we are honored to have dr glenn kellogg

dr kellogg is currently professor and

interim chair of medicinal chemistry at

the school of pharmacy at virginia

commonwealth university

dr kellogg received a bachelor degree in

chemistry from the university of new

mexico and his phd from the university

of arizona in 9

after postdoctoral training at

northwestern university he joined the

medicinal chemistry group at vcu in 9

dr kelloggs expertise is in the area of

computational chemistry

i know you will enjoy todays lecture

hello my name is glenn kellogg and im

here to talk to you about molecular

modeling and drug discovery and design

im from the department of medicinal

chemistry at virginia commonwealth

university and ive been working in this

area for nearly 0 years

my goal here is to give you a little bit

of the background and

and how it works

behind it and then give you some leading

references and

ideas that you can take forward

into your research

so uh

lets start with the simple

the simple idea of theories models and

predictions

a theory is something that relies on

first principles to make predictions

while madonna

models are more empirical nature and

they will include known data and fitting

it to

to a model

both can be used to predict properties

and thats thats a key thing

is that properties are what were

interested in because thats enables us

to talk about

drug discovery

drug activity

and so forth

if we have a more basic understanding

which means of the theory

we can improve our theory and we can get

better predictions

if we have more experimental data we can

get better models

and that also leads to better

predictions

so what is a molecular model well

it is just about any time you try and

represent a molecule because theyre too

tiny for us to see

we

try to draw them we try and uh

make representations of them that show

how what their shape is and so forth and

even xray crystal structures are models

of molecules so anytime we try and use

some sort of visualization weve created

a molecular model

so what properties are we talking about

well there are molecular properties and

those are properties that

reside on a single molecule uh and some

of these which may or may not be

important for

for drug discovery would be like the

boiling point of a molecule its melting

point its log p its molecular weight

molar reflectivity the highest occupied

molecular orbital energy etc etc theres

many many of these

structure is a molecular property now

some of these might be atomistic and

what does atomistic mean well atomistic

means that it can predict it it can be

predicted just from the atoms involved

so what is example that well one would

be molecular weight we could predict

the

uh molecular molecule

exactly by knowing the

what the

chemical formula of the molecule is and

what the atomic weights are

other these like the homo energy or so

forth would be very difficult to predict

atomistically

so the other type of properties are what

im going to call intermolecular and

those are derived from the interaction

of two molecules and that would like be

binding constants free energy of binding

anything like that and these are the

sort of things that lead us to drug

activity because its its the

its the interaction between two

molecules that lead to a

uh

some sort of biological effect

now these are not likely to be atomistic

although some

aspects of that might show up

there are a number of drug discovery and

drug design paradigms and it turns out

that they kind of follow those two

patterns of

single properties and intermolecular

properties so the first paradigm is

called ligandbased drug discovery

and that uses the properties of the

molecules themselves

and then from those we can develop

models of activity and we can assist in

designing new active lead compounds

clinical candidates and ultimately leads

to new drugs

the important part of this is that

we could do this

in the absence of knowing what the

receptor is we would just do it based on

what the molecules themselves are

telling us in terms of their activities

the other paradigm is structurebased

drug discovery

and that uses the properties and

structures of molecules as theyre bound

to their receptors

and from that we can identify the

molecular features that are ideal for

binding and by inference then ideal for

activity this is actually the root of

whats called the lock and key model for

how for for drug discovery that every

every lock which is like a protein or

receptor has a key which is the ideal

molecule

so uh what kind of techniques do we have

for molecular modeling well

there are some that are related to

building structures and structures means

structures of small molecules structures

of proteins

structures of intermolecular

interactions

im going to go through these

in variable level of detail the first

one is quantum mechanics

uh well talk about that in a minute but

thats the the

uh hamiltonian and so forth that you

probably

saw

in maybe a physical chemistry course or

perhaps you were afraid of and didnt

take a physical chemistry course but

its its the hardcore uh

quantum mechanics approach to to

molecular structure

the next level is what we call

molecular mechanics and this includes

things like md which is a shorthand

notation for molecular dynamics

and another

model which is uh can be used to develop

the

uh

threedimensional structures of

molecules is called homology or

comparative modeling and thats a model

where we we look at the structure of

known molecules and use those to predict

the structure of unknown molecules or

proteins

were also going to talk about a couple

ligandbased methods

for drug

drug more or less design not so much

drug discovery uh and we will use that

for um

well use two two types of

methods there uh

qsar

and its in its offshoot dqscr

and well look at virtual screening

of databases where we have a template we

have a known molecule were looking for

something similar to it

then well move on and talk about

structure based methods

again we can use a virtual screening

approach but in this case we use the the

structure of the target the binding site

as the as the template to try and

determine a

a

ideal molecule to match it and well

talk about docking and scoring

so quantum mechanics well thats the the

hamiltonian h site was esai where psi is

the

is the uh schroedinger equation

uh

there is um

thats good and important about this we

can determine electronic structure we

can determine electronic properties

we can determine electroorbitals

ionization energies etc but on the other

hand because

were interested in fairly large

molecules

especially if were interested in

proteins

that make the qm methods almost

impossible theyre just not enough

computer power available for us to use

it so

its impossible for proteins and very

very likely and practical for larger

drug molecules if you want to do if you

want any level speed like if youre

trying to screen a large number of

molecules

however i must say that as the speed and

cost of computers

keeps improving uh we are gradually

moving more and more towards a qm based

methodology for drug discovery

so lets talk more about molecular

mechanics the very basic idea of

molecular mechanics is that we treat

molecules as physical entities

as in things you can actually feel touch

and hold

and they are therefore governed by

conventional which means the nonquantum

newtonian laws of physics

so lets say we want to calculate the

energy of a molecule well there are

right there a variety of ways to do that

we could go back to

what you probably saw in freshman

chemistry where

you counted how many double bonds it has

how many single bonds it has and add it

all up and get some sort of energy but

another way to do that which is useful

for our goals is to

take the total energy molecule and

divide it into

the different types of

of

motions and

instabilities that a molecule might have

so well have a stretching energy well

have a bending energy

an oop which means out of plane ill

talk more about that in a minute we have

torsion energies van der waals energies

and columbic energies

in summary the total energy of molecules

are some of the various energy terms

that represent the various stresses and

strains on that molecule

so lets start with a stretch

we have a diatomic molecule here uh two

the two atoms are represented by the

simplest physical entity we can think of

atom atoms are going to be represented

by just balls like billiard balls or

golf balls or whatever you want

and the distance between them is given

by r

and is its drawn here r is a stick

representing the distance between them

but we we we know a little bit more

about molecules and that we know that

molecules are not

rigid that theres various motions on

them in fact a better representation

would be to put a spring between those

two

those two atoms between those two balls

and if we do that we can use the

we can use the

hookes law as represented here in the

bottom as a

as an equation that represents

the energy

of the spring as a function of distance

between the atoms and thats whats

shown by this graph here

hookes law

is actually has an r squared term in it

such that

such that it gives a

parabolic representation of energy so at

the the lowest energy part right down

here at the bottom

is when this the spring is at rest and

if you try and stretch a spring make it

longer youre going

its going to cost you energy and if you

try and compress the spring to try and

make it shorter its also going to cost

you energy so we can represent

the

the force between two atoms

in terms of

how much

work or energy it would take to move

that distance

in terms of hookes law now whats what

makes hookes law especially useful for

this is that this is very very simple

mathematics its just a square term and

we can

do that on a computer

lightning fast

now lets

change over to a

triatomic molecule we again have

three balls

represented here and we have two sticks

but now we have a new feature we have an

angle between those two bonds

and

we

could go through the process of using

sines and cosines to calculate energies

but a trick was developed quite a long

while ago to make this simpler why dont

we pretend theres another

spring

a fake spring between those two atoms

and now we can use hookes law again

we use hookes law to represent the

change in the angle theta as a function

of distance

and that would give us again the very

simple mathematics we we really are

craving in order to make this very fast

now i mentioned the outer plane earlier

that is what happens when we have four

atoms three of them will automatically

be in a plane because thats the

definition of plane as any three

points in space represent a plane and a

fourth is

either in that plane or out of that

plane thats what oop means out of plane

and there would be an energy associated

with that again

we could use

fairly complicated trigonometry to get

the answer but

the simple pragmatic approach is to

measure the distance between the plane

and that atom as d

and again use a simple hookes law term

to calculate the energy of that

lastly in terms of these sorts of terms

is whats called the torsion angle the

torsion angle is the

is the

rotation of these two atoms

with respect to these two atoms so it

swings around this

this uh bond here now if you look at

this it first may seem counterintuitive

but none of the none of the bond lengths

have changed none of the actual three at

none of the actual three atom

angles have changed were only changing

the

were only changing this

angle here and

the

importance of this is that that changes

the shape of the molecule a tremendous

amount having it stretched out like this

or compressed like this

changes the shape of the molecule

tremendous amount

unfortunately weve run out of tricks

that we can use to

simplify the math and we were stuck with

this fairly complicated equation

to calculate the energy of a torsion

angle

now the torsion angle turns out to be

important because it affects

the shape of molecules i mentioned and

the shape of the molecule leads us to

worrying about what is the

what is the level of interaction between

two atoms and this gives us whats

called the van der waals term or the

london force term or theres a variety

of other names that are used

so the van der waals

in simple

in simple terms and this is not

precisely physically

but this is not precisely what a

physicist would say but

if

what is happening is theres an

attraction between the nucleus on one

and the electrons on another

and this causes the electron cloud to

get pushed around a little bit so that

it tries to tries to

optimize that interaction so theres a

little bit of push and a little bit of

pull here

so it it influences each other and at a

distance where the electron cloud of one

is

optimally placed near the new the

nucleus of the other

and vice versa that is called the van

der waals distance and they are just

just touching each other so this little

animation

sort of shows how that happens

and

the optimum distance is called the van

der waals distance

this

graph shows what it looks like

mathematically

the important thing i said about the

the torsion angle really comes into play

because theres nothing in anything

weve shown you so far that present

prevents two

atoms from actually passing through each

other and this is one of the purposes of

whats called the van der waals term

and thats evdw

in simple terms and this is this isnt

physicist approved

the the van der waals terminal rises

because because two atoms each have

possibly charged nuclei and negatively

charged electrons and theres a little

bit of attraction between the positively

charged nucleus of one and the negative

charge negatively charged electrons to

the other and it causes an attraction

force and eventually reach a point down

here at the bottom

where they are

just touching each other

perfect kissing distance we might say

and that is whats called the van der

waals

distance or the ven and thats what the

van der waals radius for each atom is

derived from so this graph

shows the

energy of this as a function of distance

the

lowest

energy here is where the two

uh

the two atoms are just in this perfect

kissing position here but if if i

pull them apart from each other

ultimately were going to its going to

take a little bit of energy to pull them

apart and its going to

reach a point

at infinite distance where theres no

energy

attraction or repulsive between the two

molecules and thats and thats as you

might expect as they get farther farther

away from each other they

uh the energy dissipates however

and this is the more important part is

if you try and push them too close to

each other the energy rises very rapidly

until it reaches a point where its

extremely repulsive because were trying

to cr were trying to smash those two

two nuclei together or smash the

electrons together and that is what

course prevents the two atoms from

touching each other the the term

the mathematical term for the van der

waals force is shown here and this is

also called the leonard jones term and

it just is a simple mathematical

function that has a a power and a

power of representing the uh

the

distances between the atoms and uh this

mathematical function gives the shape as

shown here

now the next term the columnic term is a

is a charge based term which represents

the

uh

electrostatic energy between two atoms

and lets say we have one atom its got

a q charge and the other atom has a q

charge

uh columbic

energy is given as this now if those two

atoms are of the same charge thats a

repulsive force and theyre going to

push away from each other and if theyre

of different sign

uh there its an attractive force and

theyre going to pull towards each other

now where did the charges on these atoms

come from

you know are molecules neutral why are

their charges well

molecules are neutral but there are

differences in

electronegativity between atoms as you

know the electric electronegativity

increases to the right and increases as

you move up the periodic table so more

electronegative atoms have a tendency to

be far to the right or or at the top of

the periodic table so using those those

terms the electronegativity terms and

and factoring in a little bit of

polarization energy

each atom

generates a

or each each atom possesses a partial

charge

that is is relative to the other

atoms in the molecule

so on the left here we have just a

simple benzene molecule and you see that

the carbons are a little bit

electronegative with minus 00 but

each of the hydrogens is a little bit

electropositive plus 00

altogether when we add up all these

terms of course the total is zero the

molecule is neutral but each of the

atoms have a partial charge now as we

change the type of molecule in something

more complicated

we find for example in this molecule

that the the the partial charge

on the carbons reaches its maximum at

the meta position

relative to the other carbons because

the methyl methoxy carbonyl group is

electron withdrawing

and here the sensing means our electron

releasing the

the the carbon charge

the c charge

reaches its maximum at the ortho and

para positions here

in here

okay so

we put all this together

we have whats called a force field

and that describes how to calculate the

potential energy of a molecule

the bond lengths the bond angles the

atomic radii all those things we talked

about are taken from high resolution

crystal structures of small molecules

and from proteins so we we know what the

an ideal carboncarbon bond length is

when an ideal carbon hydrogen bond

length is and so forth and we use that

to help calibrate our terms

and the force constants for hookes law

the stretches and the bends are taken

from infrared and similar experiments so

we we experimentally calibrate the

functions the mathematical functions we

we have in our in our force field so we

put this all together when we have

information for every conceivable

sort of uh bond and angle and

interaction we would have a force field

that could calculate in principle the

energy of any molecule

now that that is useful in itself but it

becomes more useful because we can use

it to predict the energy of a molecule

and thats whats called optimization or

minimization

so there are two terms

energy minimization and structure

optimization they are

used somewhat interchangeably its a

simultaneous process

by definition the most optimal structure

is the one that has the lowest energy

so

it means that if we can

keep

getting a lower energy structure we must

be getting a more optimal structure

and but also by definition

these are performed at zero degrees

kelvin in vacuum so theres were

eliminating some other competing uh

uh competing energy terms now there are

all kinds of permutations on the

algorithms and im not going to talk

about the algorithms theyre mostly to

increase the speed and increase the

ability to guess ahead the basic idea of

all the

structure optimization protocols is that

if the energy is the is decreasing then

the structure is improving

so as long as the energy is going down

were getting a better and better

structure unfortunately however its not

quite so simple

there are something called

local minima and global minima so for

example

this is a

naive representation of a cyclohexane

drawn as a as a hexagon

we all know thats not really true it

actually exists

in two forms

the boat form

and the chair form

and uh that

affects the energy so

if its a simple structure

therell be a starting point of our

minimization were going to

take a

a

incorrect structure and try and create a

correct structure by minimizing the

energy so as the energy is going down

here

were getting a better and better and

better better structure until we reach a

minimum

ideally thered be only one minimum but

in reality there can be more than one or

there can be many of them so

going back to our cyclohexane here lets

say we have

a badly drawn cyclohexane

and we

minimize our energy minimize our energy

and then we reach this point here this

local minimum

which happens to correspond

unfortunately

to the boat form a cyclohexane

how do we get down here well because the

rules that i described on the previous

slide we cant because

the rule is that if the energy is going

down we must be doing the right thing

there is no way to go from here

to here by only going down we have to

actually go up a little bit and then

back down

so this is whats called local minimum

the absolute correct structure is at the

global minimum and that would be in this

case would be the

the chair form of cyclohexane

so how do we get out of these local

minima well

two major approaches and a variety of

other approaches but the two major

approaches are one to do an exhaustive

search

we would calculate the energies of all

the possibilities and then pick the one

that has lowest energy that in principle

should always work

and the one that has lowest energy would

be the one thats at the global minimum

so for example

we could vary the torsion angles of

rotatable bonds and ill show example on

that in the next slide if we rotate our

rotatable bonds by about degrees each

each turn

the number of structures we would have

to calculate would be to the n where

n is the number of bonds where i have to

rotate now if n is one or two or three

its not a big deal uh we can calculate

a few hundred a few thousand structures

without any problem at all but if n gets

up to be 0 or something like that

then were talking about

billions or more structures in which

case its not long not going to be a

simple mathematical problem but actually

a much more complicated problem

so the other approach is called

molecular dynamics

in in molecular dynamics we add some

heat

remember i said it was at zero degrees

kelvin but now we would add some heat in

wed add some time and let the molecule

kind of shake its way out of this local

minimum

and thats what you do with larger

molecules so first looking at

the exhaustive search approach

heres a

long straight molecule

but it has a variety of rotatable bonds

and we may or may not have it at the

optimum

confirmation so

lets look at these bonds well

well

what would happen if we rotate around

this bond

well this would spin the flooring around

and it wouldnt change anything so that

it may be wrote it might be rotatable

but it doesnt change the structure at

all so that ones really not very

interesting

what about this one well in this case

we would

we would take that

the

benzene ring to the left there

and

spin it in and out of the plane so that

by rotating around this bond we would we

would change this energy somewhat

now what about this one well this one

would swing that whole group to the left

around and change the structure by quite

a bit

and this one would do a similar sort of

thing

as with this

as with this as with this all those

things would be changing the structure

of the molecule

what about this one well in this case we

would swing that

methyl group there at the end around

like a propeller and yes that would

change the structure but not as

significant as some of the other ones

what about this one

what if we spin this one around

well

it depends on what detail youre

interested in if its a

if youre really interested in like

where the electron pair is on an nh

group then spinning this around would

change its structure by quite a bit

because electron pair would spin around

as you know its the

nh is a

is like a two uh

a twolegged stool

so we would we would change a structure

what about spinning around here well

maybe if youre really interested in

where those three hydrogens are on the

methyl

that might

cause a structure change but thats

probably not that important so in this

particular case we have between seven

and nine rotatable bonds depending on

which ones you consider important the

seven

would be the green ones and the

eighth and ninth would be the orange and

yellow ones

heres molecular dynamics its uh its a

whole

major topic so im just going to touch

on with this one slide but the idea is

you take a molecule you put it

ideally in a box of solvent as shown

down here on the lower right

and you add some heat to the molecule

and

it heats up starts moving around a

little bit

and you can

perhaps make it change its confirmation

depending on how much heat you add how

long you let it go and so forth but the

idea is if you heat it up and cool it

down heat it up and cool it down a

number of times you might find it in a

different conformation thats

energetically better than the one you

started in

okay im going to change topics now to

to ligandbased drug discovery now weve

kind of gone over all the approaches we

might use in order to build molecules

now lets think about how we can relate

this this

the

structure molecules to their activities

and qscr is a technique developed in the

late sixties by uh corbin hanch and tuja

fujita uh

and really its kind of really simple

idea

activity is a function of structure

it seems obvious now but

its not so obvious because

you have to find a way to represent

structure in a meaningful

and mathematically accessible way

electromechanics

as we discussed previously gives us

access to a number of mathematical

constructs that relate to structure

and

those are all useful but

theyre all physics remember what i said

when we started electromechanics is that

were trying to treat our chemicals as

being

physical physical objects so

wheres the chemistry and thats

where

thats where qsr is important is because

it it allows us to add some chemistry

so in the fujita analysis

they focused on molecular properties

more than structure descriptors

they use the pi constant which is a

the law which is related to the log p

for a one octanol water uh

mixture

are layers and thats

commonly called hydrophobicity they

looked at the ham and sigma for

electronic effects and the f e s

parameter for steric effects and the

equation they came up with shown in the

middle here

log of over ic0 is equal to

some constant a times pi plus some

constant b times sigma and plus some

constant c times the tap parameter plus

d

and they found that they could predict

the ic0 of a variety of molecules

reasonably well if they were able to

if they were able to

uh calibrate their function with a with

a good collection of known structures

but all the descriptors they used are

empirical theyre all based on

measurements of other molecules

and there arent any descriptors that

describe or explain the molecular

structure and shape of molecules

so that a

a an arms race a source developed in

order to come up with more descriptors

so theres the first group here which

are the

physical chemical properties which are

the empirical ones uh

in addition to the hash fujita

parameters people found the melting

point was sometimes useful solubility

was useful

structure was used

uh in a couple simple ways one is the

molecular weight a larger molecule will

behave differently than a smaller

molecule but also you could count the

number of hydrogen bond donors the

number of hydrogen bond acceptors and so

forth and use that as a as a descriptor

uh people looked into using quantum

mechanical terms like the

the highest occupied electron energy in

related and then uh

another whole branch came out of this

called the topographical or graph theory

descriptors

for example the kieron hall

set was used

this led to what which is really a

simple idea a qsar table so

on the left here our left column are the

compounds and well just call them a b c

d or whatever

uh and their their biological measured

biological activities are y a y b y c y

d

and the scripture one descriptor two

descriptor three descriptor four

has values for each of these now some of

these would be log p some of these might

be the kieran hall descriptors they

could be

homo energies or any number of things

are

are added here as our descriptors

and

then we try and solve this equation of y

the activity is equal to the function of

descriptor one to script two descriptor

threes and so forth in order to come up

with an equation

that

leads us to be able to predict

the activity of a b c d etc so the more

data you put into your qsar

the better your predictions will be

this is the base this is the basis of

the whats called onedimensional

sometimes twodimensional qsir

so

the trick was of course having

activity data for for a large number of

molecules being able to

collect

or

or calculate descriptors for each of

those

molecules

and uh

then

do some sort of statistical modeling in

order to in order to

create a

a iq sar a qsr equation

and you can use multilinear regression

partially squares genetic algorithms

simulated annealing neural networks

support vector machines and a number of

other approaches to create

this

this equation

now if youre

if your equation is if your equation you

create from your known data is good then

youre in a position where you can

actually start predicting the results

for unknown molecules

but they have to be more or less in the

same set of molecules that youre

looking at so you cant jump from

a series of molecules that are active at

one

activate one receptor with one kind of

one sort of

core

uh template or or

or a

core template or core structure and then

throw in another completely different

thing because those sort of descriptors

are not gonna work as well

so this led to a

different approach which was called d

qsar

in which we represented molecules not by

their

molecular properties but by their shape

and their shape was represented as

fields

so in this particular this particular

paradigm you would take a molecule

shown here put it in a threedimensional

box of grid points and its represented

by all these little dots all over the

place

and

each of those grid points would

represent sort of a little test molecule

or test atom

that would measure its

feelings towards

the molecule so if this

test atom was sitting here right root

right close to this oh group it would

feel that hey im near an oh group and i

behave this way but if youre if its

sitting over here it would not feel much

of anything because its over in the

corner of the box it wouldnt give much

a response at all

so this was the basis of the first three

qsr techniques which is called kumpha

and that included ways to calculate two

types of fields one was a

steric field which is used using the van

der waals potential

and the second was the electrostatic

field using the columnic potential

so it looks something like this uh the

columbic term

uh

would

sample each point in space with respect

to electrostatic charge

basically using the columbic equation i

showed you earlier

and

that would represent the charge effect

of that molecule throughout space and

it would look something like this

the steric field on the other hand

is using the van der waals

van der waals term or the leonard jones

potential term which i showed and that

drops off really rapidly

once youre outside the mount outside

the

the van

outside the van der waals radius of an

atom so it kind of has this almost

shrink wrap look to it so it tells you

when youre inside the molecule it tells

you when youre outside the molecule

and doesnt have much in between and

thats why it looks this looks like a

molecule thats been shrink wrapped

and again its done the same way you

just use this equation

overall its used as this equation

overall space with a set of grid points

so if you do that

we can

calculate a

qsar

in this case we have our compounds like

before

we have our biological activities like

before but instead of descriptor d d

d like i had before we have

the value of

the steric field

at point

or

0 or whatever just a

a

a

point in space and we would

fill in this table with all the points

in space so instead of there being just

0 or descriptors we will have

hundreds of thousands of descriptors

each representing the value

of the

of a grid point in space

and we would solve it

again is why the activity is a function

of these points in space

and activity would again be

a function of of those points in space

and this this is the basis of d qsr and

its turned out to be a rather rather

useful way to look at structures now

what are the

what can we do with that well we can use

statistics to drive a model

activity as a function of those points

in space

we have to use crossvalidation and ill

show you that in just a minute

that enables us to determine the

internal predictiveness of the model and

then we will validate our model with an

external test set to make sure that its

true

and we can also learn spatial

relationships by making special maps

of activity to learn where

in space and where on a molecule is the

most

likely place to change to change

the most likely place to modify the

molecule to change its activity

so just briefly cross validation in case

you havent seen this term before

because we are using so many descriptors

in dq sar we cant rely on a standard

rsquare to give us much useful

information in fact our squared on d

qsr is always

very high like 09 to 099 so we have to

do crossvalidation so what

crossvalidation is is that well lets

say we have molecules in our data set

a through z

and we want to calculate

the validation or the cross validate r

squared which is also called the q

squared for this molecule what we would

do first is we take we leave number a

out we wouldnt we would leave it out of

the model it would use b through z to

calculate a model

for the activity and from that model we

would predict a

and then we would do it again by leaving

b out or leaving c out and this is why

its sometimes called leave one out

so all those predictions for a through z

would then be compared against the

actual value

the actual measured values for that

in order to calculate whats called the

q squared so the better the q squared

means the better the model is at

predicting

uh the structure predicting the activity

of the structures in the data set

and the other advantage of d qscr is

that we dont have to

use a limited number of descriptors we

can use

many different sorts of fields we can

use the hydrophobic fields we can use

uh topological fields and a variety of

things and

when you have a collection of fields

like this it enables you to

finetune your

your dq dqsar for the type

of molecules you have like if you know

your molecules are

are correlating pretty high with it with

hydrophobicity you would use a

hydrophobic field in order to

differentiate that

so i want to make a couple points here

one is that q squared the

statistical

measure of the field

is a

important factor but you also need to

think about

the interpretation if youre using

fields that

are

difficult to understand thats not going

to help you design new compounds because

at the bottom

the bottom line is that were doing all

this because we want to design new

compounds

with new new and improved activities

and

so for example the hydrophobic field

type is based on chemistry its not

based on physics it might have more

chemical intuition than than just the

steric or electrostatic fields so its

important to consider how interpretable

it is

in addition to

how good the statistics are

so theres pros and cons of using d

qsar

one advantage and i alluded to this but

lets talk about a little more detail

because

these property fields are independent of

the specific backbone structure of the

molecules we dont have to stick with

the same backbone we dont have to use

the same template for the molecule we

can actually move from one to another

this is whats called scaffold hopping

that allows us to expand our our data

set to a variety of related molecules

not just

close families we can learn which part

of the molecules are good for activity

or bad for activity and that gives us a

clue on how to design appropriate

analogs

and we can tailor our fields to the

characteristics of the data set

but the cons are is that

we have to have

because its a because its a

ligandbased method were not paying any

we dont know we usually dont know

anything about the structure of the

receptor thats or

our enzyme its binding to we have to

presume what we know what the active

conformation is and there quite often

can be a large number of poses for a

molecule or shapes or conformations from

molecule that are very similar

and it requires us to be able to take a

collection of molecules and

overlap them in a way that they

again that we presume that they are

going to be binding at the active site

and the way to start that is if you

create a pharmacophore pharma iv would

be the collection of most important uh

most important features of the molecule

this can be sometimes nonintuitive

because its

its difficult to say what the most

important part of the molecule is if you

dont know anything about how it binds

but if you have some

if you have a good activity on a rigid

analog that can really help

and lastly and i think this is an

important point

which i sort of mentioned is already is

that

chemistry

can easily be lost if you focus too much

on the statistics and thats

thats not the point the product

of any sort of molecular modeling

exercise is to develop

new molecules not

statistics

okay so most computational chemistry

especially in medicinal chemistry is

dressed up qsar

the key thing of qsar was that the

binding efficacy and all those things

can be correlated with structure

so

uh

that

is the basis of

nearly everything else that happens in

in in

in drug discovery and drug design is

really its a form of qsar so im going

to talk a little bit about uh ligand

structure based

pharma farmwork

database searches and target based

searches and im going to talk about

docking and scoring

as well and i should also point out

theres a

a

method from jorgensens group that

uses a linear response method for

calculating free energy of binding that

is surprisingly qsar like

so virtual screening uh is a terminology

thats used to

discover molecules uh its

really

maybe where how weve

how were going to make qsr achieve the

goals that we wanted to have

so virtual screening the terminology was

came out of vertex with pat walters matt

stahl and mark marco

and late late 990s and definition is

thats a its the use of a high

performance computing method to analyze

large databases of chemical compounds in

order to identify possible candidates

so why would we do that well its a

filter

we can reduce the size of a chemical

library to do physical screening

very easily from millions to thousands

no one wants to be in the lab screening

a million compounds but you might

convince them to look at a few hundred

it will increase the likelihood of

finding good compounds

so if we

throw away all the ones that just arent

going to work

and just look at the ones that might

work we would get better hit rates

if if you look if you screen a library

of a million compounds without doing any

sort of

prescreening

uh your likelihood of you likely to

getting anything worthwhile is far less

than one percent if you do some good

virtual screening you can get that up to

five or ten percent so its

its worth doing

we can uh perform some sort of analysis

before an assay is even established

and

probably the most

interesting and cool thing is that you

can actually

figure out whats worth making before

you before you make it

now we targets out there are increasing

rapidly theres many many new targets

every year

and were going to need more and more

computational methods in order to

survey all those targets

so theres a variety of sources of small

molecule structures

uh the c c c d c which is a cambridge

structural database has a has a

a

a database of small molecules the nci

has about a half million compounds that

you can look at

and and some of those will even make

available for

testing in your lab

theres a pub ken

library and theres a library

virtual library called zinc which comes

out of uh

comes out of uh

ucsf uh and has

0 million and grows by the day it could

be more today i dont know uh lots of

compounds where they give you the

structure and they tell you how to buy

it

but perhaps most importantly for uh

for a lot of people is that the

pharmaceutical companies

have large private libraries or

compounds that have been part of

previous projects

when they are doing a cancer study and

the compound doesnt work they dont

throw it away they keep it because you

never know what it could be useful for

something else later and so those are

the crown jewels of every every drug

company

when one drug company buys another drug

company

one of the things theyre after is their

library of compounds

so

lets use our d databases in two

different ways this is the first way

thats the ligandbased method

lets say our competitor or a drug

company our competitor has an active

compound and theyre making billions on

it

we want to see if we can get something

in the same in the same uh

the same space that will allow us to

make billions too so we take our our

competitors compound we find out what

structure is and we determine the

pharmaco

and by

using the the spatial arrangement of the

components of the firearm before we can

form a query

and then we can apply our query to a

database and look for hits

and then then we go on to the chemist

that well show to the chemist and see

if he can make

leads and analogs and so forth so heres

an example of that

molecule up here in the upper left

is our reference molecule our competitor

has this and theyre just killing us in

the marketplace

we

look at it we say okay its we think

its hot because

its got this hydrophobic ring up here

in the left the cyclohexel ring uh

we like this heterocycle here

which is

and we like this uh

oxygen over here so those are going to

be our three pharmacare points here

here and here

uh and uh the green ones are the

hydrophobic hydrophobic groups the blue

are hbond acceptors and the yellow are

aromatics

and this is a representation of this

in terms of those three data those three

pharmaco points

we measure distance between them because

thats important

its nearly seven angstroms between

these two

nearly seven between these two and about

three between these two

and that take away the molecule were

left with this

we expand out the distances a little bit

because we we know that its a little

flexible and doesnt need to be that

tight

and we use this as our as our query

now we

take this query and apply it to our

database

we look at this

we look at this molecule well its got

the same features its got a green one a

yellow and a blue one but the distance

are all wrong so well throw it away

we look at this one

the distances are wrong its got the

wrong features were going to throw this

one away but here we look at this

molecule

its got the distances its got the

the

uh the features

and this is a potential this is a

potential molecule that might lead us to

a

competitive molecule in this in this uh

in this space now

when our competitor patented this

this uh

molecule

they patented the molecule not the

farmworker for they cant patent

pharmaco so if we could come up with a

molecule that has a different structure

but still has the same pharmaco

we are free to develop it

thats whats shown here

now the second way we can do this is to

use uh

d d databases in a structurebased way

we will take the active site of a

of a protein or

or

a protein or a receptor or enzyme or

whatever

and

use that to define determine what we

think the farmworker for should be and

again well then create a query well

apply the query to our database to drive

hits and then well have our chemists

look at those and see if we can make

some lead compounds now this is a kind

of contrived example but

this is the methotrexate binding site

here

and

theres a feature on the on the binding

site shown here which is a uh

which is a which is an acceptor

that means our ideal molecule should be

a donor

here this feature on the on the far on

the binding site is a is aromatic so we

want to try and get a pie pie stacking

thing going here and we will

look for something that also has an

aromatic group in this region in space

now these

these two

features on the molecule

on this molecule here are

donors which means we want to have

acceptors on our molecule

so uh lets see what we get well

surprisingly enough the

the compound that matches the best to

the methotrexate binding site is

methotrexate

uh

which like i said its contrived uh but

we have it shows the features that we

want so first

we needed something that was an acceptor

and a donor and thats shown

as shown as this

theres our heres our donor

because the the site has an acceptor we

want a donor

we wanted to have something

that was

aromatic theres a big nice aromatic

ring here in the middle

and we also wanted

to

have a um

a couple acceptors

which are shown as here and here so

surprisingly enough methotrexate binds

at the methotrexate binding site but

thats the principle behind it the

principle is with that you look at an

active site find the features of the

molecule that fit most ideally and then

do a database search to find molecules

that match those features

okay were going to move on a little bit

now into docking and scoring thats the

next phase of

of drug discovery if you find a

collection of molecules that meet the

pharmaco requirements next thing you

want to do is if if you do have a

structure of your protein that you are

going to try and dock to or bind to you

would want to see if you how well you

can

how well you can bind your

putative ligands into that molecule

so

uh theres two steps docking and scoring

theyre theyre

used together but there are two separate

steps

the first step is bringing together

the models of our receptor and our

ligands in three dimensions

this turns out to be pretty easy there

are lots of ways to do this

uh many algorithms are available so its

not a thats not a problem the hard part

is to score the interactions

which means predict how well those

molecules are going to bind

and what their activity will be

and

have it

be relatively

accurate and be able to relate different

molecules with respect to other

molecules

ie rank them

so

there are many many methods of docking

algorithms

available

this lists

seven of them or so

i point out

we wrote a chapter that is in burgers

medicinal chemistry

the more or less the bible medicinal

chemists

and you can look at that im not going

to go into i dont have time to do all

these so im just going to talk about

the first one the simplest one briefly

but if youre curious about the other

methods look in this chapter

so theres point complementarity which

im going to talk about theres

approaches based on distance geometry

which is used in the doc method that um

the kunst group has developed

theres exhaustive and systematic

methods uh theres something called

incremental

construction which is part of the

program called flex x

theres ways to use molecular dynamics

in this

uh theres some genetic algorithm

methods and thats those are used both

in the gold program and the autodoc

program and uh

are some combination methods where you

use different

approaches or you have a consensus where

you use

multiple approaches and find the one

that

most agreed to

so just briefly point complementarity

this this is a this is a very old

approach because it requires actually

rigid structures

so if we have a

uh

our molecule our molecule here

we will turn it into a set of points

and uh

use those points in space and then the

pluses are where its uh

positively

positively charted on a donor then the

minuses are where its an acceptor and

the os are where its hydrophobic wed

match that to our receptor which is

shown here which has the same sort of

thing and find

find a way to make it fit the best so we

move this over here

shift it up shift it down shift it

forward shift it back until it makes the

best match of the pluses and minuses and

os and so forth and then we call that

docked

uh

and again this is a rigid approach and

of course

we know molecules arent rigid we know

the receptors arent rigid

so the other approaches which are more

complicated

add flexibility

among other features

so

the lessons of docking are that its

its

its pretty easy to find possible poses

for for a legging in a docking site its

a geometry problem and you can always

solve it exhaustively if you try every

possibility you always one of those will

be the right one

but you want to do this in a rapid and

reliable way you dont want to

you dont want to find

and score confirmations that arent

likely to be any value you might as well

throw those away as quickly as possible

you also have to make sure that if

youre not sampling everything that

youre that you are sampling the true

pose

and we have to

pay attention to the fact that were not

just ranking the poses at the end we

also need to

have some decision support that goes on

as part of the process

we dont want to

have to score 0 0 billion

possibilities at the end wed rather we

want to throw away the ones

as quickly and easily as possible that

arent going to make it to the to being

an a correct answer

so the scoring functions that we have

have to balance

speed which

so that we can uh do it rapidly as well

as their accuracy

so in scoring functions we

map a rather abstract concept which is

the measure of a binding force to a

numeric value

and the whole point of this is to rank

one ligands pose to another

we can apply scoring functions in

multiple ways

but the two ways that are most important

are during pose generation

where

the more approximate methods and

algorithms may be good enough we dont

need to

determine to this fifth decimal place

something that isnt going to work we

just want we want to determine does it

have a chance if it doesnt lets throw

it away

and at the end we want to we want to

evaluate the final poses the ones that

weve passed through our filters

and here were more interested in

accuracy

but we have to

again pay attention to how many we we

have to score if were going to score

five or six accuracy is probab

accuracy

at high level is worthwhile before

scoring million thats not worthwhile

so in terms of building a scoring

function theres one equation we have to

pay attention to and that is

that gives free energy we always have to

pay attention to whether or not a a

scoring function gives us a free energy

value because thats the currency

of

of interactions

so theres a variety of classes and

of functions that are available theres

force field methods which are

using again newtonian electromechanics

like we talked about before

there are summary empirical methods

where

we add terms from observation

for example

a commonly used one is the hydrophobic

contact surface area

we simulate hydrophobics by seeing how

much

hydrophobic hydrophobic contact there is

between the receptor and the uh

the compound binding

there are methods that are wholly

empirical

and thats

sort of like the qsar approach we have

training sets and we derive descriptors

in order to develop empirical methods

and there are knowledgebased methods

and this uses a potential means force

potentials of mean force approach

that we

take known structures for example

crystal structures of

known protein ligand complexes and

develop rule sets of

of

how often does a carbon

oxygen bond approach a hydrophobic group

or those sorts of things and we can

develop these rule sets and that gives

us another approach to calculate methods

and i point out again that our our

chapter in burger has a detailed

description of scoring functions

but theres some lessons that come out

of all this and again theres much that

can be said about this

uh theres some tradeoffs the first

tradeoff is speed versus accuracy

its a classic example of how

you want two things at the same time you

cant really have them

you cant be infinitely fast and

infinitely accurate it just is not going

to happen we can keep trying and people

do but its just not going to happen

youve got to youve got to determine

where on the continuum between speed and

accuracy you need to be if you need to

be really accurate youre going to have

to be willing to accept the fact that

its going to take a long time to score

it

if you need to be really really fast you

have to accept the fact that

its not going to be the most accurate

method

another tradeoff again involving

accuracy is generality versus accuracy

if you want a universally

applicable function

that will work on every possible

type of protein ligand complex that

exists

its not going to be that accurate

as one that you

created and calibrated for the problem

at hand so if you want to look at

if you want to look at hiv protease

and ligands bound to hiv you created

you create a you create or calibrate a

scoring function that works for that

not expect one that was calibrated on

some other protein to work as well

the third thing that comes out of this

is that most scoring functions available

commercially are calibrated to reproduce

the crystallographic structure of the

complex

they calibrate them so that

you can it will recreate

uh a known crystallographic structure

and its not

calibrated to

to

reproduce accurately the free energy of

binding

so even though its it seems intuitive

that

if you get the crystallographic

structure right you should get the

energy right it doesnt always work that

way and you can find many uh cal many

scoring functions give very poor

predictions of binding but they do

reproduce crystal structures

okay and now i want to say a little bit

about the hydrophobic effect water

and drug discovery and design that

exploits water

and partly because weve recently

written a perspective for journal

medicinal chemistry on this but also i

think this is the last thing we really

need to look at and think about

the hydrophobic effect

is a very major driving force in protein

structure ligand binding

in fact it might be the most important

force

in these things because water is

everywhere and its influencing all

interactions even when you

are thinking that its not

so

the hydrophobic effect seems to be a lot

like a van der waals interaction in fact

some people say oh well the van der

waals

van der waals term will take care of the

hydrophobics

and and of course thats enthalpic but

it isnt always what it seems because

there are other things going on you have

to go back to remember what the origin

of the hydrophobic effect

its more related to the

rearrangements of water and their and

their possible motion

than it is to an attraction each each

water

in a

biological system seeks to form better

hydrogen bonds with other waters and

polar species so

it actually turns out the hydrophobic

effect is entropic not in dalpic

so just a simple cartoon to represent

this on the left

is uh just a collection of water

molecules remember that

uh liquid water

the average each water is on average

bound to

uh two other waters so making its

making uh or three other waters im

sorry making three hydrogen bonds some

are making four so were making two but

on average in liquid water its about

three

uh

when a molecule is present within uh

within

a water and and dissolve for example

this molecule here

uh the wall

the hydrogen bonding of the water

to that molecule

the functional groups of that water

are

it and make it fit in now this

particular molecule has a

has a methylene group in it and nothing

is hydrogen bonding to that but

because this particular molecule is

making quite a few interactions with

water

it is

pretty fully soluble there now

lets

change up a little bit

now we have

two

very hydrophobic molecules that are

are in the water

they are

caged by the water the water is

not

trying

to interact with them its trying to get

away from them

and it sort of encapsulates them by

turning away as much as it can to create

a pocket and this is actually the

hydrophobic effect because

water and oil dont mix just like just

like your mother told you and

the

small even on a micro scale

this oily molecule does not want to be

sidewise in water so what water will do

in order to

to improve its situation

is it will

push away

from the from this uh these two what

these two uh hydrophobic molecules is

shown by the purple arrows which has a

tendency to look like the the two green

molecules are approaching each other

to create something like this

so thats the hydrophobic effect theres

there appears to be a hydrophobic

attraction between these two hydrophobic

molecules but in fact its its not

really there its the fact that the

water molecules didnt want to be

associated with it that makes it appear

as as though theres a hydrophobic

interaction

so um

one

uh

thing comes out of water is how useful

is it in drug discovery and drug design

so im gonna

just briefly present two scenarios one

is a drug discovery scenario

so

is in drug discovery

were talking about

finding a new molecule to fit into a

space

that didnt have a drugtype molecule in

it before

so if that if there are some water

molecules that are isolated in

hydrophobic pockets

theyll it wont vary that water

molecule will not make very many

interactions because it doesnt have

anything to interact with just like

before just like in the previous slide

so this is probably a very highly

entropic situation

and that kind of water will be very

easily displaced to bulk its going to

go into the bulk of the water with the

with with its

uh with its other water molecules its

friends and it will

create new

new hydrogen bonds and also

become less entropic

so these new interactions

are going to be enthalpically favorable

but this creates a loss of entropy

because

it was in tropic when it was sitting

here in the hydrophobic pocket now its

not

but if the water is part of a cluster

that has a variety of interwater

interactions

it becomes a little more difficult to

figure out whats going on

so

if the cluster has inner water

interactions then its more

enthalpically favored and theres less

entropy

and also when we

push that water molecule out

then the gain entropy will not be as not

be as as large

and thats whats shown

here so on the right we have our

a collection of three water molecules

that are in a in a hydrophobic pocket

the green here represents just the fact

its hydrophobic the red are

hydrogen bond donors the blue or

hydrogen bond acceptors theres three

water molecules being around here some

of them theyre interacting with each

with each other this one these two are

this one might not be but theyre in

motion thats represented by the little

arrows

and

if were going to occupy this space with

a drugtype molecule probably one thats

pretty hydrophobic

we want these water molecules to to come

out of there and join in with the bulk

so is this an enthalpic or entropic

process well its a little bit of both

and its a little confusing

exactly

what is

whether its mostly enthalpic or mostly

entropic

so another scenario this is a drug

design scenario its an optimization

scenario in this case we have we have a

molecule already

uh

thats in our in our in our active site

but we want to improve it

and we maybe are nearby a water molecule

that could possibly uh

pop if we can displace it we could

possibly improve our

our our binding of our compound so if

its in a tightly bound water and if

its

the water is tightly bound in a polar

pocket

its going to be really hard to get out

its then pelvically favored where it is

and when we displace it the bulk

its not going to increase in entropy

very much if at all so

the only way were going to

get much

energy out of this is if we can find a

large entropic game

so we might get some by displacing it to

bulk but its

its

its not going to be very much probably

our our most likely

result is that its less enthalpically

favorable to be in bulk

however

if the water is less locked in place

it has a more much more complicated

energetic profile both the enthalpy

entropy could be affected either

favorably unfavorably by its

displacement so

this is a

a little bit of a scenario here this is

a very famous case

of

hiv protease this is the

hiv one protease in the in the unbound

form

there are a number of water molecules in

the active site and theyre labeled

different ways

water 00 sits here between the two

the two aspartates and

its the its at the cadillac center

that one that ones gotta go

uh and bis and the in the primes

over here on the other side

are the same

are

going to stick around because they help

support the the binding of the of the

ligand

these

waters that are labeled in red

letters a b c d up to g those are going

to be bumped out because theyre just

sterically in the wrong place

and lastly theres water 0 which

sticks down here at the bottom its got

a blue label on it because

it actually is very stable where it is

so the first class of molecules that

were developed

looks something like this

and they as i said bumped out water 00

bumped out all the

the lettered

the lettered

water molecules

which were sterically in in the way

it kept water 0 because thats very

energy energetically favored and the

and best

water molecules

were

support the interaction because they

bridge between the

the ligand

molecule and the protein

now a lot of people said okay lets see

if we can make 0 disappear

because if we make 0 disappear we will

get energy because of its because of its

entropic release

and we will gain the energy of making

new hydrogen bonds so the second

generation of hiv and protease

inhibitors

did just that they found a way to

replace the water with functional groups

that

were able to to

to bind to the isoleucines

and

they were found to be

somewhat

more effective than the than the ones

that didnt have that

however if you counted the energy of the

water

in the previous case where there was

water 0 there it worked out to

essentially a wash

okay talk about one more thing here

enthalpy entropy compensation

and this is this is one thing that makes

drug design very frustrating

is that water often plays a role in

enthalpy

uh enthalpy entropy compensation by

mitigating the effects of structural

changes that were designed to optimize

drugs binding so in the top case here we

have a molecule its got a methyl group

here

and we

are trying to

uh optimize it and we say when we see

that if we can put a uh

something over here that will bind with

this polar region in red here

maybe an oh group we would

improve the binding of this molecule and

we would have a

a

a better binding constant for our

molecule

so

if we do that

as shown here

look what happens well we do get better

enthalpy for sure because

it is

making a new hydrogen bond that we

didnt have before but on the other hand

were losing some entropy internal

entropy we had we had before so this is

more this is more in tropically favored

this is more metallically favored so a

change

gives us

the desired uh increase

decrease in enthalpy but also reduces

entropy so that the change in free

energy delta delta g is approximately

zero

now in the bottom case

uh

the decision was made okay lets see if

we can up

lets see if we can we got a hydrophobic

group here if we can maybe push this

water molecule out

we would get a we would get a better

binding uh

a better binding compound uh

like this

so we

added a methyl group

here

and pushed this water molecule out here

and we were interested to see if this

would be a better would be a better uh

a better a better energetic profile well

we did

unfortunately

increase

the

the

the enthalpy

but we also increased the entropy and

again our delta delta g is around zero

so the enthalpy entropy compensation is

an important factor that needs to be

considered in

drug design and theres many frustrating

examples in literature on that

so im going to finish up here

these are the main points uh

out of this talk uh

i think i

tried to show you that the computational

tools are powerful powerful adjuncts to

experiment but their underlying

principles and limitations must be

understood if the users can be truly

effective using them too many people

treat computational tools as black boxes

they dont go

to the trouble of understanding how they

work so i tried to go through a little

bit about the

the

physics and chemistry behind these

things i certainly didnt have enough

time to go into a lot of detail on that

but

maybe

uh you have an appreciation for

how it works and youll be cautious one

of the things that we emphasize in our

in our courses that we teach to our

students at vcu

is that

you have to always question

computational results

the computational results do not

ever give you an answer

but their supporting evidence

can be useful and sometimes the most

useful thing comes out of that it gives

you visualization you can actually see

the sort of things that you are

are trying to achieve and see why or why

not it might work or might not work

in my view the best computational

experiments are those that suggest great

wet experiments

if we

through our computational work give

someone an idea to create

an experiment that

answers a question because they

saw through visualization or whatever

what might happen thats a real success

for us

i tried to point out that qsar is the

qsr is the keystone principle that has

enabled

many more modern tools like virtual

screening docking scoring etc because it

its all based on the simple principle

that

the activity is a function of structure

always when youre looking at

any kind of thing where youre

calculating

scores or calculating energies gibbs

free energy is the currency you need to

pay attention it gives free energy pay

attention to both the enthalpy and the

entropy

the accurate estimates of delta g

require consideration of everything

especially the water molecules and the

roles

so id like to thank nih for inviting me

to give this presentation i

invite uh

students who take this course to write

me if they have questions or contact the

the

the coordinator of the of the course and

uh

and thank you for your attention

good night