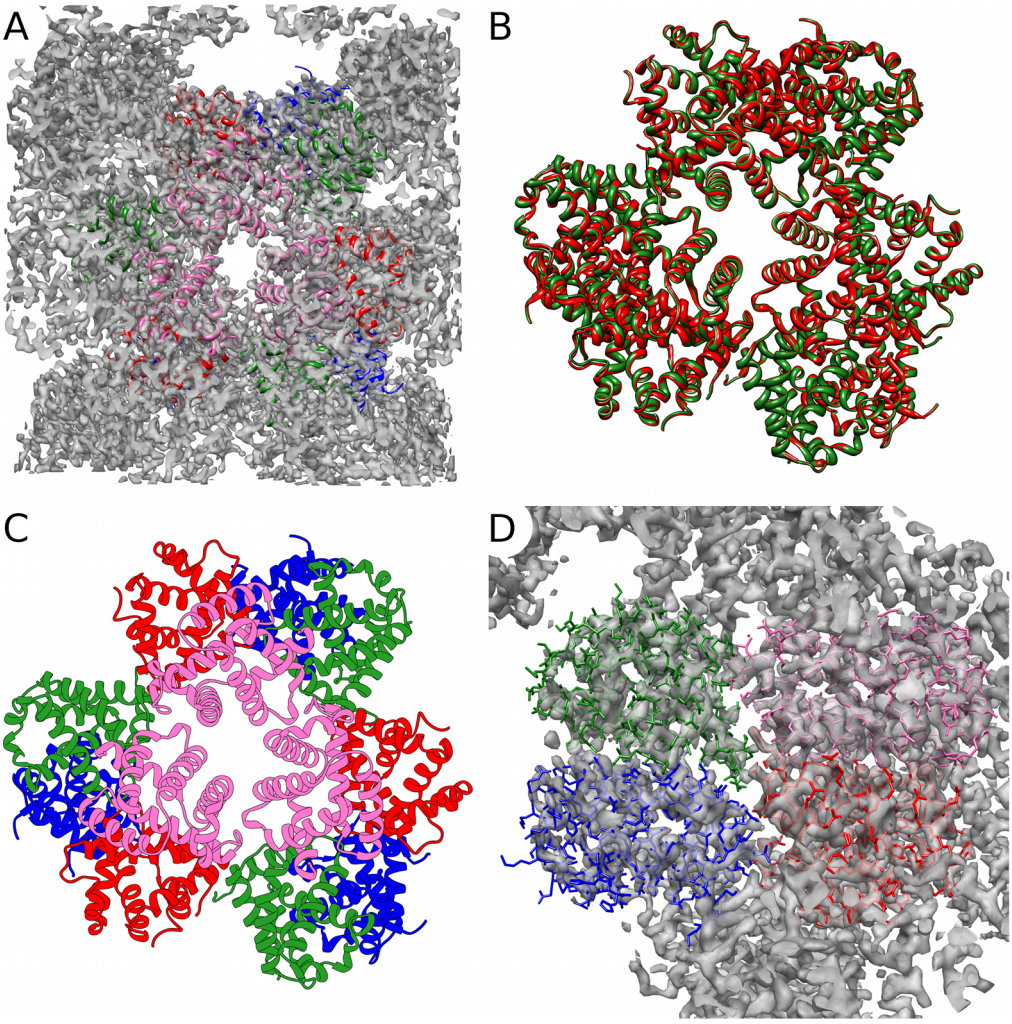


Figure 5.13 – Docking: Erythrocruorin
The placements are created by docking the atomic structures of the four globin subunits of the Lumbricus terrestris erythrocruorin to the experimentally determined X-ray map. Panel A shows the placements of the subunits in the unit cell while B depicts a superposition of the computed placements with the structure deposited in the wwPDB. In panel C, the docking result is colored by subunit exemplifying that the assembly is a trimer of tetramers. In panel D, the fit of one tetramer into the electron density map is shown. (© A. Griewel)

The map description of the experimental maps is generated by first detecting all keypoints and subsequently discarding those keypoint, which do not touch an atom of the deposited structure. In this way, a registration using solely keypoints from the experimental map is performed while suppressing matches to partially depicted molecules from neighboring unit cells, which effectively minimizes the run time used for matching. The description of the experimental X-ray crystallography map `1X75` consists of 6 274 keypoints and 35 048 descriptors. For the experimental map of `1AB4`, 6 735 keypoints with 38 670 descriptors are computed. The docking of the maps is performed in less than 5 min using 4 min for computing the map descriptions and another 32 s for matching. When using all detected keypoints without discarding those that do not touch the atomic structure, the complete docking is performed in 10 min.

Figure 5.14 shows the computed placements for registering the two experimental X-ray crystallography maps `1X75` — the source map — and `1AB4`. In panel A of the figure, a superposition of the molecules as schematic drawing is shown. The source map is shown in red and the two correct placements of `1AB4` are displayed in blue and cyan. The dimerization domain and the connecting helices are superposed, while CcdB of `1X75` and the DNA binding and structural domain of `1AB4` do not find counterparts in the other map. Figure 5.14 B and Figure 5.14 C display the same placement and superpose the electron densities of `1X75` and `1AB4`. Figure 5.14 D shows a clipped region of the map, which is rotated by 90° around a horizontal axis depicting solely the dimerization domain and the connecting helices.

5.2.2.4. Hemoglobins

One of the main functions of hemoglobin is the transport of oxygen. The most common mammalian hemoglobin is contained in erythrocytes and comprises four subunits — two α and two β chains. Each of the subunits contains a heme group, whose iron atom facilitates oxygen binding and interacts with the polypeptide chain through the imidazole ring of a histidine. Oxygen binding in hemoglobin is cooperative, which is facilitated by small conformational changes in the protein chain that effectively allow for the efficient uptake of oxygen in the lung and the delivery of oxygen to the tissue. A competitive binder to the oxygen molecule is carbon monoxide, which binds approximately 230 times stronger to hemoglobin than oxygen and forms the complex carboxyhemoglobin. [135, 198, 22]

The molecular weight of an assembled hemoglobin tetramer is approximately 66 kDa. The α and β subunits are structurally similar and resemble the myoglobin fold [36], even though the amino acid sequence of the α and β chains differs significantly. Furthermore, the sequence of equal subunits differs between

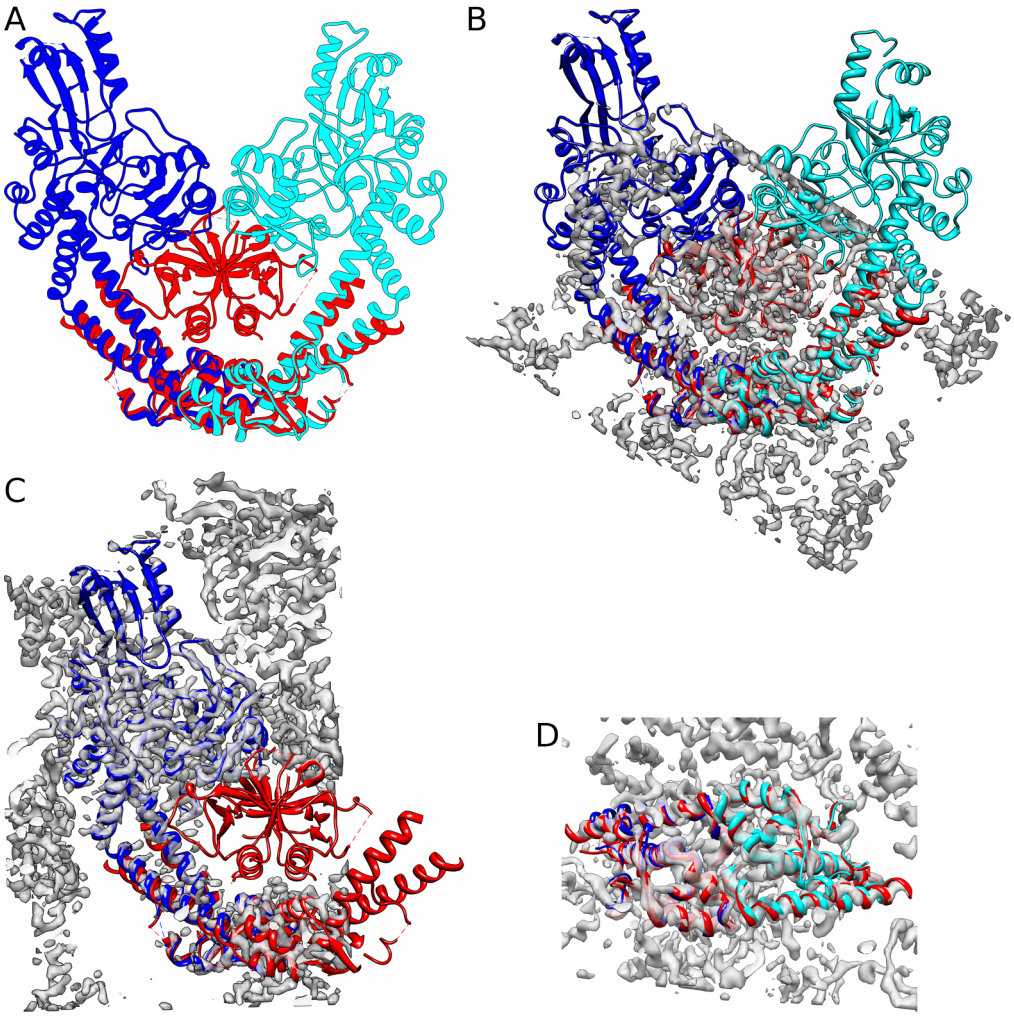


Figure 5.14 – Registration: Gyrase
Two experimental X-ray crystallography maps depicting parts of DNA gyrase, have been registered and the resulting placements are shown. In A, only the corresponding atomic models are drawn for clearer view. The model of the source map `1X75` is colored red and the two identified placements of `1AB4` are colored blue and cyan. The superposed parts correspond to the dimerization domain and the connecting helices of DNA gyrase. The structural and DNA binding domain in the blue and cyan structures are not superposed. This is also true for the red proteins, which correspond to the toxin CcdB. In panel B, the experimental electron density of `1X75` is superposed to the placements, while in panel C the electron density for the left instance of `1AB4` is shown. In D, the superposed segments of the proteins along the electron density map of `1X75` are shown in a view that is rotated by 90° around a horizontal axis. (© A. Griewel)

		2DN3		2D5X	
		α	β	α	β
2DN3	α	–	45	87	44
	β	43	–	42	83
2D5X	α	87	44	–	45
	β	42	83	44	–

Table 5.3 – Hemoglobin sequence identity
Sequence identities in percent between the α and β chains of human (2DN3) and equine (2D5X) hemoglobin.

species and this difference grows with evolutionary distance [135]. Here, two experimental X-ray crystallography electron density maps of human (2DN3 [256]) and equine (2D5X [376]) carboxyhemoglobin are registered, which each depict the complex of one α and one β subunit in a similar conformation. The sequence identity between the depicted polypeptide chains in these maps was computed using the Smith–Waterman algorithm [309] and is shown in Table 5.3. While the same chains between the two species have similarities of more than 83 %, the identity between the chains is less than 45 %.

The map of 2DN3 shows a dimer of the α and β subunit of human carboxyhemoglobin. It has a sampling interval of 0.4 Å, a resolution of 1.25 Å, and comprises 2.8×10^6 voxels. The second map depicts a dimer of α and β subunit of equine hemoglobin at a resolution of 1.45 Å and a sampling interval of 0.5 Å using 1.7×10^6 voxels. A description is computed for both maps and restricted to keypoints and descriptors that touch the corresponding atomic structure as described for DNA gyrase in Section 5.2.2.3. Here, 2DN3 with both subunits is considered as source map. The two subunits of equine hemoglobin 2D5X are docked separately to the source map, since their relative orientation differs slightly from that shown in 2DN3. For the source map, 2 431 keypoints and 16 574 descriptors are computed. The first target map comprises the α subunit of equine carboxyhemoglobin and 1 167 keypoints with 7 647 descriptors; the second target map depicts the β subunit from 2D5X with 1 269 keypoints with 8 545 descriptors. The docking is performed in less than 10 min using 8 min for creating the map description and 2 min for matching.

The placements resulting from the registration are displayed in Figure 5.15. Panel A of the figure shows that the two subunits of equine carboxyhemoglobin resemble the conformation of human hemoglobin. In panel B, the two maps are superposed and show that the unit cells of the two crystals are different. Here, several densities seem to be unregistered, which is due to the different unit cells. In panel C, a clipped superposition of low-pass filtered versions of

the two maps is shown, which demonstrates the good fit in the relevant central region of the unit cell. The last panel, Figure 5.15 D, shows the superposition of the atomic detail α subunit of `2D5X` and the electron density map of `2DN3`. The view is clipped and shows that the atoms of equine carboxyhemoglobin match the electron density of human carboxyhemoglobin well.

5.2.3. Summary

The performed experiments demonstrate the capability of *siseek* to successfully register experimental electron density maps. The run time, including computation of the map description and matching, ranges from 4–88 min depending on the maps' size. The resulting placements have minimal deviation from the given reference placements and therefore the placements can effectively be used for biophysical studies. Thus, *siseek* can be utilized for quickly computing a registration of intermediate and high resolution electron density maps.

First, it was shown that *siseek* successfully calculates the correct placements for atomic structures in high resolution cryo-EM maps. This was demonstrated using six case studies based on experimental electron density maps of GroEL and the Methanococcus Maripaludis chaperonin as well as rotavirus particle 6 and papillomavirus structural protein L1. These experiments also show that it is possible to handle very large maps such as the icosahedrally averaged map of the papillomavirus structural protein L1, which consist of more than 22.2×10^6 voxels. Subsequently, the docking of atomic structures to experimental X-ray maps was demonstrated on acetyl-CoA synthetase and erythrocruorin. The correct positions of the proteins have been identified correctly for all subunits. Furthermore, it was shown that *siseek* identifies also proteins if they are only partially depicted as in the map of acetyl-CoA synthetase. Additionally, the case study of erythrocruorin showed that it is also possible to successfully identify small subunits in larger maps.

In the last two experiments, experimental X-ray maps are registered. In the first case study, two maps showing domains of DNA gyrase in similar conformation are registered. *siseek* successfully computes a registration of the maps, which superposes the similar domains correctly. Since these maps also comprise other protein domains, the experiment shows that the registration of two experimental electron density maps can also be based on similar sub-volumes depicting similar domains. In the second case study, maps of human- and equine carboxyhemoglobin are registered. These proteins have a sequence identity of less than 90 % in the corresponding chains and the successful registration of the maps shows that also proteins with non-identical but similar sequence and

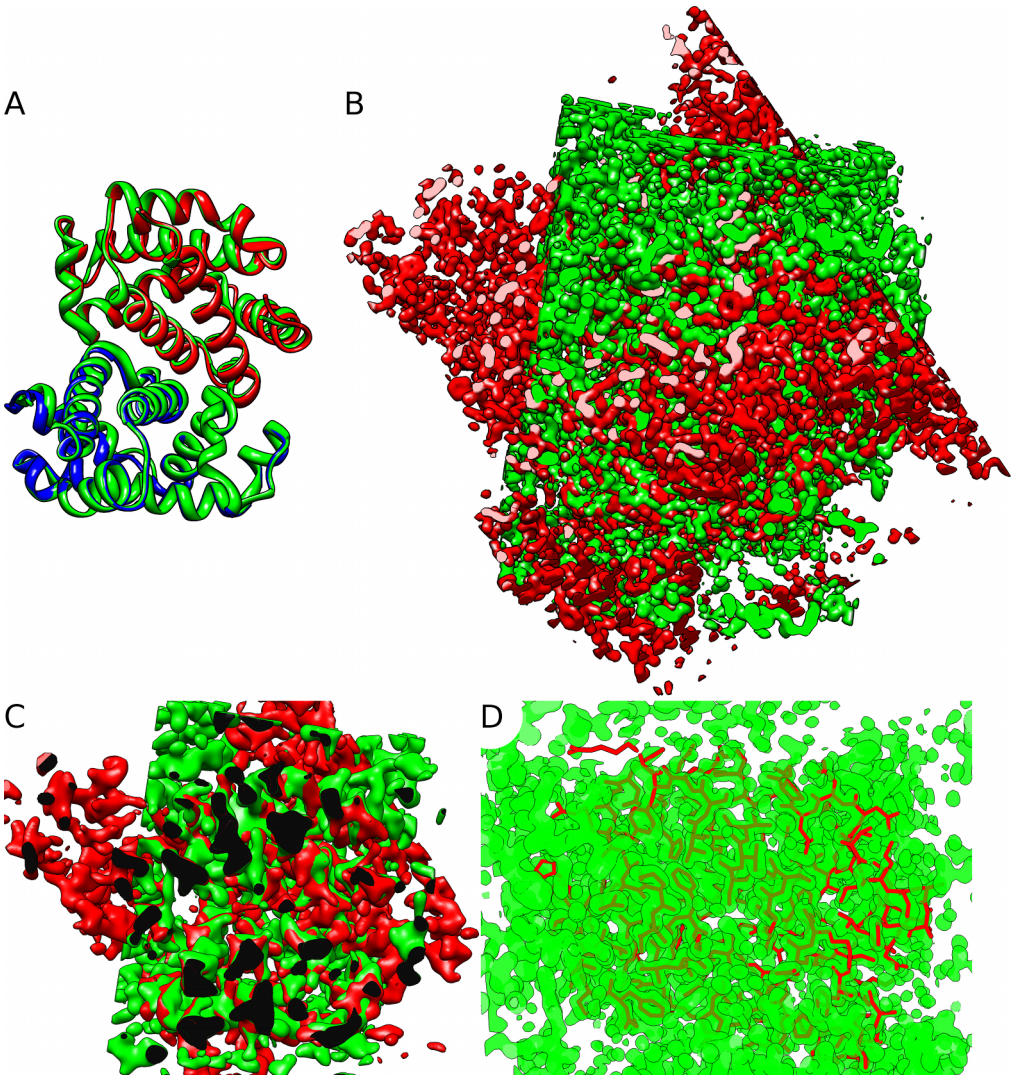


Figure 5.15 – Registration: Hemoglobin
The registration was created using the two experimental X-ray crystallography maps of 2DN3 and 2D5X, which depict human and equine hemoglobin. In panel A, the registration of the α (red) and β (blue) chains of 2D5X to 2DN3 is shown using the corresponding atomic models. Panel B depicts the registered electron density maps. A smaller, clipped version of the same view as in B is shown in C for low-pass filtered versions of the maps. The black spots depict clipped density and the superposition of the densities can be seen for the registered central part of the maps. In D, the placement of the α chain of equine hemoglobin in the experimental electron density of human hemoglobin exemplifies the agreement between the two structures. (© A. Griewel)

conformation can be registered using *siseek*. All figures regarding the utilized maps, their descriptions, and the timings are summarized in Table 5.4.

ID	T	R (Å)	GSI (Å)	X	Y	Z	#	RSI (Å)	X	Y	Z	#	# K	# D	Time (min)
5001	C	4.2	1.06	200×	200×	200	8.0	1.10	193×	193×	193	7.2	12 764	65 499	10
1XCK	S			72×	87×	72	0.5		70×	84×	70	0.4	997	7 012	
5137	C	4.3	1.33	192×	192×	192	7.1	1.13	227×	227×	227	11.7	20 963	166 226	13
3LOS	S			71×	65×	77	0.4		84×	77×	91	0.6	899	5 637	
5140	C	8.0	1.33	240×	240×	240	13.8	2.10	153×	153×	153	3.6	2 439	16 987	4
3IYF	S			75×	66×	91	0.5		48×	42×	58	0.1	149	797	
1461	C	3.8	1.23	106×	122×	134	1.7	1.00	131×	151×	166	3.3	10 291	70 751	8
1QHD	S			77×	84×	54	0.3		95×	104×	67	0.7	910	6 134	
5155	C	4.9	1.24	285×	287×	272	22.2	1.10	320×	323×	306	31.6	93 735	732 659	88
3IYJ	S			94×	77×	87	0.6		106×	87×	98	0.9	909	6 481	
5156	C	4.2	1.24	118×	122×	119	1.7	0.94	155×	160×	156	3.9	7 930	62 141	7
3IYJ	S			69×	93×	67	0.4		91×	122×	88	1.0	1 132	7 744	
1PG4	X	1.75	0.6	163×	167×	202	5.5	0.46	213×	219×	264	12.3	37 002	231 871	54
1PG4	S			119×	114×	144	2.0		156×	149×	189	4.4	4 380	32 631	
1X9F	X	2.6	0.9	137×	144×	102	2.0	0.68	181×	190×	135	4.6	16 233	92 922	13
[1X9F:A]	S			54×	58×	60	0.2		72×	77×	80	0.4	556	3 746	
[1X9F:B]	S			59×	66×	57	0.2		78×	88×	76	0.5	590	3 484	
[1X9F:C]	S			59×	64×	59	0.2		78×	85×	78	0.5	558	3 488	
[1X9F:D]	S			51×	64×	57	0.2		68×	85×	76	0.4	621	3 634	
1X75	X	2.8	0.9	79×	72×	91	0.5	0.73	97×	89×	112	1.0	6 274	35 048	5
1AB4	X	2.8	0.9	74×	79×	122	0.7		91×	97×	150	1.3	6 735	38 670	
2DN3	X	1.25	0.4	158×	127×	140	2.8	0.33	193×	155×	171	5.1	2 431	16 574	10
2D5X A	X	1.45	0.5	123×	112×	126	1.7		162×	148×	166	4.0	1 167	7 647	
2D5X B	X	1.45	0.5	123×	112×	126	1.7	0.33	162×	148×	166	4	1 269	8 545	

Table 5.4 – Registration results
The table lists the properties of the maps used in the registration experiments. It shows the ID of the map (EMDataBank entries are identified by their number only), the type of the map (S for synthetic, X for X-ray, and C for cryo-EM) and the map resolution. Column 4–8 tabulate the genuine sampling interval (GSI), the number of voxels in the X, Y, and Z dimension and the approximate total amount of voxels in millions. The following columns 9–13 list the sampling interval (RSI) and the corresponding number of voxels in millions after resampling. The last three columns specify the number of detected keypoints (# K) and descriptors (# D), and list the run times of the dockings in minutes.

5.3. Molecule Recognition

In this proof of concept study, *siseek* is used to identify the content of an electron microscopy map. For this application, a database of reference protein structures is assembled, which comprises common patterns observed in proteins. For all reference structures, a map description is computed and saved in a distributed database. The depicted protein structure(s) in the *query* map are then identified by comparing all *query* descriptors to the descriptors in the database and analyzing the results. In the following, the choice and content of the reference set are explained and the properties of the created database are summarized. Then, the computer setup is detailed. Finally, a proof of concept of the method is given using selected synthetic-, cryo-EM-, and X-ray maps. Here, the 1-NN′ and also the LR scoring schemes are assessed. Subsequently, ways to reduce the run time needed by the approach are discussed. Finally, all findings are summarized.

Reference Set

There are three predominant classifications of proteins, which generally disassemble the biomolecules into their domains and categorize them. The FSSP [146] (Families of Structurally Similar Proteins) is assembled automatically and relies solely on the structural alignment of proteins. A second catalog is provided by SCOP [247, 66, 6, 7] (Structural Classification Of Proteins), which is a mainly manually curated database of protein domains. The third classification is called CATH [254] (Class, Architecture, Topology, Homologous superfamiliy) and employs a semi-automated procedure to classify protein domains according to a hierarchic scheme. While the three databases agree on most assignments, they differ in details which is mainly due to divergences in domain assignments [133, 179].

In this experiment, the largely accepted SCOP database is employed for the identification of molecules that are depicted in electron density maps. Representatives for classes in SCOP are readily accessible through internet resources as explained below. The catalog is curated mainly manually and based on evolutionary, functional, and structural similarity criteria. Protein domains are assigned to groups, which adhere to a hierarchic system. The root of the hierarchy forms the *class* level, which is assigned on the basis of the secondary structure content and its topological organization. The more specific *fold* level is based on the 3D arrangement of secondary structure elements but does not consider evolutionary relatedness. Proteins in one fold are required to have secondary structures in the same arrangement with the same topological con-

nections. Each fold is further subdivided into *superfamilies*. Members of one superfamily share structural and also functional features but lack clear evidence of an evolutionary relationship. The *family* level groups those proteins, for which a clear evolutionary relatedness can be deduced. Families are further subdivided into *proteins*, which cluster highly similar sequences, and *species*, which is the most specific level and makes distinctions based on the expression system. The current version 1.75 of SCOP comprises 110 800 domains, which are grouped in 3 902 families, 1 962 superfamilies, 1 195 folds, and seven classes[1].

In SCOP, structures are classified at different levels: from the abstract class-level down to the family- and the species level. To enable an efficient molecule recognition application, a sufficient level of detail for the domains to be included in the reference database must be selected. Using only one representative for an abstract folding motive — such as a superfamiliy representative — is not sufficient for the identification of protein structures. This can be seen by the investigation of proteins that are contained in one fold class: Even though these share an abstract similarity in the relative orientation of secondary structure motives, the actual sequence of the protein and the specific conformation differ considerably. This renders similarity searching without the knowledge of the contained molecule impossible.

To select a sufficient level of detail, the findings from the registration experiments in Section 5.2 are summarized here. The studies of DNA gyrase and acetyl-CoA synthetase demonstrated that it is possible to identify partially depicted proteins. The case study of erythrocruorin showed that it is also possible to successfully identify small subunits in larger maps. The case study of human- and equine carboxyhemoglobin, which have a sequence identity of less than 90 % in the corresponding chains, showed that proteins with similar sequence and conformation can be registered successfully. This exemplifies that it is possible to employ a set of reference domains, which does not contain highly similar domains. This has two advantages: On the one hand, the redundancy in the database is reduced. Therefore, less descriptors are stored and the run time is lowered. On the other hand, it allows for a scoring based on the distinctiveness of a match, which is only possible if the protein structures contained in the reference set are unique. A second requirement on the set of reference domains is that its structures cover all available conformations of the contained protein domains. Complete movements of domains do not destroy the correct matching, since *siseek* is able to detect intact subunits. The quality of the matches will, however, degrade if the intrinsic conformation of the domain changes by, e. g., changes in side-chain conformations.

[1] The SCOP classification mentions eleven classes out of which four are “not true classes”.

Based on these observations, a subset of the SCOP database, which is filtered according to sequence identity, is chosen as reference set. The sequences and structures for these domains are provided by the ASTRAL compendium [54, 55, 39], which also devises sequences for multi-chain domains. This set of protein domains comprises only the highest quality representative for each identified cluster of similar sequences. Here, the quality is measured according to the resolution, the R-factor, and the stereochemical accuracy of the protein domain. In the following, the ASTRAL SCOP set in version 1.75 with protein domains of less than 95 % sequence identity is utilized. This set comprises 16 712 distinct domains, which have been created using coordinates from the wwPDB. For structures containing multiple models, which mostly correspond to NMR structures, only the first model is utilized.

The results of the registration experiments were computed using a 84-dimensional descriptor as outlined in Section 4.4.3 on page 130. For molecule recognition, however, a descriptor with 1 134 dimensions is utilized. This descriptor was chosen by inspecting the feature vector distance distributions shown in Appendix A.6.2 on page 216. These distributions show that the high dimensional descriptor, which also covers a larger map volume, separates true from false matches more clearly. This result was confirmed by performing preliminary studies with feature vectors of dimensionality 84 and 228, which both showed worse performance than the 1 134 dimensional descriptor on both synthetic and experimental maps. However, the high dimensional feature vector is also more susceptible to noise as discussed later on.

For all atomic structures in the reference set, a synthetic map is created using a resolution of 1.73 Å and a sampling interval of 0.5 Å. This level of detail is finer than the average resolution of electron density maps in the wwPDB and allows for the effective matching of query descriptors, and thereby enables the molecule recognition. For each synthetic map, a complete map description is determined and saved in a database. This includes the information on the detected keypoints, the orientation histograms, and also all descriptors as explained in Section 3.6. In total, the database contains 16 614 925 keypoints and 128 022 590 descriptors, which are created according to parameter set 2 listed in Table 4.7. This results in a database with a total size of 1.29 TiB with the utilized library [145].

Computer Setup

Due to its size, the database is distributed on a compute cluster. This decreases the run time and allows for the partitioning of the data. Using a distributed file system, the reference domains are stored at a central location and allocated

#	CPU Freq. (GHz)	# Cores	Cache (KiB)	Main Memory (GiB)
2	2.33	8	6 144	31.4
1	2.53	8	12 288	62.9
10	2.53	16	12 288	31.4
1	2.53	8	12 288	31.4
3	2.53	16	12 288	62.9

Table 5.5 – Properties of the utilized compute cluster
The compute cluster consists of 17 machines, which are equipped with Intel Xeon processors. The table lists the number of (#) computers with a specific CPU frequency (freq.), number of cores, cache and main memory size.

to different cluster nodes. The database partitions are created locally on every node, and results are again reported to a central repository. The access to the generated results is transparent, i. e., independent of the cluster node the result is computed on. This is facilitated by a system, which accounts for the allocation of references to cluster nodes and also for the association of results to reference domains.

The cluster consists of 17 computers, which are summarized in Table 5.5. The cluster is shared with other users and jobs are scheduled by a queuing system, which allocates jobs to machines based on priorities. The set of reference domains is split in 1 088 partitions that are assigned to specific cluster nodes. Thus, each node is given 64 partitions, which comprise on average the descriptions of 15 domains from the reference set. These database partitions are queried using descriptors computed from the given map. The query for each partition is independent of the processing of other partitions, and therefore the queries are carried out in a concurrent manner with no order specified on their execution. When calculating the final result, all processes on all nodes must be finished and the best scoring matches are reported to a central repository.

Query Setup

The experimental setup is analyzed using test maps. The set of test maps includes two experimental X-ray crystallography maps, `2D5X` and `1X75`, which depict protein domains that are comprised in the reference database. Furthermore, two protein structures that are represented by highly similar domains are used in the queries — `1UMO` and `2DN2`. Synthetic maps of the proteins and also experimental maps from X-ray crystallography are used for querying the reference

ID	Ref.	Type	Res.	Description
1UMO	[76]	S		Cytoglobin
2DN2	[256]	S		Doxygenated hemoglobin
148L	[189]	X	1.9	T4 Lysozyme
1GFL	[373]	X	1.9	Green fluorescent protein
1PBD	[295]	X	2.3	Hydroxylase
1T5H	[127]	X	2.0	4-chlorobenzoyl-coenzyme A ligase
1UMO	[76]	X	2.59	Cytoglobin
1W0E	[359]	X	2.8	Cytochome P450
1X75	[73]	X	2.8	DNA gyrase in complex with CcdB
2CG9	[3]	X	3.1	Heat shock protein 90
2D5X	[376]	X	1.45	Carboxyhemoglobin
2DN2	[256]	X	1.25	Doxygenated hemoglobin
2EWA	[347]	X	2.1	Protein kinase
EMD-1461	[379]	C	3.8	Rotavirus particle 6
EMD-5001	[219]	C	4.2	GroEL

Table 5.6 – Molecule recognition test maps
The table lists the maps used for assessing the molecule recognition setup. Besides the ID of the map, the reference (Ref.), the type of the map (S for synthetic, X for X-ray crystallography, and C for cryo-EM), the resolution (Res.), and a short description are tabulated.

database. This allows for the analysis of the differences between querying with a synthetic map and an experimental map. Eventually, seven experimental maps from X-ray crystallography — 1PBD, 1T5H, 148L, 1GFL, 1W0E, 2CG9, and 2EWA — and two high resolution maps from cryo-EM — EMD-1461 and EMD-5001 — are employed in the experiment. All proteins included in the test set are listed in Table 5.6.

Two scoring schemes for interpreting matches of query- and reference keypoints — 1-NN′- and LR scoring — are introduced in Section 3.6. In the following, the parameterization of 1-NN′ scoring and the results for querying the reference database using this scoring scheme are discussed. Then, a parameterization for LR scoring is determined and the results of the queries are analyzed accordingly. Subsequently, a comparison of the two scoring schemes summarizes the findings.

1-NN′ Scoring Scheme

For the 1-NN′ scoring scheme, a threshold τ_{NN} is utilized to discard indistinct keypoint matches. These are matches, in which the best matched keypoint and the second best matched keypoint have similar distance to the query keypoint. The threshold equals the distinctiveness criterion and therefore $\tau_{\mathrm{NN}} = 0.9$ is used. The result lists of the performed queries are shown in Table 5.7, where all reported reference domains are listed with their ID[1] and corresponding score.

For the experimental map `2D5X`, the correct matching α and β domains are identified as the top two hits. Furthermore, various α and β domains of other hemoglobins are reported as matches for `2D5X`. Thus, the molecule recognition successfully identified the domains depicted in the map, and also reports matches to other instances of the *globin* family. For `1X75`, the A chain, which corresponds to the connecting helices and the dimerization domain of DNA gyrase, is correctly identified as first hit. The second hit is `[3VUB:A_]`$_{\texttt{SCOP}}$ [216], which is the reference domain for CcdB and also comprised in the map of `1X75`[2]. However, the number of votes is low, which is due to the low resolution of the density map.

In a next step, the synthetic and experimental maps of human hemoglobin `2DN2` and human cytoglobin `1UMO` are used as queries. The synthetic maps for these proteins are created using the same settings as for the reference database, i. e., a sampling interval of 0.5 Å and a resolution of 1.73 Å.

For the synthetic map of `2DN2`, only correctly matching domains from the *globin* family are reported. The first-ranked domain is `[1FHJ:A_]`$_{\texttt{SCOP}}$ [86], which is the α chain of hemoglobin of chrysocyon brachyurus — the maned wolf. It has a sequence identity of less than 90 % with `[2DN2:A]` and is depicted in the aquo-met conformation. This conformation resembles the map created by `2DN2` more closely than the protein structure of `2DN3`, which depicts human hemoglobin bound to carbon monoxide. For the experimental map of `2DN2`, only four matches are reported. These correspond to the α chains of the already mentioned proteins `1FHJ` and `2DN3` as well as two other hemoglobin chains. The scores of these references are low and only for 6 keypoints distinctive matches were found. The β domain depicted in `2DN2` is not matched. This is due to the additional noise in the map, but also to the fact that several hemoglobin structures are comprised in the reference database, which limits the applicability of the distinctiveness score.

For the synthetic map of `1UMO` the correctly matching domain `[1URV:A_]`$_{\texttt{SCOP}}$ [76] is ranked first and has a considerably higher score than the other matches.

[1] The notation of identifiers in this work is explicated in Appendix A.2 on page 199.

[2] See also Section 5.2.2.3 on page 163.

(X) 148L	
1P5C:A_	8
1SWY:A_	7
2F2Q:A1	7
1P37:A_	6
1T8F:A_	6
1JTM:A_	6
1L64:A_	5
146L:A_	4
157L:A_	1

(X) 1GFL	
1MYW:A_	65
1H6R:A_	58
1OXD:A_	42
1KP5:A_	41
2F01:A1	1
1GGX:A_	1

(X) 2D5X	
2D5X:B1	58
2D5X:A1	47
1JEB:A_	7
2QSS:A1	7
2DN3:B1	7
2DN3:A1	6
2QSS:B1	5
1FHJ:A_	5
1QPW:A_	4
1OUT:B_	2
1WMU:A_	2
1WMU:B_	2
1FHJ:B_	2
1SHR:B_	1
1A4F:B_	1
1QPW:B_	1
1HBR:B_	1
3BJ1:B1	1
1A4F:A_	1
1CG5:A_	1

(S) 1UMO	
1URV:A_	514
1A9W:E_	6
1UC3:A_	3
2V1F:A1	3
1O5Z:A1	2
1CQX:A1	1
1HBG:A_	1
1XQ5:A_	1
1EOV:A2	1
1ECD:A_	1
1DJX:B1	1
1QGO:A_	1
1W98:B1	1
1V93:A_	1
1G2U:A_	1
1JBO:B_	1
2B9V:A2	1
1NKP:B_	1
1JL7:A_	1
1R6F:A_	1

(S) 2DN2	
2D5X:A1	189
2QSS:A1	143
2D5X:B1	96
2QSS:B1	62
1WMU:A_	61
1I3D:A_	54
1FHJ:A_	49
2DN3:A1	47
1FHJ:B_	33
3D1K:B1	28
1CG5:A_	22
2DN3:B1	20
1WMU:B_	20
1SHR:B_	20
2H8F:B1	17
1V4W:B_	16
1JEB:A_	13
1JEB:B_	13
3D1K:A1	13
1CH4:A_	13

(X) 1PBD
none

(X) 1T5H	
3CW9:A1	96
1PG4:A_	1
1MDB:A_	1

(X) 1UMO
none

(X) 1W0E	
1TQN:A_	35

(X) 1X75	
1X75:A1	2
3VUB:A_	1

(X) 2CG9	
1USU:A_	2

(X) 2DN2	
2DN3:A1	2
1FHJ:A_	2
1QPW:A_	1
1HBR:B_	1

(X) 2EWA	
2GFS:A1	10
1PME:A_	1
1Q8Y:A_	1

(C) EMD-1461
none

(C) EMD-5001	
1YNJ:D1	7
2NPP:B1	2
1SMY:C_	1

Table 5.7 – Molecule recognition results: 1-NN′ scoring scheme
The table shows the result lists using the 1-NN′ scoring scheme. Each query map is denoted by its ID and an identifier of its type (S for synthetic, X for X-ray crystallography, C for cryo-EM). All lists are truncated to 20 entries.

For the experimental map of 1UMO, however, no matches are reported. This is due to the properties of the experimental map, which has a low resolution of 2.59 Å and a high R-factor of 0.212. The low resolution prevents the detection of small-scale keypoints and thus fewer keypoints for matching with the database are available. The high R-factor shows that there are differences between the experimental and the synthetic map, which is another reason for the failure.

For the maps 148L, 1GFL, 1T5H, 1W0E, 2CG9, and 2EWA domains from the correct SCOP families are identified. For 148L, 1W0E, and 2EWA all matches stem from the correct protein level — i. e., the next, more specific level after the family level. The map of 1GFL clearly matches four domains, which are all instances of the *green fluorescent protein*. The latter domains are similar in fold, but do not belong to the group of *green fluorescent proteins*. The IDs correspond to a red fluorescent protein, which also belongs to the family of *fluorescent proteins*, and streptavidin, which also shares the beta barrel fold of 1GFL. The map of 1T5H is clearly matched to its counterpart 3CW9 [272], which differs from 1T5H by a hinge movement of the domains and therefore by the relative orientation of the two comprised domains. Thus, the two domains do not differ in their internal structure and are therefore clearly matched in contrast to 1UMO. Furthermore, 1PG4 [128] and 1MDB [230] are reported as matches for 1T5H, which also belong to the family of *acetyl-CoA synthetase-like* proteins. For the map of 2CG9, the correctly matching domain of $\texttt{[1USU:A\_]}_{\texttt{SCOP}}$ [235] is identified by one hit.

For the experimental maps of 1PBD and EMD-1461, no matches are reported. This is due to the lower resolution of the maps in combination with higher levels of noise. Thus, the correctly matching descriptors have a relatively large distance to the query descriptors and are not identified as distinctly matching. The reference domains reported for EMD-5001 are incorrect and based on matches of descriptors of larger scale. The SCOP subdivides proteins into domains and therefore generally small-scale descriptors are computed. Thus — using this database — it is more probable that a large-scale match is distinct because fewer large-scale descriptors exit.

The scores reported in Table 5.7 show that the amount of votes is relatively low. This is, on the one hand, due to noise in the recorded data and also the noise, which is introduced by allowing small changes in the conformation. The low scores are, on the other hand, also due to the combination of the SCOP database with the 1-NN′ scoring scheme. The 1-NN′ scoring scheme requires distinctive reference matches and if the reference database comprises structures twice — as is the case for, e. g., hemoglobins or [1X75:A] — matches can not be identified as distinct.

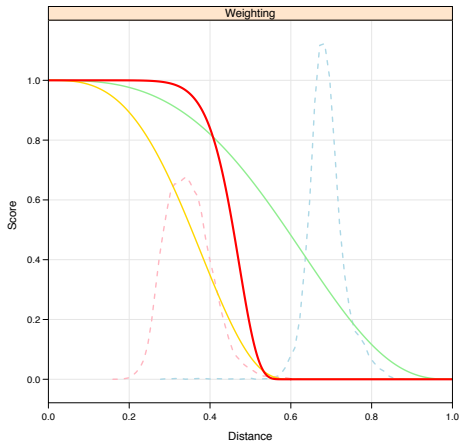
LR Scoring Scheme

Based on these observations, the performance of an alternative weighting method is assessed. The LR scoring scheme is inspired by local regression and employs a finite support weighting function. Thus, all feature vectors within a distance τ_{LR} of the query are regarded according to their calculated weight. The weights of all keypoint matches are summed up, and thereby matching references are identified.

For the LR scoring scheme, the parameters τ_{LR} and λ_{LR} need to be determined, which define the shape of the weight function. Subsequently, the query results can be examined. In Figure 5.16 the relative distribution of true- and decoy matches as introduced in Section 4.4.3 is shown along the tri-cube function. From this diagram it is clear, that the tri-cube function is not sufficient to separate true from false matches, because is assigns diminished weights already to true matches. Thus, the support of the tri-cube function is adapted to the interval $[0; 0.6]$, where $\tau_{\mathrm{LR}} = 0.6$ is the absolute threshold that was already specified in Table 4.7 for discriminating true from false matches based on the absolute value of the calculated distance. Again, this function assigns diminished values to true matches, because the slope of the adapted tri-cube function is not sufficiently steep. Thus, the slope of the function is adapted using an exponent of $\lambda_{\mathrm{LR}} = 10$.

The results of the LR scoring are listed in Table 5.8 and Table 5.9. In contrast to the 1-NN′ scoring scheme, the lists have real valued scores and do not always have a clear cut-off. Using the LR scoring scheme, the correctly matching domains carboxyhemoglobin `2D5X` are identified successfully and rank first with scores that are twice as high as for the following matches. The remaining top twenty matches are also hemoglobin domains except for rank 16 and 17, which correspond to domains that also have large amounts of α-helical structure. For DNA gyrase and CcdB depicted in the experimental map of `1X75`, the three references ranked first are the correctly identified domains of `[3VUB:A_]`$_{\texttt{SCOP}}$, `[1AB4:A_]`$_{\texttt{SCOP}}$ [48], and `[1X75:A1]`$_{\texttt{SCOP}}$. However, all matches for this map are very small yielding total scores of 0.0107, 0.0043, and 0.0021. This shows that most correct matches have a feature vector distance that is only slightly less than 0.6. Again, this low matching score can be attributed to the low resolution and high R-factor of `1X75`.

For the synthetic map of `1UMO`, the correctly matching domain of `[1URV:A_]`$_{\texttt{SCOP}}$ is ranked first with a large score of more than 500, while the remaining matches in the list have a total score of less than 108. These scores are very high in comparison to the scores observed for experimental maps, which can be attributed to the depiction of the surrounding solvent. The synthetic



$$t(\eta) = \begin{cases} \left(1-\eta^3\right)^3 & \text{if } |\eta| \leq 1 \\ 0 & \text{else} \end{cases}$$

$$t_{[0;0.6]}(\eta) = \begin{cases} \left(1-\left(\frac{\eta}{0.6}\right)^3\right)^3 & \text{if } |\eta| \leq 0.6 \\ 0 & \text{else} \end{cases}$$

$$w(\eta) = \begin{cases} \left(1-\left(\frac{\eta}{0.6}\right)^{10}\right)^{10} & \text{if } |\eta| \leq 0.6 \\ 0 & \text{else} \end{cases}$$

Figure 5.16 – Parameterization of the LR scoring scheme
The plot shows the weighting of feature vector matches in the LR scoring scheme. The relative distributions of true- and decoy matches as introduced in Section 4.4.3 are shown as dashed lines in light pink and light blue respectively. The tri-cube weight function t is drawn as solid line in green and a tri-cube function $t_{[0;0.6]}$ with a support in the interval $[0; 0.6]$ is shown as solid line in yellow. The utilized weight function w is plotted as solid red line. For all true matches, w assigns weights close to one. The value of w decreases continuously with larger feature vector distance and assigns a weight of zero to decoy matches. (© A. Griewel)

(X) 148L	
1SWY:A_	22.099
1L64:A_	21.576
157L:A_	20.465
1JTM:A_	19.666
2F2Q:A1	18.542
146L:A_	17.777
1T8F:A_	17.340
1P37:A_	14.012
1P5C:A_	13.438
2BPT:A1	2.738
1NVU:S_	2.690
1W27:A_	2.430
1RT8:A_	2.408
1JDH:A_	2.212
1SU7:A_	2.065
1DOX:A2	1.965
1YQS:A1	1.918
1HZ4:A_	1.905
2RDZ:A1	1.890
1CI9:A_	1.867

(X) 1GFL	
1H6R:A_	21.042
1KP5:A_	16.023
1OXD:A_	15.878
1MYW:A_	15.061
1GGX:A_	0.142
1UIS:A_	0.070
1MOU:A_	0.044
1G4M:A2	0.034
1XQM:A_	0.028
1MF7:A_	0.015
1UYN:X_	0.012
2RH7:A1	0.011
1XDP:A2	0.010
1GL4:A1	0.010
2B4W:A1	0.009
1UYR:A2	0.008
3LKF:A_	0.007
1JZ8:A4	0.007
1PJX:A_	0.007
1N2S:A_	0.007

(X) 1PBD	
1KOI:A1	0.002
1W8O:A3	0.002
1HDH:A_	0.001
1FL2:A2	0.001
2FCT:A1	0.000
1K7H:A_	0.000
1VYB:A_	0.000
1NKG:A3	0.000
1UKC:A_	0.000
2QP8:A1	0.000
1FLG:A_	0.000
1HPL:A2	0.000
1T0B:A_	0.000
1JOH:A3	0.000
1NME:,1	0.000
1FWX:A2	0.000
2EBS:A1	0.000
1DKL:A_	0.000
3PNP:A_	0.000
1JUH:A_	0.000

(X) 1T5H	
3CW9:A1	4.701
1RY2:A_	0.014
1LCI:A_	0.009
3CO7:A2	0.005
1RRH:A1	0.004
1PG4:A_	0.001
1JLW:A2	0.001
1YAO:A1	0.001
2ALR:A_	0.001
1MLA:A1	0.000
1MDB:A_	0.000
3ENB:A1	0.000
2GG2:A1	0.000
1XRT:A2	0.000
1CI9:A_	0.000
1JKG:A_	0.000
2D5B:A1	0.000
1SR8:A_	0.000
1T47:A2	0.000
1SU7:A_	0.000

(X) 1UMO	
1URV:A_	0.000

(X) 1W0E	
1TQN:A_	0.176

(X) 1X75	
3VUB:A_	0.011
1AB4:A_	0.004
1X75:A1	0.002
2PNW:A1	0.000
1G9F:A_	0.000
1M47:A_	0.000
1YQ2:A3	0.000
1NE8:A_	0.000
2FCT:A1	0.000
2VK9:A1	0.000
1VAV:A_	0.000
1BHG:A3	0.000
1Y2O:A1	0.000
1HQT:A_	0.000
1LJ5:A_	0.000
1YQ2:A5	0.000
1PDZ:A1	0.000
1HZ4:A_	0.000
2A1H:A1	0.000
1WZL:A3	0.000

(X) 2CG9	
none	

(X) 2D5X	
2D5X:B1	59.220
2D5X:A1	52.512
2QSS:B1	29.429
2QSS:A1	25.535
1WMU:B_	24.857
1FHJ:A_	19.420
1WMU:A_	18.657
1FHJ:B_	17.516
2DN3:B1	16.702
2DN3:A1	16.016
1SHR:B_	15.722
1A4F:B_	15.510
1QPW:A_	13.220
1CH4:A_	12.900
1I3D:A_	12.271
2BPT:A1	11.748
1XU9:A_	10.662
3D1K:A1	10.600
1V4W:A_	10.554
1JEB:A_	10.015

(X) 2DN2	
2QSS:A1	0.596
2DN3:A1	0.427
1SVD:M1	0.288
1WMU:A_	0.273
1QPW:A_	0.236
2V3Q:A1	0.231
1VKF:A_	0.220
1BR2:A2	0.219
2CAR:A1	0.199
1JEB:A_	0.198
2OK5:A1	0.186
1CG5:A_	0.173
1TUA:A2	0.169
1KI9:A_	0.168
1R8K:A_	0.162
1LUC:A_	0.159
1PDA:A1	0.155
1QNT:A1	0.153
1KU1:A_	0.153
1XTT:A1	0.145

Table 5.8 – Molecule recognition results: LR scoring scheme (I)
The table shows the first part of the result lists using the LR scoring scheme. Each query map is denoted by its ID and an identifier of its type (S for synthetic, X for X-ray crystallography, C for cryo-EM). All lists are truncated to 20 entries.

(X) 2EWA		(C) EMD-1461		(C) EMD-5001	(S) 1UMO		(S) 2DN2	
2GFS:A1	1.212	1QHD:A2	0.173	*none*	1URV:A_	543.075	2D5X:A1	606.407
1PME:A_	1.085	1W6S:A_	0.025		2BPT:A1	107.409	2QSS:A1	584.394
1UKH:A_	0.600	1LRW:A_	0.017		1CQX:A1	85.214	1I3D:A_	494.661
1GZ8:A_	0.311	2AD6:A1	0.017		1DOW:A_	80.975	1FHJ:A_	468.825
1OB3:A_	0.278	2COH:A1	0.014		2OV9:A1	74.472	2D5X:B1	452.944
1Q8Y:A_	0.249	1RAO:A2	0.012		2J7Y:A1	74.058	2QSS:B1	451.930
1Q5K:A_	0.159	1QBA:A3	0.012		2H8F:B1	73.078	1WMU:A_	441.592
3BLH:A1	0.149	1SU7:A_	0.010		1F5N:A1	71.653	2DN3:A1	401.654
1UNL:A_	0.117	2NTO:A2	0.010		1FEW:A_	71.055	2H8F:B1	397.686
1CM8:A_	0.104	2QFR:A2	0.008		1V4W:A_	70.796	1FHJ:B_	387.076
1UA2:A_	0.056	1A4M:A_	0.007		2H8P:C1	70.402	1WMU:B_	383.731
3BQC:A1	0.022	1ECE:A_	0.007		1UED:A_	68.718	1XQ5:A_	354.235
1BLX:A_	0.018	1H1N:A_	0.006		2D4C:A1	68.349	1CH4:A_	338.526
1JKS:A_	0.017	1OGY:A2	0.006		2UUI:A1	68.046	3BJ1:B1	332.216
1TKI:A_	0.013	1G01:A_	0.005		1W7J:A2	67.352	3D1K:B1	322.952
1O6Y:A_	0.010	1AUI:A_	0.005		2DI4:A1	66.764	1V4W:A_	322.811
1U59:A_	0.008	1UUQ:A_	0.005		1RE5:A_	66.739	1SHR:B_	318.035
1XBB:A_	0.006	1FWX:A2	0.005		1WA5:C_	64.516	1CG5:A_	308.005
1YWN:A1	0.005	1RBL:A1	0.004		1L2P:A_	64.038	1JEB:A_	305.592
1LUF:A_	0.004	1CXL:A4	0.004		2AY1:A_	63.035	2DN3:B1	298.004

Table 5.9 – Molecule recognition results: LR scoring scheme (II)
The table shows the second part of the result lists using the LR scoring scheme. Each query map is denoted by its ID and an identifier of its type (S for synthetic, X for X-ray crystallography, C for cryo-EM). All lists are truncated to 20 entries.

maps are not distorted by noise, and therefore the surrounding of the protein does not have density. This causes smaller distances between the feature vectors that generally also comprises solvent. For the experimental map of 1UMO, one matching descriptor of the corresponding cytoglobin 1URV is reported, which has a large distance and therefore yields a weight close to zero. Thus, even though one correct match is reported, the docking itself has failed.

For the synthetic map of deoxy hemoglobin 2DN2, the first twenty matches all belong to the correct *globin* family. For the experimental map of this protein complex, the first two matches identify references that correspond to α chains of human and bovine hemoglobin. The reported, total scores are less than one, which indicates that the distances of correctly matched descriptors were close to the threshold $\tau_{\mathrm{LR}} = 0.6$.

The content of the experimental electron density maps of 148L, 1GFL, and 1T5H are identified with a significantly higher vote for the matching protein domains than for other domains. For 1PBD, 1W0E, 2EWA, and EMD-1461 matching domains are also ranked first. However, their total score is low and indicates that all descriptor matches have a large distance only slightly smaller than 0.6. The correctly matching reference domain of 1K0I [356] for the hydroxylase 1PBD is ranked first, but has a marginal score of 0.0023. The second best match for this

map is not similar to 1PBD and has a score 0.0015, which is close to the score of 1K0I. For 1W0E, the correctly matching domain of 1TQN [374] is identified as sole match. The reference list of the protein kinase depicted in 2EWA ranks the representative of the protein, 2GFS [114], first. Other entries in the list are also members of the SCOP family *protein kinases, catalytic subunit*. For the cryo-EM map of EMD-1461 the distal H-domain $\texttt{[1QHD:A2]}_{\texttt{SCOP}}$ [228], which mainly comprises β-strands, is identified in rank one with a score significantly higher than for all other references. The proximal D-domain $\texttt{[1QHD:A2]}_{\texttt{SCOP}}$ is not among the list of the first twenty matches. This is due to the large amount of noise in the D-domain, which can be seen in the lower part of Figure 5.8.

Using the LR scoring scheme, no matches for 2CG9 and EMD-5001 are identified. This means that no descriptor matches with distance of less than 0.6 have been found, and therefore no votes were assigned to the reference domains.

Both scoring scheme have different advantages and drawbacks. LR scoring does not require unique entries in the database since it is not based on distinctive matches. Therefore, this scoring would also be applicable for searching a redundant database of protein structures, such as the set of all available electron density maps. 1-NN′ scoring, however, is more discriminative and also reports if no distinct matches are found. LR scoring generally returns a list of matches that requires interpretation.

To secure the result, two methods for post-processing are proposed. On the one hand, the entries in the reported reference list can be analyzed for consistency. In reference lists without pronounced matches in rank one, the consistency of the votes can be analyzed. This would, e. g., in the case of the protein kinase 2EWA assure the relationship to the *protein kinases, catalytic subunit* family. On the other hand, the matching algorithm can be employed for assuring the content of the result list. Since all keypoints, orientation histograms, and descriptors are saved in the database, a calculation of reference domain placements with respect to the query map is readily available. Using the implemented clustering and scoring, these placements can be assessed for validity by analyzing the consistency of the placements.

Run Time

The accumulated run times of the performed queries are listed in Table 5.10 and show that the current setup requires large amounts of computing time. For most queries, the required run time is larger than 100 days on a single processor — for 2DN2 the runtime is even longer than one year. This is, on the one hand, due to the high dimensionality of the feature vector but, on the other hand, also due to the number of descriptors in the database, which is larger than 128 million

ID	Type	# Keypoints	# Descriptors	Time (days)
1UMO	S	2 426	18 645	232
2DN2	S	4 399	32 051	395
1W0E	X	2 587	13 732	13
2EWA	X	4 362	23 381	127
148L	X	4 503	29 306	131
1GFL	X	7 617	45 908	235
1PBD	X	4 111	24 005	149
1T5H	X	5 246	31 012	160
1UMO	X	3 263	18 852	15
1X75	X	6 247	35 045	24
2CG9	X	4 643	24 094	24
2D5X	X	7 182	47 173	266
2DN2	X	10 525	71 220	415
EMD-1461	C	10 291	70 751	178
EMD-5001	C	12 764	65 499	125

Table 5.10 – Molecule recognition: Map properties and run time
The table lists the properties of the performed queries. Maps are identified by their ID and type (S for synthetic, X for X-ray, and C for cryo-EM). For each map, the number of (#) detected keypoints and descriptors is listed along the required run times, which are specified in days of computing time on a single processor.

entries. Furthermore, the run time of the query also depends on the resolution of the map. If the resolution of the map is low, as for e. g. 1X75 and 2CG9, only keypoints at larger scale are detected. These are compatible to fewer keypoints in the database and thus fewer comparisons are performed.

The high run times make the search in the current setup impractical. However, the problem of similarity searching is reduced to computing distances between feature vectors, which is a well studied problem in computer science. The problem of comparing feature vectors belongs to the class of *embarrassingly parallel problems* [93], which are problems that are easily separated into parallel tasks. Recent advances in parallel computing hardware have allowed for speeding up nearest neighbor searching 300-fold using one single graphics card [105, 106, 201, 202, 165]. Therefore, most of the performed queries can be performed in less than one day of computing using such hardware. Further speedups can be achieved by partitioning the database and allocating the computations to several computers with the corresponding hardware. It can be expected, that the capabilities of parallel processing hardware will improve even more in the next years, which would allow for using the molecule recognition in practice.

Summary

This section showed an implementation of the molecule recognition setup as described in Section 3.6 on page 94. The reference database is equipped with map descriptions computed from synthetic maps of the atomic structures included in the 95 % sequence identity filtered set of SCOP domains. These maps are created at a resolution of 1.73 Å using a sampling interval of 0.5 Å. The database is queried with map descriptions determined from synthetic, X-ray crystallography, and cryo-EM maps and the resulting matches are evaluated using two different scoring schemes.

For synthetic maps, the related protein domains are identified clearly as matches in both scoring schemes. The performance of clearly identifying molecules in experimental maps, however, was not as clear. Even though the matching domains are identified in the first rank for nearly all maps, they are assigned marginal scores. This indicates that the distance between the determined feature vectors for correctly matched descriptors are close to randomly matched descriptors. This is due to several reasons. First of all, the differences in the experimental query map and the synthetic maps of the reference domains can differ significantly. This is due to differences in resolution and general discrepancies between the maps, as measured by the R-factor. These discrepancies prevent the exact location of keypoints in the map. Without exactly located keypoints no descriptor match can be performed. Additionally, the high dimensionality of the feature vector impedes a clear classification in the presence of noise. This is a well studied effect, known as curse of dimensionality [20, 28, 140].

Even though molecule recognition is not yet regularly applicable, it is a promising application. To make *siseek* applicable in practice, the large run time of the approach need to be addressed, by either relying on a smaller set of reference domains, using a descriptor with fewer dimensions, or by employing state of the art hardware for parallel processing. Furthermore additional tests of feature vectors and their correspondences between experimental and synthetic maps are required to improve the scoring schemes. Eventually, means for post-processing and thus assuring correct matches need to be analyzed. Approaches for addressing these issues have been proposed that can also be used in this application [218]. If these issues can be solved, molecule recognition with *siseek* will enable practical, novel means of research in the field of molecular biology and enable large-scale comparisons of genuine electron density maps.

5.4. Summary

In this chapter, *siseek* was used in various experimental setups. In the first section, the general capabilities of *siseek* have been investigated by docking subunits to synthetic maps of their corresponding complexes. The validation showed that the subunits are successfully located in high and intermediate resolution electron density maps, even if noise is present. However, *siseek* is more prone to resolution lowering than the docking tool ADP_EM, which is able to identify the correct placements of subunits even in low resolution maps.

Subsequently, *siseek* was employed to determine registrations of experimental and synthetic maps. First, atomic structures of subunits were docked to experimental cryo-EM maps of their complexes. In six case studies on high and intermediate resolution cryo-EM maps correct superpositions were determined. Subsequently, atomic structures were registered to their corresponding X-ray crystallography maps. This application demonstrated that *siseek* is capable of detecting partially depicted molecules. Eventually, two experimental maps acquired by X-ray crystallography have been registered. The experiments showed that also segments of proteins are sufficient to calculate correct registrations. Furthermore, the docking of maps depicting similar but non-identical proteins was assessed using experimental maps of equine and human carboxyhemoglobin. The experiment showed that small differences in protein sequence are tolerable for the registration if the overall conformation of the proteins is highly similar.

Eventually, molecule recognition was tested using various synthetic and experimental query maps. In the performed test, the reference database consists of the ASTRAL SCOP set version 1.75 with protein domains of less than 95 % sequence identity [54, 55, 39]. Two scoring schemes for interpreting the results are proposed, which both have advantages and drawbacks. However, the method is able to successfully identify matching subunits for most of the performed queries. The molecule recognition setup is a proof of concept and not yet regularly applicable in practice — especially due to large run times. Ways for addressing the identified challenges have been proposed in order to facilitate further research towards this promising application.

6. Conclusion

This work presents an approach for similarity searching in electron density maps coined *siseek*, which has been implemented in a software system. The first part of this chapter comprises a summary of the methods used in *siseek*. Subsequently, the key findings of the validation and the applicability to experimental data are reviewed. This is followed by an outlook, which highlights ways for future research opportunities.

The methods presented in this work are based upon the scale-invariant feature transform (SIFT) and implemented in the software system *siseek*, which computes an abstract representation for macromolecular electron density maps in three steps. First, it detects scale-invariant keypoints, which mark blobs of various size in the analyzed map. Second, keypoints are assigned discrete orientations based on the gradient in the local neighborhood. Third, a descriptor is calculated for each orientation. The size of the descriptor depends on the scale of the keypoint and captures information on the gradient in the local neighborhood of the keypoint. Each descriptor yields a high dimensional feature vector, which is used to compare the keypoints' neighborhoods. The set of the keypoints, orientations, and descriptors yields a map description, which is utilized for the registration of 3D electron density maps. By computing the Euclidean distance between the calculated feature vectors of source- and target map, the similarity of the local neighborhoods can be analyzed. Based on the most similar descriptors, registrations of the maps are computed using the known location and orientation of the descriptors. Additionally, the map description can be used for molecule recognition, i. e., for determining the content of a macromolecular electron density map. In this application, the descriptors of a query map are compared to a set of reference protein descriptors, and the content of the map is identified based on a voting scheme.

The presented work is based upon the SIFT and extends this method to 3D, differing from other approaches. *siseek* relies on the identical keypoint detection algorithm as the SIFT, as scale-space theory is equally applicable to 3D space. It introduces new methods for assigning discrete orientations to a keypoint based on the local gradient field. These methods handle all degrees of freedom of 3D space properly, since they employ a uniform geodesic grid sampling the sphere surface. Furthermore, a geodesic index has been developed, which facilitates

the efficient processing of gradient vectors. Descriptors, which are calculated for each orientation, have been designed inspired by the 2D SIFT descriptor, but allowing for several parameters to be set. This enables a detailed analysis of the descriptor properties, which is carried out in the validation.

A validation of all stages of *siseek* — keypoint detection, orientation assignment, and descriptor calculation — was performed separately in several experiments based on synthetic maps of proteins. During these tests, optimal parameters for all objects used in the method have been determined, which allows for the identification of similarities in both synthetic and experimental maps. The basic keypoint detection experiments show that repeatability rates of up to 74 % are achieved. Equally, the repeatability of orientation assignment to keypoints was assessed yielding a rate of more than 90 %. Also, the utilized descriptors and feature vectors were assessed in detail. The descriptor parameters and their effect on the calculated distances was investigated. Based on these finding, reasonable parameters were selected that successfully discriminate true from false matches with a recall of more than 97 % and a precision more than of 88 % in the performed experiments.

The performance of *siseek* was first assessed by registering synthetic and experimental electron density maps. First, synthetically generated maps of different resolutions and signal-to-noise ratios are used. It was shown that *siseek* successfully docks atomic structures to maps as low as 7.5 Å resolution with an RMSD of less than 1 Å. For these tasks *siseek* requires less time than other computer programs geared to solving the same problem. In the performed experiments, *siseek* was on average an order of magnitude faster since it relies only on the content and the resolution of the map and not on the provided sampling interval. Thus, the experiments on synthetic maps showed that *siseek* is capable of successfully identifying similar volumes in high and intermediate resolution maps. The failures in low resolution maps are mainly due to the keypoint detection step, which fails if structural detail is insufficiently depicted in the maps.

In the results chapter, typical experimental maps acquired by cryo-EM and X-ray crystallography are used. In the first experiments, atomic structures have been docked to high resolution cryo-EM maps with resolutions as low as 8 Å. For all maps, the correct position of the subunits was determined showing that *siseek* performed well on the test cases. For X-ray crystallography maps, it was shown that also partially depicted atomic structures are successfully identified. Thus, it is possible to register maps based on similar sub-volumes, rather than having to consider the complete electron density map. Eventually, two experimental X-ray crystallography maps were registered. *siseek* successfully identified similar

subvolumes and computed registrations accordingly. This application allows for the analysis of similarities and differences between the maps. The alignments can reveal inconsistencies between two experimental maps and thereby highlight differences between the measured data. This is also true for a registration of a synthetic map with an experimentally determined X-ray crystallography map.

In a second experimental setup, a molecule recognition experiment was carried out. In this test, the content of experimental electron density maps is matched to a database of reference structures, which consists of the ASTRAL SCOP set version 1.75 with protein domains of less than 95 % sequence identity [54, 55, 39]. For molecule recognition, a high dimensional feature vector is selected, which allows for the identification of similar subvolumes. The database is queried using the map description of the electron density map. Subsequently, relevant references are identified in a voting procedure, which is based on the computed descriptor matches. *siseek* was tested using synthetic and experimental maps. For the noiseless synthetic maps, the correctly matching domains are identified clearly. For experimental maps, however, the votes were not as clear. This is due to the noise in the maps and also to the high dimensional feature vector, which especially susceptible to noise. However, for the large majority of the maps, the correct domains were identified. Besides this, the application requires large computing times. An approach for accelerating the computation through indexing and parallel computation has been proposed. Based on a more efficient setup, further research can allow for a more detailed analysis of the application. Thus, it is conceivable that further research and progress in hardware will facilitate an efficient molecule recognition setup.

Reasons for success and failure have been carefully analyzed in order to characterize the performance of *siseek*. In summary, the objective of developing a method for similarity searching in macromolecular electron density maps and demonstrating the applicability to experimental data has been reached. This was shown in two experimental scenarios: electron density map registration and molecule recognition. For the first application, electron density map registration, *siseek* can be used on intermediate and high resolution electron density maps. It was demonstrated in experiments that intermediately sized proteins are reliably identified for resolutions as low as 7.5 Å, even if noise is present. However, the performance of *siseek* on low resolution electron density maps remains limited and therefore docking to these maps cannot be supported. Furthermore, a proof of concept for molecule recognition in macromolecular electron density maps using *siseek* has been presented. The current implementation of this application requires large run times and was validated only on a small set of proteins and therefore further research on this application is needed.

In the chapter on results of this work, several applications of *siseek* have been outlined. However, there are several topics worth investigating with respect to future research, which are summarized in the following. These include, on the one hand, ways for improving the method itself and thus improving the results of registration and molecule recognition. On the other hand, new applications of *siseek* are listed since the method can, in theory, also be used for registering other types of 3D images.

- In the last section of the chapter on results, a setup for molecule recognition in electron density maps was proposed. The current implementation of the method requires large run times and is therefore impracticable. Fortunately, most of the calculations carried out in *siseek* belong to the class of embarrassingly parallel problems [93]. This does not only hold for feature vector comparison but also for computations performed during keypoint detection, orientation assignment, and descriptor calculation. All these computations can be accelerated largely using modern parallel processing hardware as explained in the discussion of the molecule recognition experiment.

- Currently, *siseek* uses a command line interface in combination with a static molecular viewer [269]. Based on the modularized implementation of *siseek*, an incorporation of the software into interactive molecular viewers is possible with little effort. This would, e. g., enable the integration of the calculated placements with other bioinformatics programs and allow for an interactive adjustment of the results. Using a highly efficient, optimally hardware-supported implementation for the parallel calculations of *siseek*, significantly lower run times can be expected allowing for an interactive interpretation of the macromolecular electron density maps.

- Both descriptor parameter sets yield high dimensional feature vectors — the descriptor used for registration is 84-dimensional and the descriptor used for database searching even has 1 134 dimensions. A further analysis of the descriptor properties may give new insights into frequently observed structures in proteins. Using dimensionality reduction [170], the true information content of the feature vectors can be analyzed and smaller feature vectors could be used. Furthermore, a closer analysis of the descriptor distinctiveness with respect to the large amount of reference feature vectors is worth of investigation: Based on the sampled distance and distinctiveness distributions introduced in Section 4.4.3 on page 130, it is possible to develop an algorithm for automatically determining optimal descriptor parameters and matching thresholds. Given this algorithm — in combination

with a high-performance setup for descriptor comparison — a method for scanning a wide range of descriptor parameters can be developed. This, in turn, allows for specifying objectives, e. g., minimization of the run time or the maximization of the recall and precision rate, and then to select the parameter set which best fits these objectives.

– The map descriptions are employable not only for similarity searching in a pair of synthetic- and experimental map but also for a pair of two experimental maps, as demonstrated by registering X-ray crystallography maps of DNA gyrase and hemoglobin in Section 5.2. Using this finding for molecule recognition, a direct search of structural features in the experimental data can be facilitated by setting up the reference database with experimental maps. One application of this setup is the identification of similar subvolumes between the experimental maps. This application would allow for new findings in the field of biophysics by identifying the similarities and differences of electron density maps in an automated manner. Furthermore, it could allow for the identification of given protein motives in the set of experimental reference maps. This can be facilitated by designing the sought structural motive in silicio and using it as query protein. In comparison to the atom-based classifications of protein structures, this application operates on the experimentally measured electron density maps and does not rely on atomic interpretations.

– Further investigations of the stages of the method are going to be a topic for future research. For orientation assignment, high repeatability rates have been achieved, and for the computed descriptors a high discriminative power was shown. However, for keypoint detection only repeatability rates of up to 74 % have been achieved. Keypoint detection lies at the heart of the *siseek* since keypoints are used for identifying salient features and form the basis for orientation assignment as well as descriptor computation. Increasing this rate will improve the recognition performance and decrease the required run time.

– Further investigations of the resolution model in structural biology can lead to a higher success rate. *siseek* relies on real space, i. e., spatial features of electron density maps. Synthetic maps are created using a standard modeling for resolution lowering as described in Section 3.1, i. e., the convolution with a Gaussian function. This treatment of resolution is only sufficient if the resolution is lowered by independent isotropic movements of each atom. However, especially for maps with low resolutions this is not the case, but large, concerted motions of protein subunits are responsible

for the lack of high frequency information. Furthermore, the resolvability of spatial features in different regions of a map may differ, e. g., if certain parts of the protein are more rigid, while other parts show more flexibility and are thus less well resolved. On the one hand, the scale-space approach in combination with the knowledge of protein structure features could allow for the development of a new criterion to estimate the resolvability of objects in certain regions of a map. An experimental setup for assessing this criterion requires a thorough simulation pipeline, which also supplies ground truth, and proteins with known internal kinetics. On the other hand, *siseek* would benefit from a detailed description of local resolvability. This description would allow for the application of locally adapted filters and therefore enable a more thorough detection of keypoints on experimental maps.

– A major challenge of structural biology is going to be the integration of data from different sources and scales. Today, it is possible to analyze the atomic structure of macromolecules and their assemblies on a fine scale using methods like X-ray crystallography or cryo-EM. Furthermore, methods for the analysis of the spatial organization of molecules in complete cells on a larger scale are an active topic of current research. Combining information from these scales will yield new insights into the mechanisms underlying life. A major opportunity of the scale-space approach lies in its intrinsic handling of objects of different size. However, the currently recorded whole-cell images are extremely noisy and have low resolution from the standpoint of molecular biology. Based on the tremendous advancements in the field of cryo-EM and the overall improvement of imaging techniques, a belief in cell images that are high in both quality and resolution seems justified. Considering the high complexity of current electron density maps, which depict solely single molecular structures, it can be foreseen that these maps are going to be very large in size and will elude themselves from manual interpretation. Therefore, software for image analysis at different scales, such as *siseek*, will be needed to aid the expert scientist in the analysis of experimental data, and also for the automated interpretation of data.

– In this work, *siseek* is applied to data acquired in X-ray crystallography and cryo-EM experiments. In principle, *siseek* could equally be applied to maps acquired by new techniques such as, e. g., X-ray free electron lasers (XFEL). Furthermore, *siseek* is based on the generic SIFT and therefore, in theory, not limited to macromolecular electron density maps but

generally applicable to 3D image data. It will be interesting to assess new parameterizations for applying the method to other 3D image data — such as medical images — and to assess *siseek*'s performance on objects other than macromolecules.

The presented method *siseek* provides means for similarity searching in electron density maps by applying state of the art techniques from the field of image analysis. It allows for detecting similar regions in high- and intermediate resolution electron density maps using less run time than other programs, and thus enables the efficient identification of correspondences in maps. *siseek* computes a map representation, which consists of keypoints, orientations, and descriptors and thereby captures the content of a map. It explicitly addresses both, X-ray crystallography and cryo-EM maps, and supplies new means for their interpretation by combining the principles of scale-space theory with the resolution measure found in both X-ray crystallography and cryo-EM. *siseek* allows for the docking of atomic structures to experimental cryo-EM and X-ray crystallography maps and provides means for the registration of two maps depicting similar proteins. The latter allows for the comparison of two experimental maps and thus supplements the means for analyzing the macromolecular electron density maps. Thus, *siseek* can be used for registering intermediate and high resolution electron density maps enabling novel ways of research in the fields of bioinformatics as well as molecular and structural biology.

A. Appendix

A.1. Mathematical Notation

Three-dimensional vectors and voxels are denoted by bold symbols

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{A.1}$$

The Euclidean norm of a vector is denoted as

$$|\mathbf{x}| = \sqrt{x^2 + y^2 + z^2} \tag{A.2}$$

The Euclidean distance of two vectors $\mathbf{x}$ and $\mathbf{y}$ is calculated as

$$|\mathbf{x} - \mathbf{y}| = \left| \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \right| = \sqrt{\sum_{i=1}^{3} (x_i - y_i)^2} \tag{A.3}$$

The angle $\sphericalangle$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is calculated as

$$\sphericalangle(\mathbf{x}, \mathbf{y}) = \arccos\left(\frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|} \right) \tag{A.4}$$

Intervals are denoted using square brackets and semicolons. Thus, the closed interval between a and b is denoted by $[a; b]$, the left half-open interval by $]a; b]$, the right half-open interval by $[a; b[$, and the open interval by $]a; b[$. Sets are denoted using curly braces and semicolons such as $\{a; b; c\}$.

The two functions $\arg\min(f(x))$ and $\arg\max(f(x))$ yield the set of local minima and maxima of the function f. In this work, arg min and arg max are used to identify voxels in images that have an intensity value that is either larger or smaller than the voxels in the 6-neighborhood. This neighborhood includes all voxels that have a distance of one sampling interval to the considered voxel.

The partial derivatives with respect to multi-dimensional functions need to be calculated for the detection of keypoints. The notation for derivatives is defined in Section 2.1 on page 11.

A.2. Atomic Structure and Electron Density Map Identifiers

Throughout this work, atomic structures are referenced by their wwPDB [23] identifier consisting of four alphanumeric, capital characters such as `4DFR`. Specific chains of a structure contained in a wwPDB file are referenced using the

chain identifiers such as `[1AW5:A]` or `[1L1F:A-F]`. Furthermore, SCOP [247, 66, 6, 7] domains are identified by their wwPDB ID, a character identifying the chain, and another character identifying the domain along a subscript "SCOP" such as `[3VUB:A_]`$_{\texttt{SCOP}}$. CATH [254] domains are identified by the wwPDB ID, the chain ID, and a two-digit domain number along a subscript "CATH" such as `[1B25:A01]`$_{\texttt{CATH}}$. In lists that relate only to one database the subscripts CATH and SCOP may be dropped.

Electron density maps are referenced by their corresponding identifiers. For synthetic maps and experimental X-ray crystallography maps, the wwPDB identifier is used. Entries from the EMDataBank [196] are referenced with their 4-digit number and the prefix "EMD" such as `EMD-5001`.

A.3. Laplacian of Gaussian

The Laplacian of Gaussian in 3D results from the application of the Laplacian operator to the Gaussian function and is therefore the sum of all unmixed second-order partial derivatives of the Gaussian function. The 3D normalized isotropic Gaussian function with standard deviation σ for a variable $\mathbf{x} = (x, y, z)^T$ is defined as

$$G(\mathbf{x}; \sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} \cdot e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \tag{A.5}$$

The first-order partial derivative of Equation A.5 with respect to one of the three dimensions $a \in \{x, y, z\}$ is given in Equation A.6 and the unmixed second-order partial derivative in Equation A.7.

$$G_a(\mathbf{x}; \sigma) = \frac{\partial G}{\partial a}(\mathbf{x}; \sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} \cdot \left(-\frac{a}{\sigma^2}\right) \cdot e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \tag{A.6}$$

$$G_{aa}(\mathbf{x}; \sigma) = \frac{\partial^2 G}{\partial a^2}(\mathbf{x}; \sigma) = \frac{1}{(2\pi)^{\frac{3}{2}}\,\sigma^5} \cdot \left(\frac{a^2}{\sigma^2} - 1\right) \cdot e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \tag{A.7}$$

The Laplacian ∇^2 of the 3D isotropic Gaussian function is defined as the sum of the three unmixed second-order partial derivatives

$$\nabla^2 G(\mathbf{x}; \sigma) = \sum_{a \in \{x,y,z\}} \frac{\partial^2 G}{\partial a^2}(\mathbf{x}; \sigma) = \frac{1}{(2\pi)^{\frac{3}{2}}\,\sigma^5} \cdot \left(\frac{|\mathbf{x}|^2}{\sigma^2} - 3\right) \cdot e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} \tag{A.8}$$

A.4. Geodesic Grid Properties

A geodesic grid is created by subdividing an icosahedron. The geodesic grid of subdivision level 0 corresponds to the genuine icosahedron, which comprises 12 vertices, 30 edges, and 20 faces. When subdividing a given geodesic grid — called the genuine grid in the following — new vertices are created at the center of each genuine edge. These newly created vertices are translated outwards along a straight line between the icosahedron center and the vertex so that they lie on the circumsphere of the genuine icosahedron. The new vertices are connected to the genuine vertices that lie on the same genuine edge. Furthermore, all new corners that lie on a genuine face are connected by an edge. This splits each face into four new faces during subdivision as shown in Figure 3.6 on page 75.

Using this description, recurrence relations for the number of vertices V_g, the number of edges E_g, and the number of faces F_g for subdivision level $i \in \{0; 1; 2; \ldots\}$ are defined in Equation A.9, Equation A.10, and Equation A.11.

$$V_\mathrm{g}(0) = 12 \qquad V_\mathrm{g}(i) = V_\mathrm{g}(i-1) + E_\mathrm{g}(i-1) \mid i > 0 \tag{A.9}$$

$$E_\mathrm{g}(0) = 30 \qquad E_\mathrm{g}(i) = 2 \cdot E_\mathrm{g}(i-1) + 3 \cdot F_\mathrm{g}(i-1) \mid i > 0 \tag{A.10}$$

$$F_\mathrm{g}(0) = 20 \qquad F_\mathrm{g}(i) = 4 \cdot F_\mathrm{g}(i-1) \mid i > 0 \tag{A.11}$$

Closed forms for these relations are

$$V_\mathrm{g}(i) = 4^i \cdot 10 + 2 \qquad E_\mathrm{g}(i) = 4^i \cdot 30 \qquad F_\mathrm{g}(i) = 4^i \cdot 20 \tag{A.12}$$

They have been derived in three steps. For the number of faces F_g, the closed form in Equation A.13 is obtained directly from the recurrence relation in Equation A.11.

$$F_\mathrm{g}(i) = 4^i \cdot 20 \tag{A.13}$$

The closed form for number of edges E_g is derived in Equation A.14. It depends on both the number of faces F_g and the number of edges E_g in the preceding subdivision level $i - 1$.

$$\begin{aligned} E_\mathrm{g}(i) &= 2 \cdot E_\mathrm{g}(i-1) + 3 \cdot F_\mathrm{g}(i-1) \\ &= 2^2 \cdot E_\mathrm{g}(i-2) + 2^1 \cdot 3 \cdot F_\mathrm{g}(i-2) + 2^0 \cdot 3 \cdot F_\mathrm{g}(i-1) \\ &= 2^k \cdot E_\mathrm{g}(i-k) + 3 \cdot \sum_{j=0}^{k-1} 2^j \cdot F_\mathrm{g}(i-j-1) \\ &= 2^i \cdot E_\mathrm{g}(0) + 3 \cdot \sum_{j=0}^{i-1} 2^j \cdot F_\mathrm{g}(i-j-1) \end{aligned} \tag{A.14}$$

$$
\begin{aligned}
&= 2^i \cdot E_{\mathrm{g}}(0) + 3 \cdot \sum_{j=0}^{i-1} 2^j \cdot 4^{i-j-1} \cdot 20 \\
&= 2^i \cdot 30 + 60 \cdot \sum_{j=0}^{i-1} 2^{2i} \cdot 2^{-j} \cdot 2^{-2} \\
&= 2^i \cdot 30 + 15 \cdot 4^i \cdot \sum_{j=0}^{i-1} \left(\frac{1}{2}\right)^j \\
&= 2^i \cdot 30 + 15 \cdot 4^i \cdot \frac{1 - \left(\frac{1}{2}\right)^i}{1 - \frac{1}{2}} \\
&= 2^i \cdot 30 + 30 \cdot 4^i \cdot \left(1 - \left(\frac{1}{2}\right)^i\right) \\
&= 30 \cdot \left(2^i + 4^i - 4^i \cdot \left(\frac{1}{2}\right)^i\right) \\
&= 30 \cdot \left(2^i + 4^i - 2^i\right) \\
&= 4^i \cdot 30
\end{aligned}
$$

A closed form for the number of vertices V_{g} in subdivision level i can be calculated using Euler's formula [69], which specifies the relationship between the number of faces, edges, and vertices in convex polyhedra. Using the defined functions, Euler's formula is given in Equation A.15.

$$2 = V_{\mathrm{g}}(i) - E_{\mathrm{g}}(i) + F_{\mathrm{g}}(i) \tag{A.15}$$

Thus, the number of vertices V_{g} in subdivision level i can be calculated as shown in Equation A.16.

$$
\begin{aligned}
V_{\mathrm{g}}(i) &= 2 + E_{\mathrm{g}}(i) - F_{\mathrm{g}}(i) \\
&= 2 + 4^i \cdot 30 - 4^i \cdot 20 \\
&= 4^i \cdot 10 + 2
\end{aligned} \tag{A.16}
$$

A list of the number of vertices, edges, and faces in the geodesic grid for subdivision levels 0–8 is found in Table A.1. Furthermore, the table lists the minimal and maximal angles that are subtended by the geodesic grid edges. A plot of these values is found in Figure 3.7 on page 76.

Level	#Corners	#Edges	#Faces	Min. ∢	Max. ∢
0	12	30	20	63.44	63.44
1	42	120	80	31.72	36.00
2	162	480	320	15.86	18.70
3	642	1 920	1 280	7.93	9.44
4	2 562	7 680	5 120	3.97	4.73
5	10 242	30 720	20 480	1.98	2.37
6	40 962	122 880	81 920	0.99	1.18
7	163 842	491 520	327 680	0.50	0.59
8	655 362	1 966 080	1 310 720	0.25	0.30

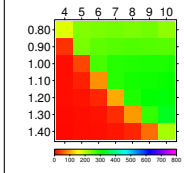
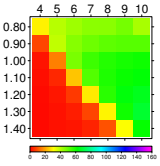
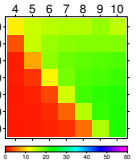
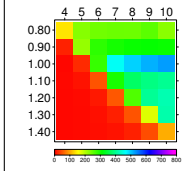
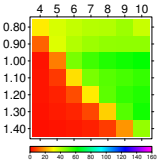
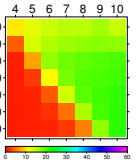
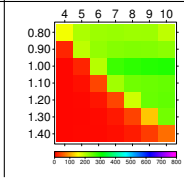
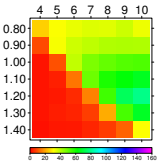
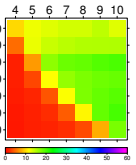
Table A.1 – Geodesic grid properties
The number of (#) corners, edges, and faces is listed for subdivision level 0–8 of the geodesic grid. Furthermore, the minimal and maximal angle that is subtended by an edge of the grid is specified in degree.

A.5. Keypoint Repeatability

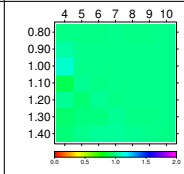
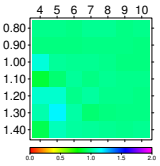
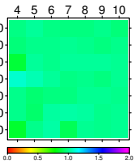
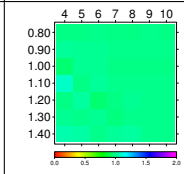
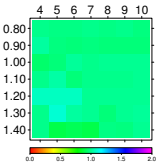
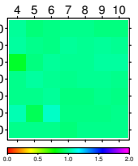
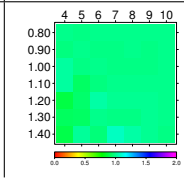
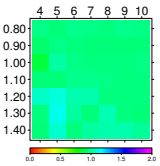
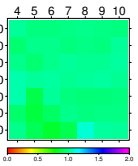
The following pages show the achieved repeatability rates for all experiments detailed in Section 4.2 on page 106 and Section 4.3 on page 115. The heat maps are rainbow colored and show the detected number of keypoints, the excess ratio, the repeatability by distance, and the repeatability by orientation for various internal parameters in different external conditions. An "A" in the diagrams denotes a value that is above the scale. The internal parameters are the number of intervals s for each octave, which is shown on the horizontal axis of each heat map, and the initial width of the point spread function $\{\sigma_0\}_{\text{vox}}$, which is shown on the vertical axis of each heat map. Furthermore, different initial voxel spacings (VS) and resolutions are assessed, as indicated by the rows and columns of the table. The SNR is varied as indicated by the headings of each diagram. Further details on the experiments are found in Chapter 4 on page 101

Without Added Noise

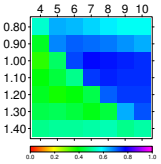
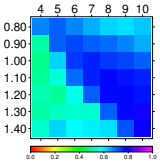
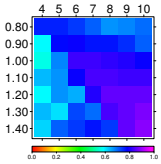
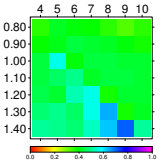
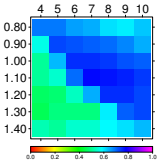
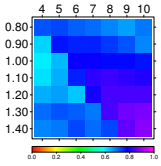
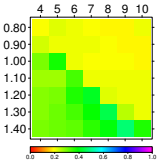
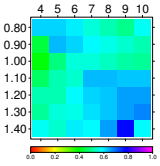
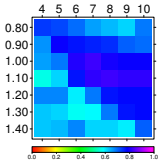
Number of Keypoints (no added noise)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

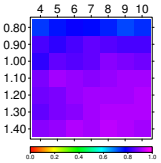
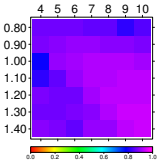
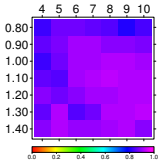
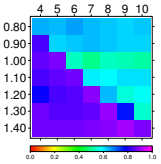
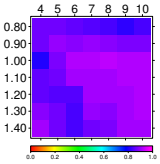
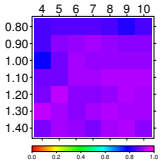
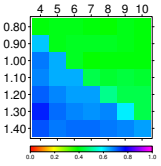
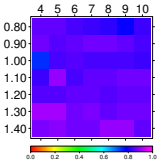
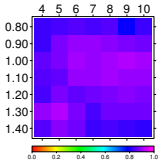
Excess Ratio (no added noise)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Distance (no added noise)

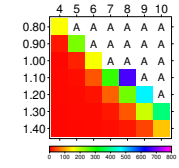
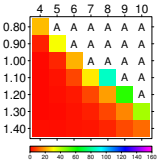
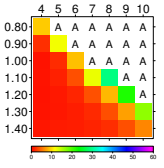
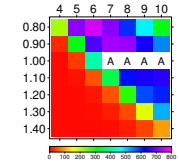
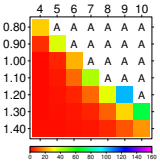
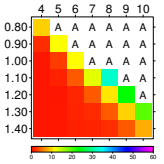
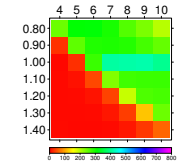
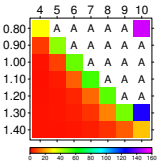
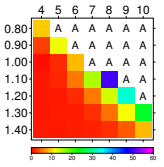
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Orientation (no added noise)

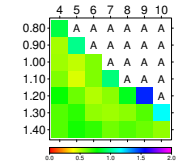
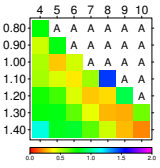
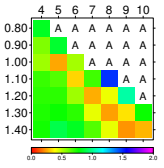
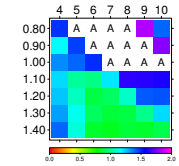
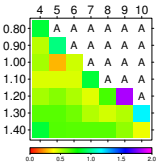
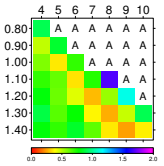
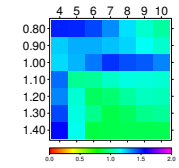
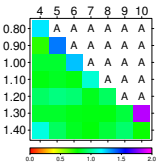
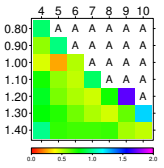
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Signal to Noise Ratio 5.0

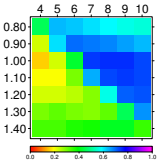
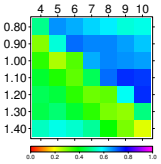
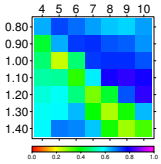
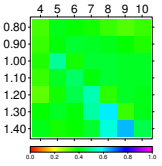
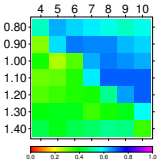
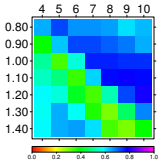
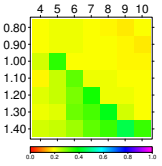
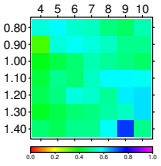
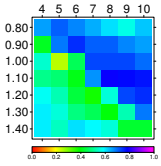
Number of Keypoints (SNR 5.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

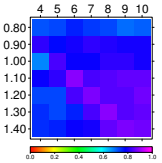
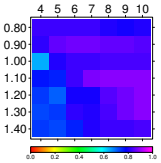
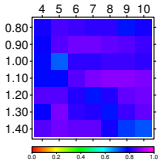
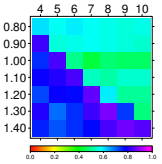
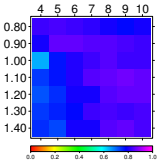
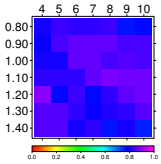
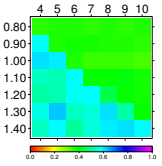
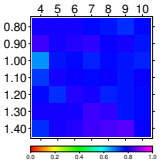
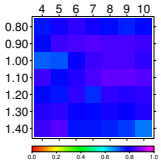
Excess Ratio (SNR 5.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Distance (SNR 5.0)

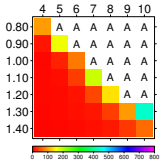
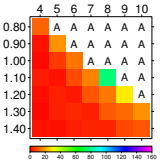
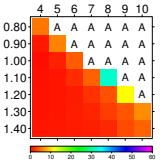
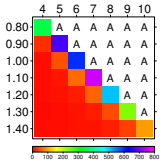
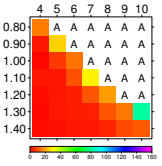
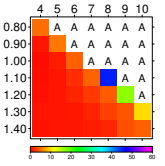
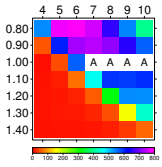
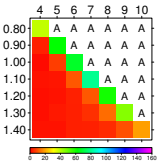
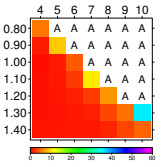
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Orientation (SNR 5.0)

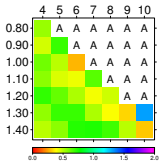
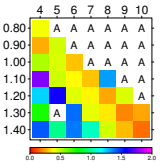
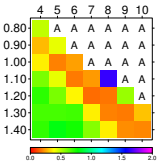
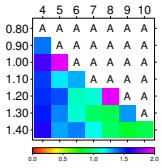
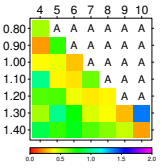
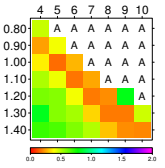
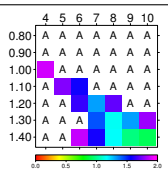
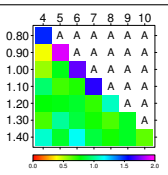
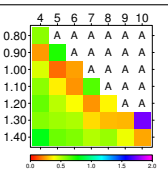
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Signal to Noise Ratio 2.0

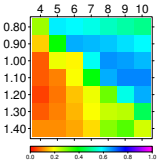
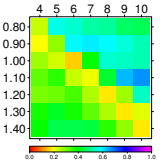
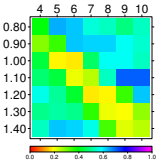
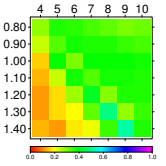
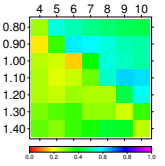
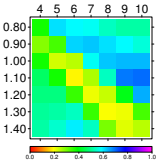
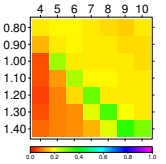
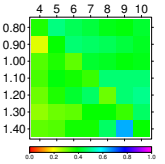
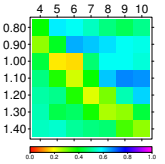
Number of Keypoints (SNR 2.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

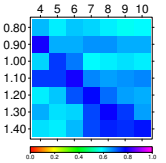
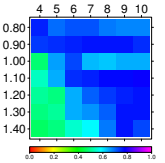
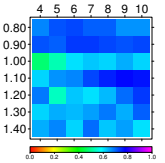
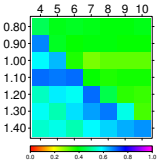
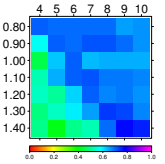
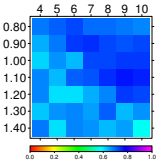
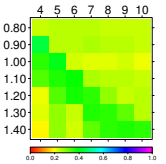
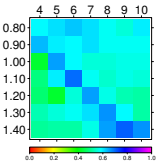
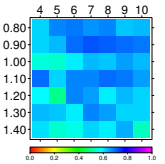
Excess Ratio (SNR 2.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Distance (SNR 2.0)

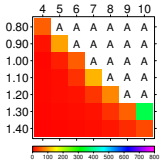
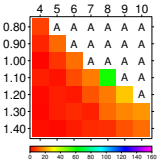
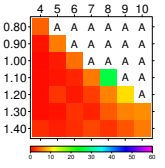
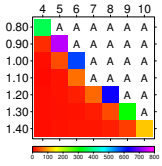
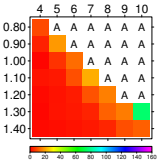
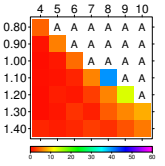
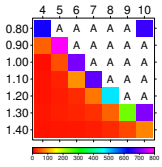
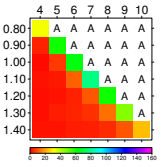
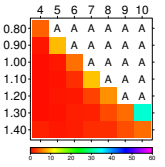
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Orientation (SNR 2.0)

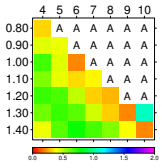
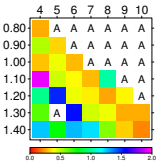
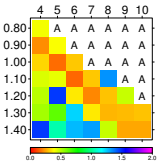
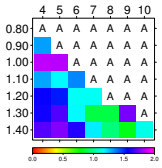
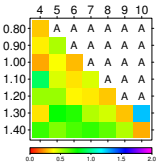
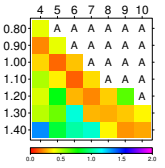
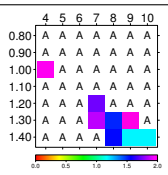
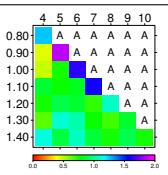
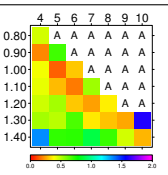
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Signal to Noise Ratio 1.0

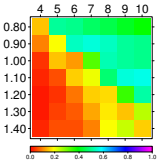
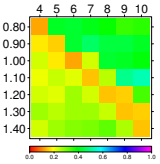
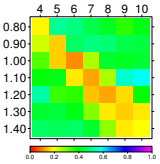
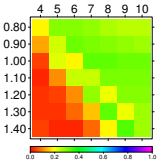
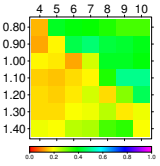
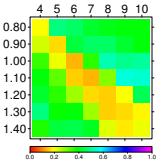
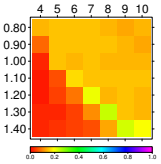
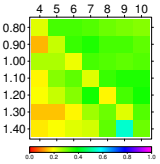
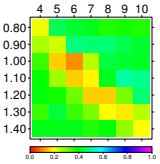
Number of Keypoints (SNR 1.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

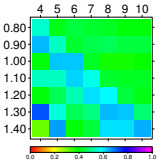
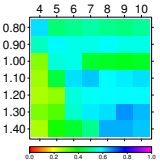
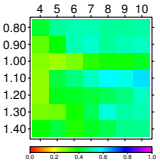
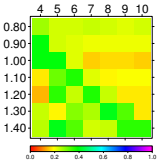
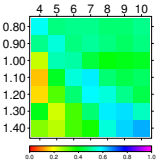
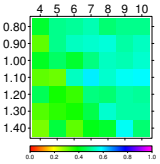
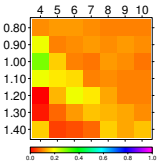
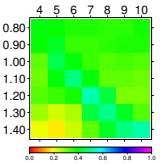
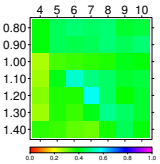
Excess Ratio (SNR 1.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Distance (SNR 1.0)

	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

Repeatability by Orientation (SNR 1.0)

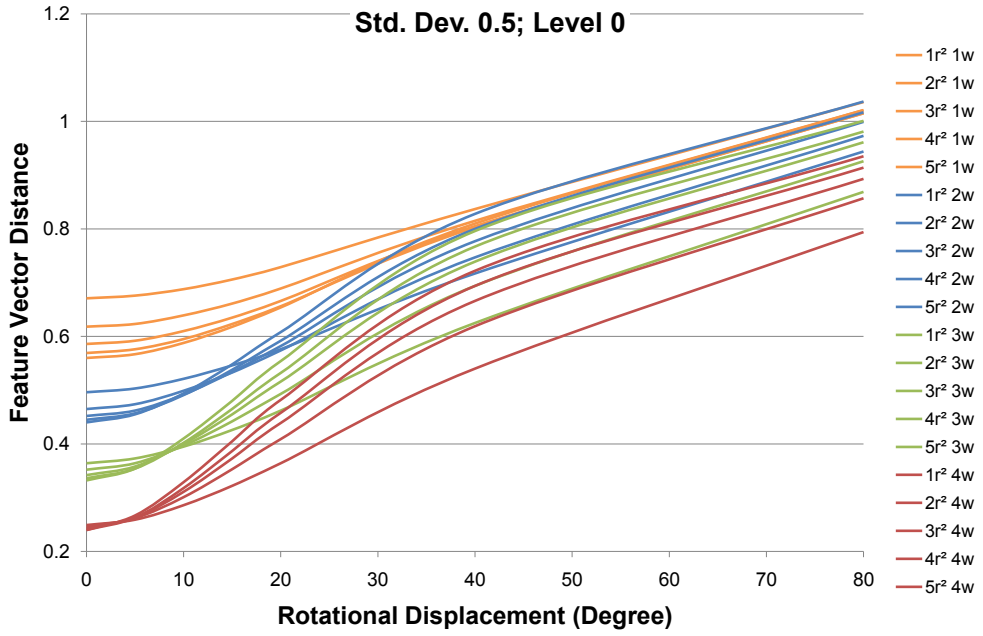
	Resolution 3.5 Å $\sigma = 1$ Å	Resolution 6.9 Å $\sigma = 2$ Å	Resolution 10.4 Å $\sigma = 3$ Å
VS 1 Å			
VS 2 Å			
VS 3 Å			

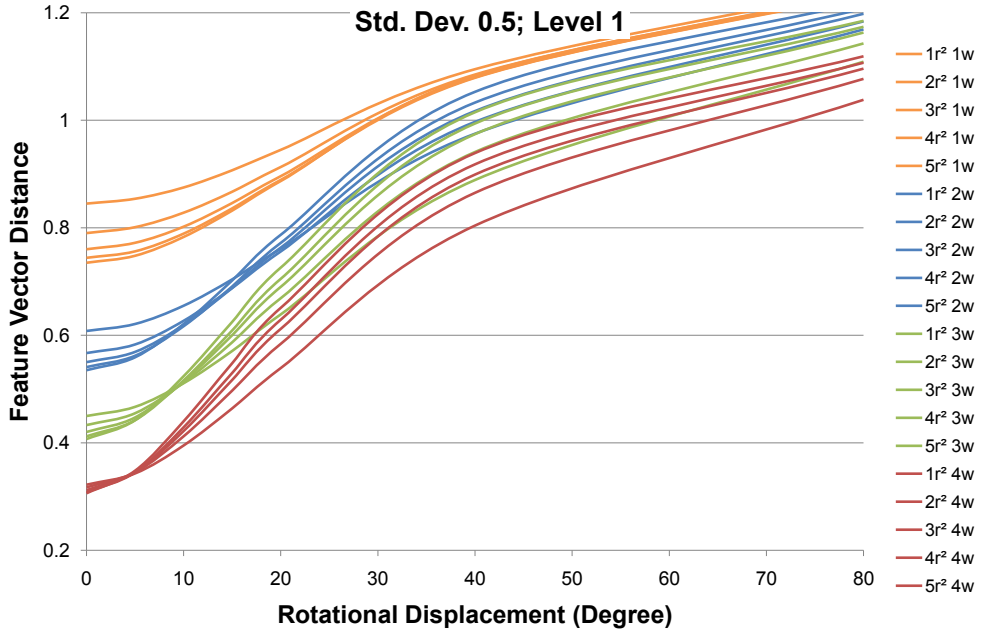
A.6. Descriptor Analysis

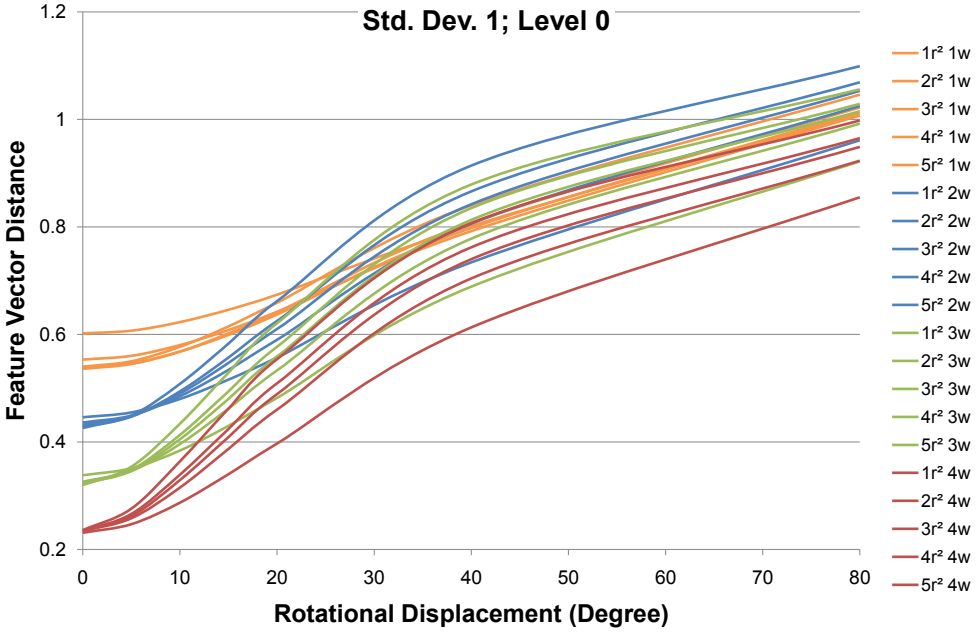
In the following, diagrams and tables are listed that supplement the analysis of the descriptor properties with respect to different parameter sets.

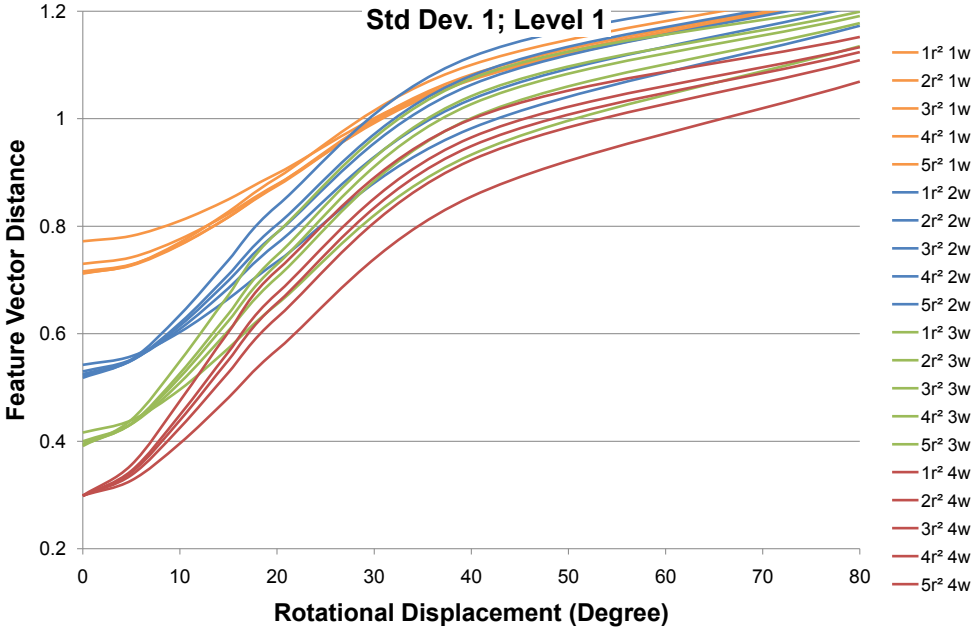
A.6.1. Robustness to Distortions

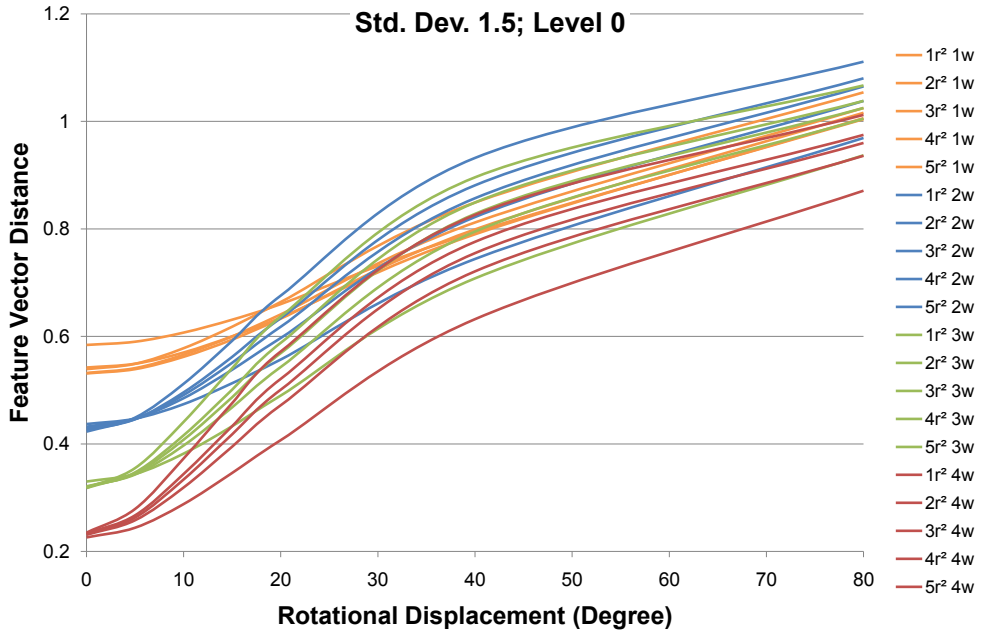
The change in feature vector distance under random displacement was analyzed with respect to the amount of rotation in Section 4.4.1 on page 120. Here, additional diagrams showing the measured distance when using different values for the weighting σ_d of the gradient vectors are shown. The value for σ_d and g are shown in the diagram headings. The width of the desriptor δ is indicated by the color of the line and denoted as w in the legend. The value for r^2, denoted as r^2 in the legend, is not indicated in the plot for clearer view. However, it can be determined by the slope of the curve: The steeper the curve, the larger r.

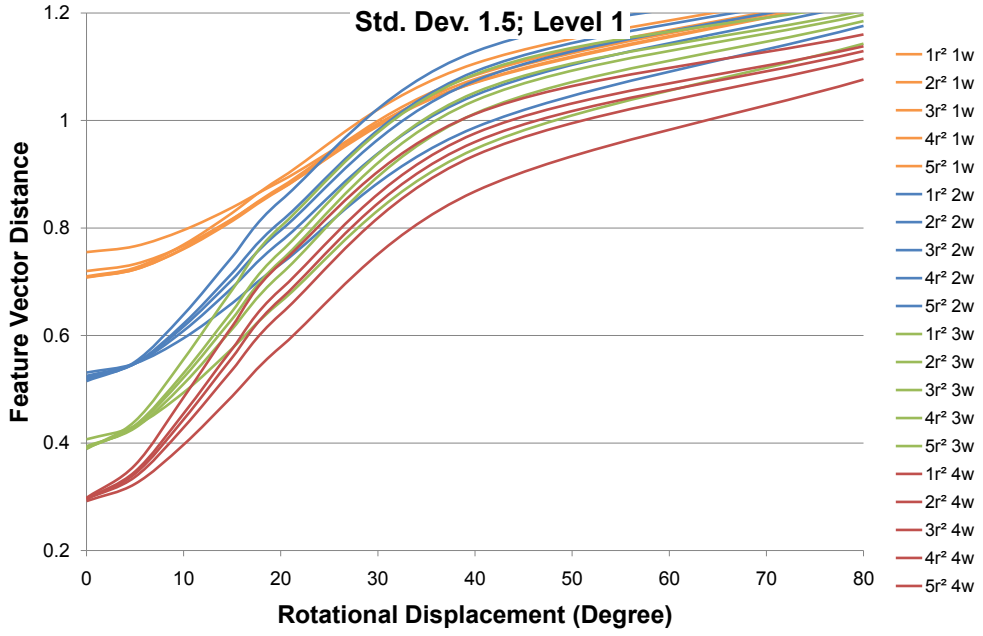

Std. Dev. 0.5; Level 0
Feature Vector Distance
Rotational Displacement (Degree)
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
1r² 1w
2r² 1w
3r² 1w
4r² 1w
5r² 1w
1r² 2w
2r² 2w
3r² 2w
4r² 2w
5r² 2w
1r² 3w
2r² 3w
3r² 3w
4r² 3w
5r² 3w
1r² 4w
2r² 4w
3r² 4w
4r² 4w
5r² 4w


Std. Dev. 0.5; Level 1
Feature Vector Distance
Rotational Displacement (Degree)
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
1r² 1w
2r² 1w
3r² 1w
4r² 1w
5r² 1w
1r² 2w
2r² 2w
3r² 2w
4r² 2w
5r² 2w
1r² 3w
2r² 3w
3r² 3w
4r² 3w
5r² 3w
1r² 4w
2r² 4w
3r² 4w
4r² 4w
5r² 4w


Std. Dev. 1; Level 0
Feature Vector Distance
Rotational Displacement (Degree)
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
1r² 1w
2r² 1w
3r² 1w
4r² 1w
5r² 1w
1r² 2w
2r² 2w
3r² 2w
4r² 2w
5r² 2w
1r² 3w
2r² 3w
3r² 3w
4r² 3w
5r² 3w
1r² 4w
2r² 4w
3r² 4w
4r² 4w
5r² 4w


Std Dev. 1; Level 1
Feature Vector Distance
Rotational Displacement (Degree)
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
1r² 1w
2r² 1w
3r² 1w
4r² 1w
5r² 1w
1r² 2w
2r² 2w
3r² 2w
4r² 2w
5r² 2w
1r² 3w
2r² 3w
3r² 3w
4r² 3w
5r² 3w
1r² 4w
2r² 4w
3r² 4w
4r² 4w
5r² 4w

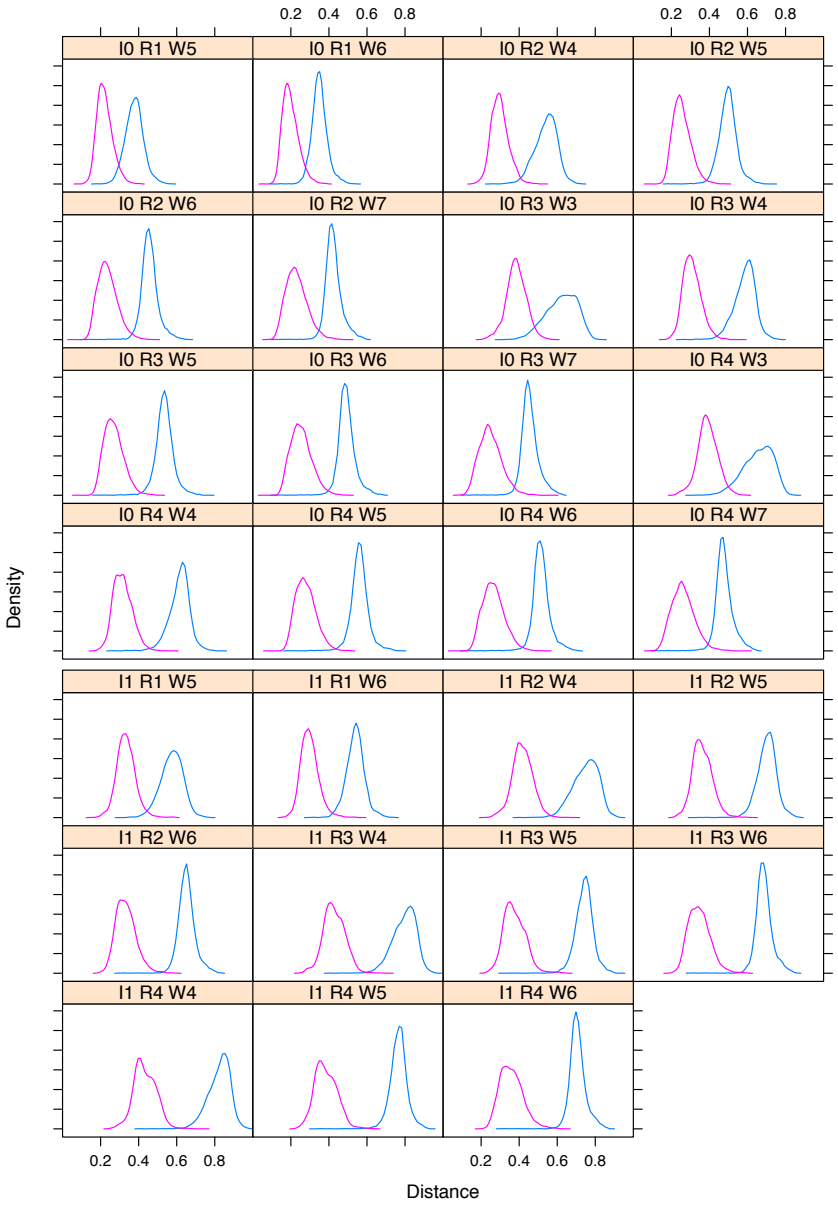

Std. Dev. 1.5; Level 0
Feature Vector Distance
Rotational Displacement (Degree)
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
1r² 1w
2r² 1w
3r² 1w
4r² 1w
5r² 1w
1r² 2w
2r² 2w
3r² 2w
4r² 2w
5r² 2w
1r² 3w
2r² 3w
3r² 3w
4r² 3w
5r² 3w
1r² 4w
2r² 4w
3r² 4w
4r² 4w
5r² 4w


Std. Dev. 1.5; Level 1
Feature Vector Distance
Rotational Displacement (Degree)
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
1r² 1w
2r² 1w
3r² 1w
4r² 1w
5r² 1w
1r² 2w
2r² 2w
3r² 2w
4r² 2w
5r² 2w
1r² 3w
2r² 3w
3r² 3w
4r² 3w
5r² 3w
1r² 4w
2r² 4w
3r² 4w
4r² 4w
5r² 4w

A.6.2. Classification Performance

The classification performance of the feature vector was analyzed in detail for all parameter combinations determined in the parameter preselection detailed in Section 4.4.2 on page 123. On the next page, the relative distributions of true and decoy feature vector distances is shown in several plots for all parameter sets. As before, the pink line corresponds to the distribution of true matches, while the blue line corresponds to the distribution of decoy matches. On following page, a table lists a summary of the properties of the distributions. The table gives exact values for the minimum (Min.), 25 % quantile (25%), the median (Med.), the mean, the 75 % quantile (75%), the maximum (Max.), and the standard deviation (Std. Dev.) of the distributions. The following two pages give the same information — plots and a table — for the distribution of the sample distinctiveness values as explained in Section 4.4.3 on page 130.

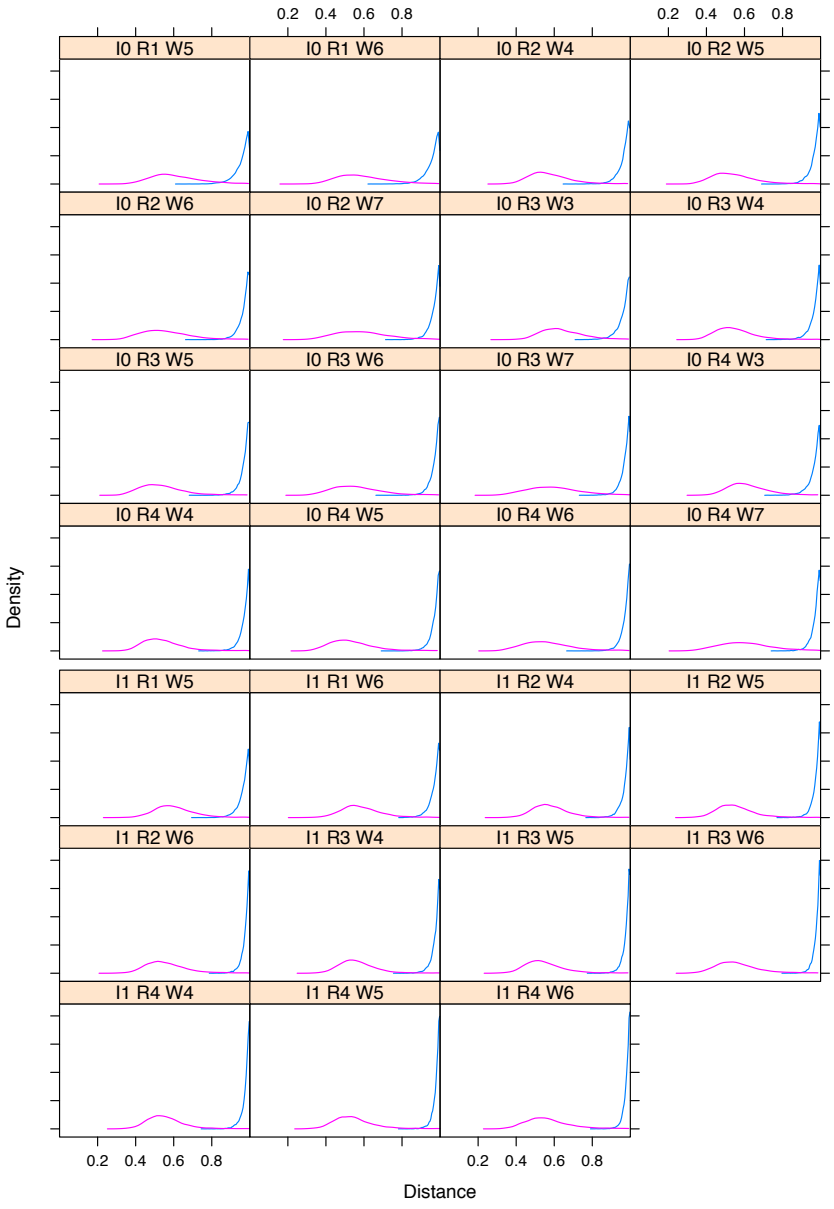
Sample distribution plots for distance values



Sample distribution properties for distance values

Title	Min.	25 %	Med.	Mean	75 %	Max.	Std. Dev.
I0 R1 W5 decoy	0.172	0.346	0.378	0.377	0.407	0.574	0.048
I0 R1 W5 query	0.081	0.193	0.217	0.222	0.246	0.409	0.041
I0 R1 W6 decoy	0.104	0.325	0.348	0.350	0.373	0.550	0.041
I0 R1 W6 query	0.053	0.168	0.193	0.199	0.223	0.392	0.042
I0 R2 W4 decoy	0.249	0.504	0.547	0.540	0.582	0.723	0.059
I0 R2 W4 query	0.152	0.266	0.295	0.298	0.325	0.528	0.046
I0 R2 W5 decoy	0.177	0.468	0.497	0.496	0.523	0.734	0.048
I0 R2 W5 query	0.081	0.221	0.250	0.255	0.284	0.488	0.047
I0 R2 W6 decoy	0.106	0.432	0.455	0.458	0.479	0.667	0.045
I0 R2 W6 query	0.053	0.199	0.231	0.237	0.268	0.484	0.052
I0 R2 W7 decoy	0.110	0.398	0.419	0.424	0.445	0.602	0.045
I0 R2 W7 query	0.080	0.190	0.224	0.230	0.264	0.498	0.056
I0 R3 W3 decoy	0.311	0.567	0.630	0.622	0.686	0.820	0.080
I0 R3 W3 query	0.199	0.352	0.384	0.385	0.419	0.585	0.052
I0 R3 W4 decoy	0.251	0.553	0.593	0.586	0.623	0.775	0.057
I0 R3 W4 query	0.160	0.275	0.304	0.308	0.338	0.569	0.047
I0 R3 W5 decoy	0.176	0.509	0.535	0.535	0.561	0.778	0.049
I0 R3 W5 query	0.078	0.231	0.264	0.269	0.301	0.510	0.051
I0 R3 W6 decoy	0.111	0.467	0.489	0.492	0.515	0.691	0.047
I0 R3 W6 query	0.057	0.214	0.250	0.254	0.288	0.501	0.055
I0 R3 W7 decoy	0.112	0.430	0.452	0.456	0.479	0.630	0.047
I0 R3 W7 query	0.084	0.207	0.243	0.249	0.285	0.575	0.059
I0 R4 W3 decoy	0.311	0.602	0.663	0.655	0.715	0.843	0.078
I0 R4 W3 query	0.209	0.357	0.389	0.390	0.424	0.592	0.053
I0 R4 W4 decoy	0.254	0.584	0.620	0.614	0.648	0.840	0.055
I0 R4 W4 query	0.164	0.280	0.312	0.316	0.348	0.583	0.050
I0 R4 W5 decoy	0.179	0.536	0.560	0.561	0.585	0.788	0.049
I0 R4 W5 query	0.083	0.239	0.274	0.279	0.311	0.510	0.053
I0 R4 W6 decoy	0.107	0.490	0.513	0.515	0.538	0.717	0.048
I0 R4 W6 query	0.057	0.225	0.261	0.266	0.301	0.539	0.057
I0 R4 W7 decoy	0.112	0.451	0.473	0.477	0.500	0.654	0.048
I0 R4 W7 query	0.088	0.218	0.257	0.261	0.298	0.591	0.060
I1 R1 W5 decoy	0.304	0.537	0.577	0.574	0.615	0.774	0.059
I1 R1 W5 query	0.147	0.301	0.331	0.333	0.363	0.589	0.050
I1 R1 W6 decoy	0.291	0.510	0.540	0.539	0.568	0.746	0.048
I1 R1 W6 query	0.155	0.267	0.295	0.299	0.326	0.572	0.048
I1 R2 W4 decoy	0.399	0.705	0.756	0.749	0.799	0.924	0.067
I1 R2 W4 query	0.217	0.383	0.416	0.418	0.453	0.692	0.054
I1 R2 W5 decoy	0.310	0.672	0.704	0.702	0.733	0.872	0.051
I1 R2 W5 query	0.212	0.327	0.359	0.363	0.396	0.626	0.052
I1 R2 W6 decoy	0.291	0.625	0.650	0.652	0.675	0.835	0.046
I1 R2 W6 query	0.188	0.294	0.328	0.333	0.365	0.597	0.054
I1 R3 W4 decoy	0.404	0.756	0.803	0.795	0.841	0.963	0.062
I1 R3 W4 query	0.247	0.389	0.424	0.428	0.467	0.709	0.057
I1 R3 W5 decoy	0.312	0.714	0.744	0.741	0.769	0.936	0.049
I1 R3 W5 query	0.227	0.335	0.370	0.376	0.412	0.645	0.058
I1 R3 W6 decoy	0.294	0.662	0.684	0.687	0.709	0.864	0.046
I1 R3 W6 query	0.189	0.305	0.342	0.346	0.381	0.599	0.057
I1 R4 W4 decoy	0.406	0.787	0.832	0.822	0.864	0.981	0.059
I1 R4 W4 query	0.248	0.393	0.427	0.433	0.475	0.739	0.059
I1 R4 W5 decoy	0.315	0.741	0.767	0.765	0.791	0.941	0.047
I1 R4 W5 query	0.226	0.342	0.379	0.384	0.424	0.637	0.058
I1 R4 W6 decoy	0.295	0.682	0.703	0.707	0.728	0.885	0.046
I1 R4 W6 query	0.203	0.314	0.354	0.359	0.398	0.635	0.061

Sample distribution plots for distinctiveness values



Sample distribution properties for distinctiveness values

Title	Min.	25 %	Med.	Mean	75 %	Max.	Std. Dev.
I0 R1 W5 decoy	0.625	0.943	0.972	0.959	0.988	1.000	0.040
I0 R1 W5 query	0.272	0.517	0.588	0.609	0.685	0.999	0.129
I0 R1 W6 decoy	0.636	0.943	0.972	0.960	0.988	1.000	0.039
I0 R1 W6 query	0.225	0.494	0.571	0.591	0.670	1.000	0.133
I0 R2 W4 decoy	0.659	0.955	0.977	0.968	0.991	1.000	0.033
I0 R2 W4 query	0.303	0.499	0.558	0.576	0.636	0.999	0.112
I0 R2 W5 decoy	0.701	0.959	0.979	0.970	0.991	1.000	0.030
I0 R2 W5 query	0.244	0.462	0.527	0.545	0.607	0.997	0.118
I0 R2 W6 decoy	0.673	0.959	0.979	0.970	0.992	1.000	0.030
I0 R2 W6 query	0.235	0.463	0.539	0.557	0.633	1.000	0.131
I0 R2 W7 decoy	0.724	0.960	0.980	0.971	0.992	1.000	0.028
I0 R2 W7 query	0.246	0.486	0.577	0.589	0.674	1.000	0.139
I0 R3 W3 decoy	0.723	0.952	0.977	0.966	0.990	1.000	0.034
I0 R3 W3 query	0.322	0.561	0.625	0.640	0.707	0.999	0.112
I0 R3 W4 decoy	0.725	0.960	0.980	0.971	0.992	1.000	0.030
I0 R3 W4 query	0.293	0.477	0.535	0.553	0.606	1.000	0.110
I0 R3 W5 decoy	0.692	0.963	0.981	0.973	0.992	1.000	0.028
I0 R3 W5 query	0.266	0.454	0.522	0.539	0.600	0.999	0.121
I0 R3 W6 decoy	0.673	0.963	0.982	0.973	0.992	1.000	0.027
I0 R3 W6 query	0.254	0.467	0.546	0.563	0.638	0.999	0.132
I0 R3 W7 decoy	0.742	0.964	0.982	0.974	0.992	1.000	0.027
I0 R3 W7 query	0.254	0.500	0.587	0.599	0.684	0.999	0.139
I0 R4 W3 decoy	0.718	0.957	0.979	0.968	0.991	1.000	0.032
I0 R4 W3 query	0.349	0.544	0.603	0.618	0.678	0.996	0.107
I0 R4 W4 decoy	0.741	0.963	0.982	0.973	0.993	1.000	0.028
I0 R4 W4 query	0.276	0.466	0.526	0.543	0.598	0.997	0.110
I0 R4 W5 decoy	0.701	0.964	0.982	0.974	0.993	1.000	0.026
I0 R4 W5 query	0.271	0.453	0.517	0.537	0.597	0.999	0.121
I0 R4 W6 decoy	0.675	0.965	0.983	0.975	0.993	1.000	0.026
I0 R4 W6 query	0.267	0.472	0.550	0.567	0.640	1.000	0.132
I0 R4 W7 decoy	0.749	0.966	0.983	0.975	0.993	1.000	0.025
I0 R4 W7 query	0.273	0.506	0.593	0.606	0.689	0.999	0.138
I1 R1 W5 decoy	0.706	0.957	0.979	0.969	0.991	1.000	0.031
I1 R1 W5 query	0.280	0.530	0.590	0.602	0.661	0.998	0.107
I1 R1 W6 decoy	0.794	0.960	0.980	0.971	0.992	1.000	0.028
I1 R1 W6 query	0.252	0.514	0.571	0.585	0.644	0.999	0.109
I1 R2 W4 decoy	0.775	0.966	0.983	0.975	0.993	1.000	0.025
I1 R2 W4 query	0.283	0.514	0.569	0.582	0.634	0.996	0.100
I1 R2 W5 decoy	0.779	0.969	0.985	0.978	0.994	1.000	0.023
I1 R2 W5 query	0.284	0.484	0.540	0.554	0.605	1.000	0.105
I1 R2 W6 decoy	0.795	0.971	0.986	0.979	0.994	1.000	0.021
I1 R2 W6 query	0.259	0.483	0.544	0.561	0.619	0.997	0.113
I1 R3 W4 decoy	0.764	0.969	0.985	0.977	0.994	1.000	0.024
I1 R3 W4 query	0.293	0.499	0.553	0.566	0.617	1.000	0.101
I1 R3 W5 decoy	0.781	0.973	0.986	0.980	0.995	1.000	0.021
I1 R3 W5 query	0.287	0.478	0.533	0.551	0.604	1.000	0.110
I1 R3 W6 decoy	0.804	0.974	0.987	0.981	0.995	1.000	0.019
I1 R3 W6 query	0.293	0.483	0.546	0.563	0.622	1.000	0.116
I1 R4 W4 decoy	0.752	0.972	0.986	0.979	0.995	1.000	0.022
I1 R4 W4 query	0.298	0.489	0.543	0.556	0.605	1.000	0.101
I1 R4 W5 decoy	0.788	0.975	0.988	0.982	0.995	1.000	0.019
I1 R4 W5 query	0.287	0.474	0.532	0.547	0.599	0.999	0.109
I1 R4 W6 decoy	0.797	0.976	0.988	0.983	0.995	1.000	0.018
I1 R4 W6 query	0.287	0.487	0.549	0.567	0.627	1.000	0.120

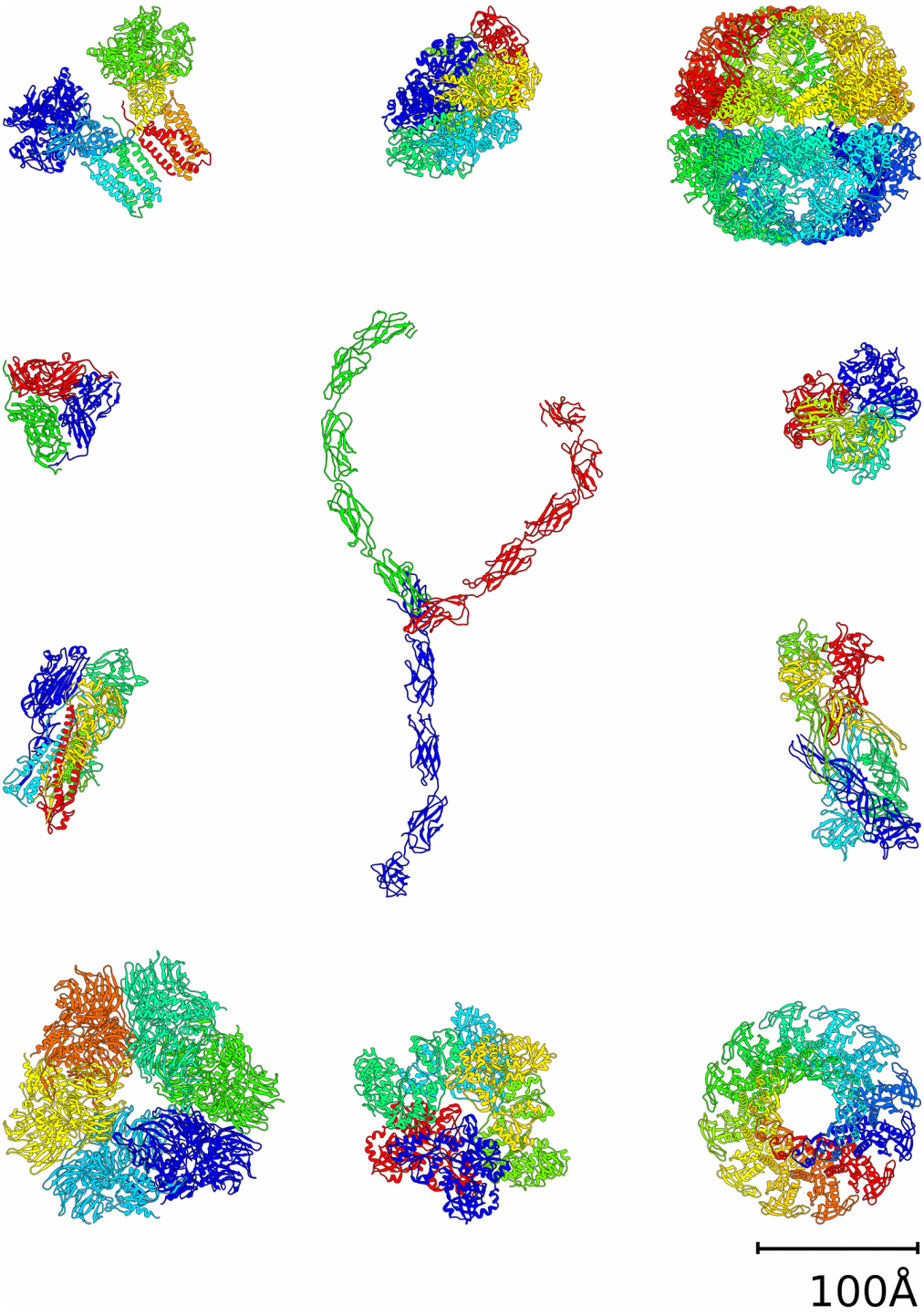
A.7. Test Set

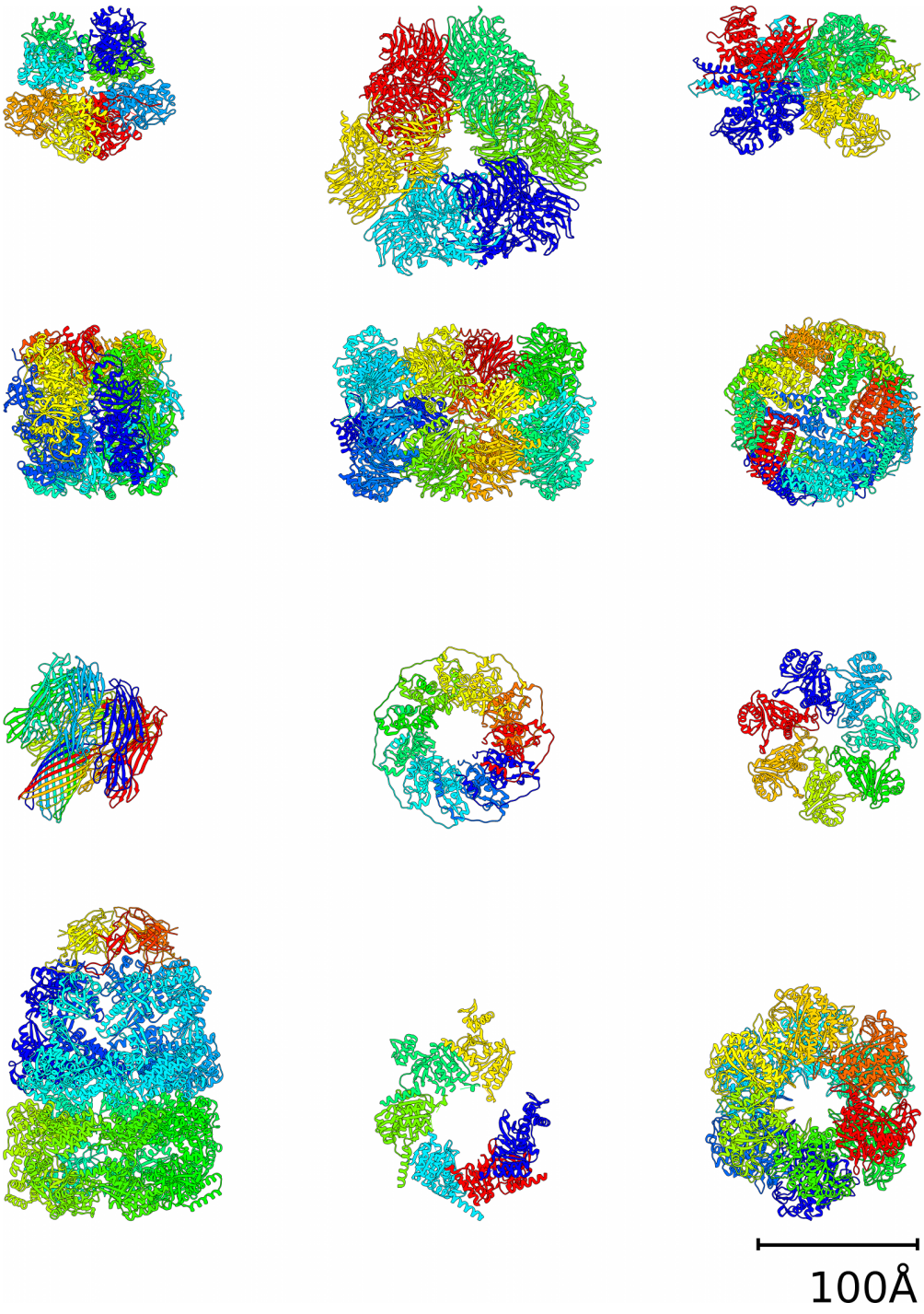
The following figures show the test set complexes and monomers utilized in Section 5.1. There are in total 23 complexes, which contain 234 copies of 35 distinct monomers.

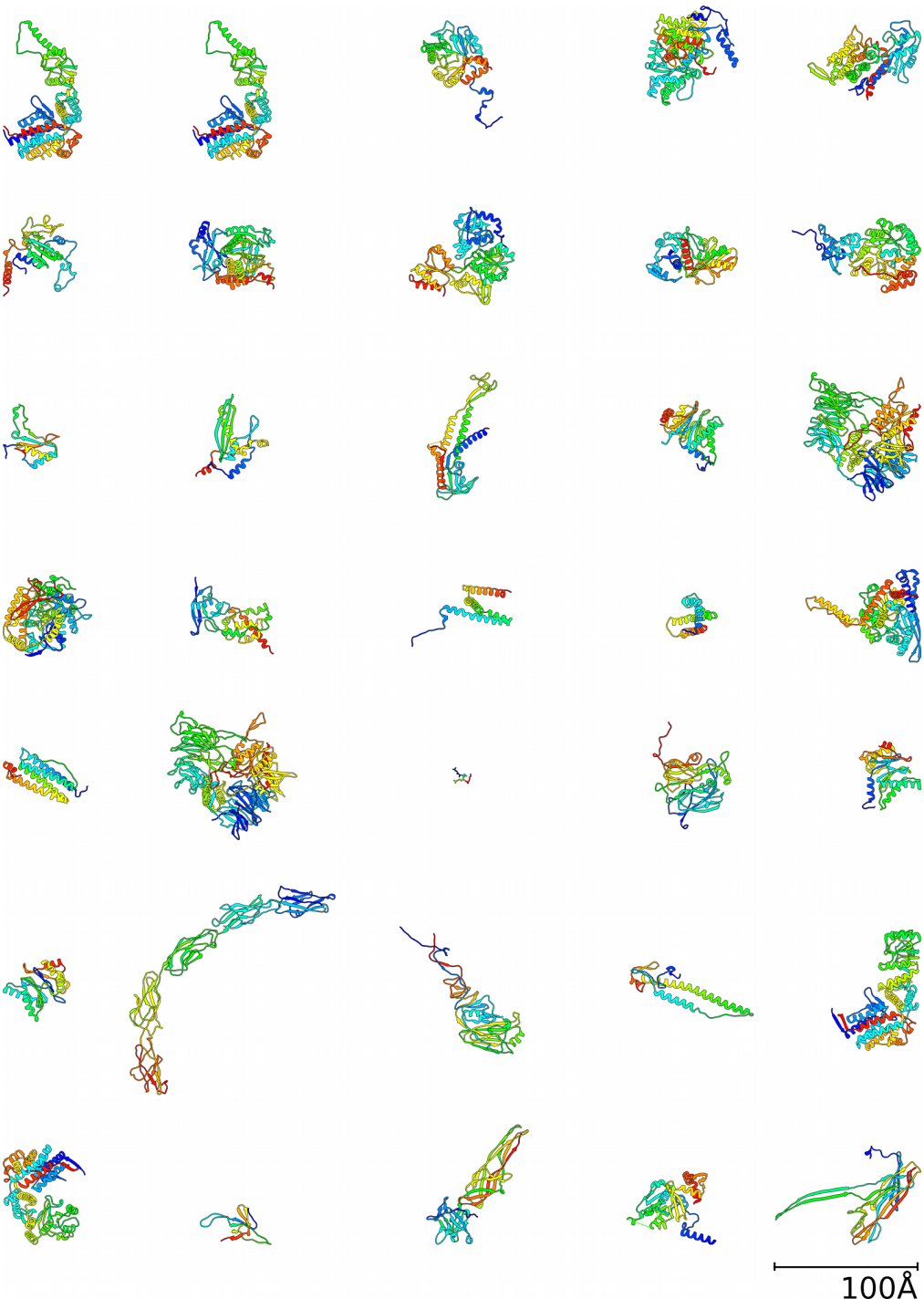
The first page shows — from left to right and top to bottom — the complexes 1KF6, 1E6V, 1A6D, 1NIC, 1Q5B (middle), 1GD1, 1RUZ, 1W3A, 1N6D, 1G8G, 1IJG.

The second page shows the complexes 1AW5, 1K32, 1L1F, 1GK8, 1PMA, 1MFR, 7AHL, 1H2I, 1J2P, 1SX4, 1XMV, 1FPY.

The third page shows the distinct monomers [1A6D:A], [1A6D:B], [1AW5:A], [1E6V:AD], [1E6V:BE], [1E6V:CF], [1FPY:A-L], [1G8G:AB], [1GD1:OPQR], [1GK8:ACEG], [1GK8:IKMO], [1H2I:A-K], [1IJG:A-L], [1J2P:A-G], [1K32:A-F], [1KF6:AM], [1KF6:BN], [1KF6:CO], [1KF6:DP], [1L1F:A-F], [1MFR:A-X], [1N6D:A-F], [1N6D:G-L], [1NIC:A], [1PMA:AC-O], [1PMA:12BP-Z], [1Q5B:ABC], [1RUZ:HJL], [1RUZ:IMK], [1SX4:A-G], [1SX4:H-N], [1SX4:O-U], [1W3A:A], [1XMV:A], [7AHL:A-G].


100Å


100Å


100Å

A.8. Utilized Computer Progams

Creating this work was supported by several computer programs. The software *siseek* is written in C++ [317], using the standard C++ and the boost[1] library. Furthermore, the external visualization program FlexV[2], the database SQLite[3] [145], and a library providing an implementation of an R-tree[4] [132] have been used. The code of the software system was written mainly in vi[5] using the version control system git[6].

This work was written using LaTeX and the reference manager JabRef[7]. Statistical evaluations were carried out using R[8] [267], Microsoft Excel 2007[9], and google documents[10]. Drawings were created in Microsoft Powerpoint 2007[11] and Adobe Illustrator CS4 14.0.0[12]. Molecular drawings were created using the computer programs UCSF Chimera[13] [262] and FlexV[14] in combination with POV-ray[15].

A.9. Software Architecture and Implementation

siseek is interactive and can be controlled through a command line interface (CLI). The available commands are listed in the following Appendix A.9.1. The implementation of *siseek* is split into several libraries, which encapsulate specific functionality. An overview of the software architecture is discussed in Appendix A.9.2.

A.9.1. Command Line Interface

The functionality of *siseek* can be invoked by issuing commands, which include the following:

[1] http://www.boost.org
[2] http://www.biosolveit.de/flexv
[3] http://www.sqlite.org
[4] http://www2.research.att.com/~marioh/spatialindex
[5] http://www.vim.org
[6] http://git-scm.com
[7] http://jabref.sourceforge.net
[8] http://www.r-project.org
[9] http://office.microsoft.com/en-us/excel
[10] http://docs.google.com
[11] http://office.microsoft.com/en-us/powerpoint
[12] http://www.adobe.com/products/illustrator
[13] http://www.cgl.ucsf.edu/chimera
[14] http://www.biosolveit.de/flexv
[15] http://www.povray.org

SELOUTP, INFO, INFO_TIME, INFO_DOCK, INFO_SOL, SETPARAM

These commands allow for the interaction of the user with the internal status of objects and the program. SETPARAM can be used for changing internal parameters such as the current scoring function. The INFO* commands yield information on all *molecular objects* — i. e., molecules or maps — and the output of calculations can be redirected using SELOUTP.

READMAP, WRITMAP, READPDB, WRITEPDB, WRITEPLM, CLEAR

Atomic structures in PDB file format [364] can be read using READPDB, while READMAP reads in density maps in Situs [365] format. CLEAR can be used to delete all molecular objects. The WRITE* commands allow for saving molecular objects and placements in files.

BLUR, INJECT, COPY, SET_THRS, RESAMPLE, FILTER_MAP, NOISE

These commands allow for altering molecular objects. COPY creates a copy of the given molecular object, while BLUR creates a synthetic map for an atomic structure. INJECT associates an atomic structure to a map and thereby allows for restricting the map description to relevant parts. SET_THRS alters the isosurface threshold of the map, RESAMPLE resamples the map using trilinear interpolation, and NOISE adds white Gaussian noise to a map. FILTER_MAP applies a filter to a map — available filters include the sampled Gaussian, the discrete Gaussian, the first partial derivative of the Gaussian, the Sobel, the Harris, the Difference of Gaussians, the Laplacian of Gaussian, and the sinc filter.

SIFT, DOCK, RESCORE, CLUSTER, OPTIMIZE

These commands enable the registration of maps. SIFT calculates a map description for a given molecular object and DOCK allows for docking two molecular objects. The resulting placements can be RESCOREd, CLUSTERed, and OPTIMIZEd using the provided commands. For these commands, one of the scoring functions detailed in Section 3.5 can be selected using the SETPARAM command (number of keypoint matches, weighted number of keypoint matches, correlation of the maps, sum of the densities interpolated at the atom positions, number of atoms enclosed in the current isosurface).

DB/ADD, DB/ADD_BULK, DB/CREATE_IDX, DB/LOAD, DB/EXTRACT

For molecule recognition, a database storing map descriptions has been implemented. The interaction with this database is provided by the DB/* commands. They allow for inserting map descriptions in the database (ADD and ADD_BULK), the creation of an index on the row ID, and the extraction of map descriptions from the database (LOAD and EXTRACT).

DRAW, DRAWSLICE, DRAWPLM

The calculated placements can be analyzed interactively using the visualization software FlexV [269]. DRAW allows for the depiction of molecular objects DRAWSLICE draws slices of maps, and DRAWPLM is dedicated to displaying molecular objects in the positions that were calculated by the registration.

A.9.2. Software Architecture

The software is implemented in C++ [317] and divided into libraries as shown in Figure A.1. All but the external libraries were genuinely implemented by the author of this work.

The external libraries are used throughout *siseek* and include the C++ standard template library [236], the boost library [386], an R-tree [132], and a database [145]. Furthermore, the stand-alone external display program FlexV [269] is used to interact with the user.

The basis of *siseek* is the math library, which supplies functionality for 3D linear algebra. It includes representations for a 3D vector, a point, a matrix, a rotation, a 3D box, an icosahedron, and a sphere. It also comprises the implementation of the 3D orientation histogram and the geodesic grid introduced in Section 3.3.1 on page 73 and Section 3.3.2 on page 77. Additionally, methods for the efficient computation of the RMSD [270] between two sets of points and means for creating random 3D vectors are provided. Furthermore, methods for computing the 3D Delaunay triangulation of a set of points are included, which were used in a first prototype [124]. The math library also provides an abstract implementation of clustering algorithms, which allow for analyzing any set of objects given a suitable metric on the objects. Furthermore, an abstract implementation of the Monte Carlo optimization method is provided. The math library also contains a graph representation and an implementation of an algorithm for finding maximal cliques in these graphs [40].

The libraries containing map and molecule representations both rely on the math library. The molecule library provides basic functionality for representing

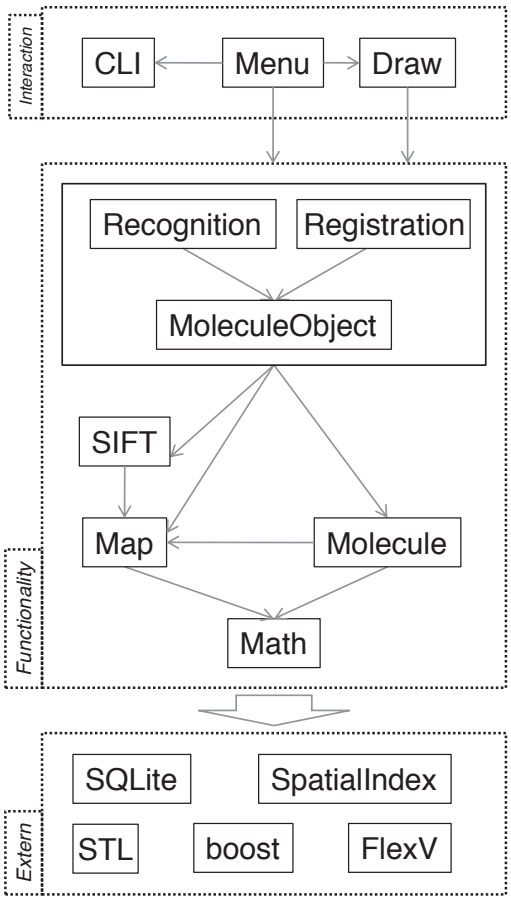


Figure A.1 – Software architecture
siseek is implemented in C++ [317] and divided into libraries. It uses the external C++ standard template library and the boost library [386], which provide basic functionality. Furthermore, the external display program FlexV [269], a database [145] (SQLite) and an R-tree implementation [132] (SpatialIndex) are included. The foundation for the functionality of the software is formed by the map and molecule library, which both rely on the math library. Map representations are generated by the SIFT library, which solely relies on the map and math library. The functionality for registration and recognition is based on the MoleculeObject, which is a facade providing and enriching access to objects of the map and molecule libraries. User interaction with *siseek* is facilitated through a menu, which is based on an abstract implementation of a command line interface. Three-dimensional scenes can be exported by the drawing library and either be displayed by the external program FlexV or other programs. (© A. Griewel)

molecular structures, i. e., atoms along their properties and their connectivity. Furthermore, it contains the functionality for creating synthetic maps. The map library provides an efficient representation of electron density maps sampled on a cubic grid. It enables efficient filtering by exploiting filter properties such as separability [115]. Since all libraries are modular and encapsulate functionality, it is possible to easily use libraries of *siseek* in other software. This has been done with the map library, which forms the basis of DoGSite [348], a software for the detection of active sites of proteins.

The SIFT library encapsulates functionality for the automated computation of map descriptions. It relies solely on the map and math libraries and enables the computation of map representations. Furthermore, it provides the functionality for saving SIFT objects in a database. All SIFT functionality except the orientation histogram and geodesic index are located in this module. The latter two are generic objects, which can well be used in other modules and are thus situated in the math library.

Registration and recognition both rely on the MoleculeObject, which is a facade encapsulating access to the relevant objects of the map, molecule, and SIFT libraries. Both registration and recognition solely use functionality from the SIFT library, however, the MoleculeObject also provides access to the basic objects map and molecule for, e. g., displaying solutions in the external viewer.

The user interacts with the software through a menu, which is based on an abstract implementation of a CLI. Using the draw library, it is possible to display molecules and maps using the visualization program FlexV, which has been altered for efficiently displaying macromolecular objects, i. e., proteins and electron density maps. For this purpose, an implementation of the marching cubes algorithm [215] was added to the software, which allows for displaying isosurfaces of electron density maps.

Acknowledgements

I would like to thank Matthias Rarey for his support and encouragement as well as for always believing in my skills. Thanks for giving me the freedom to carry out not only the research presented in this work but also work in the fields of computer science, chemoinformatics, and drug design. Thanks also for giving me the opportunity of teaching various courses at the University of Hamburg (UHH), which was a lot of fun. Equally, I would like to thank H. Siegfried Stiehl for showing deep interest in this work, and for finding the time to give valuable comments on the draft of this thesis, although being heavily involved as vice president of the UHH. Thanks also to Stefan Birmanns, University of Texas, for welcoming me in his group in Houston, TX, and for valuable discussions.

My thanks extend also to my teachers that lay the basis for my work in the interdisciplinary field of bioinformatics. For once, Stefan Kurtz and Andrew Torda, Center for Bioinformatics Hamburg (ZBH), and all others involved in my education at the UHH. Thanks to Willy Wriggers, Cornell University, NY, for valuable discussions and helpful comments. I would also like to extend my thanks to Thomas Huber, University of Queensland, and Paul March, University of New South Wales, for inviting me for projects to their labs. In particular, my thanks go to my teachers at school who woke my interest in the field Mr. Schriek, Mr. Hahn, Mr. Rasche, and Mr. Kiko, Mariengymnasium and Ursulinengymnasium Werl, and also Mr. Schondelmeyer and Mr. Milam, Belgrade High School, MT.

Thanks to Jochen Schlosser for being a great friend and office mate, as well as for the work in our joint projects. Thanks to Andrea Volkamer for the great time, our fruitful project in pocket detection, and for showing the generic applicability of the implemented libraries. My special thanks go to Thilo Mende, University of Bremen, Stefanie Gräwe, Berhard Nocht Institute, and Till Kothe, who I got to know during my studies and ever since have been great friends and discussion partners in both private and scientific matters.

To all staff and students of the ZBH: thanks for the wonderful time. My particular thanks go to those colleagues who showed interest in this work and gave helpful comments including, besides the already mentioned, Tobias Lippert, Karen Schomburg, Sascha Urbaczeck, Jörg Degen, Juri Pärn, and Stefan Bietz. Thanks to Christian Rhein and Jörn Adomeit for the computer support, and

thanks to Viola Hingst, Natascha Dönges, and Melanie Geringhoff for helping me in all administrative matters.

My deepest thanks for the greatest support one can wish for goes to my family. Thank you, Roman, for being always there for me. Thanks for being patient, supportive and encouraging, for making me happy and for making me content. Thank you Eva and Ferdi for your unconditional and enduring support and for being great friends, too. Thank you Klara, for all you did for me.

References

[1] E. T. Adman, J. W. Godden, and S. Turley. The Structure of Copper-nitrite Reductase from Achromobacter cycloclastes at Five pH Values, with NO_2^- bound and with Type II Copper Depleted. *J Biol Chem*, 270(46):27458–27474, 1995. 141

[2] F. Alber, F. Förster, D. Korkin, M. Topf, and A. Sali. Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem*, 77:443–477, 2008. 50, 61

[3] M. M. U. Ali, S. M. Roe, C. K. Vaughan, P. Meyer, B. Panaretou, P. W. Piper, C. Prodromou, and L. H. Pearl. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, 440(7087):1013–1017, 2006. 176

[4] S. Allaire, J. J. Kim, S. L. Breen, D. A. Jaffray, and V. Pekar. Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, Los Alamitos, CA, 2008. IEEE Computer Society. 27

[5] F. H. Allen. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B Struct Crystallogr Cryst Chem*, 58(Pt 3 Pt 1):380–388, 2002. 44

[6] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–D229, 2004. VII, 172, 200

[7] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, 36(Database issue):D419–D425, 2008. VII, 172, 200

[8] M. Ankerst, G. Kastenmüller, H. P. Kriegel, and T. Seidl. 3D Shape Histograms for Similarity Search and Classification in Spatial Databases. In

R. Güting, D. Papadias, and F. Lochovsky, editors, *Proceedings of the 6th International Symposium on Advances in Spatial Databases*, volume 1651 of *Lecture Notes in Computer Science*, pages 207–226, Berlin, Germany, 1999. Springer. 18

[9] P. E. Anuta. Spatial registration of multispectral and multitemporal digital imagery using fast Fourier transform techniques. *IEEE Trans Geosci Electron*, 8(4):353–368, 1970. 18

[10] M. Baker. Whole-animal imaging: The whole picture. *Nature*, 463(7283):977–980, 2010. 1, 19, 38, 50, 61

[11] M. L. Baker, M. R. Baker, C. F. Hryc, and F. Dimaio. Analyses of sub-nanometer resolution cryo-EM density maps. *Methods Enzymol*, 483:1–29, 2010. 50, 61

[12] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognit*, 13(2):111–122, 1981. 17

[13] C. Bartolucci, D. Lamba, S. Grazulis, E. Manakova, and H. Heumann. Crystal structure of wild-type chaperonin GroEL. *J Mol Biol*, 354(4):940–951, 2005. 150

[14] J. D. Bauman, K. Das, W. C. Ho, M. Baweja, D. M. Himmel, A. D. Clark, D. A. Oren, P. L. Boyer, S. H. Hughes, A. J. Shatkin, and E. Arnold. Crystal engineering of HIV-1 reverse transcriptase for structure-based drug design. *Nucleic Acids Res*, 36(15):5083–5092, 2008. 103

[15] H. Bay and Tuytelaars. SURF: Speeded Up Robust Features. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Proceedings of the Ninth European Conference on Computer Vision*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417, Berlin, Germany, 2006. Springer. 25

[16] P. R. Beaudet. Rotationally Invariant Image Operators. In *Proceedings of the International Joint Conference on Pattern Recognition*, pages 579–583, 1978. 14

[17] M. Beck, J. A. Malmström, V. Lange, A. Schmidt, E. W. Deutsch, and R. Aebersold. Visual proteomics of the human pathogen Leptospira interrogans. *Nat Methods*, 6(11):817–823, 2009. 1, 19, 50, 61

[18] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. In H. Garcia-Molina, editor, *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data*, pages 322–331, New York, NY, 1990. ACM. 80, 93

[19] J. S. Beis and D. G. Lowe. Shape Indexing Using Approximate Nearest-Neighbour Search in High-Dimensional Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1000–1006, Los Alamitos, CA, 1997. IEEE Computer Society. 95

[20] R. E. Bellman. *Adaptive control processes: a guided tour*. Princeton University Press, Princeton, NJ, 1961. 96, 119, 186

[21] S. Berchtold, D. A. Keim, and H. P. Kriegel. The X-tree: An Index Structure for High-Dimensional Data. In *Proceedings of 22nd International Conference on Very Large Data Bases*, pages 28–39. Morgan Kaufmann Publishers Inc., 1996. 95

[22] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*. Biochemistry. W. H. Freeman, New York, NY, 6$^{\text{th}}$ edition, 2007. 29, 165

[23] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, 10(12):980–980, 2003. `http::// www.wwpdb.org`. VII, 2, 34, 149, 199

[24] H. M. Berman. The Protein Data Bank: a historical perspective. *Acta Crystallogr A Found Crystallogr*, 64(Pt 1):88–95, 2008. 1, 37, 50

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, 2000. `http://www.pdb.org`. 34

[26] J. D. Bernal and D. Crowfoot. X-ray photographs of crystalline pepsin. *Nature*, 133(3369):794–795, 1934. 42

[27] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 112(3):535–542, 1977. 34

[28] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When Is "Nearest Neighbor" Meaningful? In C. Beeri and P. Buneman, editors, *Database*

Theory — ICDT'99, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235, Berlin, Germany, 1999. Springer. 96, 119, 186

[29] S. Birmanns, M. Rusu, and W. Wriggers. Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes. *J Struct Biol*, 173(3):428–435, 2011. 51, 54, 60

[30] S. Birmanns and W. Wriggers. Interactive fitting augmented by force-feedback and virtual reality. *J Struct Biol*, 144(1–2):123–131, 2003. 54, 60

[31] M. Bleichenbacher, S. Tan, and T. J. Richmond. Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. *J Mol Biol*, 332(4):783–793, 2003. 103

[32] M. Bomans, K.-H. Hohne, U. Tiede, and M. Riemer. 3-D segmentation of MR images of the head for 3-D display. *IEEE Trans Med Imaging*, 9(2):177–183, 1990. 13

[33] G. E. P. Box and M. E. Muller. A note on the generation of random normal deviates. *Ann Math Stat*, 29(2):610–611, 1958. 105

[34] K. Braig, P. D. Adams, and A. T. Brünger. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. *Nat Struct Biol*, 2(12):1083–1094, 1995. ii, 31, 67, 69

[35] C. I. Brändén and T. Alwyn Jones. Between objectivity and subjectivity. *Nature*, 343:687–689, 1990. 38

[36] C. I. Brändén and J. Tooze. *Introduction to Protein Structures*. Introduction to Protein Structure Series. Garland Pub., New York, NY, 2nd edition, 1999. 152, 156, 165

[37] H. Brandstetter, J. S. Kim, M. Groll, and R. Huber. Crystal structure of the tricorn protease reveals a protein disassembly line. *Nature*, 414(6862):466–470, 2001. 141

[38] J. Brecher. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure Appl Chem*, 80(2):277–410, 2008. 29

[39] S. E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–256, 2000. 174, 187, 191

[40] C. Bron and J. Kerbosch. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun ACM*, 16(9):575–577, 1973. 17, 227

[41] L. G. Brown. A survey of image registration techniques. *ACM Comput Surv*, 24(4):325–376, 1992. 9

[42] M. Brown and D. G. Lowe. Invariant Features from Interest Point Groups. In P. L. Rosin and A. D. Marshall, editors, *Proceedings of the British Machine Vision Conference 2002*, pages 253–262, Cardiff, United Kingdom, 2002. British Machine Vision Association. 24, 71, 72

[43] M. Brown and D. G. Lowe. Recognising Panoramas. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1218–1227, Los Alamitos, CA, 2003. IEEE Computer Society. 24

[44] C. B. Buck, N. Cheng, C. D. Thompson, D. R. Lowy, A. C. Steven, J. T. Schiller, and B. L. Trus. Arrangement of L2 within the papillomavirus capsid. *J Virol*, 82(11):5190–5197, 2008. 156

[45] Bureau International des Poids et Mesures. *The International System of Units (SI)*. Paris, France, 8$^{\text{th}}$ edition, 2006. 28

[46] P. Burt and E. Adelson. The Laplacian pyramid as a compact image code. *IEEE Trans Commun*, 31(4):532–540, 1983. 23

[47] P. J. Burt. Fast filter transform for image processing. *Comput Graph Image Process*, 16(1):20–51, 1981. 22

[48] J. H. M. Cabral, A. P. Jackson, C. V. Smith, N. Shikotra, A. Maxwell, and R. C. Liddington. Crystal structure of the breakage-reunion domain of DNA gyrase. *Nature*, 388(6645):903–906, 1997. 163, 180

[49] W. Cai, X. Shao, and B. Maigret. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J Mol Graphics Modell*, 20(4):313–328, 2002. 19

[50] M. S. Campo. Bovine papillomavirus: old system, new lessons? In M. S. Campo, editor, *Papillomavirus research: from natural history to vaccines and beyond*, chapter 23. Caister Academic Press, Wymondham, England, 2006. 155

[51] J. Canny. A Computational Approach To Edge Detection. *IEEE Trans Pattern Anal Mach Intell*, 8(6):679–698, 1986. 13

[52] H. Ceulemans and R. B. Russell. Fast Fitting of Atomic Structures to Low-Resolution Electron Density Maps by Surface Overlap Maximization. *J Mol Biol*, 338(4):783–793, 2004. 54, 60

[53] P. Chacon and W. Wriggers. Multi-resolution contour-based fitting of macromolecular structures. *J Mol Biol*, 317(3):375–384, 2002. 18, 52, 53, 60, 148

[54] J.-M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL Compendium in 2004. *Nucleic Acids Res*, 32(Database issue):D189–D192, 2004. 174, 187, 191

[55] J.-M. Chandonia, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. ASTRAL compendium enhancements. *Nucleic Acids Res*, 30(1):260–263, 2002. 174, 187, 191

[56] C. Chaudhry, A. L. Horwich, A. T. Brunger, and P. D. Adams. Exploring the structural dynamics of the E.coli chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states. *J Mol Biol*, 342(1):229–245, 2004. 141

[57] W. Cheung and G. Hamarneh. N-SIFT: N-Dimensional Scale Invariant Feature Transform For Matching Medical Images. In *Proceedings of the 4^{th} IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 720–723, Los Alamitos, CA, 2007. IEEE Computer Society. 26

[58] W. Cheung and G. Hamarneh. n-SIFT: n-dimensional scale invariant feature transform. *IEEE Trans Image Process*, 18(9):2012–2021, 2009. 26

[59] W. Chiu, M. L. Baker, W. Jiang, M. Dougherty, and M. F. Schmid. Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure*, 13(3):363–372, 2005. 38

[60] J.-H. Cho, S. Sato, E. Y. Kim, H. Schindelin, and D. P. Raleigh. Highly Cooperative Interactions in the Denatured State of a Globular Protein. `http://www.ebi.ac.uk/pdbe`, PDB ID 2HBA, 2007. 103

[61] W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *J Am Stat Assoc*, 74(368):829–836, 1979. 96

[62] W. S. Cleveland and S. J. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J Am Stat Assoc*, 83(403):596–610, 1988. 96

[63] Y. Cong, M. L. Baker, J. Jakana, D. Woolford, E. J. Miller, S. Reissmann, R. N. Kumar, A. M. Redding-Johanson, T. S. Batth, A. Mukhopadhyay, S. J. Ludtke, J. Frydman, and W. Chiu. 4.0-Å resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. *Proc Natl Acad Sci USA*, 107(11):4967–4972, 2010. 150

[64] Y. Cong and S. J. Ludtke. Single particle analysis at high resolution. *Methods Enzymol*, 482:211–235, 2010. 19, 149

[65] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 1983. 29

[66] L. L. Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 30(1):264–267, 2002. VII, 172, 200

[67] J. M. Cook. Rational formulae for the production of a spherically symmetric probability distribution. *Math Comput*, 11:81–82, 1957. 93, 102, 120

[68] R. B. Corey and L. Pauling. Molecular Models of Amino Acids, Peptides, and Proteins. *Rev Sci Instrum*, 24(8):621–627, 1953. 29

[69] H. S. M. Coxeter. *Regular Polytopes*. Dover Books on Advanced Mathematics. Dover Publications, Mineola, NY, 3rd edition, 1973. 202

[70] J. L. Crowley and A. C. Parker. A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform. *IEEE Trans Pattern Anal Mach Intell*, 6(2):156–170, 1984. 22, 23

[71] J. L. Crowley and A. C. Sanderson. Multiple resolution representation and probabilistic matching of 2-D gray-scale shape. *IEEE Trans Pattern Anal Mach Intell*, 9(1):113–121, 1987. 23

[72] R. A. Crowther. The fast rotation function. In M. G. Rossmann, editor, *The molecular replacement method: a collection of papers on the use of non-crystallographic symmetry*, International Science Review Series, pages 173–178. Gordon and Breach, New York, NY, 1972. 52

[73] M.-H. Dao-Thi, L. V. Melderen, E. D. Genst, H. Afif, L. Buts, L. Wyns, and R. Loris. Molecular basis of gyrase poisoning by the addiction toxin CcdB. *J Mol Biol*, 348(5):1091–1102, 2005. 163, 176

[74] S. A. Darst, N. Opalka, P. Chacon, A. Polyakov, C. Richter, G. Zhang, and W. Wriggers. Conformational flexibility of bacterial RNA polymerase. *Proc Natl Acad Sci USA*, 99(7):4296–4301, 2002. 49, 51

[75] A. M. Davis, S. A. St-Gallay, and G. J. Kleywegt. Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov Today*, 13(19–20):831–841, 2008. 38

[76] D. de Sanctis, S. Dewilde, A. Pesce, L. Moens, P. Ascenzi, T. Hankeln, T. Burmester, and M. Bolognesi. Crystal Structure of Cytoglobin: The Fourth Globin Type Discovered in Man Displays Heme Hexa-coordination. *J Mol Biol*, 336(4):917–927, 2004. 176, 177

[77] M. A. DePristo, P. I. W. de Bakker, and T. L. Blundell. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure*, 12(5):831–838, 2004. 38

[78] R. Diamond. A Real-Space Refinement Procedure for Proteins. *Acta Crystallogr A Crystal Phys, Diffraction, Theor and General Crystallogr*, 27(5):436–452, 1971. 65

[79] L. Ding, A. Goshtasby, and M. Satter. Volume image registration by template matching. *Imag Vision Comput*, 19(12):821–832, 2001. 17, 18

[80] L. Ditzel, J. Löwe, D. Stock, K. O. Stetter, H. Huber, R. Huber, and S. Steinbacher. Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. *Cell*, 93(1):125–138, 1998. 141, 150

[81] G. J. Doherty and H. T. McMahon. Mediation, modulation, and consequences of membrane-cytoskeleton interactions. *Annu Rev Biophys*, 37:65–95, 2008. 29

[82] O. Dror, K. Lasker, R. Nussinov, and H. Wolfson. EMatch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies. *Acta Crystallogr D Biol Crystallogr*, 63(Pt 1):42–49, 2007. 55, 60

[83] P. T. Erskine, N. Senior, S. Awan, R. Lambert, G. Lewis, I. J. Tickle, M. Sarwar, P. Spencer, P. Thomas, M. J. Warren, P. M. Shoolingin-Jordan, S. P. Wood, and J. B. Cooper. X-ray structure of 5-aminolaevulinate dehydratase, a hybrid aldolase. *Nat Struct Biol*, 4(12):1025–1031, 1997. 141

[84] P. Evans and A. McCoy. An introduction to molecular replacement. *Acta Crystallogr D Biol Crystallogr*, 64(Pt 1):1–10, 2008. 51

[85] F. Fabiola and M. S. Chapman. Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure*, 13(3):389–400, 2005. 50

[86] V. Fadel, F. Canduri, J. R. Olivieri, A. L. S. Smarra, M. F. Colombo, G. O. Bonilla-Rodriguez, and W. F. de Azevedo. Crystal structure of hemoglobin from the maned wolf (Chrysocyon brachyurus) using synchrotron radiation. *Protein Pept Lett*, 10(6):551–559, 2003. 177

[87] S. Falke, F. Tama, C. L. Brooks, E. P. Gogol, and M. T. Fisher. The 13 Å Structure of a Chaperonin GroEL-Protein Substrate Complex by Cryo-electron Microscopy. *J Mol Biol*, 348(1):219–230, 2005. 150

[88] M. Ferrant, S. K. Warfield, A. Nabavi, F. A. Jolesz, and R. Kikinis. Registration of 3D Intraoperative MR Images of the Brain Using a Finite Element Biomechanical Model. In S. L. Delp, A. M. DiGioia, and B. Jaramaz, editors, *Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 1935 of *Lecture Notes in Computer Science*, pages 19–28, Berlin, Germany, 2000. Springer. 10

[89] B. Fischer and J. Modersitzki. Ill-posed medicine – an introduction to image registration. *Inverse Prob*, 24(3):034008, 2008. 9

[90] E. Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges*, 27(3):2985–2993, 1894. 28

[91] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM*, 24(6):381–395, 1981. 17

[92] J. M. Fitzpatrick, D. L. G. Hill, and C. R. M. Jr. Image Registration. In J. Beutel and M. Sonka, editors, *Handbook of Medical Imaging: Medical Image Processing and Analysis*, volume 2 of *Press Monograph Series*, chapter 8, pages 447–514. SPIE Press, Bellingham, WA, 2000. 10

[93] I. Foster. *Designing and building parallel programs: concepts and tools for parallel software engineering.* Addison-Wesley, Boston, MA, 1995. 185, 192

[94] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, New York, NY, 2006. 1, 2, 3, 37, 45, 61, 65

[95] J. Frank. Single-particle reconstruction of biological macromolecules in electron microscopy – 30 years. *Q Rev Biophys*, 42(3):139–158, 2009. 45, 51, 149

[96] J. Frank and R. K. Agrawal. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, 406(6793):318–322, 2000. 49

[97] S. Frantz. *Local and Semi-Global Approaches to the Extraction of 3D Anatomical Landmarks from 3D Tomographic Images*. Akademische Verlagsgesellschaft Aka GmbH, Berlin, Germany, 2001. 14

[98] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans Pattern Anal Mach Intell*, 13(9):891–906, 1991. 18

[99] F. Förster, B.-G. Han, and M. Beck. Visual proteomics. *Methods Enzymol*, 483:215–243, 2010. 1, 10, 19, 50, 61

[100] W. Förstner. A feature based correspondence algorithm for image matching. *Int Arch Photogram Rem Sens*, 26(3):150–166, 1986. 15

[101] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Prococeedings of the ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281–305, 1987. 15

[102] I. S. Gabashvili, R. K. Agrawal, C. M. Spahn, R. A. Grassucci, D. I. Svergun, J. Frank, and P. Penczek. Solution structure of the E. coli 70S ribosome at 11.5 Å resolution. *Cell*, 100(5):537–549, 2000. 51, 61

[103] J. M. Galvez and M. Canton. Normalization and shape recognition of three-dimensional objects by 3D moments. *Pattern Recognit*, 26(5):667–681, 1993. 18

[104] S. J. Gamblin, L. F. Haire, R. J. Russell, D. J. Stevens, B. Xiao, Y. Ha, N. Vasisht, D. A. Steinhauer, R. S. Daniels, A. Elliot, D. C. Wiley, and J. J. Skehel. The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science*, 303(5665):1838–1842, 2004. 141

[105] V. Garcia, E. Debreuve, and M. Barlaud. Fast k nearest neighbor search using GPU. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, Los Alamitos, CA, 2008. IEEE Computer Society. 95, 185

[106] V. Garcia and F. Nielsen. Searching High-Dimensional Neighbours: CPU-Based Tailored Data-Structures Versus GPU-Based Brute-Force Method. In A. Gagalowicz and W. Philips, editors, *Computer Vision/Computer Graphics CollaborationTechniques*, volume 5496 of *Lecture Notes in Computer Science*, pages 425–436. Springer, Berlin, Germany, 2009. 95, 185

[107] J. I. Garzon, J. Kovacs, R. Abagyan, and P. Chacon. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*, 23(4):427–433, 2007. 19, 55, 60, 140, 142, 146

[108] S. H. Gellman. Introduction: Molecular Recognition. *Chem Rev*, 97(5):1231–1232, 1997. 94

[109] A. Gennerich and R. D. Vale. Walking the walk: how kinesin and dynein coordinate their steps. *Curr Opin Cell Biol*, 21(1):59–67, 2009. 61

[110] H. S. Gill and D. Eisenberg. The crystal structure of phosphinothricin in the active site of glutamine synthetase illuminates the mechanism of enzymatic inhibition. *Biochemistry*, 40(7):1903–1912, 2001. 141

[111] R. J. Gillespie and R. S. Nyholm. Inorganic stereochemistry. *Q Rev Chem Soc*, 11(4):339–380, 1957. 28, 73

[112] R. M. Glaeser. Cryo-electron microscopy of biological nanostructures. *Phys Today*, 61(1):48, 2008. 45

[113] T. D. Goddard, C. C. Huang, and T. E. Ferrin. Visualizing density maps with UCSF Chimera. *J Struct Biol*, 157(1):281–287, 2007. 54, 60

[114] D. M. Goldstein, T. Alfredson, J. Bertrand, M. F. Browner, K. Clifford, S. A. Dalrymple, J. Dunn, J. Freire-Moar, S. Harris, S. S. Labadie, J. L. Fargue, J. M. Lapierre, S. Larrabee, F. Li, E. Papp, D. McWeeney, C. Ramesha, R. Roberts, D. Rotstein, B. S. Pablo, E. B. Sjogren, O.-Y. So, F. X. Talamas, W. Tao, A. Trejo, A. Villasenor, M. Welch, T. Welch, P. Weller, P. E. Whiteley, K. Young, and S. Zipfel. Discovery of S-[5-amino-1-(4-fluorophenyl)-1H-pyrazol-4-yl]-[3-(2,3-dihydroxypropoxy)phenyl]methanone (RO3201195), an orally bioavailable and highly selective inhibitor of p38 MAP kinase. *J Med Chem*, 49(5):1562–1575, 2006. 184

[115] R. C. Gonzalez and R. E. Woods. *Digital image processing.* Pearson Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2008. 2, 12, 18, 23, 105, 108, 229

[116] I. Gordon and D. Lowe. What and Where: 3D Object Recognition with Accurate Pose. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 67–82, Berlin, Germany, 2006. Springer. 24

[117] J. Gore, Z. Bryant, M. D. Stone, M. Nöllmann, N. R. Cozzarelli, and C. Bustamante. Mechanochemical analysis of DNA gyrase using rotor bead tracking. *Nature*, 439(7072):100–104, 2006. 163

[118] A. Goshtasby. Registration of images with geometric distortions. *IEEE Trans Geosci Remote Sens*, 26(1):60–64, 1988. 10

[119] A. A. Goshtasby. *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications.* Wiley-Interscience, Hoboken, NJ, 2005. 2, 9, 11, 13, 15, 17, 18, 61

[120] M. Goto. Crystal Structure of T.th.HB8 Branched-Chain Amino Acid Aminotransferase Complexed with 4-Methylvaleric Acid. `http://www.ebi.ac.uk/pdbe`, PDB ID 2EIY, 2007. 103

[121] W. Grabarse, F. Mahlert, S. Shima, R. K. Thauer, and U. Ermler. Comparison of three methyl-coenzyme M reductases from phylogenetically distant organisms: unusual amino acid modification, conservation and adaptation. *J Mol Biol*, 303(2):329–344, 2000. 141

[122] R. P. Grant, D. Neuhaus, and M. Stewart. Structural basis for the interaction between the Tap/NXF1 UBA domain and FG nucleoporins at 1 Å resolution. *J Mol Biol*, 326(3):849–858, 2003. 103

[123] A. Griewel, O. Kayser, J. Schlosser, and M. Rarey. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. *J Chem Inf Model*, 49(10):2303–2311, 2009. 28

[124] A. Griewel and M. Rarey. Computational Reconstruction of Macromolecular Assemblies. In U. H. E. Hansmann, J. Meinke, S. Mohanty, and O. Zimmermann, editors, *Proceedings of the NIC Workshop From Computational Biophysics to Systems Biology*, volume 36 of *NIC Series*, pages 121–123, Jülich, 2007. John von Neumann Institute for Computing. 227

[125] M. Groll, H. Brandstetter, H. Bartunik, G. Bourenkow, and R. Huber. Investigations on the maturation and regulation of archaebacterial proteasomes. *J Mol Biol*, 327(1):75–83, 2003. 141

[126] A. M. Gulick. Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem Biol*, 4(10):811–827, 2009. 160

[127] A. M. Gulick, X. Lu, and D. Dunaway-Mariano. Crystal structure of 4-chlorobenzoate:CoA ligase/synthetase in the unliganded and aryl substrate-bound states. *Biochemistry*, 43(27):8670–8679, 2004. 176

[128] A. M. Gulick, V. J. Starai, A. R. Horswill, K. M. Homick, and J. C. Escalante-Semerena. The 1.75 Å crystal structure of acetyl-CoA synthetase bound to adenosine-5′-propylphosphate and coenzyme A. *Biochemistry*, 42(10):2866–2873, 2003. 160, 179

[129] I. Gutsche, L. O. Essen, and W. Baumeister. Group II chaperonins: new TRiC(k)s and turns of a protein folding machine. *J Mol Biol*, 293(2):295–312, 1999. 150, 151

[130] A. Guttman. R-trees: a dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, volume 14, pages 47–57, New York, NY, 1984. ACM. 93

[131] Y. Ha, D. Shi, G. W. Small, E. C. Theil, and N. M. Allewell. Crystal structure of bullfrog M ferritin at 2.8 Å resolution: analysis of subunit interactions and the binuclear metal center. *J Biol Inorg Chem*, 4(3):243–256, 1999. 141

[132] M. Hadjieleftheriou. Spatial Index Library. `http://libspatialindex.github.com`, Accessed October 2011. 93, 225, 227, 228

[133] C. Hadley and D. T. Jones. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, 7(9):1099–1112, 1999. 172

[134] J. V. Hajnal, D. J. Hawkes, and D. L. G. Hill. *Medical image registration*. Biomedical engineering series. CRC Press, Boca Raton, FL, 2001. 2, 10

[135] R. C. Hardison. A brief history of hemoglobins: plant, animal, protist, and bacteria. *Proc Natl Acad Sci USA*, 93(12):5675–5679, 1996. 165, 167

[136] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, volume 15, pages 147–151, 1988. 15

[137] T. Hartkens, K. Rohr, and H. S. Stiehl. Performance of 3D differential operators for the detection of anatomical point landmarks in MR and CT images. In K. M. Hanson, editor, *Medical Imaging 1999: Image Processing*, volume 3361 of *Proceedings of SPIE*, pages 32–43, San Diego, CA, 1999. SPIE Press. 15, 101, 106

[138] T. Hartkens, K. Rohr, and H. S. Stiehl. Evaluation of 3D operators for the detection of anatomical point landmarks in MR and CT images. *Comput Vision Image Understand*, 86(2):118–136, 2002. 15, 101, 106

[139] F. U. Hartl and M. Hayer-Hartl. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295(5561):1852–1858, 2002. 28, 150

[140] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, Berlin, Germany, 2nd edition, 2009. 96, 97, 119, 186

[141] J. Hattne and V. S. Lamzin. Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps. *Acta Crystallogr D Biol Crystallogr*, D64(8):834–842, 2008. 18

[142] W. He, P. Cowin, and D. L. Stokes. Untangling desmosomal knots with electron tomography. *Science*, 302(5642):109–113, 2003. 141

[143] T. L. Heath. *The Thirteen Books of the Elements*. Dover Publications, Mineola, NY, 1956. 74

[144] J. Heyd and S. Birmanns. Immersive structural biology: a new approach to hybrid modeling of macromolecular assemblies. *Virt Reality*, 13(4):245–255, 2009. 54, 60

[145] D. R. Hipp. SQLite. `http://www.sqlite.org`, Accessed October 2011. 95, 174, 225, 227, 228

[146] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. A database of protein structure families with common folding motifs. *Protein Sci*, 1(12):1691–1698, 1992. VI, 94, 172

[147] P. V. C. Hough. Method and means for recognizing complex patterns. US patent no. 3069654, December 1962. 14

[148] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Trans Inf Theory*, 8(2):179–187, 1962. 18

[149] Y. Hu, S. Faham, R. Roy, M. W. Adams, and D. C. Rees. Formaldehyde ferredoxin oxidoreductase from Pyrococcus furiosus: the 1.85 Å resolution crystal structure and its mechanistic implications. *J Mol Biol*, 286(3):899–914, 1999. 103

[150] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans Pattern Anal Mach Intell*, 15(9):850–863, 1993. 16

[151] J. Illingworth and J. Kittler. A survey of the Hough transform. *Comput Vision Graph Image Proc*, 44(1):87–116, 1988. 14, 17

[152] International Union of Pure and Applied Chemistry (IUPAC) and International Union of Biochemistry Joint Commission on Biochemical Nomenclature. Nomenclature and Symbolism for Amino Acids and Pepties. *Pure Appl Chem*, 56(5):595–624, 1984. 28

[153] T. M. Iverson, C. Luna-Chavez, L. R. Croal, G. Cecchini, and D. C. Rees. Crystallographic studies of the Escherichia coli quinol-fumarate reductase with inhibitors bound to the quinol-binding site. *J Biol Chem*, 277(18):16124–16130, 2002. 141

[154] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Comput Aided Des*, 37(5):509–530, 2005. 18

[155] G. J. Jensen, editor. *Cryo-EM Part A Sample Preparation and Data Collection*, volume 481 of *Methods in Enzymology*. Academic Press, Heidelberg, Germany, 2010. 45

[156] G. J. Jensen, editor. *Cryo-EM, Part B: 3-D Reconstruction*, volume 482 of *Methods in Enzymology*. Academic Press, Heidelberg, Germany, 2010. 45

[157] W. Jiang, M. L. Baker, S. J. Ludtke, and W. Chiu. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J Mol Biol*, 308(5):1033–1044, 2001. 53, 60

[158] W. Jiang and S. J. Ludtke. Electron cryomicroscopy of single particles at subnanometer resolution. *Curr Opin Struct Biol*, 15(5):571–577, 2005. 38, 149

[159] C. C. Jolley, S. A. Wells, P. Fromme, and M. Thorpe. Fitting Low-Resolution Cryo-EM Maps of Proteins Using Constrained Geometric Simulations. *Biophys J*, 94(5):1613–1621, 2008. 51, 65

[160] T. A. Jones and M. Kjeldgaard. Electron-density map interpretation. *Methods Enzymol*, 277:173–208, 1997. 38

[161] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A Found Crystallogr*, 32(5):922–923, 1976. 16

[162] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A Found Crystallogr*, 34(5):827–828, 1978. 16

[163] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. 29, 31

[164] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci USA*, 89(6):2195–2199, 1992. 18

[165] K. Kato and T. Hosino. Solving k-Nearest Neighbor Problem on Multiple Graphics Processors. In M. Parashar and R. Buyya, editors, *Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pages 769–773, Los Alamitos, CA, 2010. IEEE Computer Society. 95, 185

[166] G. Katona, R. C. Wilmouth, P. A. Wright, G. I. Berglund, J. Hajdu, R. Neutze, and C. J. Schofield. X-ray structure of a serine protease acyl-enzyme complex at 0.95-Å resolution. *J Biol Chem*, 277(24):21962–21970, 2002. 103

[167] T. Kawabata. Multiple Subunit Fitting into a Low-Resolution Density Map of a Macromolecular Complex Using a Gaussian Mixture Models. *Biophys J*, 95(10):4643–4658, 2008. 51

[168] V. Kaynig, B. Fischer, E. Muller, and J. M. Buhmann. Fully automatic stitching and distortion correction of transmission electron microscope images. *J Struct Biol*, 171(2):163–173, 2010. 24

[169] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In L. Kobbelt, P. Schröder, and H. Hoppe, editors, *Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, pages 156–164, Aire-la-Ville, Switzerland, 2003. Eurographics Association. 19

[170] Y. Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II 506–513, Los Alamitos, CA, 2004. IEEE Computer Society. 25, 192

[171] J. Keeler. *Understanding NMR Spectroscopy*. John Wiley & Sons Inc., Hoboken, NJ, 2nd edition, 2010. 34

[172] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666, 1958. 35

[173] R. Khayat, G. C. Lander, and J. E. Johnson. An automated procedure for detecting protein folds from sub-nanometer resolution electron density. *J Struct Biol*, 170(3):513–521, 2010. 57, 60

[174] J. S. Kim, M. Groll, H. J. Musiol, R. Behrendt, M. Kaiser, L. Moroder, R. Huber, and H. Brandstetter. Navigation inside a protease: substrate selection and product exit in the tricorn protease from Thermoplasma acidophilum. *J Mol Biol*, 324(5):1041–1050, 2002. 141

[175] A. R. Kinjo, R. Yamashita, and H. Nakamura. PDBj Mine: design and implementation of relational database interface for Protein Data Bank Japan. *Database (Oxford)*, 2010:baq021, 2010. 34

[176] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. 93

[177] L. Kitchen and A. Rosenfeld. Gray-level corner detection. *Pattern Recognit Lett*, 1(2):95–102, 1982. 14

[178] G. J. Kleywegt, M. R. Harris, J. Y. Zou, T. C. Taylor, A. Wahlby, and T. A. Jones. The Uppsala Electron-Density Server. *Acta Crystallogr D*

Biol Crystallogr, 60(12):2240–2249, 2004. `http://eds.bmc.uu.se`. 3, 45, 149, 160

[179] P. Koehl. Protein Structure Classification. In K. B. Lipkowitz, T. R. Cundari, and V. J. Gillet, editors, *Reviews in Computational Chemistry*, volume 22 of *Reviews in Computational Chemistry*, chapter 1, pages 1–56. John Wiley & Sons Inc., Hoboken, NJ, 2006. 29, 94, 172

[180] J. J. Koenderink. The structure of images. *Biol Cybern*, 50(5):363–370, 1984. 20, 21

[181] W. L. Koltun. Space filling atomic units and connectors for molecular models. US patent no. 3170246, February 1965. 29

[182] D. E. Koshland. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci USA*, 44(2):98–104, 1958. 28

[183] J. A. Kovacs, P. Chacón, Y. Cong, E. Metwally, and W. Wriggers. Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr D Biol Crystallogr*, 59(Pt 8):1371–1376, 2003. 55, 57, 60

[184] J. A. Kovacs and W. Wriggers. Fast rotational matching. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 8):1282–1286, 2002. 55, 60

[185] N. Kresge, V. D. Vacquier, and C. D. Stout. 1.35 and 2.07 Å resolution structures of the red abalone sperm lysin monomer and dimer reveal features involved in receptor binding. *Acta Crystallogr D Biol Crystallogr*, 56(Pt 1):34–41, 2000. 103

[186] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley. C_{60}: Buckminsterfullerene. *Nature*, 318(6042):162–163, 1985. 28, 73

[187] J. J. Kuffner. Effective Sampling and Distance Metrics for 3D Rigid Body Path Planning. In *Proceedings of the IEEE Conference on Robotics and Automation*, volume 4, pages 3993–3998, Los Alamitos, CA, 2004. IEEE Computer Society. 115

[188] F. Kuhl, G. Crippen, and D. Friesen. A Combinatorial Algorithm For Calculating Ligand-Binding. *J Comput Chem*, 5(1):24–34, 1984. 17

[189] R. Kuroki, L. H. Weaver, and B. W. Matthews. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. *Science*, 262(5142):2030–2033, 1993. 176

[190] A. R. Kusmierczyk and J. Martin. Nucleotide-dependent protein folding in the type II chaperonin from the mesophilic archaeon Methanococcus maripaludis. *Biochem J*, 371(Pt 3):669–673, 2003. 150

[191] S. Lanzavecchia, F. Cantele, and P. L. Bellon. Alignment of 3D structures of macromolecular assemblies. *Bioinformatics*, 17(1):58–62, 2001. 18

[192] K. Lasker, O. Dror, R. Nussinov, and H. Wolfson. Discovery of Protein Substructures in EM Maps. In R. Casadio and G. Myers, editors, *Algorithms in Bioinformatics*, volume 3692 of *Lecture Notes in Computer Science*, pages 423–434, Berlin, Germany, 2005. Springer. 55, 60

[193] K. Lasker, O. Dror, M. Shatsky, R. Nussinov, and H. J. Wolfson. EMatch: discovery of high resolution structural homologues of protein domains in intermediate resolution cryo-EM maps. *IEEE/ACM Trans Comput Biol Bioinform*, 4(1):28–39, 2007. 55, 60

[194] K. Lasker, M. Topf, A. Sali, and H. J. Wolfson. Inferential Optimization for Simultaneous Fitting of Multiple Components into a CryoEM Map of their Assembly. *J Mol Biol*, 388(1):180–194, 2009. 51

[195] C. L. Lawson. Unified data resource for cryo-EM. *Methods Enzymol*, 483:73–90, 2010. 49

[196] C. L. Lawson, M. L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, G. van Ginkel, B. Devkota, I. Lagerstedt, S. J. Ludtke, R. H. Newman, T. J. Oldfield, I. Rees, G. Sahni, R. Sala, S. Velankar, J. Warren, J. D. Westbrook, K. Henrick, G. J. Kleywegt, H. M. Berman, and W. Chiu. EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res*, 39(Database issue):D456–D464, 2011. `http://www.emdatabank.org`. V, 3, 49, 149, 200

[197] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In A. Hoppe, S. Barman, and T. Ellis, editors, *Proceedings of the British Machine Vision Conference*, volume 2, pages 959–968, Kingston, United Kingdom, 2004. British Machine Vision Association. 25

[198] A. L. Lehninger, D. L. Nelson, and M. M. Cox. *Lehninger Principles of Biochemistry*, volume 4. W. H. Freeman, New York, NY, 2005. 165

[199] H. Li and D. G. Thanassi. Use of a combined cryo-EM and X-ray crystallography approach to reveal molecular details of bacterial pilus assembly by the chaperone/usher pathway. *Curr Opin Microbiol*, 12(3):326–332, 2009. 51, 61

[200] X. Li, H. Lee, J. Wu, and E. Breslow. Contributions of the interdomain loop, amino terminus, and subunit interface to the ligand-facilitated dimerization of neurophysin: crystal structures and mutation studies of bovine neurophysin-I. *Protein Sci*, 16(1):52–68, 2007. 103

[201] S. Liang, Y. Liu, C. Wang, and L. Jian. A CUDA-based parallel implementation of K-nearest neighbor algorithm. In *Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 291–296, Los Alamitos, CA, 2009. IEEE Computer Society. 95, 185

[202] S. Liang, C. Wang, Y. Liu, and L. Jian. CUKNN: A parallel implementation of K-nearest neighbor on CUDA-enabled GPU. In *Proceedings of the IEEE Youth Conference on Information, Computing and Telecommunication*, pages 415–418, Los Alamitos, CA, 2009. IEEE Computer Society. 95, 185

[203] L. M. Lifshitz and S. M. Pizer. A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Trans Pattern Anal Mach Intell*, 12(6):529–540, 1990. 21

[204] J.-H. Lin and T. Clark. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J Chem Inf Model*, 45(4):1010–1016, 2005. 19

[205] T. Lindeberg. Scale-space for discrete signals. *IEEE Trans Pattern Anal Mach Intell*, 12(3):234–254, 1990. 21, 114

[206] T. Lindeberg. *Discrete Scale-Space Theory and the Scale-Space Primal Sketch*. PhD thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden, 1991. 22

[207] T. Lindeberg. Junction detection with automatic selection of detection scales and localization scales. In *Proceedings of the IEEE Conference on Image Processing*, volume 1, pages 924–928, Los Alamitos, CA, 1994. IEEE Computer Society. 22

[208] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *J Appl Statist*, 21(2):225–270, 1994. 21, 23, 24, 26

[209] T. Lindeberg. On the axiomatic foundations of linear scale-space: Combining semi-group structure with causality vs. scale invariance. In J. Sporring and M. Nielsen, editors, *Gaussian Scale-Space Theory: Proc. PhD School*

on Scale-Space Theory, chapter 6, pages 75–98. Kluwer Academic Publishers, London, United Kingdom, 1997. 21

[210] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *Int J Comput Vision*, 30(2):117–154, 1998. 22, 69

[211] T. Lindeberg. Feature detection with automatic scale selection. *Int J Comput Vision*, 30(2):79–116, 1998. 22

[212] T. Lindeberg. Principles for automatic scale selection. In B. Jähne, H. Haussecker, and P. Geissler, editors, *Handbook on Computer Vision and Applications*, volume 2, pages 239–274. Academic Press, Boston, MA, 1999. 22

[213] T. Lindeberg. Scale-Space. In B. Wah, editor, *Encyclopedia of Computer Science and Engineering*, volume IV, pages 2495–2504. John Wiley and Sons, Hoboken, NJ, 2009. 3, 19, 22, 97

[214] S. Lindert, P. L. Stewart, and J. Meiler. Hybrid approaches: applying computational methods in cryo-electron microscopy. *Curr Opin Struct Biol*, 19(2):218–225, 2009. 38, 50, 61

[215] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In M. C. Stone, editor, *Proceedings of the 14^{th} Annual Conference on Computer Graphics and Interactive Techniques*, pages 163–169, New York, NY, 1987. ACM. 38, 229

[216] R. Loris, M. H. Dao-Thi, E. M. Bahassi, L. V. Melderen, F. Poortmans, R. Liddington, M. Couturier, and L. Wyns. Crystal structure of CcdB, a topoisomerase poison from E. coli. *J Mol Biol*, 285(4):1667–1677, 1999. 177

[217] D. G. Lowe. Object Recognition from Local Scale-Invariant Features. In B. Werner, editor, *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume II, pages 1150–1157, Los Alamitos, CA, 1999. IEEE Computer Society. 18, 23, 63, 66, 97

[218] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int J Comput Vision*, 60(2):91–110, 2004. 3, 23, 24, 25, 63, 71, 73, 90, 96, 97, 101, 186

[219] S. J. Ludtke, M. L. Baker, D.-H. Chen, J.-L. Song, D. T. Chuang, and W. Chiu. De novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure*, 16(3):441–448, 2008. 50, 51, 148, 150, 176

[220] J. Löwe, D. Stock, B. Jap, P. Zwickl, W. Baumeister, and R. Huber. Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 Å resolution. *Science*, 268(5210):533–539, 1995. 141

[221] J. B. Maintz and M. A. Viergever. A survey of medical image registration. *Med Image Anal*, 2(1):1–36, 1998. 10

[222] J. M. Mancheño, H. Tateno, I. J. Goldstein, M. Martínez-Ripoll, and J. A. Hermoso. Structural analysis of the Laetiporus sulphureus hemolytic pore-forming lectin in complex with sugars. *J Biol Chem*, 280(17):17251–17259, 2005. 141

[223] T. C. Marlovits, T. Kubori, M. Lara-Tejero, D. Thomas, V. M. Unger, and J. E. Galán. Assembly of the inner rod determines needle length in the type III secretion injectisome. *Nature*, 441(7093):637–640, 2006. 61

[224] T. C. Marlovits, T. Kubori, A. Sukhan, D. R. Thomas, J. E. Galán, and V. M. Unger. Structural insights into the assembly of the type III secretion needle complex. *Science*, 306(5698):1040–1042, 2004. 61

[225] D. Marr and E. Hildreth. Theory of edge detection. *Proc R Soc London, Ser B*, 207(1167):187–217, 1980. 13

[226] M. T. Mason. *Mechanics of Robotic Manipulation*. Intelligent robots and autonomous agents. MIT Press, Cumberland, RI, 2001. 26

[227] M. Mathieu, I. Petitpas, J. Navaza, J. Lepault, E. Kohli, P. Pothier, B. V. Prasad, J. Cohen, and F. A. Rey. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. *EMBO J*, 20(7):1485–1497, 2001. 152

[228] M. Mathieu, I. Petitpas, J. Navaza, J. Lepault, E. Kohli, P. Pothier, B. V. Prasad, J. Cohen, and F. A. Rey. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. *EMBO J*, 20(7):1485–1497, 2001. 184

[229] A. Maxwell. DNA gyrase as a drug target. *Biochem Soc Trans*, 27(2):48–53, 1999. 163

[230] J. J. May, N. Kessler, M. A. Marahiel, and M. T. Stubbs. Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc Natl Acad Sci USA*, 99(19):12120–12125, 2002. 179

[231] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read. Phaser crystallographic software. *J Appl Crystallogr*, 40(Pt 4):658–674, 2007. 55, 60

[232] A. J. McCoy, R. W. Grosse-Kunstleve, L. C. Storoni, and R. J. Read. Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr*, 61(Pt 4):458–464, 2005. 55, 60

[233] A. D. McNaught, A. Wilkinson, M. Nic, J. Jirat, B. Kosata, and A. Jenkins. *IUPAC. Compendium of Chemical Terminology (the "Gold Book") and XML on-line corrected version.* Blackwell Scientific Publications, Oxford, United Kingdom, 2nd edition, 1997. `http://goldbook.iupac.org`. 27, 28

[234] S. G. Megason and S. E. Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 2007. 1, 50, 61

[235] P. Meyer, C. Prodromou, C. Liao, B. Hu, S. M. Roe, C. K. Vaughan, I. Vlasic, B. Panaretou, P. W. Piper, and L. H. Pearl. Structural basis for recruitment of the ATPase activator Aha1 to the Hsp90 chaperone machinery. *EMBO J*, 23(6):1402–1410, 2004. 179

[236] S. Meyers. *Effective STL: 50 specific ways to improve your use of the standard template library.* Addison-Wesley Professional Computing Series. Addison-Wesley, Boston, USA, 2001. 227

[237] K. Mikolajczyk and C. Schmid. Scale & Affine Invariant Interest Point Detectors. *Int J Comput Vision*, 60(1):63–86, 2004. 22, 101

[238] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans Pattern Anal Mach Intell*, 27(10):1615–1630, 2005. 18, 25, 101

[239] T. Minamino, K. Imada, and K. Namba. Molecular motors of the bacterial flagella. *Curr Opin Struct Biol*, 18(6):693–701, 2008. 61

[240] K. Mitra and J. Frank. Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps. *Annu Rev Biophys Biomol Struct*, 35(1):299–317, 2006. 51, 61

[241] J. Modersitzki. *Numerical Methods for Image Registration.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, NY, 2004. 9

[242] P. J. Mohr, B. N. Taylor, and D. B. Newell. CODATA recommended values of the fundamental physical constants: 2006. *Rev Mod Phys*, 80(2):633–730, 2008. 28

[243] R. J. Morris, R. J. Najmanovich, A. Kahraman, and J. M. Thornton. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355, 2005. 19

[244] M.-H. Mousa, R. Chaine, S. Akkouche, and E. Galin. Toward an efficient triangle-based spherical harmonics representation of 3D objects. *Comput Aided Geom Des*, 25(8):561–575, 2008. 19

[245] K. Murakami, T. Yasunaga, T. Q. P. Noguchi, Y. Gomibuchi, K. X. Ngo, T. Q. P. Uyeda, and T. Wakabayashi. Structural basis for actin assembly, activation of ATP hydrolysis, and delayed phosphate release. *Cell*, 143(2):275–287, 2010. 32

[246] G. E. Murphy, J. R. Leadbetter, and G. J. Jensen. In situ structure of the complete Treponema primitia flagellar motor. *Nature*, 442(7106):1062–1064, 2006. 61

[247] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, 1995. VII, 94, 172, 200

[248] N. Nandhagopal, A. A. Simpson, J. R. Gurnon, X. Yan, T. S. Baker, M. V. Graves, J. L. V. Etten, and M. G. Rossmann. The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. *Proc Natl Acad Sci USA*, 99(23):14758–14763, 2002. 103

[249] J. Navaza. AMoRe: an automated package for molecular replacement. *Acta Crystallogr A Found Crystallogr*, 50(2):157–163, 1994. 56, 60

[250] J. Navaza, J. Lepault, F. A. Rey, C. Alvarez-Rúa, and J. Borge. On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 10 Pt 2):1820–1825, 2002. 56, 60

[251] D. Ni, Y. P. Chui, Y. Qu, X. Yang, J. Qin, T.-T. Wong, S. S. H. Ho, and P. A. Heng. Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT. *Comput Med Imag Graph*, 33(7):559–566, 2009. 26

[252] D. Ni, Y. Qu, X. Yang, Y. Chui, T.-T. Wong, S. Ho, and P. Heng. Volumetric Ultrasound Panorama Based on 3D SIFT. In D. Metaxas, L. Axel, G. Fichtinger, and G. Székely, editors, *Medical Image Computing and Computer-Assisted Intervention*, volume 5242 of *Lecture Notes in Computer Science*, pages 52–60, Berlin, Germany, 2008. Springer. 26

[253] M. Niemeijer, M. K. Garvin, K. Lee, B. van Ginneken, M. D. Abràmoff, and M. Sonka. Registration of 3D spectral OCT volumes using 3D SIFT feature point matching. In J. P. W. Pluim and B. M. Dawant, editors, *Medical Imaging 2009: Image Processing*, volume 7259 of *Proceedings of SPIE*, pages I 1–8. SPIE Press, 2009. 26

[254] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH—a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997. V, 94, 102, 172, 200

[255] K. Ozawa, Y. Takayama, F. Yasukawa, T. Ohmura, M. A. Cusanovich, Y. Tomimoto, H. Ogata, Y. Higuchi, and H. Akutsu. Role of the aromatic ring of Tyr43 in tetraheme cytochrome c_3 from Desulfovibrio vulgaris Miyazaki F. *Biophys J*, 85(5):3367–3374, 2003. 103

[256] S.-Y. Park, T. Yokoyama, N. Shibayama, Y. Shiro, and J. R. H. Tame. 1.25 Å resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. *J Mol Biol*, 360(3):690–701, 2006. 167, 176

[257] J. Patriarche and B. Erickson. A review of the automated detection of change in serial imaging studies of the brain. *J Digit Imaging*, 17(3):158–174, 2004. 10

[258] A. L. Patterson. A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals. *Phys Rev*, 46(5):372–376, 1934. 52

[259] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA*, 37(4):205–211, 1951. 29

[260] P. A. Penczek. Resolution measures in molecular electron microscopy. *Methods Enzymol*, 482:73–100, 2010. 48, 65

[261] J. Pesavento, S. Crawford, M. Estes, and B. Venkataram Prasad. Rotavirus Proteins: Structure and Assembly. In P. Roy, editor, *Reoviruses:*

Entry, Assembly and Morphogenesis, volume 309 of *Current Topics in Microbiology and Immunology*, pages 189–219. Springer, Berlin, Germany, 2006. 152

[262] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612, 2004. 54, 60, 225

[263] G. Pintilie, J. Zhang, W. Chiu, and D. Gossard. Identifying Components in 3D Density Maps of Protein Nanomachines by Multi-scale Segmentation. In *Proceedings of the IEEE/NIH Life Science Systems and Applications Workshop*, pages 44–47, Los Alamitos, CA, 2009. IEEE Computer Society. 56, 60

[264] G. D. Pintilie, J. Zhang, T. D. Goddard, W. Chiu, and D. C. Gossard. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J Struct Biol*, 170(3):427–438, 2010. 56, 60, 65

[265] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, and R. Quick. The open science grid. *J Phys: Conference Series*, 78(1):012057, 2007. 58

[266] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C Example Book: The Art of Scientific Computing*. Cambridge University Press, New York, NY, 1992. 53

[267] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. `http://www.R-project.org/`. 225

[268] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:95–99, 1963. 43

[269] M. Rarey. `http://www.biosolveit.de/flexv`, Accessed October 2011. 192, 227, 228

[270] M. Rarey, S. Wefing, and T. Lengauer. Placement of Medium Sized Molecular Fragments into Active Sites of Proteins. *J Comput Aided-Mol Des*, 10(1):41–54, 1996. 91, 227

[271] R. J. Read. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr*, 57(Pt 10):1373–1382, 2001. 55, 60

[272] A. S. Reger, R. Wu, D. Dunaway-Mariano, and A. M. Gulick. Structural characterization of a 140 degrees domain movement in the two-step reaction catalyzed by 4-chlorobenzoate:CoA ligase. *Biochemistry*, 47(31):8016–8025, 2008. 179

[273] G. Rhodes. *Crystallography made crystal clear: a guide for users of macromolecular models*. Complementary Science Series. Academic Press / Elsevier, London, United Kingdom, 3rd edition, 2006. 1, 2, 37, 38, 41, 51, 61, 65

[274] K. Rohr. Modelling and identification of characteristic intensity variations. *Imag Vision Comput*, 10(2):66–76, 1992. 15, 26

[275] K. Rohr. Localization properties of direct corner detectors. *J Math Imaging Vision*, 4(2):139–150, 1994. 15, 26

[276] K. Rohr. On 3D differential operators for detecting point landmarks. *Imag Vision Comput*, 15(3):219–233, 1997. 14

[277] K. Rohr. *Landmark-Based Image Analysis: Using Geometric and Intensity Models*. Computational Imaging and Vision. Kluwer Academic Publishers, London, United Kingdom, 2001. 14

[278] K. Rohr, H. S. Stiehl, S. Frantz, and T. Hartkens. Performance Characterization of Landmark Operators. In R. Klette, H. S. Stiehl, M. Viergever, and K. Vincken, editors, *Performance Characterization in Computer Vision*, volume 17 of *Computational Imaging and Vision*, pages 285–297. Kluwer Academic Publishers, London, United Kingdom, 2000. 15, 101, 106

[279] A. M. Roseman. Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Crystallogr D Biol Crystallogr*, 56(Pt 10):1332–1340, 2000. 53, 60, 65

[280] M. G. Rossmann. Fitting atomic models into electron-microscopy maps. *Acta Crystallogr D Biol Crystallogr*, 56(Pt 10):1341–1349, 2000. 53, 60

[281] M. G. Rossmann, R. Bernal, and S. V. Pletnev. Combining electron microscopic with X-ray crystallographic structures. *J Struct Biol*, 136(3):190–200, 2001. 53, 60, 93

[282] M. G. Rossmann, M. C. Morais, P. G. Leiman, and W. Zhang. Combining X-ray crystallography and electron microscopy. *Structure*, 13(3):355–362, 2005. 50, 61

[283] E. Rosten and T. Drummond. Machine Learning for High-Speed Corner Detection. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision - ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 430–443, Berlin, Germany, 2006. Springer. 15

[284] W. E. Royer, K. Strand, M. van Heel, and W. A. Hendrickson. Structural hierarchy in erythrocruorin, the giant respiratory assemblage of annelids. *Proc Natl Acad Sci USA*, 97(13):7107–7111, 2000. 161

[285] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–324, 2004. 50

[286] M. Rusu and S. Birmanns. Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions. *J Struct Biol*, 170(1):164–171, 2010. 51

[287] M. Rusu, S. Birmanns, and W. Wriggers. Biomolecular pleiomorphism probed by spatial interpolation of coarse models. *Bioinformatics*, 24(21):2460–2466, 2008. 51

[288] F. A. Sadjadi and E. L. Hall. Three-dimensional moment invariants. *IEEE Trans Pattern Anal Mach Intell*, 2(2):127–136, 1980. 18

[289] E. B. Saff and A. B. J. Kuijlaars. Distributing many points on a sphere. *Math Intell*, 19(1):5–11, 1997. 73, 74

[290] M. Saha, M. Levitt, and W. Chiu. MOTIF-EM: an automated computational tool for identifying conserved regions in CryoEM structures. *Bioinformatics*, 26(12):i301–i309, 2010. 56, 60

[291] K. Sahr, D. White, and A. J. Kimerling. Geodesic Discrete Global Grid Systems. *Cartograph Geograph Inf Sci*, 30(2):121–135, 2003. 73, 74

[292] A. Sali, R. Glaeser, T. Earnest, and W. Baumeister. From words to literature in structural proteomics. *Nature*, 422(6928):216–225, 2003. 1, 50, 61

[293] D. Sarkar. *Lattice: multivariate data visualization with R*. Use R! Springer, Berlin, Germany, 2008. 130

[294] D. Saupe and D. Vranić. 3D Model Retrieval with Spherical Harmonics and Moments. In B. Radig and S. Florczyk, editors, *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, volume 2191 of *Lecture Notes in Computer Science*, pages 392–397, Berlin, Germany, 2001. Springer. 18, 19

[295] H. A. Schreuder, A. Mattevi, G. Obmolova, K. H. Kalk, W. G. Hol, F. J. van der Bolt, and W. J. van Berkel. Crystal structures of wild-type p-hydroxybenzoate hydroxylase complexed with 4-aminobenzoate,2,4-dihydroxybenzoate, and 2-hydroxy-4-aminobenzoate and of the Tyr222Ala mutant complexed with 2-hydroxy-4-aminobenzoate. Evidence for a proton channel and a new binding mode of the flavin ring. *Biochemistry*, 33(33):10161–10170, 1994. 39, 92, 176

[296] G. F. Schröder, A. T. Brunger, and M. Levitt. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure*, 15(12):1630–1641, 2007. 51

[297] P. Scovanner, S. Ali, and M. Shah. A 3-Dimensional SIFT Descriptor and its Application to Action Recognition. In *Proceedings of the 15th International Conference on Multimedia*, pages 357–360, New York, NY, 2007. ACM. 26

[298] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, pages 517–524, New York, NY, 1968. ACM. 76

[299] H. Shi and P. B. Moore. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: a classic structure revisited. *RNA*, 6(8):1091–1105, 2000. 32

[300] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the 1994 IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, Los Alamitos, CA, 1994. IEEE Computer Society. 15

[301] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004. 37

[302] X. Siebert and J. Navaza. UROX 2.0: an interactive tool for fitting atomic models into electron-microscopy reconstructions. *Acta Crystallogr D Biol Crystallogr*, 65(Pt 7):651–658, 2009. 56, 60

[303] A. A. Simpson, P. G. Leiman, Y. Tao, Y. He, M. O. Badasso, P. J. Jardine, D. L. Anderson, and M. G. Rossmann. Structure determination of the head-tail connector of bacteriophage phi29. *Acta Crystallogr D Biol Crystallogr*, 57(Pt 9):1260–1269, 2001. 141

[304] M. R. Singleton, L. M. Wentzell, Y. Liu, S. C. West, and D. B. Wigley. Structure of the single-strand annealing domain of human RAD52 protein. *Proc Natl Acad Sci USA*, 99(21):13492–13497, 2002. 141

[305] T. Skarzyński, P. C. Moody, and A. J. Wonacott. Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from Bacillus stearothermophilus at 1.8 Å resolution. *J Mol Biol*, 193(1):171–187, 1987. 141

[306] P. Slomka and R. Baum. Multimodality image registration with software: state-of-the-art. *Eur J Nucl Med Mol Imaging*, 36:44–55, 2009. 10

[307] R. Smith and B. Carragher. Software tools for molecular microscopy. *J Struct Biol*, 163(3):224–228, 2008. 45

[308] S. M. Smith and J. M. Brady. SUSAN – A New Approach to Low Level Image Processing. *Int J Comput Vision*, 23(1):45–78, 1997. 15

[309] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981. 167

[310] T. J. Smith, T. Schmidt, J. Fang, J. Wu, G. Siuzdak, and C. A. Stanley. The structure of apo human glutamate dehydrogenase details subunit communication and allostery. *J Mol Biol*, 318(3):765–777, 2002. 141

[311] L. Song, M. R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J. E. Gouaux. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science*, 274(5294):1859–1866, 1996. 141

[312] H. Stark, P. Dube, R. Lührmann, and B. Kastner. Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. *Nature*, 409(6819):539–542, 2001. 51, 61

[313] A. C. Steven and W. Baumeister. The future is hybrid. *J Struct Biol*, 163(3):186–195, 2008. 50, 61

[314] I. Stokes-Rees and P. Sliz. Protein structure determination by exhaustive search of Protein Data Bank derived databases. *Proc Natl Acad Sci USA*, 107(50):21476–21481, 2010. 58, 60

[315] L. C. Storoni, A. J. McCoy, and R. J. Read. Likelihood-enhanced fast rotation functions. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 3):432–438, 2004. 55, 60

[316] K. Strand, J. E. Knapp, B. Bhyravbhatla, and W. E. Royer. Crystal structure of the hemoglobin dodecamer from Lumbricus erythrocruorin: allosteric core of giant annelid respiratory complexes. *J Mol Biol*, 344(1):119–134, 2004. 161

[317] B. Stroustrup. *The C++ Programming Language*, volume 3. Addison-Wesley, Boston, MA, 2000. 225, 227, 228

[318] K. Suhre, J. Navaza, and Y. H. Sanejouand. NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr D Biol Crystallogr*, 62(Pt 9):1098–1100, 2006. 51

[319] D. I. Svergun. Small-angle X-ray and neutron scattering as a tool for structural systems biology. *Biol Chem*, 391(7):737–743, 2010. 34

[320] F. Tama, O. Miyashita, and C. L. Brooks. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J Mol Biol*, 337(4):985–999, 2004. 65

[321] F. Tama, O. Miyashita, and C. L. Brooks. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J Struct Biol*, 147(3):315–326, 2004. 51

[322] F. Tama, W. Wriggers, and C. L. Brooks. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J Mol Biol*, 321(2):297–305, 2002. 51

[323] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, Boston, MA, 2006. 131

[324] R. K.-Z. Tan, B. Devkota, and S. C. Harvey. YUP.SCX: coaxing atomic models into medium resolution electron density maps. *J Struct Biol*, 163(2):163–174, 2008. 51

[325] P. Taylor, J. Dornan, A. Carrello, R. F. Minchin, T. Ratajczak, and M. D. Walkinshaw. Two structures of cyclophilin 40: folding and fidelity in the TPR domains. *Structure*, 9(5):431–438, 2001. 103

[326] T. C. Taylor, A. Backlund, K. Bjorhall, R. J. Spreitzer, and I. Andersson. First crystal structure of Rubisco from a green alga, Chlamydomonas reinhardtii. *J Biol Chem*, 276(51):48159–48164, 2001. 141

[327] W. D. Tolbert, J. Daugherty, C. Gao, Q. Xie, C. Miranti, E. Gherardi, G. V. Woude, and H. E. Xu. A mechanistic basis for converting a receptor tyrosine kinase agonist to an antagonist. *Proc Natl Acad Sci USA*, 104(37):14592–14597, 2007. 103

[328] M. Topf, M. L. Baker, B. John, W. Chiu, and A. Sali. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J Struct Biol*, 149(2):191–203, 2005. 55, 60, 65

[329] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali. Protein structure fitting and refinement guided by cryo-EM density. *Structure*, 16(2):295–307, 2008. 51, 65

[330] W. W. C. Topley, L. Collier, G. S. Wilson, L. H. Collier, A. Balows, and M. Sussman. *Topley & Wilson's Microbiology and Microbial Infections: Virology*. Arnold, London, United Kingdom, 1998. 73

[331] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods*, 49(2):174–180, 2009. 51

[332] C.-J. Tsai and C. Ziegler. Coupling electron cryomicroscopy and X-ray crystallography to understand secondary active transport. *Curr Opin Struct Biol*, 20(4):448–455, 2010. 50, 61

[333] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Found Trends Comput Graphics Vision*, 3(3):177–280, 2008. 14, 15

[334] T. C. Ullrich, M. Blaesse, and R. Huber. Crystal structure of ATP sulfurylase from Saccharomyces cerevisiae, a key enzyme in sulfate activation. *EMBO J*, 20(3):316–329, 2001. 141

[335] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Res*, 36(Database issue):D402–D408, 2008. 34

[336] S. Umhau, L. Pollegioni, G. Molla, K. Diederichs, W. Welte, M. S. Pilone, and S. Ghisla. The x-ray structure of D-amino acid oxidase at very high resolution identifies the chemical mechanism of flavin-dependent substrate dehydrogenation. *Proc Natl Acad Sci USA*, 97(23):12463–12468, 2000. 103

[337] M. Urschler, J. Bauer, H. Ditt, and H. Bischof. SIFT and Shape Context for Feature-Based Nonlinear Registration of Thoracic CT Images. In R. Beichel and M. Sonka, editors, *Computer Vision Approaches to Medical Image Analysis*, volume 4241 of *Lecture Notes in Computer Science*, pages 73–84, Berlin, Germany, 2006. Springer. 26

[338] A. Vagin and A. Teplyakov. MOLREP: an Automated Program for Molecular Replacement. *J Appl Crystallogr*, 30(6):1022–1025, 1997. 52, 60

[339] A. Vagin and A. Teplyakov. An approach to multi-copy search in molecular replacement. *Acta Crystallogr D Biol Crystallogr*, 56(Pt 12):1622–1624, 2000. 52, 60

[340] A. A. Vagin and M. N. Isupov. Spherically averaged phased translation function and its application to the search for molecules and fragments in electron-density maps. *Acta Crystallogr D Biol Crystallogr*, 57(Pt 10):1451–1456, 2001. 52, 60

[341] M. van Heel. Unveiling ribosomal structures: the final phases. *Curr Opin Struct Biol*, 10(2):259–264, 2000. 37, 48, 65

[342] M. van Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution. *Q Rev Biophys*, 33(4):307–369, 2000. 149

[343] M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz. A new generation of the IMAGIC image processing system. *J Struct Biol*, 116(1):17–24, 1996. 37

[344] M. van Heel and M. Schatz. Fourier shell correlation threshold criteria. *J Struct Biol*, 151(3):250–262, 2005. 48, 65

[345] S. Velankar, C. Best, B. Beuth, C. H. Boutselakis, N. Cobley, A. W. S. D. Silva, D. Dimitropoulos, A. Golovin, M. Hirshberg, M. John, E. B. Krissinel, R. Newman, T. Oldfield, A. Pajon, C. J. Penkett, J. Pineda-Castillo, G. Sahni, S. Sen, R. Slowley, A. Suarez-Uruena, J. Swaminathan, G. van Ginkel, W. F. Vranken, K. Henrick, and G. J. Kleywegt.

PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*, 38(Database issue):D308–D317, 2010. `http://www.ebi.ac.uk/pdbe`. 34

[346] J. A. Velazquez-Muriel, C. O. S. Sorzano, S. H. W. Scheres, and J.-M. Carazo. SPI-EM: towards a tool for predicting CATH superfamilies in 3D-EM maps. *J Mol Biol*, 345(4):759–771, 2005. 57, 60

[347] M. Vogtherr, K. Saxena, S. Hoelder, S. Grimme, M. Betz, U. Schieborr, B. Pescatore, M. Robin, L. Delarbre, T. Langer, K. U. Wendt, and H. Schwalbe. NMR characterization of kinase p38 dynamics in free and ligand-bound forms. *Angew Chem Int Ed Engl*, 45(6):993–997, 2006. 176

[348] A. Volkamer, A. Griewel, T. Grombacher, and M. Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J Chem Inf Model*, 50(11):2041–2052, 2010. 229

[349] N. Volkmann. Confidence intervals for fitting of atomic models into low-resolution densities. *Acta Crystallogr D Biol Crystallogr*, 65(Pt 7):679–689, 2009. 52, 60

[350] N. Volkmann and D. Hanein. Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J Struct Biol*, 125(2–3):176–184, 1999. 52, 60, 65

[351] N. Volkmann and D. Hanein. Docking of atomic models into reconstructions from electron microscopy. *Methods Enzymol*, 374:204–225, 2003. 52, 60

[352] N. Volkmann, H. Liu, L. Hazelwood, E. B. Krementsova, S. Lowey, K. M. Trybus, and D. Hanein. The structural basis of myosin V processive movement as revealed by electron cryomicroscopy. *Mol Cell*, 19(5):595–605, 2005. 61

[353] D. V. Vranić, D. Saupe, and J. Richter. Tools for 3D-object retrieval: Karhunen-Loeve transform and spherical harmonics. In J.-L. Dugelay and K. Ros, editors, *Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing*, pages 293–298, Los Alamitos, CA, 2001. IEEE Computer Society. 18, 19

[354] D. A. Wah, C. Fernández-Tornero, L. Sanz, A. Romero, and J. J. Calvete. Sperm coating mechanism from the 1.8 Å crystal structure of PDC-109-phosphorylcholine complex. *Structure*, 10(4):505–514, 2002. 103

[355] T. Walter, D. W. Shattuck, R. Baldock, M. E. Bastin, A. E. Carpenter, S. Duce, J. Ellenberg, A. Fraser, N. Hamilton, S. Pieper, M. A. Ragan, J. E. Schneider, P. Tomancak, and J.-K. Hériché. Visualization of image data from cells to organisms. *Nat Methods*, 7(3 Suppl):S26–S41, 2010. 1, 19, 50, 61

[356] J. Wang, M. Ortiz-Maldonado, B. Entsch, V. Massey, D. Ballou, and D. L. Gatti. Protein and ligand dynamics in 4-hydroxybenzoate hydroxylase. *Proc Natl Acad Sci USA*, 99(2):608–613, 2002. 183

[357] J. C. Wang. Cellular roles of DNA topoisomerases: a molecular perspective. *Nat Rev Mol Cell Biol*, 3(6):430–440, 2002. 163

[358] R. C. Weast, M. J. Astle, and W. H. Beyer. *CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data.* CRC Press, Boca Raton, FL, 1984. 28

[359] P. A. Williams, J. Cosme, D. M. Vinkovic, A. Ward, H. C. Angove, P. J. Day, C. Vonrhein, I. J. Tickle, and H. Jhoti. Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science*, 305(5684):683–686, 2004. 176

[360] A. P. Witkin. Scale-Space Filtering. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, volume 2, pages 1019–1022, San Francisco, CA, 1983. Morgan Kaufmann Publishers Inc. 20

[361] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, 87(12):4576–4579, 1990. 150

[362] M. Wolf, R. L. Garcea, N. Grigorieff, and S. C. Harrison. Subunit interactions in bovine papillomavirus. *Proc Natl Acad Sci USA*, 107(14):6298–6303, 2010. 50, 51, 148, 156

[363] H. J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Comput Sci Eng*, 4(4):10–21, 1997. 17

[364] World Wide Protein Data Bank. *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*, 3.20$^{\text{th}}$ edition, Accessed October 2011. `http://www.wwpdb.org/docs.html`. 35, 226

[365] W. Wriggers. Using Situs for the integration of multi-resolution structures. *Biophys Rev*, 2(1):21–27, 2010. 37, 52, 60, 149, 160, 226

[366] W. Wriggers. Resolution Estimation. `http://www.biomachina.org/courses/structures/082.pdf`, Accessed October 2011. 65

[367] W. Wriggers and S. Birmanns. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J Struct Biol*, 133(2–3):193–202, 2001. 51, 52, 60, 65

[368] W. Wriggers and P. Chacón. Modeling tricks and fitting techniques for multiresolution structures. *Structure*, 9(9):779–788, 2001. 52, 60, 65, 93

[369] W. Wriggers, R. A. Milligan, and J. A. McCammon. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J Struct Biol*, 125(2–3):185–195, 1999. 52, 53, 60

[370] W. Wriggers, R. A. Milligan, K. Schulten, and J. A. McCammon. Self-organizing neural networks bridge the biomolecular resolution gap. *J Mol Biol*, 284(5):1247–1254, 1998. 52, 53, 60

[371] S. Wu, J. Liu, M. C. Reedy, R. T. Tregear, H. Winkler, C. Franzini-Armstrong, H. Sasaki, C. Lucaveche, Y. E. Goldman, M. K. Reedy, and K. A. Taylor. Electron tomography of cryofixed, isometrically contracting insect flight muscle reveals novel actin-myosin interactions. *PLoS One*, 5(9):161–193, 2010. 51, 61

[372] X. Xing and C. E. Bell. Crystal structures of Escherichia coli RecA in complex with MgADP and MnAMP-PNP. *Biochemistry*, 43(51):16142–16152, 2004. 141

[373] F. Yang, L. G. Moss, and G. N. Phillips. The molecular structure of green fluorescent protein. *Nat Biotechnol*, 14(10):1246–1251, 1996. 176

[374] J. K. Yano, M. R. Wester, G. A. Schoch, K. J. Griffin, C. D. Stout, and E. F. Johnson. The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-Å resolution. *J Biol Chem*, 279(37):38091–38094, 2004. 184

[375] S. Yin and N. V. Dokholyan. Fingerprint-based structure retrieval using electron density. *Proteins: Struct , Funct , Bioinf*, 79(3):1002–1009, 2011. 58, 59, 60

[376] T. Yokoyama, S. Neya, A. Tsuneshige, T. Yonetani, S.-Y. Park, and J. R. H. Tame. R-state haemoglobin with low oxygen affinity: crystal structures of deoxy human and carbonmonoxy horse haemoglobin bound

to the effector molecule L35. *J Mol Biol*, 356(3):790–801, 2006. 38, 167, 176

[377] J. Zhang, M. L. Baker, G. F. Schroder, N. R. Douglas, S. Reissmann, J. Jakana, M. Dougherty, C. J. Fu, M. Levitt, S. J. Ludtke, J. Frydman, and W. Chiu. Mechanism of folding chamber closure in a group II chaperonin. *Nature*, 463(7279):379–383, 2010. 40, 50, 51, 148, 151

[378] S. Zhang, D. Vasishtan, M. Xu, M. Topf, and F. Alber. A fast mathematical programming procedure for simultaneous fitting of assembly components into cryoEM density maps. *Bioinformatics*, 26(12):i261–i268, 2010. 51

[379] X. Zhang, E. Settembre, C. Xu, P. R. Dormitzer, R. Bellamy, S. C. Harrison, and N. Grigorieff. Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc Natl Acad Sci USA*, 105(6):1867–1872, 2008. 50, 148, 155, 176

[380] W. Zheng. Accurate Flexible Fitting of High-Resolution Protein Structures into Cryo-Electron Microscopy Maps Using Coarse-Grained Pseudo-Energy Minimization. *Biophys J*, 100(2):478–488, 2011. 51

[381] Z. H. Zhou. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr Opin Struct Biol*, 18(2):218–228, 2008. 38

[382] B. Zitova and J. Flusser. Image registration methods: a survey. *Imag Vision Comput*, 21(11):977–1000, 2003. 11

[383] 21-December-2010: From 7 to 70,000: The PDB Reaches a New Milestone. `http://www.wwpdb.org/news/news.html`, Accessed October 2011. 2, 35, 37

[384] RCSB PDB, Policies & References. `http://www.pdb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html`, Accessed October 2011. 37

[385] EMDB reaches 1000 entries! `http://www.emdatabank.org/thousand_emdb_entries.html`, Accessed October 2011. 49, 149

[386] `http://www.boost.org`, Accessed October 2011. 227, 228

Index