# RATIONAL DESIGN OF ANTIBODIES: FROM MECHANISMS OF SPECIFICITY TO NOVEL VACCINE STRATEGIES

By

Jordan Willis

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMICAL AND PHYSICAL BIOLOGY

August, 2014

Nashville, Tennessee

| Approved: | Date: |
|---|---|
| Benjamin Spiller, Ph.D (Chair) | |
| Christopher Aiken, Ph.D | |
| Jens Meiler, Ph.D | |
| Spyros Kalams, M.D. | |
| James Crowe, M.D. | |

## TABLE OF CONTENTS


Page

**LIST OF TABLES** . . . . . **iii**

**LIST OF FIGURES** . . . . . **iv**

**I Introduction** . . . . . **1**

I.1 The sampling and scoring methods used by the Rosetta software suite . . . 1
I.1.1 Sampling strategies for backbone degrees of freedom . . . . . 1
I.1.2 Sampling strategies for side-chain degrees of freedom . . . . . 2
I.1.3 Rosetta energy function . . . . . 2
I.1.3.1 Knowledge based centroid energy function . . . . . 2
I.1.3.2 Knowledge based all atom energy function . . . . . 3
I.2 Protein design using Rosetta . . . . . 3
I.2.1 De novo protein design . . . . . 4
I.2.2 Redesign of existing proteins . . . . . 4
I.3 Existing challenges with protein surface design . . . . . 5
I.3.1 Electrostatic energy is insufficient to predict the impact of protein surface mutations . . . . . 5
I.3.2 Computationally designed proteins frequently aggregate unless "supercharged" . . . . . 6
I.4 The history of ligand docking . . . . . 6
I.4.1 Early attempts at hand-docking ligands using physical models . . . 6
I.4.2 An overview of influential protein-ligand docking methods . . . . 7
I.4.2.1 DOCK . . . . . 7
I.4.2.2 GRID . . . . . 8
I.4.2.3 The importance of protein and ligand flexibility . . . . . 8
I.4.2.4 FlexX and GOLD . . . . . 9
I.4.2.5 Glide . . . . . 10
I.5 The history of RosettaLigand . . . . . 10
I.5.1 RosettaLigand is capable of successfully predicting binding based on comparative models . . . . . 11
I.5.2 Applications of RosettaLigand to drug discovery . . . . . 12
I.6 Computational ligand docking has inconsistent predictive power . . . . . 12
I.7 Artificial Neural Network techniques have proven valuable for extracting complex signals . . . . . 13
I.8 Over-training and over-fitting are common pitfalls in the use of **ANN!**s for pattern recognition . . . . . 13
I.8.1 Deep networks and node dropout as novel methods for improving network generalizability . . . . . 14

I.9 Using **ANN!**s to make predictions regarding drug activity is a major area of current research . . . . . 15

## LIST OF TABLES

Table Page

## LIST OF FIGURES

Figure Page

# CHAPTER I

# Introduction

## I.1 The sampling and scoring methods used by the Rosetta software suite

Rosetta is a software package for protein structure prediction and functional design. It has been applied to predict protein structures with and without the aid of sparse experimental data, perform protein-protein and protein-small molecule docking, design novel proteins, and redesign existing proteins for altered function. Rosetta allows for rapid tests of hypotheses in biomedical research which would be impossible or exorbitantly expensive to perform via traditional experimental methods. As a result, Rosetta methods have gained increasing importance in the interpretation of biological findings from genome projects, the engineering of therapeutics, probe molecules, and model systems in biomedical research.

While the Rosetta suite is capable of performing a wide range of modeling tasks, it uses a core set of sampling and scoring strategies to accomplish most of these. The majority of conformational sampling protocols in Rosetta use the Metropolis Monte Carlo algorithm to guide sampling. Gradient based minimization is often employed for last step refinement of initial models. Since each Rosetta protocol allows degrees of freedom specific for the task, Rosetta can perform a diverse set of protein modeling tasks (**?**).

### I.1.1 Sampling strategies for backbone degrees of freedom

Rosetta separates large backbone conformational sampling from local backbone refinement. Large backbone conformational changes are modeled by exchanging the backbone conformations of 9 or 3 amino acid peptide fragments. Peptide conformations are collected from the **PDB!** (**PDB!**) for homologous stretches of sequence (**?**) which capture the structural bias of the local sequence (**?**). For local refinement of protein models, Rosetta utilizes Metropolis Monte Carlo sampling of phi and psi angles calculated not to disturb the global fold of the protein. Rohl (**?**) provides a review of the fragment selection and backbone

refinement algorithms in Rosetta.

### I.1.2 Sampling strategies for side-chain degrees of freedom

Systematic sampling of side-chain degrees of freedom of even short peptides quickly becomes intractable (**?**). Rosetta drastically reduces the number of conformations sampled by usage of discrete conformations of side-chains observed in the PDB (**??**). These "rotamers" capture allowed combinations between side-chain torsion angles as well as the backbone phi and psi angles and thereby reduce the conformational space (**?**). A Metropolis Monte Carlo simulated annealing run is used to search for the combination of rotamers occupying the global minimum in the energy function (**??**). This general approach is extended to protein design by replacing a rotamer of amino acid *A* with a rotamer of amino acid *B* in the Monte Carlo step.

### I.1.3 Rosetta energy function

Simulations with Rosetta can be classified based on whether amino acid side-chains are represented by super atoms or centroids in the low-resolution mode or at atomic detail in the high-resolution mode. Both modes have optimized energy functions that have been reviewed previously by Rohl (**?**).

#### I.1.3.1 Knowledge based centroid energy function

The Rosetta low-resolution energy function treats the side-chains as centroids (**??**). This energy function models solvation, electrostatics, hydrogen bonding between beta strands, and steric clashes. Solvation effects are modeled as the probability of seeing a particular amino acid residue with a given number of alpha carbons within an amino acid dependent cutoff distance. Electrostatic interactions are modeled as the probability of observing a given distance between centroids of amino acids. Hydrogen bonding between beta strands is evaluated based on the relative geometric arrangement of strand fragments. Backbone atom and side-chain centroid overlap is penalized providing the repulsive component to

a van der Waals force. A radius of gyration term is used to model the effect of van der Waals attraction. All probability profiles have been derived using Bayesian statistics on crystal structures from the **PDB!**. The lower resolution of this centroid-based energy function smoothes the energy landscape at the expense of its accuracy. This smoother energy landscape allows structures which are close to the true global minima to maintain a low energy even with structural defects that a full atom energy function would penalize harshly.

#### I.1.3.2 Knowledge based all atom energy function

The all-atom high-resolution energy function used by Rosetta was originally developed for protein design (**??**). It combines the 6-12 Lennard Jones potential for van der Waals forces, a solvation approximation (**?**), an orientation dependent hydrogen bonding potential (**?**), a knowledge based electrostatics term, and a knowledge based conformation dependent amino acid internal free energy term (**?**). An important consideration when constructing this potential was that all energy terms are pairwise decomposable. The pairwise decomposition of each of the terms limits the total number of energy calculations to $\frac{1}{2}N(N-1)$ where $N$ is the number of atoms within the system. This limitation allows pre-computation and storage of many of these energy contributions in the computer memory which is necessary for rapid execution of the Metropolis Monte Carlo sampling strategies employed by Rosetta during protein design and atomic-detail protein structure prediction.

### I.2 Protein design using Rosetta

Protein design methods seek to determine an amino acid sequence that folds into a given protein structure or performs a given function. In this context the protein design problem of finding a sequence that folds into a given tertiary structure is also known as the "inverse protein folding problem". The RosettaDesign (**?**) algorithm is an iterative process that energetically optimizes both the structure and sequence of a protein. RosettaDesign alternates between rounds of fixed backbone sequence optimization and flexible backbone energy minimization (**?**). During the sequence optimization step, a Monte Carlo simulated

annealing search is used to sample the sequence space. Every amino acid is considered at every position in the sequence, and rotamer positions are constrained using the Dunbrack Library (**?**). After each round of Monte Carlo sequence optimization, the backbone is relaxed to accommodate the designed amino acids (**?**). The practical uses of RosettaDesign can be divided into five basic categories: Design of novel folds (**?**), redesign of existing proteins (**?**), protein interface design, enzyme design (**?**), and prediction of fibril forming regions in proteins (**?**).

### I.2.1 De novo protein design

The RosettaDesign method has been used for the *de novo* design of a fold that was not (yet) represented in the **PDB!**. A starting backbone model consisting of a five stranded beta-sheet and two packed alpha-helices was constructed with the Rosetta *de novo* protocol using distance constraints derived from a two-dimensional sketch (**?**). The sequence was iteratively designed with five simulation trials of 15 cycles each. The final sequence was expressed and the structure was determined using X-ray crystallography. The experimental structure has an **RMSD!** (**RMSD!**) to the computational design of less than 1.1 Å (**?**).

Similarly, a molecular switch which folded into a trimeric coiled coil in the absence of zinc, and a monomeric zinc finger in the presence of zinc was designed by extending RosettaDesign to simultaneously optimize a sequence in two different folds. The sequence of an existing zinc finger domain was aligned with a coiled-coil hemaglutinin domain. During the design protocol the sequence was optimized to fold into both tertiary structures (**?**).

### I.2.2 Redesign of existing proteins

When nine globular proteins were stripped of all side-chains and then redesigned using RosettaDesign the average sequence recovery was 35% for all residues (**?**). In four out of nine cases the stability of the proteins improved as measured by chemical denaturation. The structure of a redesigned human procarboxypeptidase (**PDB!** 1AYE) (**?**) was deter-

mined experimentally. RosettaDesign was then used to systematically identify mutations of procarboxypeptidase that would improve the stability of the proteins. All of the tested mutants were more stable than the wildtype protein with the top scoring mutant having a reduction of free energy of 5.2 kcal/mol (**?**).

RosettaDesign has also been used to modify the structure of existing proteins. In one study, the HisF TIM Barrel protein was selected as the basis for the design of a novel symmetric protein. The backbone structure of half the barrel was duplicated, and Rosetta was used to redesign the new structure to have both symmetric sequence and structure. The new protein, named FLR, was expressed and crystallized. The two resulting crystal structures had RMSDs of 0.49 Å and 0.87 Åto the computational prediction, demonstrating the ability of RosettaDesign to make accurate predictions of side-chain conformation and energies (**?**).

## I.3 Existing challenges with protein surface design

### I.3.1 Electrostatic energy is insufficient to predict the impact of protein surface mutations

While protein design has had frequent successes, there are outstanding challenges, particularly with respect to the design of the surfaces of soluble proteins. Solvent interactions are critical to accurately measuring the electrostatics and stability of protein surfaces (**?**). However, due to the computational complexity associated with explicit solvent modeling, implicit models are frequently used, and may not be sufficiently detailed to make accurate energetic predictions. Furthermore, the network of interactions between protein surface residues and the overall stability of the proteins are highly complex. Xiao (**?**) demonstrated that while electrostatic surface interactions are important for stability, the impact of a single mutation on experimentally determined stability frequently cannot be explained by the impact on computed electrostatic energy.

### I.3.2 Computationally designed proteins frequently aggregate unless "supercharged"

In addition to issues with stability, computationally designed proteins often have issues with aggregation. A study was performed in which RosettaDesign was used to fully redesign 10 proteins (**?**). They found that 4 of the 10 designed proteins formed insoluble aggregates at 1 mM concentration. Aggregation appears to be a general phenomenon affecting protein design, and it has been repeatedly demonstrated that "supercharging" proteins by introducing large numbers of charged surface residues (**???**) can reduce aggregation in designed proteins. However, "supercharged" proteins are infrequently seen in nature, suggesting that evolved mechanisms for retaining solubility and avoiding aggregation are more complex. Observation of the folding properties of supercharged proteins suggest that excessive charging can inhibit folding (**?**), which may have acted as an evolutionary barrier to natural supercharging.

## I.4 The history of ligand docking

### I.4.1 Early attempts at hand-docking ligands using physical models

Attempts to model and predict protein-drug interactions date began shortly after the publication of the X-ray structure of hemoglobin, with Beddell et. al. Publishing a proof of concept method for structure based drug discovery in 1976 (**?**). The method developed by Beddell relied on the manual placement of physical molecular models into a scale model of the hemoglobin electron density, which allowed the authors to identify novel compounds with millimolar activity. While the identified compounds were relatively poor by modern standards, and the method of manual placement into physical models did not provide a means of postulating mechanism of action, the authors recognized the value of the new technique, saying:

> It has been common practice to design new drugs by modifying the chemical structure of a known substance which has the desired biological properties, and this procedure has imposed severe restraints on the choice. However, it is not

necessary for the novel compounds to be related to the original substance when the structure of the receptor site is already known.

It is remarkable that this observation on the state of rational drug discovery continues to be relevant, nearly 40 years after it was originally made.

### I.4.2 An overview of influential protein-ligand docking methods

After the advent of relatively inexpensive general purpose computers in the early 1980s, the promise of accurate and rapid computational design of novel small molecules has driven a wide array of research into improved methods for predicting protein-ligand interfaces. A successful protein-ligand docking tool must solve two basic problems: sampling and scoring. To effectively solve the sampling problem, the software must be able to efficiently explore both the rigid space of the protein binding site, as well as the conformational space of both the protein and the ligand. To effectively solve the scoring problem, a score function must be developed which can rapidly distinguish between energy favorable and unfavorable conformations. Solving both of these problems has proven highly challenging, although great progress has been made.

#### I.4.2.1 DOCK

In 1982, Kuntz et al. published DOCK, one of the earliest computational tools for modeling protein-ligand interactions (**?**). DOCK used a relatively simple energy function which modeled repulsive forces as hard spheres, and a rough approximation of hydrogen bonding which favored binding positions in which hydrogen bond donor groups on the ligand were within 3-5 Å of acceptor nitrogens and oxygens on the protein backbone. In concept, the DOCK algorithm is similar to the manual placement method described by Beddell et al. above. The program uses the van der Waals radii of the protein and ligand atoms to create "space filled" representations of both the receptor pocket and the ligand. Pairs of protein and ligand spheres are then considered systematically, and the set of pairings which

minimizes sphere overlap is selected. This algorithm is driven almost entirely by shape complementarity, and effectively models the "lock and key" hypothesis of protein-ligand binding, in which a rigid protein is matched with a rigid ligand.

#### I.4.2.2 GRID

In 1984, Goodford et al published GRID, a computational method for predicting energetically favorable protein-ligand binding conformations (**?**). GRID differed from previous attempts structure based drug discovery in that it used chemical information rather than relying entirely on receptor fit. Specifically, it assessed the protein-ligand interaction using an empirical energy function consisting of a Lennard-Jones term, electrostatic term, and hydrogen bonding term. This energy function was precomputed as a 3-dimensional grid overlaid on the ligand binding site. Thus, the total score of the ligand could be rapidly assessed as the sum of the grid squares the atoms are located in. Pre-computation of the scoring grid enabled many ligand conformations and compositions to be rapidly assessed, and the addition of chemical information in addition to shape proved more effective than simply evaluating shape complementarity.

While the GRID method proved reasonably effective, several shortcomings which limited the effectiveness of the method. The physics based force field used was relatively rudimentary, and the limited set of chemical probes used to create the grid. Additionally, accurate docking into a full-atom grid based model requires a high degree of precision in the position of the protein atoms, which limits the effectiveness of such a model in cases where the accuracy of the protein structure is lower.

#### I.4.2.3 The importance of protein and ligand flexibility

In the years following the publication of DOCK and GRID, additional experimental study of protein structure began to indicate that the rigid body lock and key model was not adequate for the modeling of protein-ligand interactions. It had long been suspected (**?**) that enzymes and receptors may be flexible to accommodate the fit of small molecules (the so

called “induced fit” hypothesis), however in 1995, Nicklaus et al. (**?**) published work suggesting that small molecules also undergo substantial conformational shift on binding. This conclusion was arrived at by comparing the geometry of flexible small molecules observed bound to proteins with the geometry of the same small molecules when crystallized in the absence of a protein, or when computationally minimized using molecular mechanics. The results of this study indicated that while the conformations of rigid structures typically differed by < 0.1 Å **RMSD!** between the bound and unbound context, flexible ligands typically differed significantly, frequently by several angstroms. Furthermore, the difference in **RMSD!** between bound and unbound ligands was strongly correlated with the number of rotatable bonds in the ligand, with an $R^2$ correlation of 0.82. In response to this research, the development of newer protein-ligand docking methods began to focus on the flexibility of the system. While flexibility had previously been avoided due to the inherent increase in computational complexity associated with modeling it, these findings made it clear that flexibility was a critical component of protein-ligand interaction.

#### I.4.2.4 FlexX and GOLD

FlexX (**?**) and **GOLD!** (**GOLD!**) (**?**), are two of the early methods which attempted to model ligand flexibility. FlexX represents the ligand binding site using a set of interaction sites, which are defined as surfaces surrounding hydrogen bond donors and acceptors, metals and metal acceptors, aromatic rings, methyls and amides. An empirical scoring function is used to score ligand conformations based on the distance and angle between defined protein and ligand interaction sites. FlexX uses an incremental construction algorithm to model ligand flexibility. An initial central fragment of the ligand is placed in the binding site using an incremental construction algorithm, and the additional fragments necessary to build the entire ligand are then placed such that they can connect to the initial fragment and minimize the energy function score. **GOLD!**, on the other hand, relies on the user providing a reasonable initial position for the ligand inside the protein binding site.

From that initial position, a genetic algorithm (**?**) was used to optimize the rotation angles of both the ligand and the interacting protein side-chains. The genetic algorithm makes it possible to rapidly find a high quality local minimum without the exhaustive sampling of bond angles that had made the problem previously intractable. As a result of this new sampling technique, **GOLD!** was able to successfully recover the correct binding conformation in 71 out of 100 X-ray crystal structures in a benchmarking study.

#### I.4.2.5 Glide

In 2004, Glide (**?**) was published as a novel method for protein-ligand docking aimed at the screening of large libraries of small molecules. To improve the speed of the algorithm, Glide models the receptor site using a set of cartesian scoring grids, and keeps the receptor atoms fixed. This allows the ligand to be rapidly scored, making it possible for a large number of ligand positions to be evaluated. Glide performs a set of exhaustive searches along at cartesian grid overlaid on the receptor binding site. To reduce the amount of sampling required, the step size of the grid is reduced over the course of the search process, beginning with a 2.0 Å pitch grid. Additionally, a set of filters based on the empirical ChemScore (**?**) energy function are used to progressively filter the set of allowable binding orientations using increasingly detailed metrics. After an initial starting position is accepted, The conformational space of the ligand is exhaustively searched, and the final pose is energy minimized. The use of a grid representation for the energy function makes it possible to to screen large numbers of compounds very rapidly, making Glide a popular choice for virtual screening studies (**??**).

### I.5 The history of RosettaLigand

RosettaLigand was originally published in 2006 (**?**) as a protein-ligand docking tool based off of the previously published RosettaDock (**?**) protein-protein docking tool. The original RosettaLigand docking algorithm took advantage of the knowledge based energy function used by RosettaDock. The use of a knowledge based potential rather than a physics based

potential is advantageous as knowledge based potentials are capable of indirectly modeling effects that are difficult to model directly. Additionally, the ability of RosettaLigand to rapidly optimize protein side-chain geometry (**?**) made it possible to model protein-ligand interactions with full atomic detail. While RosettaLigand was frequently able to accurately predict the binding orientation ligands (**?**), it was unable to model backbone or ligand flexibility, which have long been suspected to be critical for protein-ligand binding (**??**). To rectify this situation, further extensions were made to RosettaLigand by Davis et al (**?**) which allowed RosettaLigand to fully consider the flexibility of all parts of both the protein and the ligand. A blind benchmarking study comparing the pose recovery performance of the 2009 version of RosettaLigand suggested that overall it performed similarly to other major ligand docking tools (**?**). A notable conclusion of this study is that while most of the tools studied have a similar performance overall, the performance in predicting docking pose for individual protein targets varies wildly. This inconstant performance between protein targets and protein docking tools is seem in other studies as well.

### I.5.1 RosettaLigand is capable of successfully predicting binding based on comparative models

One of the advantages of a knowledge based energy function is the ability to accurately model complex physical effects without a direct physical model. In principle, this, combined with the ability to model both backbone and side-chain flexibility would make RosettaLigand well suited to the docking of ligands into comparative models or other low resolution protein structures. To assess this, a benchmarking study was performed in which small molecules with known binding positions were docked into homology models generated in the **CASP!** (**CASP!**) experiment (**?**). The results of this benchmark demonstrated that in most of the tested cases, Rosetta was able to generate low energy binding positions within 2.0Å of the crystallographic binding site.

### I.5.2 Applications of RosettaLigand to drug discovery

In addition to benchmarking studies, Rosetta has been used to develop models of ligand binding in **GPCR!** (**GPCR!**)s. A comparative model of hSERT was created based on the dSERT crystal structure. S- and R-citalopram were docked into this comparative model using RosettaLigand, and the resulting predicted binding poses were used to design mutational studies to identify residues critical for S-citalopram binding. Rosetta was able to correctly predict that Y95 and E444 formed protein-ligand interactions critical to binding (**?**). Similarly, RosettaLigand was used to model the binding of Positive Allosteric Modulators in a comparative model of $mGlu_5$ (**?**). In this case, the predictions made by RosettaLigand were used to guide mutation and radioligand binding studies, the results of which were used to further refine models. These models made it possible to map out critical interactions between Positive Allosteric Modulators and the $mGlu_5$ binding site even in the absence of crystal structure information.

## I.6 Computational ligand docking has inconsistent predictive power

A common thread running through the ligand docking research described above is the difficulty of docking ligands into some proteins. For every protein-ligand method developed, some percentage of protein-ligand interfaces cannot be effectively predicted. While the predictions generated by protein-ligand docking has made some major scientific contributions to drug discovery and molecular modeling, the unreliability of the method has historically constrained its usefulness.

In 2006, a diverse set of 81 protein targets, each with a diverse set of known active and predicted inactive ligands was assembled as the DEKOIS 2.0 dataset (**?**). Glide, **GOLD!** and Autodock Vina were used to screen this dataset, and the pROC AUC enrichment for each target and each screening method was computed. The results of this benchmark showed a wide range in the predictive ability of the three screening methods. While all three docking methods had strong predictive power against some protein tar-

gets (COX2, KIF11), there were several cases in which no method had predictive power (HSP90, QPCT), and more cases in which some methods were able to make accurate predictions while others were not (COX1, ROCK-1). Furthermore, it was not possible for the authors to identify straightforward patterns to predict which protein targets could be successfully screened against and which could not. The phenomenon of structure based **vHTS!** (**vHTS!**) methods having inconsistent performance depending on the protein target has been replicated in other studies. For example, the Directory of Useful Decoys: Enhanced (DUD-E) benchmark set was screened using DOCK, and the resulting predictions exhibited similar inconsistencies to those seen in the DEKOIS 2.0 study (**?**).

### I.7 Artificial Neural Network techniques have proven valuable for extracting complex signals

Since the publication of the perceptron as a method of machine learning (**?**), **ANN!** (**ANN!**) techniques have become an area of great interest to the machine learning community. While there was much initial optimism regarding the use of **ANN!**s to learn complex tasks, early perceptron based models proved limited in their abilities (**?**), and the state of computational hardware at the time prevented **ANN!** based techniques from living up to the early optimism. In more recent years, the availability of large clusters of low cost computer hardware as lead to a renaissance in both the development and application of **ANN!** based machine learning techniques. **ANN!**s have been used for tasks such as face recognition (**?**), cancer cell identification (**?**), and drug activity classification (**?**). **ANN!**s are popular choices for these tasks due to their ability to extract the signal from complex patterns.

### I.8 Over-training and over-fitting are common pitfalls in the use of ANN!s for pattern recognition

While **ANN!** based approaches have been valuable to many fields, they are often difficult to use in practice. Due to the very large number of free parameters in a neural network, they are very prone to over-fitting. In over-fitting, the neural network effectively "memorizes"

the dataset, and becomes a model that exactly implements the set of data used for training (**?**). The consequence of overfitting is that the model will be capable of exactly reproducing the training data set, but will have no ability to make predictions beyond that. The standard method for addressing this is to use as small a network as possible, and to perform a "cross-validation", in which part of the training dataset is withheld from training and used to keep track of the network performance as training proceeds. cross-validation makes it possible to determine when over-fitting is occurring and halt training, resulting in a model that is well trained but still general.

### I.8.1 Deep networks and node dropout as novel methods for improving network generalizability

Very recently, new methods in network training have been developed to improve the generalizability of neural networks and prevent over-training. The development of inexpensive General Purpose **GPU!** (**GPU!**) hardware has made it possible for extremely large networks to be efficiently trained. Additionally, development of new training methodologies (**?**) has made it possible to train networks with very large numbers of nodes, and more than 2 layers of hidden nodes. These so-called "deep networks" appear to be capable of learning abstract features and concepts in an un-supervised fashion (**?**), and appear to exhibit the kinds of learning behaviors that were originally envisioned by the developers of early perceptron methods. Another promising and broadly applicable new method in the training of neural networks is the so-called "node dropout" method. In this method, every time a new training case is provided to the network, 50% of the nodes in the network are excluded. This has the effect of preventing nodes from becoming dependent on each other, which leads to over-training. By using node dropout, it has been possible to both conventional shallow networks and deep networks using a larger number of nodes than would normally be allowable, increasing the generalizability and performance of the models (**?**).

## I.9 Using ANN!s to make predictions regarding drug activity is a major area of current research

As a result of their properties to model complex interactions in natural systems, **ANN!** based methods are a popular choice for constructing models of drug activity and binding. In many ways, drug activity is a harder problem to solve than image recognition. Unlike images, The activity of a drug depends in large part on its conformation (**?**). A cat does not become a bobsled if it folds its legs, but active small molecules can become inactive in certain geometric conformations. To sidestep this, **ANN!** based methods are often used to make 2D ligand-based **QSAR!** (**QSAR!**) models which are trained using the 2 dimensional structures of known active and inactive small molecules without including protein structure information (**?**). While 2D descriptors do frequently outperform 3D descriptors, 3D descriptors can be made useful. By encoding 3D information in the form of a **RDF!** (**RDF!**), the 3D geometry of the small molecule is described in a way that is rotation and translation independent. Additionally, **RDF!**s encode 3 dimensional protein data as a one dimensional fingerprint, making them ideally suited as **ANN!** descriptors. **RDF!** based descriptors, in conjunction with 2D descriptors, have been used to build a **QSAR!** model capable of predicting novel active compounds (**?**), demonstrating the value of the technique as a whole.

While ligand-based **QSAR!** methods have proven valuable for predicting the activity of drugs against specific targets, these models have a fundamental limitation in that they can only be applied to the target they were trained against. Techniques exist to define and maximize the domain of applicability of these models (**?**), but they are fundamentally tuned to a specific target or subset of targets. Furthermore, the training of a ligand based **QSAR!** model requires that a set of experimentally known active and inactive compounds exist, which limits the use of ligand based methods to targets which have already been experimentally evaluated. As a result of this limitation, some recent research has focused on using **ANN!** based models to score protein-ligand docking positions. Because of the

ability of modern **ANN!** based methods to recognize very complex and noisy signals, the potential exists to develop an **ANN!** model which is capable of distinguishing between active and inactive small molecule poses even in cases where the scoring function of the docking system is unable to do so. A number of methods have been developed to do this (**?**), and while they have in general been successful (**?**), the dream of a **vHTS!** method that acts as a generally applicable model of protein-ligand binding affinity has not yet been realized, and research in the area continues.