After we have discussed the formation of protein ligand complexes in detail from an experimental and thermodynamic perspective, we will now concentrate on **how to predict protein-like and complexes computationally**. This process is commonly referred to as **docking.**

Docking is applied for two major tasks: the one is to predict the structures of protein-ligand or protein-protein complexes in detail. We will be mostly concerned with the prediction of complexes formed between small molecules and target proteins of interest. This is a major part of **structure-based drug design**.
In addition, docking calculations are applied to computationally screen compound databases on template structures to identify new active compounds (novel hit). And this computational screening process is commonly termed **structure-based virtual screening -** computational screening of compound databases on protein structures of interest.

The docking process has two major components:

1. **Posing** refers to the prediction of the binding conformation and orientation of a ligand in a protein-binding site - the so-called **binding pose.**
   When we discussed x-ray structures of complexes, binding pose was also termed the *binding mode*, if you recall. And regardless which terminology we use, binding pose or binding mode really describe the bioactive conformation of a ligand. So, in our computational docking analysis we try to predict the bioactive conformation of true ligands.
2. This leads us to the second component of the docking process which is called **scoring.** In scoring we apply energy or other fitness functions to rank small molecules with predicted binding modes according to the likelihood to be true ligands and distinguish them from false positives

Posing - conformational prediction of ligand binding modes; **Scoring** - characterizing these conformations and comparing them in quantitative terms to identify those compounds which truly have highest probability to be novel actors.

As we see later on, this also leads us to a general conundrum of docking calculations because docking will predict ligand binding modes for the majority of database compounds if they're not simply too large to fit into a given binding site. And therefore derive putative poses for many many false positives that need to be distinguished from true binders via scoring assessment of the binding modes.

Essential information for docking analysis:

1. Essential information for docking analysis first and foremost includes a **reliable 3d structure of our target protein of interest**. Ideally we would like to have a high resolution very refined experimental structure available.
   Only *if this is not the case* would one attempt to use **homology models** that are by definition less accurate than an experimentally determined structure
2. And secondly we must have **detailed knowledge of the ligand binding site** (the active site in an enzyme / the ligand binding site in a receptor) in order to identify

ligands that best fit into the site and compete with the natural substrates or ligands of our proteins of interest.

If we use computational models homology models as templates, then in the course of the docking process we carry out something like a double hypothetical analysis, because we have first a hypothetical binding site even in a homology model and then we try to develop binding hypotheses for small ligands that might be true positives in the minority of cases or false positives in the majority of cases.

Regardless of the methods that we use, the general goal of docking calculations (which is very well in accord with what we have learned about the characteristics of experimental ligand target complexes) is to achieve for a given small molecule that we dock the highest possible degree of shape complementarity and chemical complementarity.
So, in our computational docking calculations we also try to achieve high shape complementarity and high chemical interaction complementarity, because we know that this is a hallmark of experimental complexes. Here the situation is even a little bit more complicated, because achieving these high degrees of shape and interaction complementarity is an **essential but not sufficient condition for the identification of new active compounds**. It is essential because a true ligand as we know engaging in a viable binding process must have these characteristics when a complex is formed with a given binding site, however what one also observes in docking analysis is that false positives often achieve a high degree of shape complementarity and comical a chemical complementarity that is readily comparable to true positives, although they don't bind for **other reasons** because of other negative contributions of components that are involved in the binding process. **So we know that for a true ligand observing these characteristics is essential, unfortunately there will also be varying numbers of false positives that almost yield an indistinguishably good fit into a given binding site that needs to be taken into consideration, especially during the evaluation and scoring process.**

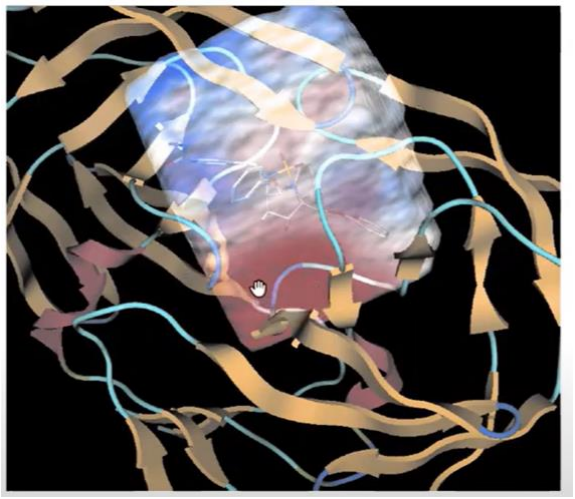
Docking methods

In addition, docking methods heavily rely on the molecular representation of the binding site that is used.

- **Full atomic representations** are rare because they are computationally inefficient and they are almost out of use at this point.
- What is much more common is to use **molecular surface representations of the binding site** or
- **Grid representations**.
  In grid representations the binding site is superimposed on evenly spaced grid and grid points store electrostatic or van der Waals interaction energies or

potentials that are obtained by calculating these interaction energies with force fields.

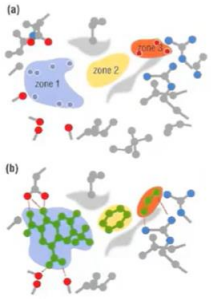
This use of grid representations and grid points that store interaction energies has its foundation in the pioneering grid program developed by peter goodford. And in order to appreciate these fundamentals for grid-based representation we will discuss the grid approach in some detail in the next few slides.
This is how a typical surface or grid representation will look like:



So we have here a surface plot of a predefined protein binding site, we have an underlying grid representation here. And as you can understand from the cube that is defined here, and on the grid representation surface in this particular case is superimposed. And what has been calculated here for the active site of hiv protease using charged probe atoms is electrostatic interaction energies and the ensuing electrostatic potential characterizing the binding side. So these types of representations will essentially be encountered whenever we define a grid representation of a given binding site. And then depending on what we probe (what type of energetic calculations we perform) we can assess electrostatic potentials (the distribution of electrostatic interaction energies or van der waals interaction energies) with respect to grid points that cover a pre-defined binding site.

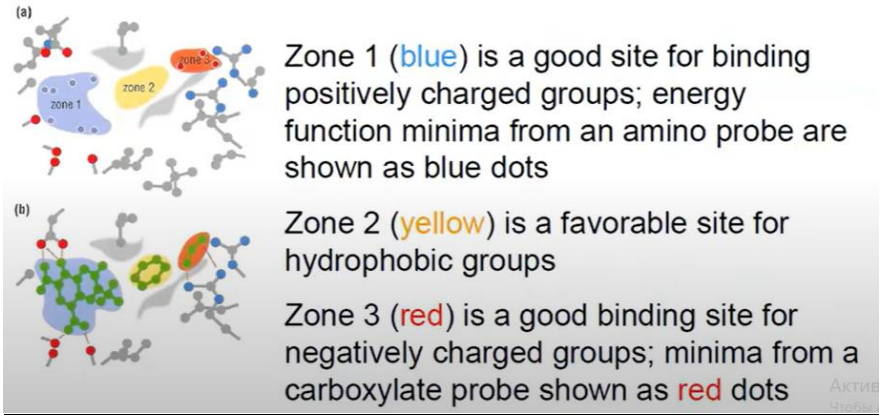
GRID Approach

So let's have a look at the fundamental grid approach that has really paved the way for grid representations of protein binding sites, that are nowadays one of the preferred representations for docking calculations.



Protein is placed on a grid, different types of probe atoms are placed on grid points, and interaction energies are calculated using a simple force field

In this particular case **the binding site is superimposed on the grid and the grid representation is then used to place probe atoms of a given radius on the grid points.** And these probe atoms (depending on what type of interactions I would like to probe with them, to assess with them) are either in the "good molecular mechanics spirit balls" that represent van der Waals interaction points or they are "charged balls", charged atoms. And such a probe is placed on each and every grid point within the binding site, and then the interaction energy is calculated with a typical force field (the type of force field we have learned about). Then the resulting (either van der Waals or electrostatic) energies are assigned to the corresponding grid parts.
Based on the stored interaction energies we can then identify regions in the binding site that are favorable for particular protein ligand interactions.



Red - oxigen; blue - nitrogen; gray - carbon.

So for example when we look at these different zones that have been delineated we can identify zones that are *favorable for engaging in charge-charge interactions* surrounded here by arginine residues. We can identify zones that are *favorable for hydrophobic interactions with for example hydrophobic groups or aromatic binding sites,* and we can identify other regions where *not only hydrophobic interactions would be favored but also specific hydrogen bonds* on the basis of coulombic interactions.
And so these different zones can be approximated and then from the combinations of grid points and then they identify in the grid programs. In the grid program regions where corresponding molecular fragments can be placed that are hypothesized to form viable interactions with this region of the binding cell.
So here we can look at these zones in more detail:

- **Red zone**: we have here regions where charge-charge interactions with complementary charges would be preferred

- **Yellow**: we have here regions where hydrophobic interactions would be entirely favorable
- **Blue**: and then we have regions where other particular interactions between opposing charges would be favored.

This could be then assessed with different molecular fragments that would engage in the formation of corresponding charge-charge interactions, involving hydrogen bonds, salt bridges and so on and so forth.
So complementarity of putative interactions is assessed on the basis of grid point energies that are calculated with probe atoms. So in the hydrofolate reductase complementary interactions here would be larger groups as indicated here (blue zone) that would be essentially positively polarized or charged and capable of interacting with these negatively polarized or charged residues by contrast here (red zone) we have a region where negative charges would be favored to interact with these arginine residues and then (yellow zone) we have an intervening hydrophobic region where in hydrophobic groups on aromatic ring would be placed.
Essentially the same type of consideration is then applied to a docking exercise except that we now don't use probe atoms to calculate the grid-based interaction energies, but actually ligand atoms. So grid representation is then used to store at each in every grid point force field interactions energies that are calculated with between ligand atoms matching to the grid points and surrounding protein atoms in the binding side. So **this is how the grid principle has been transferred and adapted for docking analysis.**

This shows a former grid representation that is created around a protein. The important aspect here is, if we look at such a large grid that is generated, we only need to focus on the grid region that narrowly includes the binding site. And this of course reduces dramatically the amount of force field based calculations that we have to carry out to sample these grid energies.

Docking Methods 1

We are essentially distinguishing two different docking approaches. The original implementation of docking methods (at times when relatively little computational power was available) is called **rigid body docking**.
Here the protein binding side is kept rigid - no conformational changes are permitted, and small molecules are docked as rigid pre-computed confirmers. And this type of rigid body docking essentially corresponds and follows to a “lock and key” type model.

Rigid Body Docking

So we are looking for a precise fit of a precomputed ligand conformation into a binding site with high shape and interaction complementarity, and this “lock and key” type interaction then strictly depends on the ability to really pre-compute bioactive conformations for true ligands. And when you think about that, the rigid body docking

approach is extremely vulnerable to false positives but also vulnerable to false negatives. So if false positive ligands (that really do not bind) are obtained in a confirmation that fits the binding site very well, they may be considered as true ligands. By contrast, if for a truly active compound pre-computing of conformations for rigid body docking has failed to approximate the bioactive conformation well, then this will inevitably result in a false negative. So rigid body docking is prone to a lot of potential error sources and yet for reasons we will discuss later on it often has succeeded and still succeeds in identifying new active compounds.
Importantly in rigid body docking we have only to consider six degrees of freedom when we carry out the conformational search. There are **three translational** degrees of freedom moving a ligand in pre-computed rigid conformation into x y or z direction and then there are **three rotational** degrees of freedom. This makes the ligand docking process computationally feasible, even if not a lot of computational power is available, and therefore the first generation of docking methods really focus on rigid body docking.

Lock-and-Key Model

And again, the “lock and key” model rules in this particular case: we look for small molecule conformers that must fit exactly into the binding side of our target protein, and are then considered as potentially active molecules with a caveat, as discussed, that we made during pre-computation of protein conformers generate false positives but also false negatives.

Docking Methods 2

Then there are the second generation docking approaches that take flexibility to varying extent into account. **Flexible ligand docking** means that we **still keep the protein template structure as a fixed entity** (so it is still considered a rigid “lock”), but, as the name applies, now **our ligands** not only have the six degrees of rotational and translational freedom associated with them but they are **also permitted to undergo internal bond rotations.** This of course renders the calculations considerably more time consuming, but under the conditions of flexible ligand docking we now circumvent the critical problem that is really error-prone and that is pre-calculating conformers as putative ligand binding modes.
So flexible ligand docking essentially takes this key variable out of the docking process, and optimizes the confirmation of a ligand and its orientation during the docking process until convergence has been reached for a particular pose.

As a further extension of the flexibility concept there's also **flexible protein and ligand docking**, this is essentially the same as above, however, in addition, predefined side chain atoms within the binding side (that are most likely to be contact residues of ligands) are permitted to move during the calculation.
Now of course we have a much larger conformational space associated with the calculations they become a lot more time consuming and experience shows (the

success rate of these different levels of docking calculations) shows that inclusion of protein binding site residues like a contact residues in flexible docking does not necessarily further improve the success rate and liability of flexible ligand docking. Although induced fit occurs, as we know, and conformational adjustments in protein binding sites happen, but they are very often happen at very small magnitude and in the context of docking calculations, the accuracy of these calculations in the presence of small conformational changes is still sufficient, if we consider a rigid template to frequently identify true ligands. But if we permit flexibility of protein residues, then local conformational changes that are observed during the docking process in protein binding sites may really be prone to producing shape errors in the binding site. They may depart from experiment and so we gain some potential accuracy by considering induced fit events on the one hand but we also again include extend possible error sources by varying the shape of protein binding sites that do not necessarily correspond to experimental reality, and confirmation ensembles that these binding sites really can adopt. So it's a trade-off. And for these reasons **the most robust and promising approach (**also computational feasible approach) for high throughput tracking of large compound databases and currently state-of-the-art continues to be **flexible ligand docking**. Taking into special consideration that induced fit conformational changes in protein binding sites are often, but certainly not always, of only small magnitude that can within the accuracy limits of docking calculations frequently be well approximated using a static template. (*Sometimes, the shape changes in protein binding sites, called induced fit, are small enough that we can just use a fixed template in docking calculations, even though it doesn't always work perfectly.*)

Flexible Ligand Docking

So flexible ligand talking is really the key to success in many cases. And in addition to our **three translation and three rotational degrees of freedom**, that we already need to address in rigid body docking, we have now **n conformational degrees of freedom along torsion angle,** so essentially any rotatable bond in our test compound adds another degree of freedom.
And accordingly for typical small molecule ligands, the conformational search process during posing can add considerable requirements in terms of computational time and it's far from being a trivial process to accommodate a flexible ligand in a rigid binding site. However, today's computational performance levels flexible ligand docking is even feasible for a very large compound databases.

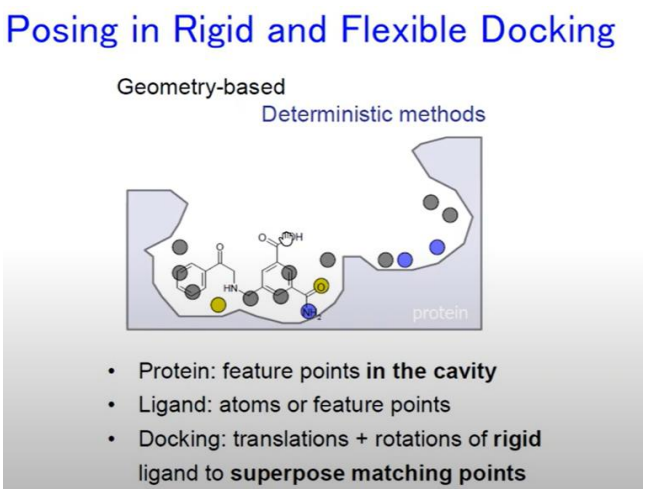
From rigid docking to flexible ligand docking and further docking events that include advertising extend protein flexibility, the computational costs dramatically increase and so both in terms of accuracy success rate and computational requirement this is currently by far the best compromise - flexible like docking.

Posing in Flexible Ligand Docking

In order to carry out the flexible ligand docking process, for posing - so for finding the most likely bound confirmation (pose - binding mode) of a given small molecule a variety of different conformational search approaches have been introduced. Any of these has distinct features during the process of conformational sampling and assessment of intermediate conformance. There's a great variety of different docking approaches, different docking algorithms that address the posing problem in algorithmically different ways.

We will limit our considerations here to few representative approaches that have been and continue to be widely used to the introduction of many different algorithmic variants.

Posing in Rigid and Flexible Docking



These approaches, that we will discuss, these different classes of approaches are essentially with modification adaptable to solving the posing problem in both rigid and flexible docking.

One school of these approaches is really **geometry based**.

So focusing on geometric features of ligand binding sites and conformational complementarity between ligand and binding site. These calculations are essentially deterministic in nature. Typically, what's done with a protein binding site is to identify feature points that represent specific features of the binding site based on the types of amino acids present in those areas. And once we have feature points, then we can also define corresponding feature points in ligand atoms, based on the chemical nature of the group.
So for example as a feature point we can have hydrophobic atoms in a hydrophobic moiety of the ligand or aromatic ring atoms, we can have hydrogen bonds, donors, polar groups and so on and so forth. So the way these feature points can be defined is variable.

Once we've identified feature points for a binding site, which is usually done automatically by categorizing residues and their chemical and lighting characteristics,

we get an abstract representation of the binding site for the ligand. And this is particularly suitable in the first instance for rigid body docking, because now we have only to carry out translations and rotations of the ligand until there is the best superposition the best match of feature points obtained. And depending on the quality of the match, the pose is accepted finally or intermediate poses are rejected.
So **feature point definition and feature point matching is a typical geometry based approach to ligand posing** and originally introduced four rigid body docking calculations.

Then another school of methods is **energy based (**force field based).

Essentially force field based even during the conformational sampling and posing process. These methods rely much more on stochastic approaches and on sampling properties. For a protein binding site, atoms that map to the surface frequently are translated into feature points. For ligands, either atoms are used for matching or as abstraction feature points. During the computational sampling process, the conformation of the ligands is then modified.

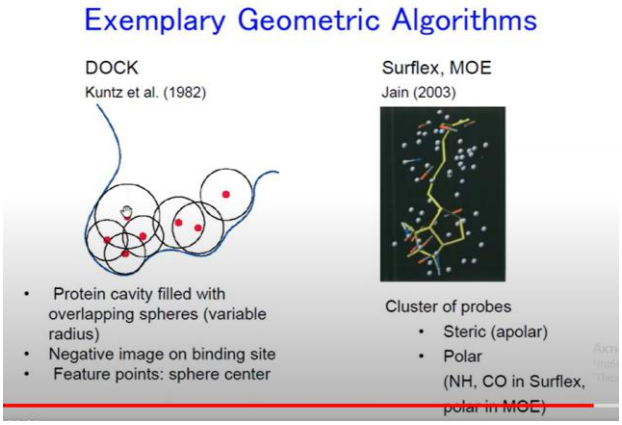
And during this modification process the protein and ligand atoms of feature points (that were defined at the beginning) are then used to calculate interaction energies using a force field and optimize these energies until a minimum has been converged at. So this energy based confirmation sampling for posing has originally been introduced as a more stochastic approach for flexible ligand docking, because it enables us to really optimize an energy function instead of searching for the best possible match of predefined feature points.

Of course both of these approaches energy based approaches and geometrical approaches are adaptable for both rigid body docking and flexible ligand docking, but they rely on different principles either a matching of feature points as precisely as possible without optimizing any fitness function or conformational assessment searching relies on the optimization of an energy function typically a molecular force field. So these are two major schools of approaches for posing.

Exemplary Geometric Algorithms

These different types of approaches are implemented in pioneering docking programs. This is the docking program with a telling name DOCK, that has really been a pioneering effort first developed in the early 1980s and the DOCK program really has spawned a lot of further developments. And to this date DOCK and its later further improved generations is one of the premier tools that is used for docking.
And so both in this approach but then also in approaches introduced about 20 years later such as the Surflex program that in a similar way has then been implemented in the molecular operating environment (MOE) that we use here (for example for our practical molecular modeling course) geometric algorithms in different variants have been implemented.

And so the pioneering approach in docking has been the DOCK approach and it really is unique, and the specific geometry based approach of the DOCK program should be remembered, because it has really been a pioneering effort in the field. So what is done here is for a given predefined binding site (active site) the cavity or pocket is filled with overlapping spheres variable radius until the best possible fill of the cavity is obtained.



Every sphere has a geometric sphere center. And once the best possible “fill” of a binding site with spheres of different radii has been obtained, the binding site is then represented by the sphere centers. So this can be rationalized as creating a negative image of the binding site: we use the spheres to map out the binding site, fill it completely, and then use the sphere centers as feature points.

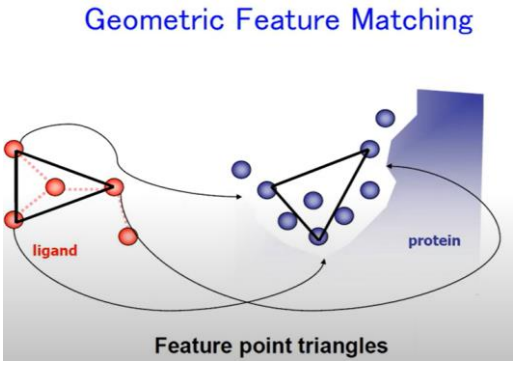
Now we have a feature point representation of the binding site that originates from the sphere centers. With the binding site represented by feature points from sphere centers, docking a ligand involves matching the ligand atoms or ligand spheres to these binding site spheres. Whether using rigid body docking or flexible docking, the goal is to find the best fit between the ligand's feature points and the binding site's feature points. This best fit then determines the pose of the ligand in the binding site.

(Shows MOE picture) Similar yet distinct implementations are as mentioned in these programs, for example here one employs clusters of probes that surround as feature points are important residues in a binding site. Also they're probes defined to represent the residue distribution at the binding site and these probes can be either steric in nature, so accounting for van der Waals /hydrophobic interactions are polar in nature and then a match of a ligand corresponding ligand spheres to these probe clusters is determined analogously to what has been done in the DOCK approach.

And the way these probes are defined and what they account for varies between different implementations, but the principle of using either **sphere centers of a negative image of a binding site** or **clusters of probes representing the binding site**, this really then is constantly applied in approaches that are based on these efforts.

Geometric Feature Matching

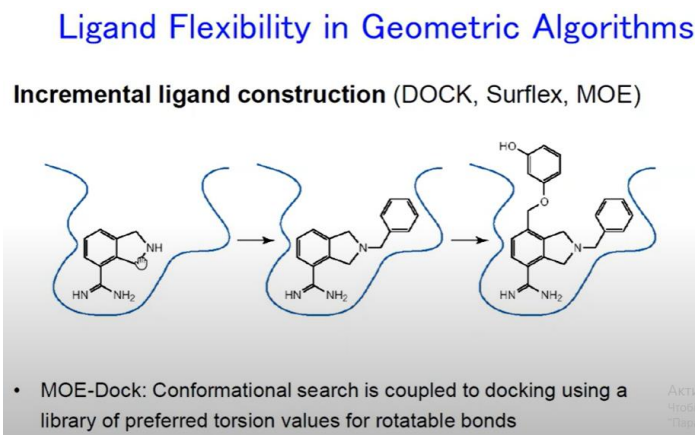
Geometric feature matching can also be facilitated algorithmically in different ways. There is of course **clique matching** that can be employed to match different ligand and protein features, but also **feature point triangles** instead of completely connected subgraphs of larger groups of features have been used for geometric feature matching.



And of course the guiding principle behind the use of triangles or cliques is to find matches between groups of complementary protein and ligand spheres and thereby really reduce the search space for these feature matching calculations. So again feature matching although it's common to these algorithms that originate from the dark approach or other geometric matching algorithms can be facilitated in different ways algorithmically distinct, but follows essentially the same principles to match groups of ligand and protein spheres to arrive ultimately at the best complementarity fit which then determines the pose.

Ligand Flexibility in Geometric Algorithms

So these geometric approaches can really be adjusted in different ways for rigid and flexible ligand docking so of course one can carry out a conformational search in a given binding site. And during the docking process, once a ligand is initially accommodated in the site, and then either explore torsion angles at different grid intervals systematically or for computational efficiency such as in the molecular operating environment MOE-Dock module that we're using in the practical then employ a library of precomputed torsion angle values for rotator by bonds to scan through and then find the best possible match.



So confirmation research of the likeness is possible. What is also feasible is incremental ligand construction. So following this approach a ligand is separated into predefined substructure types, and then the substructure types are used to incrementally build up the ligand in the binding site, starting with the largest fragment and then keeping the connectivity information, adding iteratively smaller fragments to the ligand. So this is essentially also a growing process that is used in computational programs to design ligands within a given protein binding site.

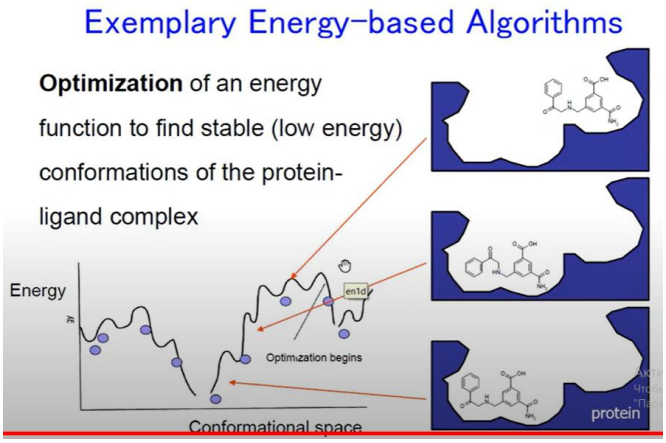
And so this incremental likened construction has the computational advantage that it reduces dramatically this the conformational search space for a given ligand because the confirmation assessment for the substructures is done independently:

- first the best fit for the largest fragment is determined then it's kept constant
- then the next one is added to it.

So this incremental addition of fragments really reduces the confirmation search space that is available for a given ligand on the basis of all its degrees of freedom. This is another variant of how to facilitate conformational searching of a ligand within a binding site.

Exemplary Energy-based Algorithms

And on the other hand the energy-based stochastic algorithms, that largely rely on conformational sampling within the binding site, try to optimize an energy function along search space. And then again a great variety of different algorithmic implementations and solutions to this energy based conformational search, and again can be carried out in different ways.



But to some extent they all follow the same principle and this is to probe with a given ligand the different regions in the binding site systematically explore different poses in the binding site and then sample the resulting interaction energies for these poses. And the process is then terminated once a sufficiently large number of putative poses has been explored and the lowest one is then taken to define the pose of a ligand.

There are also variants to this process. So for example this energy approach can also be adapted to assess ligand conformation under the conditions of incremental growth, but

here the guiding principle is really to explore as many different poses as possible and evaluate interaction energy as a quality criterion for the best possible pose.

Conformer Populations

Now last but not least a point of consideration is how different conformer populations in the process of docking are generated.
There are different ways to carry out conformational search either explicitly or through ligand growth. But there are also systematic ways to pre-generate conformer populations of ligands and assess these conformers from populations individually, and so this differs in principle from the conformational search of a ligand within a binding site. So one can also in the process of flexible ligand docking generate conformer populations and assess them and for the best possible fit.

And this also can be done in different ways algorithmically. So typically one randomly generates a starting population for random confirmation search of our ligand and then modifies the confirmations of these originally derived instances during the second stage, and then samples energies for these conformers and arrives at the final population which then can again be a starting point for a randomization step, another modification step. Again energy-based selection here it is possible to either really pre-compute ligand conformations by only looking at the internal energies of each ligand or one can even assess interaction energies at this stage of conformers within protein binding sites. When you look at this particular approach then you can sense already that algorithmically for example this could be well accomplished with an evolutionary genetic algorithm where one would encode starting population and a sort of conformational chromosomes which could be modified through torsion angle operations and then subjected again to energy assessment, generating a new generation of starting conformations, that could again be modified using the chromosomal approach of evolutionary algorithms. And then assess until one has some sort of convergence criterion reached.

So in summary in both rigid body docking in particular flexible ligand docking the posing problem can be tackled in very different ways, and if you remember the two major schools of geometry based algorithms and energy based algorithms, the former rely on feature point definition essentially in feature point matching during different algorithmic approaches and the latter on stochastic confirmation sampling (what can also be done in very different ways) and assessment of internal and interaction energies, then you have comprehended two major approaches to the posing problem. And then the next step of course will be to bridge from the posing to the scoring problem and evaluate these two components of the docking process on a relative scale.