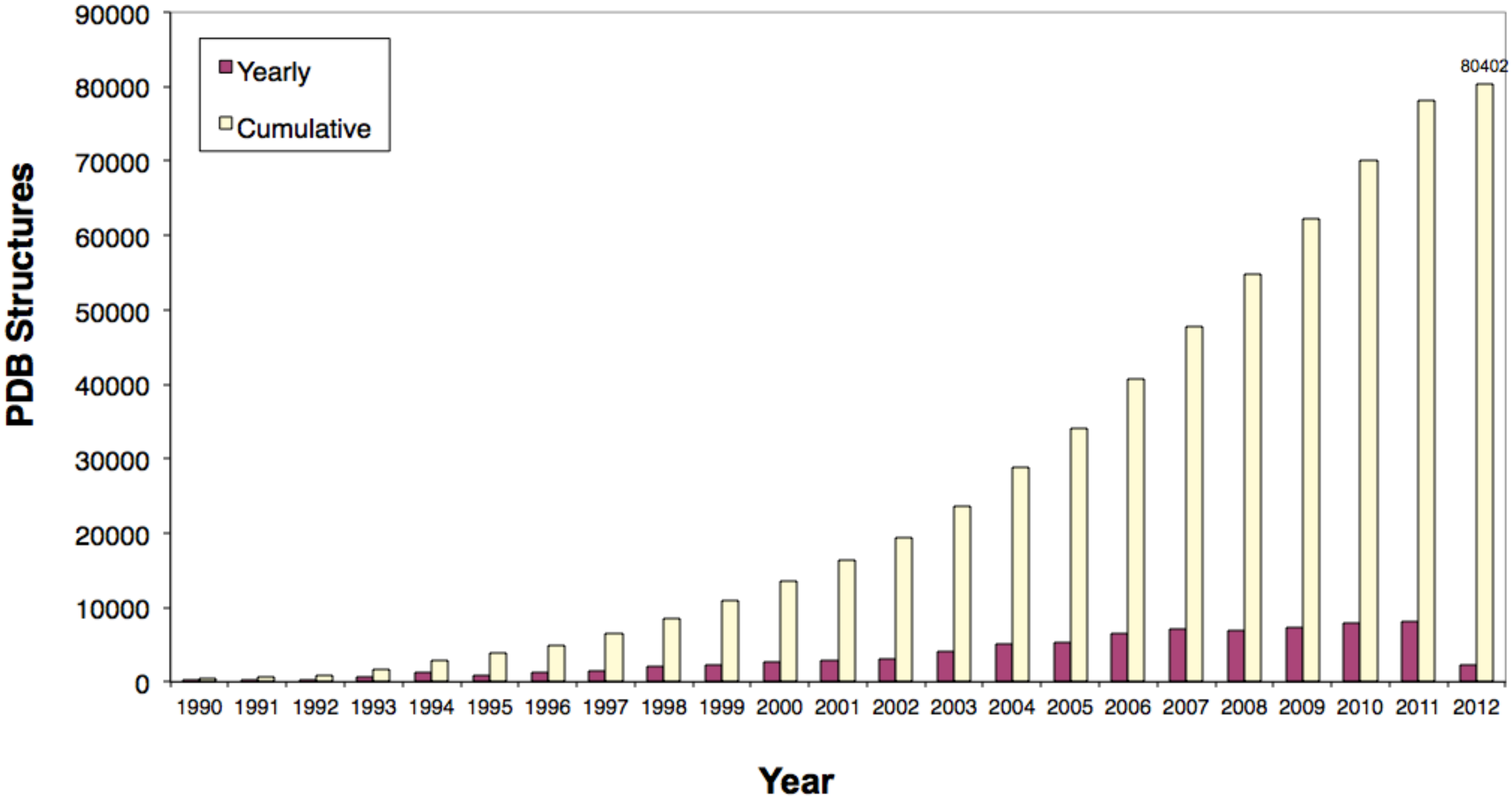
Using Site Similarity to Generate New Matter Ideas

Dr. Steven Muskal

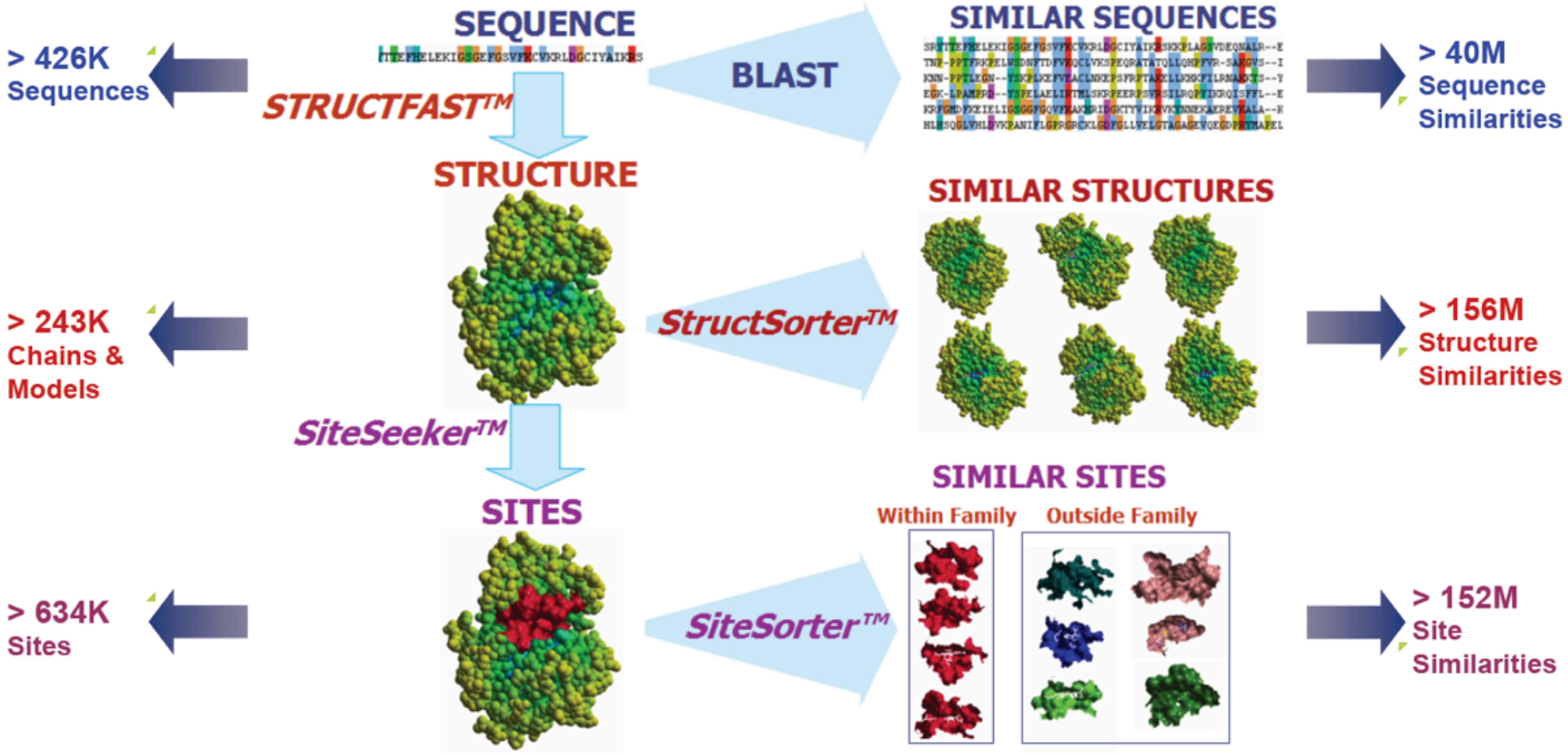
Chief Executive Officer
Eidogen-Sertanty, Inc
smuskal@eidogen-sertanty.com

Protein Structure Growth Continues

> 80K structures/co-complexes “templates” (Apr-2012)
> 650 new templates/month (>150/week)



Target Informatics Platform (TIP)



SEQUENCE

STRUCTFAST™

STRUCTURE

SiteSeeker™

SITES

BLAST

StructSorter™

SiteSorter™

SIMILAR SEQUENCES

SIMILAR STRUCTURES

SIMILAR SITES

Within Family

Outside Family

> 426K Sequences

> 243K Chains & Models

> 634K Sites

> 40M Sequence Similarities

> 156M Structure Similarities

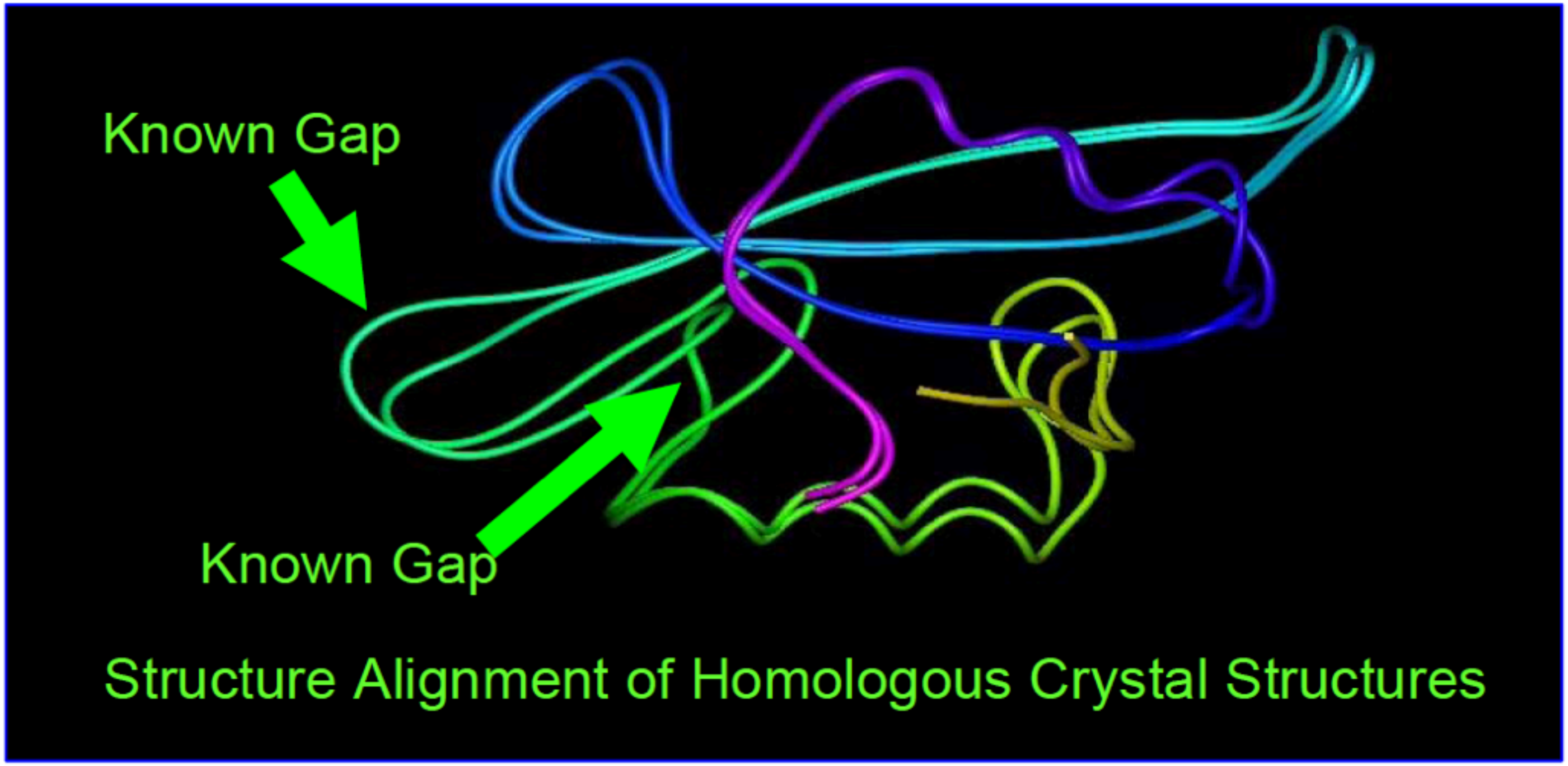
> 152M Site Similarities

- Interrogating the druggable genome with structural informatics **MolecularDiversity (2006)**
- STRUCTFAST: Protein Sequence Remote Homology Detection and Alignment Using Novel Dynamic Programming and Profile-Profile Scoring **Proteins. 2006 64:960-967**
- StructSorter: A Method for Continuously Updating a Comprehensive Protein Structure Alignment Database **J. Chem. Inf. Model. 2006, 46, 1871-1876**
- Convergent Island Statistics: A fast method for determining local alignment score significance. **Bioinformatics, 2005, 21, 2827-2831.**

STRUCTFAST

STructure Realization Utilizing Cogent Tips From Aligned Structural Templates

Basic Principle: Gaps known to exist should not be strongly penalized.



Structure Alignment of Homologous Crystal Structures

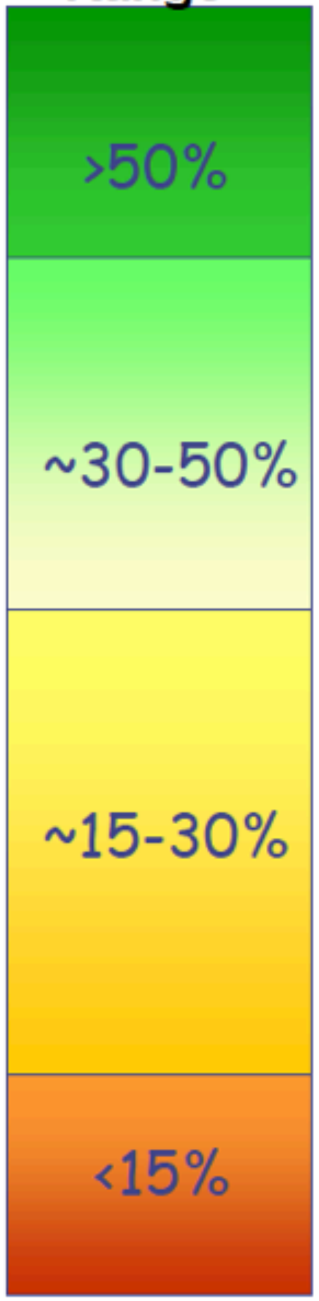
Leverages experimental structure and structural alignment data to create better alignments

- STRUCTFAST: Protein Sequence Remote Homology Detection and Alignment Using Novel Dynamic Programming and Profile-Profile Scoring **Proteins. 2006 64:960-967**
- Convergent Island Statistics: A fast method for determining local alignment score significance. **Bioinformatics, 2005, 21, 2827-2831.**

Modeling Algorithm Comparison

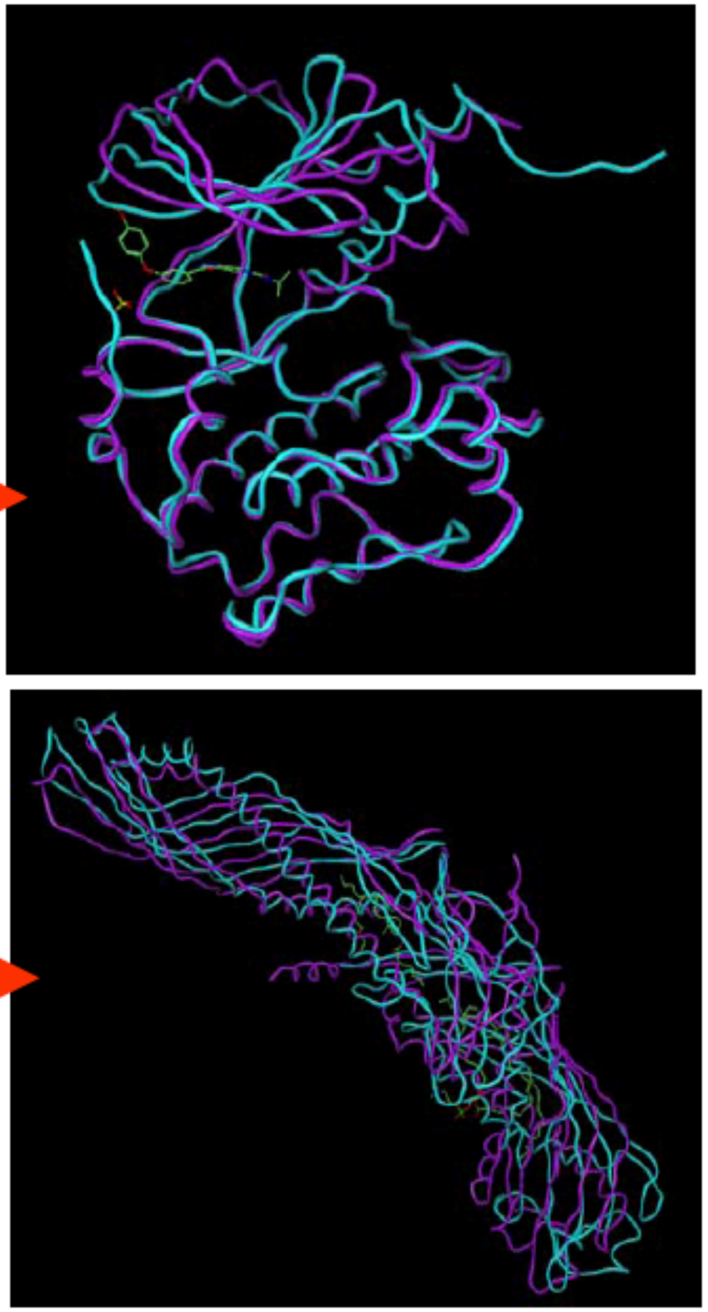
Alignment	Scoring Methods	Gap Treatment	Examples
Sequence-Sequence	BLOSUM PAM GONET	Length Proportional Affine	BLAST FASTA Smith-Waterman Needleman-Wunsch
Sequence-Profile	PSSM HMM	Affine Position-Specific	PSI-Blast HMMer
Sequence-Structure	Threading potential	Affine Position-Specific	Raptor GenThreader
Profile-Profile	Dot-product Log Average **Analytic Statistics**	Position-Specific **Structural Family-based**	3D-PSSM FFAS **STRUCTFAST**

Template Homology Range



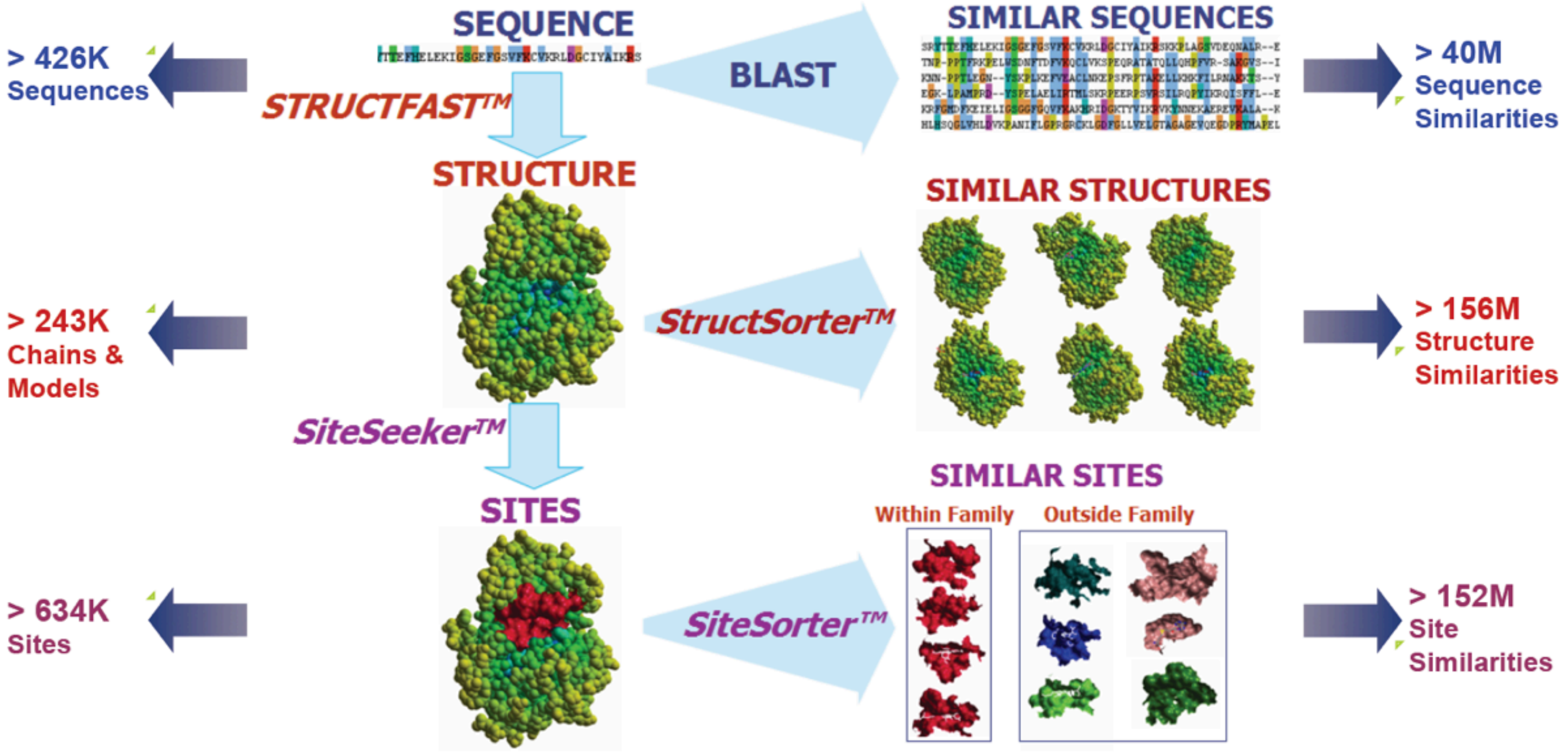
Comparison to a Client's Crystal Structures

Protein	# of residues	In house resolution	template resolution	PDB ID	% ID	# atoms fit	rmsd
Family 1	288	1.6 A	1.8 A		75%	976	0.88A
Family 2	347	1.7 A	2.3 A		87%	1148	1.23A
Family 3	285	2.5 A	2.4 A		34%	780	1.48A
Family 4	941	2.1 A	1.7 A		20%	1824	6.66A
Family 5	313	2.6 A	2.0 A		57%	1048	1.23 A
Family 6	442	2.05 A	2.1 A		95%	1376	0.87 A



Purple –client's in house
Cyan – homology model

Target Informatics Platform (TIP)



- Interrogating the druggable genome with structural informatics **MolecularDiversity (2006)**
- STRUCTFAST: Protein Sequence Remote Homology Detection and Alignment Using Novel Dynamic Programming and Profile-Profile Scoring **Proteins. 2006 64:960-967**
- StructSorter: A Method for Continuously Updating a Comprehensive Protein Structure Alignment Database **J. Chem. Inf. Model. 2006, 46, 1871-1876**
- Convergent Island Statistics: A fast method for determining local alignment score significance. **Bioinformatics, 2005, 21, 2827-2831.**

SiteSeeker

Geometric Site-Finding Algorithms Find Many Pockets

But they don't know which pockets are important!

Evolutionary Trace Approach

Can't clearly define site boundary

Not all conserved residues are functionally relevant

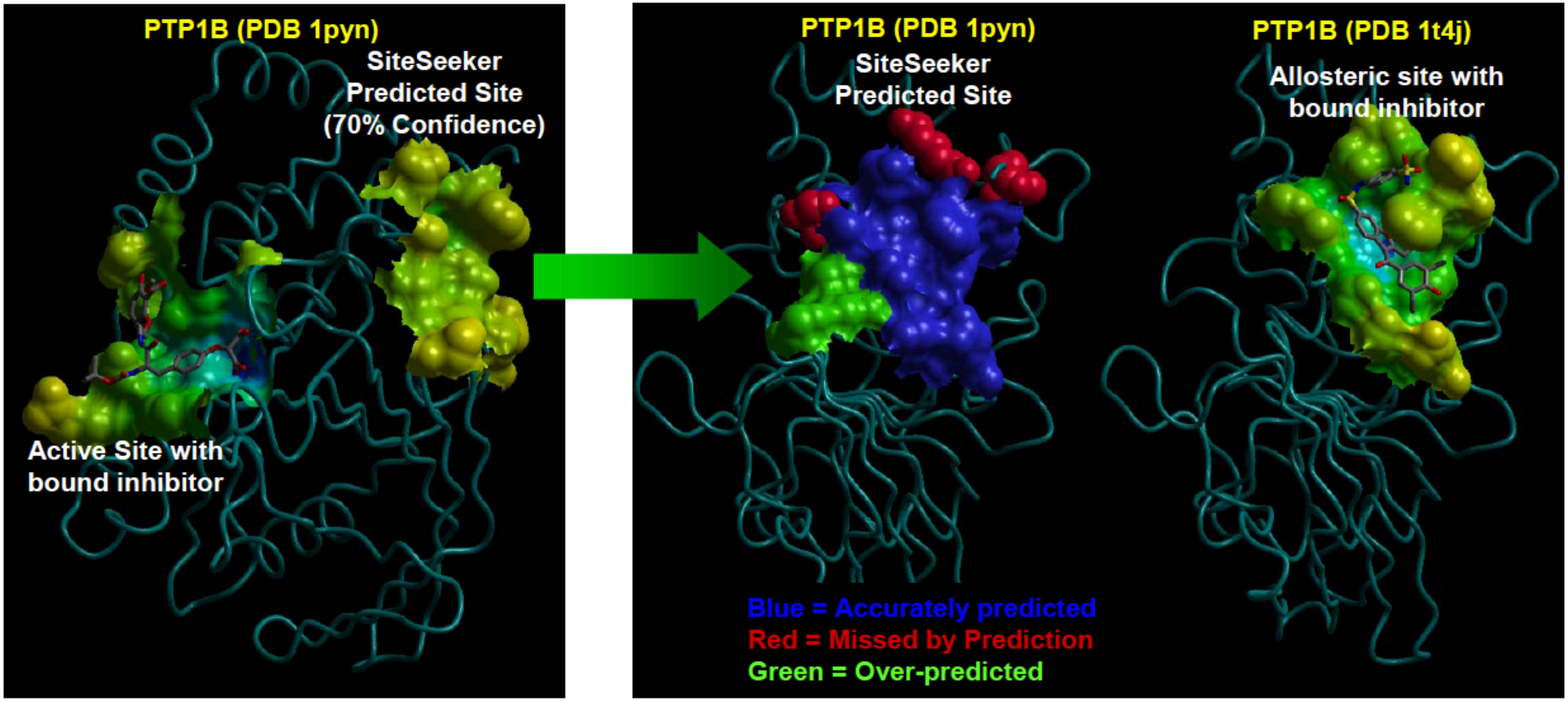
SiteSeeker combines *both* methods

Reliability & Confidence

We use proteins with apo- & co-crystal structures in the PDB to test the accuracy & reliability of method

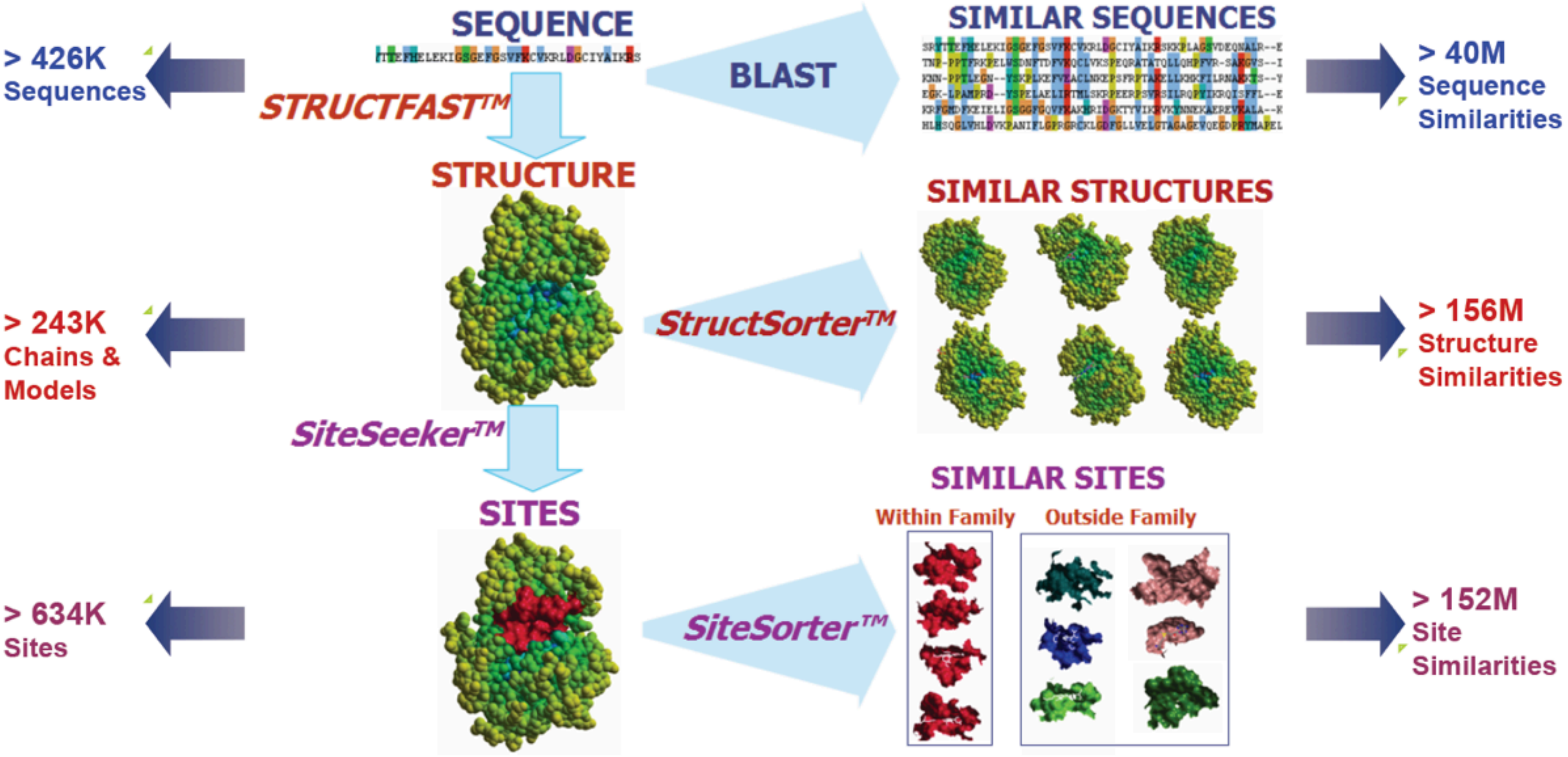
Allows us to map *SiteSeeker* score to predict confidence!
(e.g. At this *SiteSeeker* score, 80% are "real" co-crystal sites)
→ Sites with <60% confidence are not stored in TIP

SiteSeeker Example



All structures in TIP are annotated with known and predicted binding sites, along with **confidence** levels for each annotation

Target Informatics Platform (TIP)



> 426K Sequences

SEQUENCE

STRUCTFAST™

BLAST

SIMILAR SEQUENCES

> 40M Sequence Similarities

> 243K Chains & Models

STRUCTURE

StructSorter™

SIMILAR STRUCTURES

> 156M Structure Similarities

SiteSeeker™

> 634K Sites

SITES

SiteSorter™

SIMILAR SITES

Within Family **Outside Family**

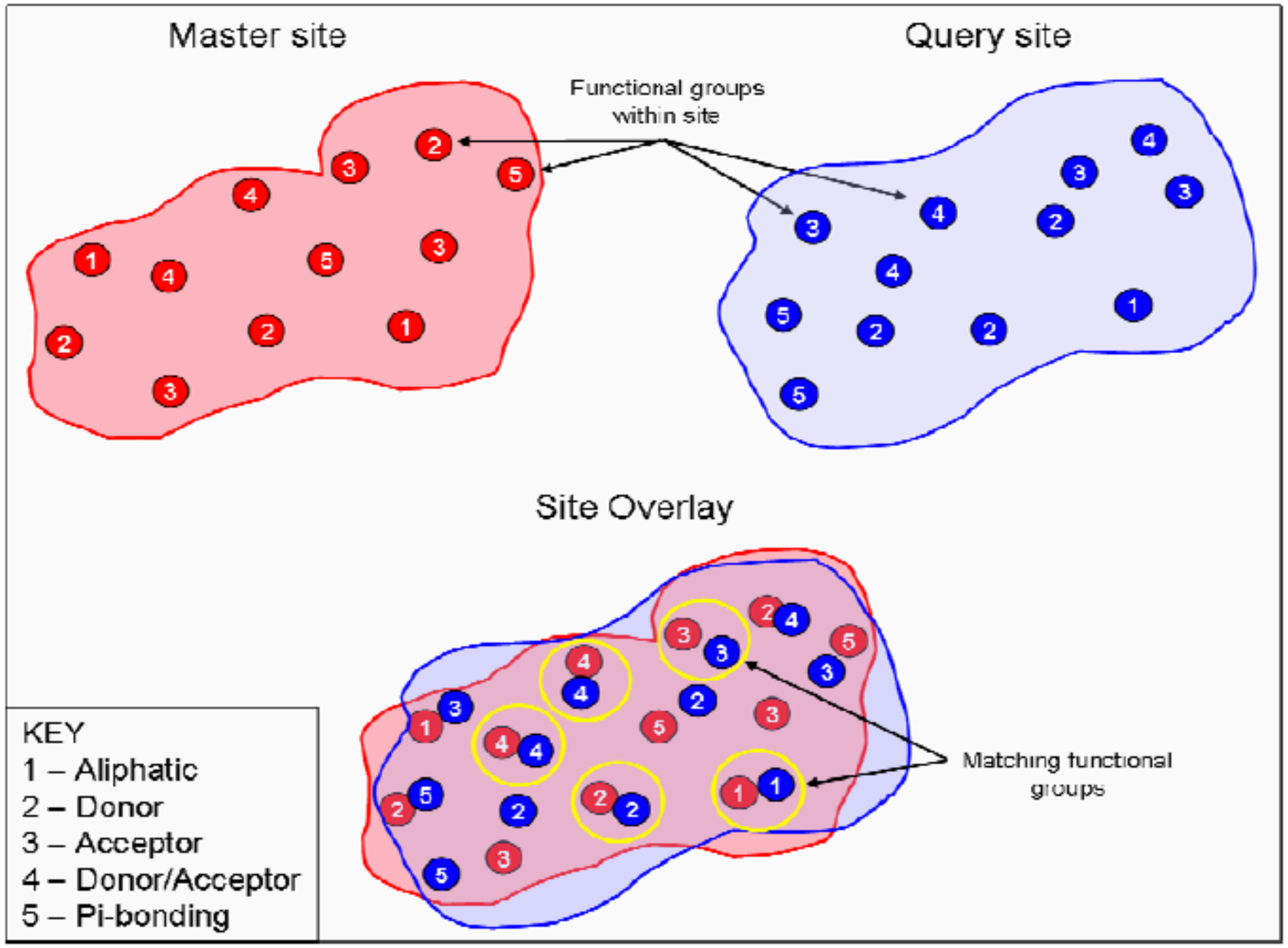
> 152M Site Similarities

- Interrogating the druggable genome with structural informatics **MolecularDiversity (2006)**
- STRUCTFAST: Protein Sequence Remote Homology Detection and Alignment Using Novel Dynamic Programming and Profile-Profile Scoring **Proteins. 2006 64:960-967**
- StructSorter: A Method for Continuously Updating a Comprehensive Protein Structure Alignment Database **J. Chem. Inf. Model. 2006, 46, 1871-1876**
- Convergent Island Statistics: A fast method for determining local alignment score significance. **Bioinformatics, 2005, 21, 2827-2831.**

SiteSorter

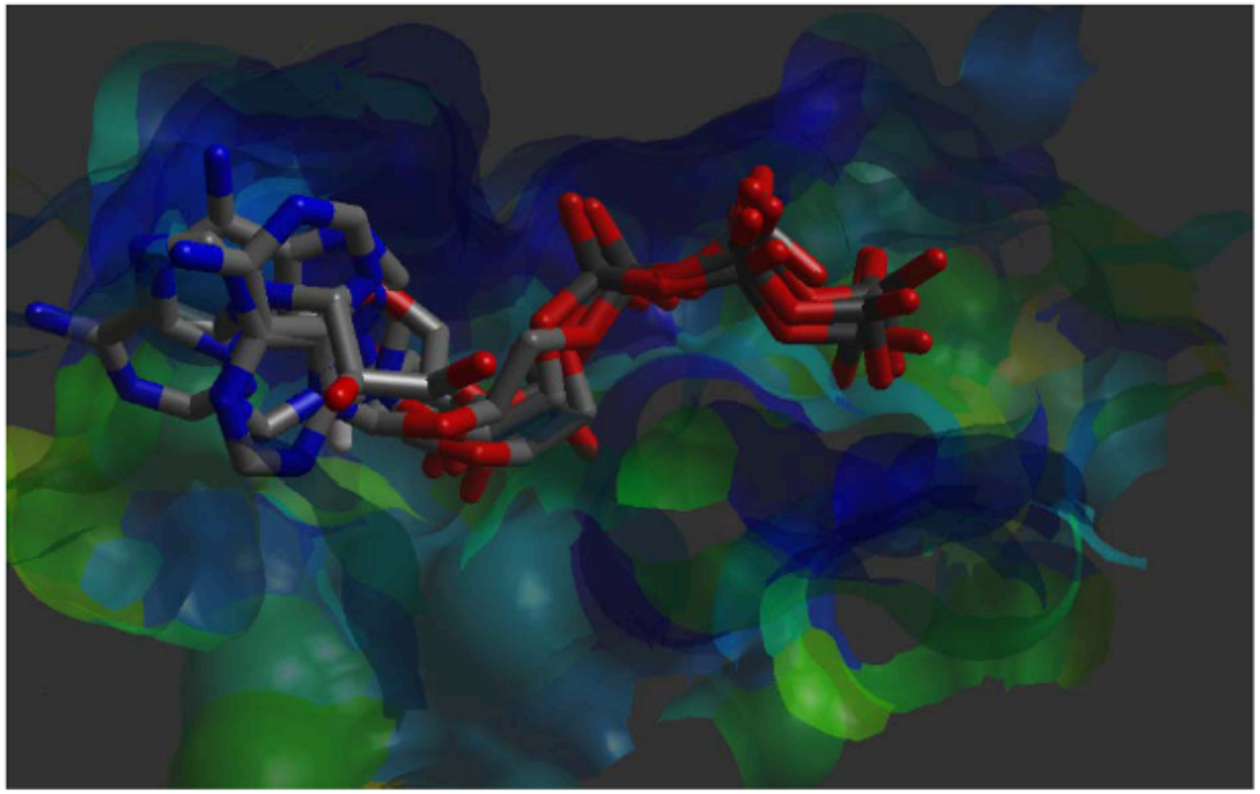
Weighted Clique Detection Algorithm

Importance of Points (Weights) Related to their Similarity

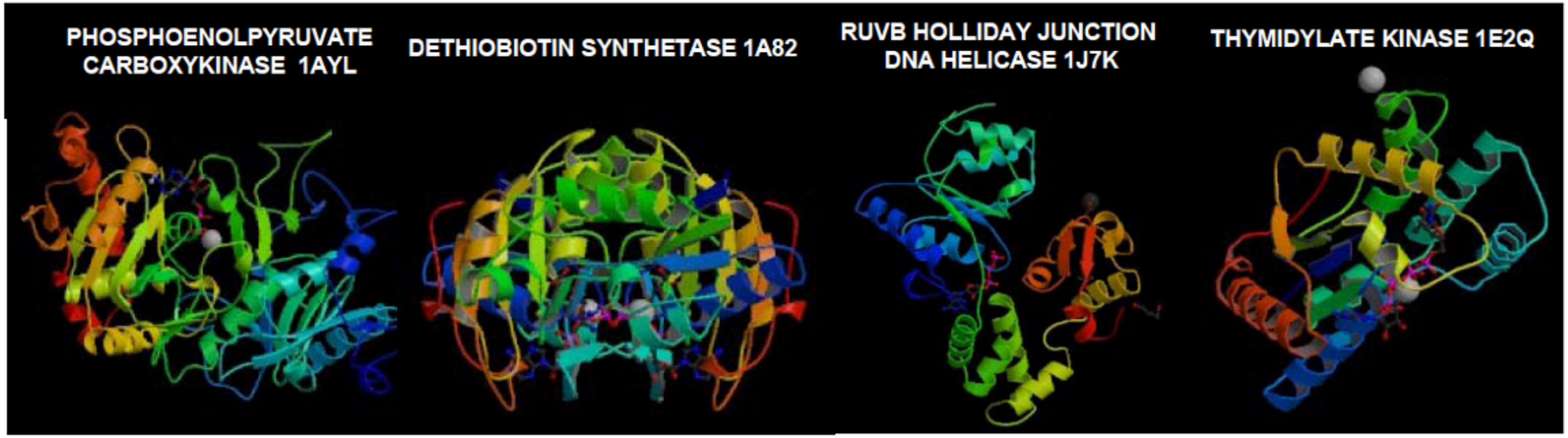


Surface Atoms Assigned One of 5 Different Chemical Characters
Matching points increase the *SiteSorter* similarity score

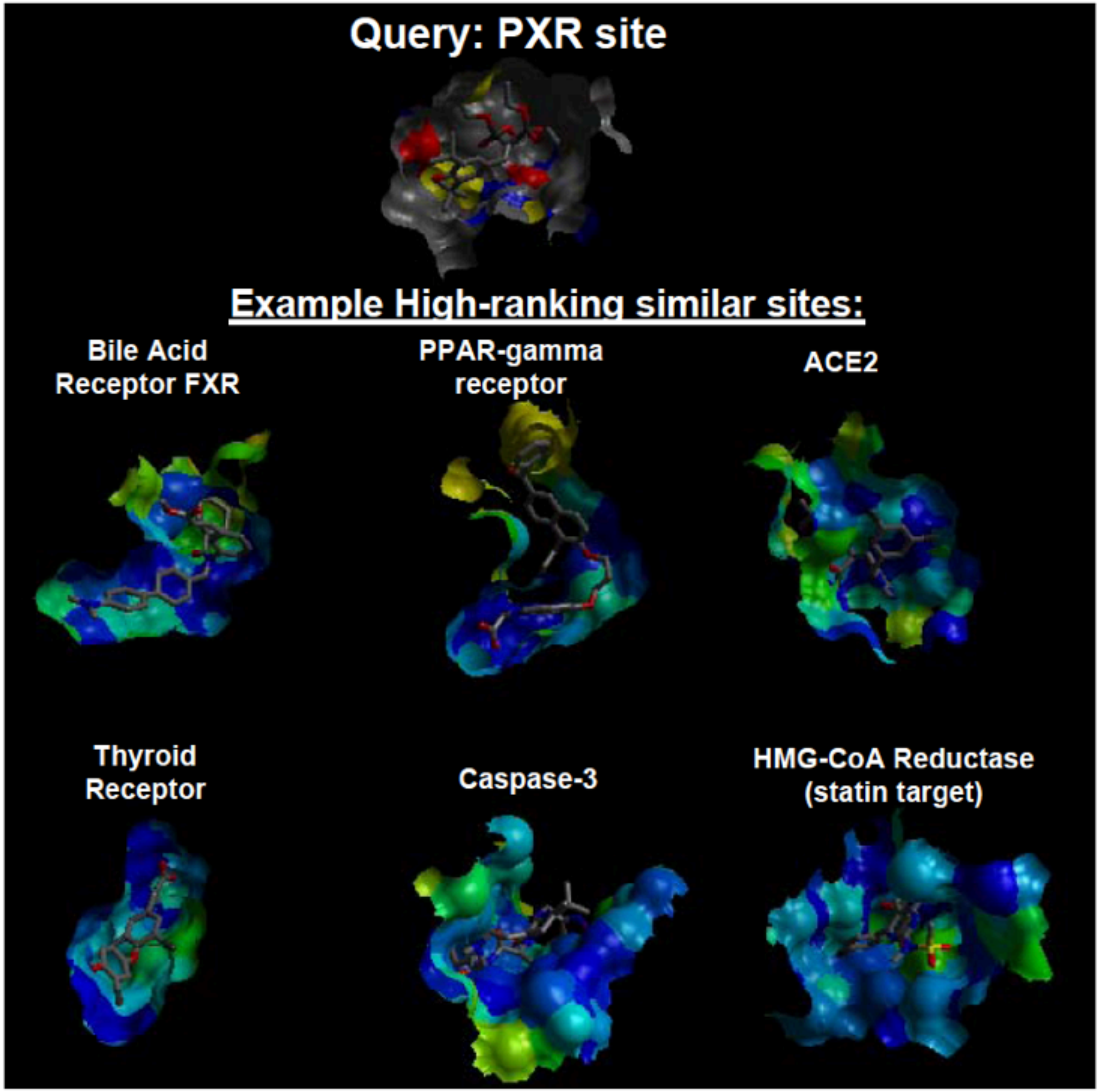
SiteSorter Example



Overlay of ATP binding sites from completely different folds

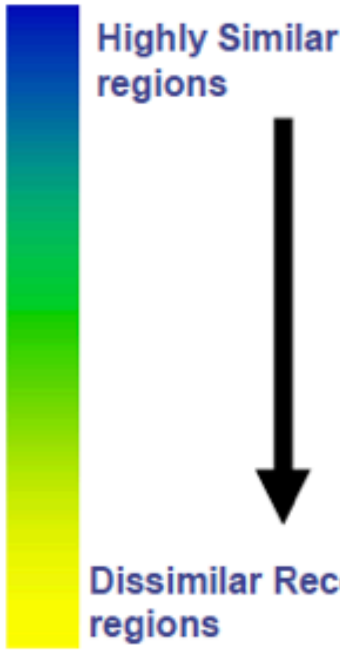


Nature Exploits Site-Similarity



Pregnane X-receptor – PXR ("sensor)" →CYP3A4 ("executioner")
<u>PXR Binds > 50% drugs</u>
Including some bile acids, statins, herbal components, a selection of HIV protease inhibitors, calcium channel modulators, numerous steroids, plasticizers and monomers, organochlorine pesticides, a peroxisome proliferator-activated receptor-äantagonist, xenobiotics and endobiotics…

Site Similarity Coloring



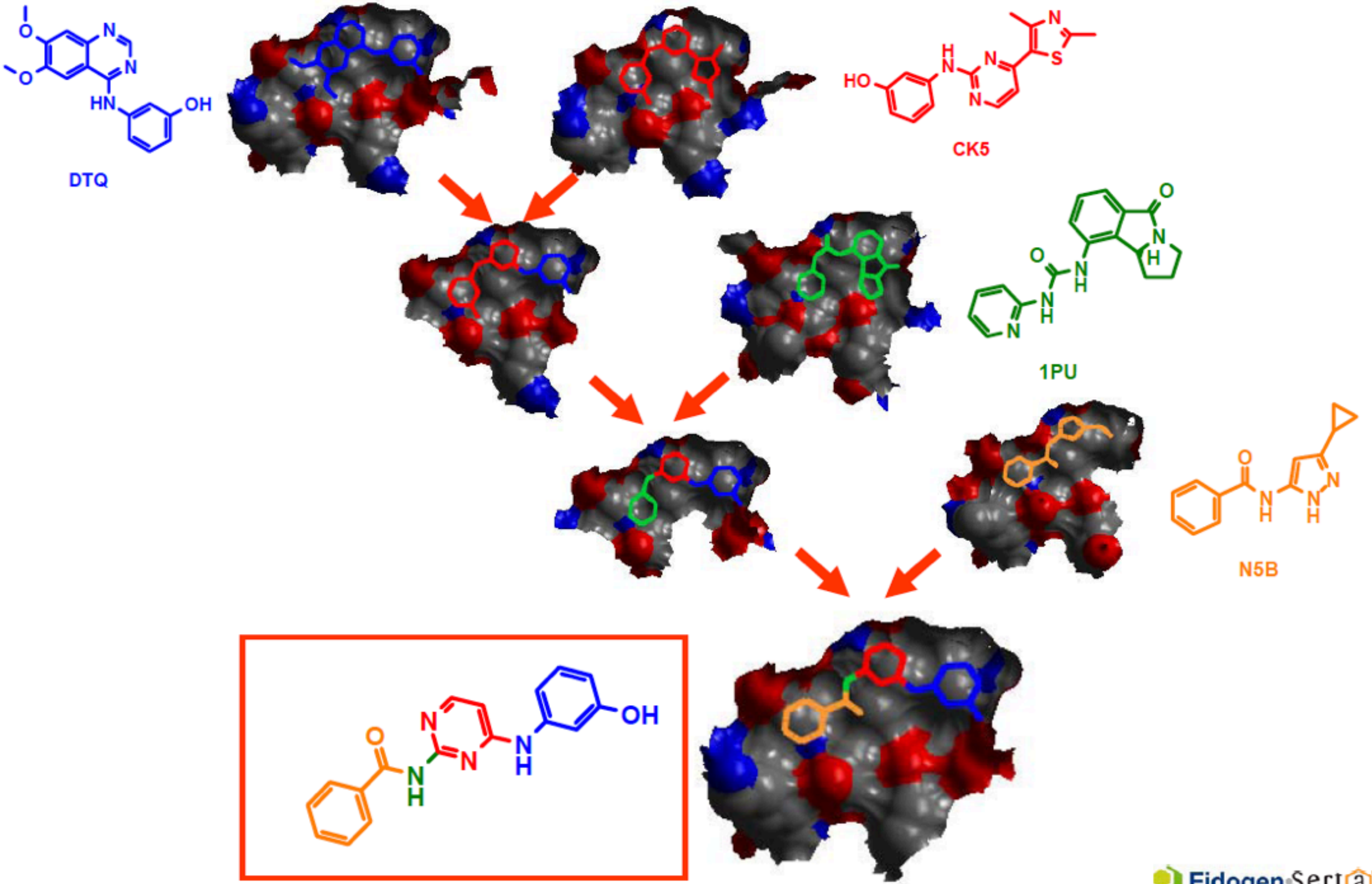
Highly Similar Receptor regions

Dissimilar Receptor regions

Eidogen-Sertanty

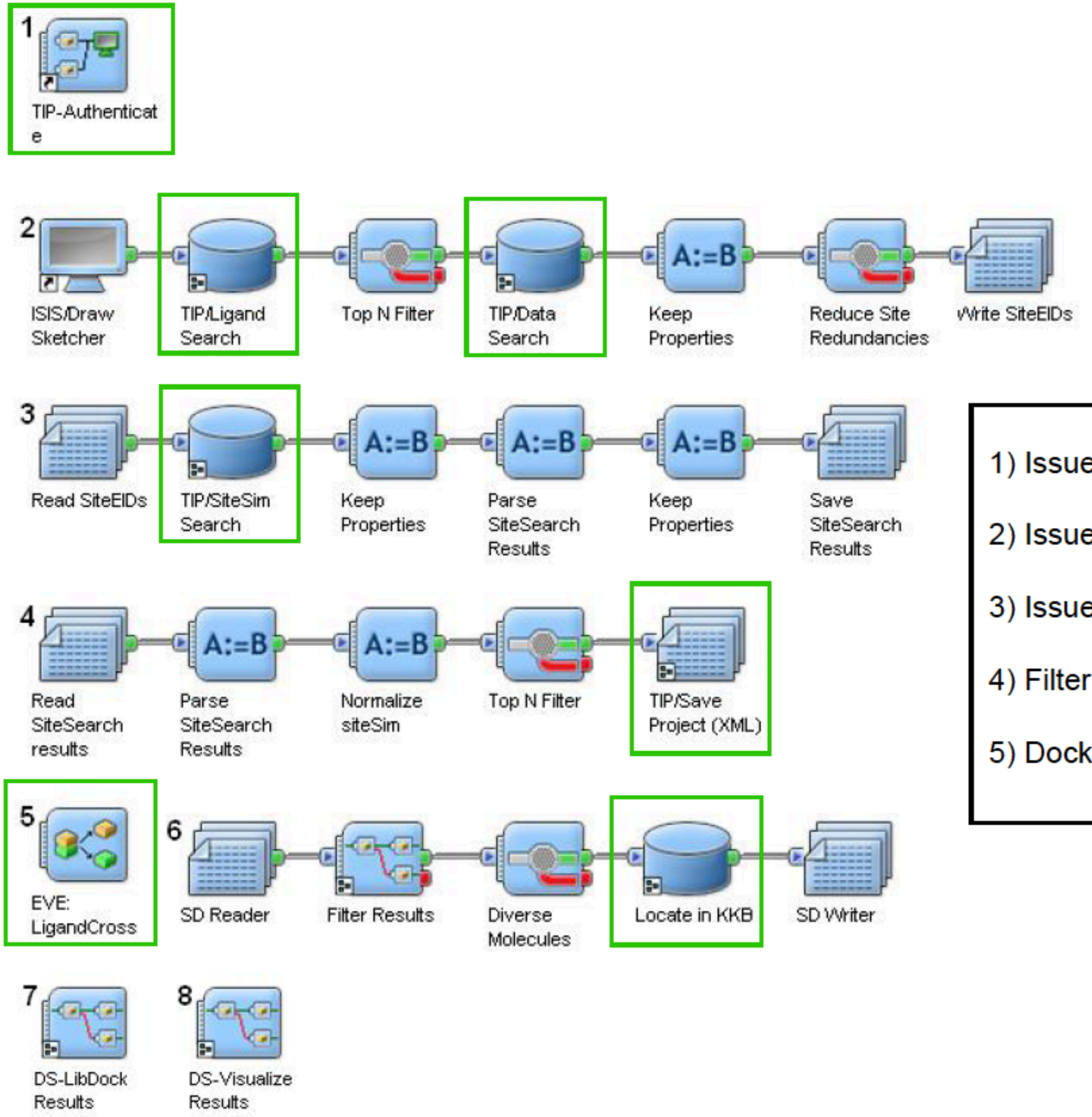
LigandCross: Shuffling Ligand Functionality

Similar to Vertex's BREED: J. Med. Chem. **47**, 2768 (2004)



Eidogen-Sertanty

LigandCross via PipelinePilot



1) Issue TIP/LigandSearch

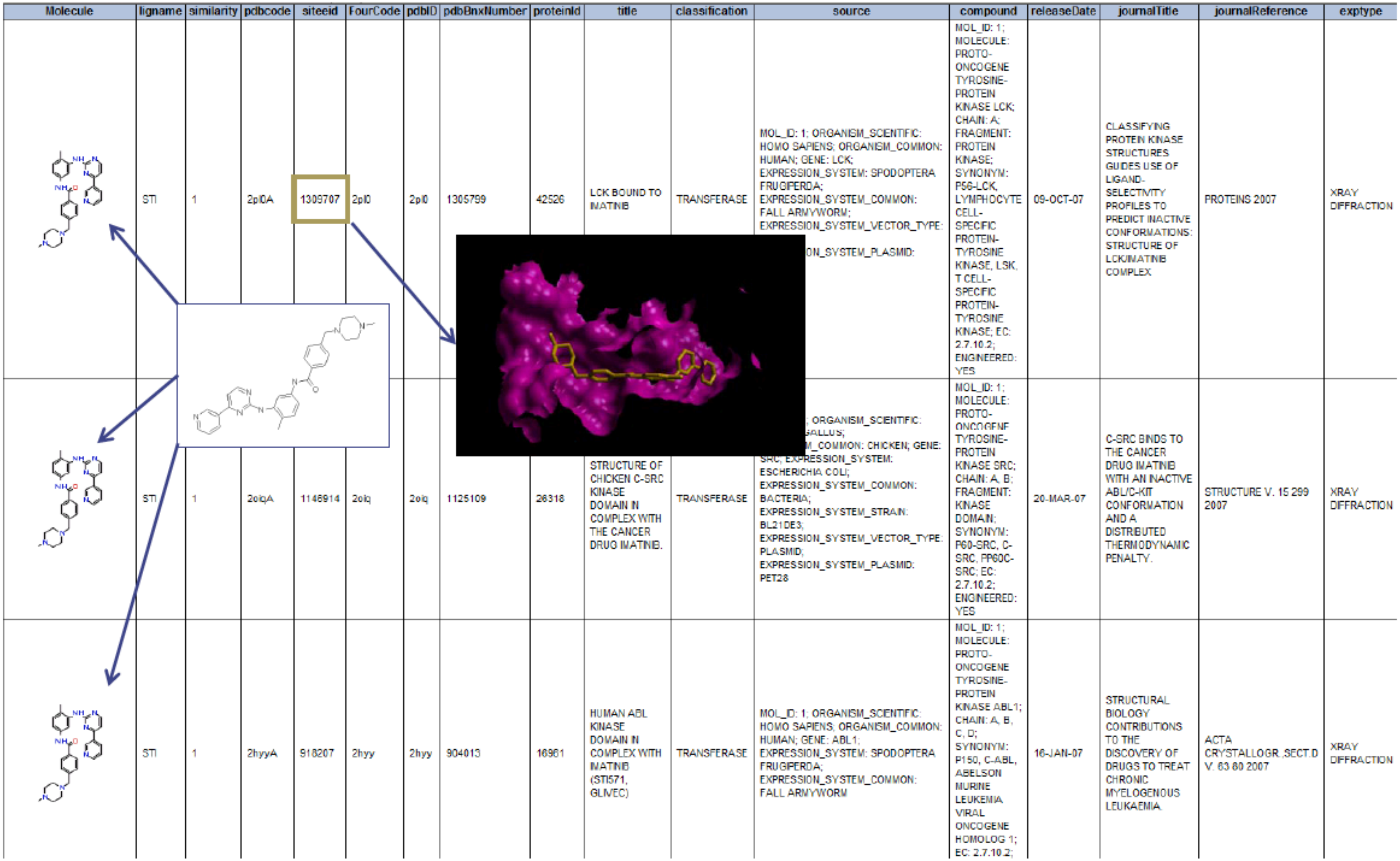
2) Issue TIP/SiteSimSearch

3) Issue LigandCross

4) Filter and locate results in KKB

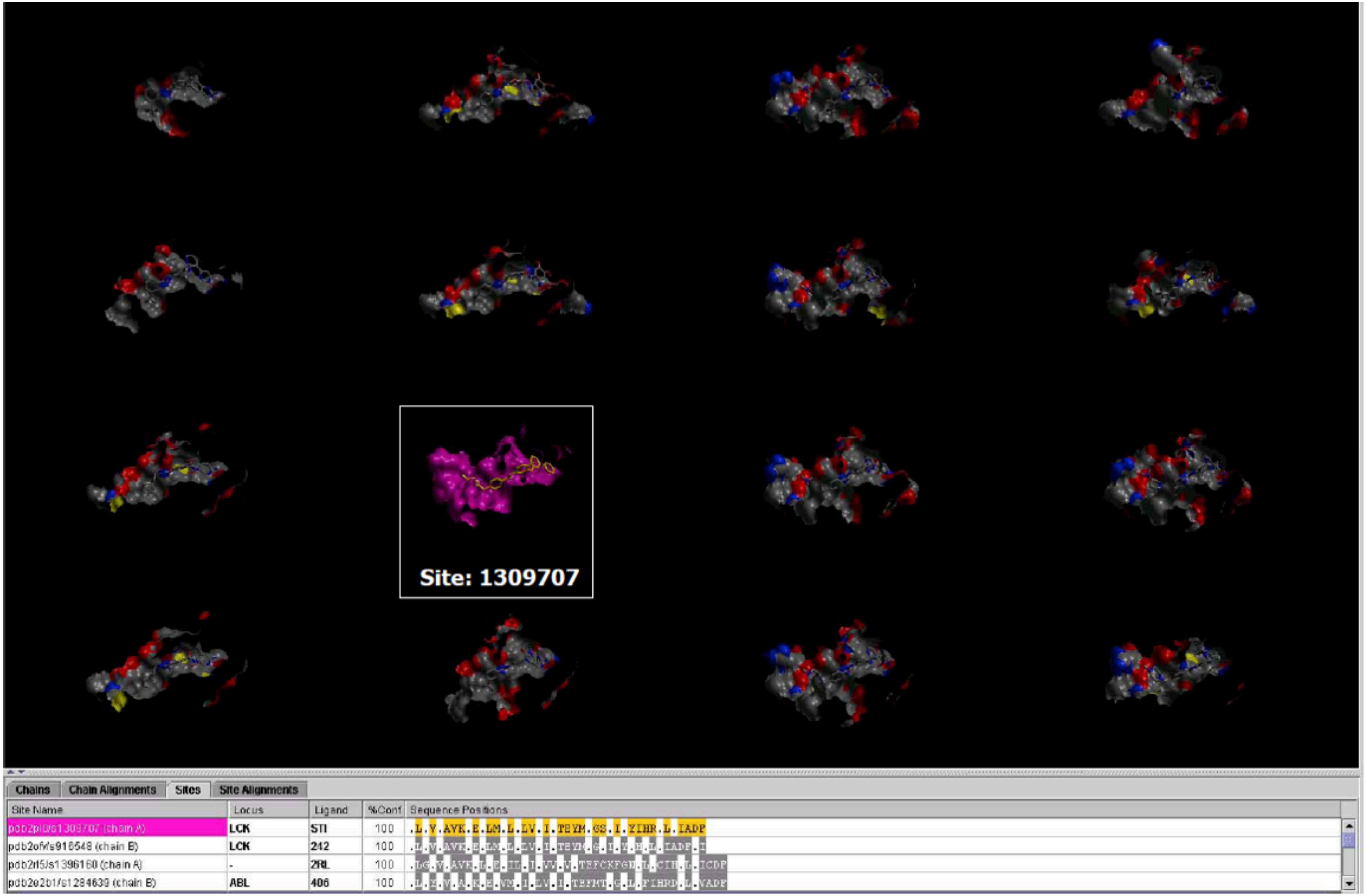
5) Dock and visualize results

Step 1: Find Co-complexes and Sites



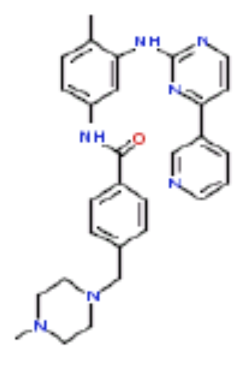
Molecule	ligname	similarity	pdbcode	siteid	FourCode	pdbID	pdbBnxNumber	proteinId	title	classification	source	compound	releaseDate	journalTitle	journalReference	exptype
	STI	1	2pl0A	1309707	2pl0	2pl0	1305799	42526	LCK BOUND TO IMATINIB	TRANSFERASE	MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; GENE: LCK; EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA; EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM; EXPRESSION_SYSTEM_VECTOR_TYPE: ON_SYSTEM_PLASMID:	MOL_ID: 1; MOLECULE: PROTO-ONCOGENE TYROSINE-PROTEIN KINASE LCK; CHAIN: A; FRAGMENT: PROTEIN KINASE; SYNONYM: P56-LCK, LYMPHOCYTE CELL-SPECIFIC PROTEIN-TYROSINE KINASE, LSK, T CELL-SPECIFIC PROTEIN-TYROSINE KINASE; EC: 2.7.10.2; ENGINEERED: YES	09-OCT-07	CLASSIFYING PROTEIN KINASE STRUCTURES GUIDES USE OF LIGAND-SELECTIVITY PROFILES TO PREDICT INACTIVE CONFORMATIONS: STRUCTURE OF LCK/IMATINIB COMPLEX	PROTEINS 2007	XRAY DIFFRACTION
	STI	1	2oiqA	1146914	2oiq	2oiq	1125109	26318	STRUCTURE OF CHICKEN C-SRC KINASE DOMAIN IN COMPLEX WITH THE CANCER DRUG IMATINIB.	TRANSFERASE	; ORGANISM_SCIENTIFIC: GALLUS; M_COMMON: CHICKEN; GENE: SRC; EXPRESSION_SYSTEM: ESCHERICHIA COLI; EXPRESSION_SYSTEM_COMMON: BACTERIA; EXPRESSION_SYSTEM_STRAIN: BL21DE3; EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID; EXPRESSION_SYSTEM_PLASMID: PET28	MOL_ID: 1; MOLECULE: PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC; CHAIN: A, B; FRAGMENT: KINASE DOMAIN; SYNONYM: P60-SRC, C-SRC, PP60C-SRC; EC: 2.7.10.2; ENGINEERED: YES	20-MAR-07	C-SRC BINDS TO THE CANCER DRUG IMATINIB WITH AN INACTIVE ABL/C-KIT CONFORMATION AND A DISTRIBUTED THERMODYNAMIC PENALTY.	STRUCTURE V. 15 299 2007	XRAY DIFFRACTION
	STI	1	2hyyA	918207	2hyy	2hyy	904013	16961	HUMAN ABL KINASE DOMAIN IN COMPLEX WITH IMATINIB (STI571, GLIVEC)	TRANSFERASE	MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; GENE: ABL1; EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA; EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM	MOL_ID: 1; MOLECULE: PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL1; CHAIN: A, B, C, D; SYNONYM: P150, C-ABL, ABELSON MURINE LEUKEMIA VIRAL ONCOGENE HOMOLOG 1; EC: 2.7.10.2;	16-JAN-07	STRUCTURAL BIOLOGY CONTRIBUTIONS TO THE DISCOVERY OF DRUGS TO TREAT CHRONIC MYELOGENOUS LEUKAEMIA	ACTA CRYSTALLOGR.,SECT.D V. 63 80 2007	XRAY DIFFRACTION

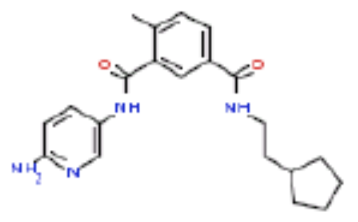
Step 2: Find Other Sites via SiteSimilarity

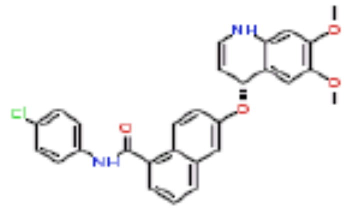


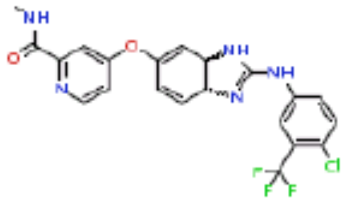
Site Name	Locus	Ligand	%Conf	Sequence Positions
pdb2pl0/s1309707 (chain A)	LCK	STI	100	.L.V.AVK.E.LM.L.LV.I.TEYM.GS.I.YIHR.L.IADP
pdb2ofv/s916548 (chain B)	LCK	242	100	.L.V.AVK.E.LM.L.LV.I.TEYM.G.I.Y.H.L.IADF.I
pdb2rl5/s1396160 (chain A)	-	2RL	100	.LG.V.AVK.L.E.IL.I.VV.V.TEFCKFGN.L.CIH.L.ICDF
pdb2e2b/s1284639 (chain B)	ABL	406	100	.L.Y.V.A.K.E.VM.I.LV.I.TEFMT.G.L.FIHRD.L.VADF

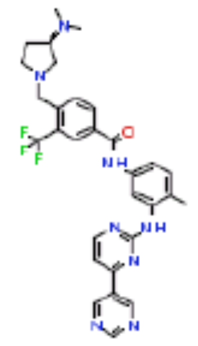
Example Ligands from Similar Sites

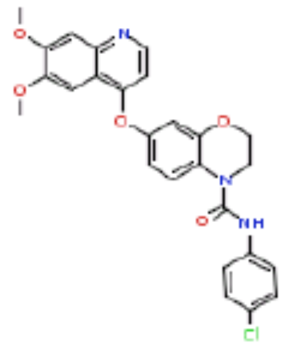

STI

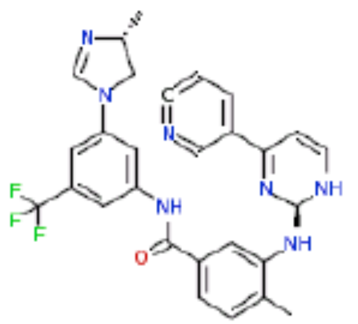

C92

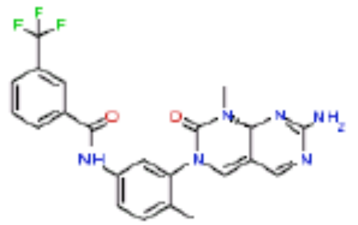

900

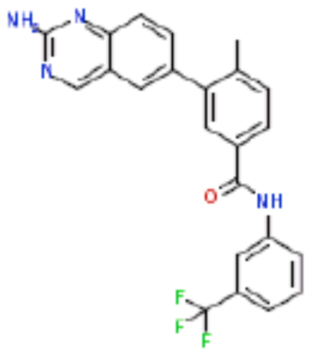

276

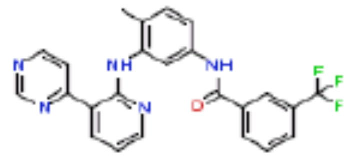

406

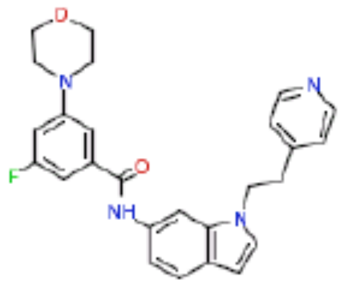

2RL

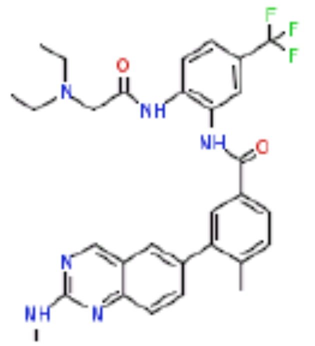

NIL

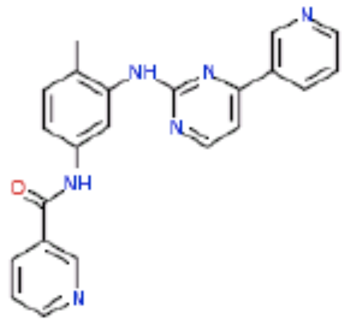

7MP

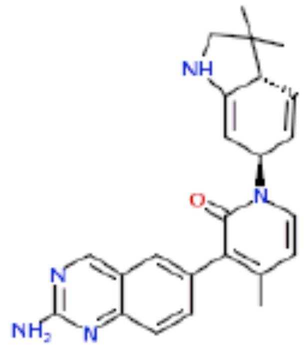

242

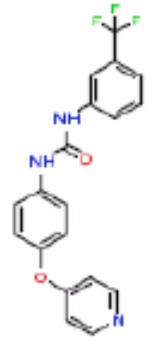

MUH

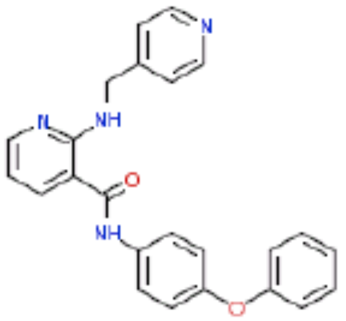

WBT

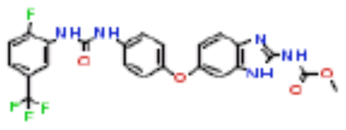

1N8

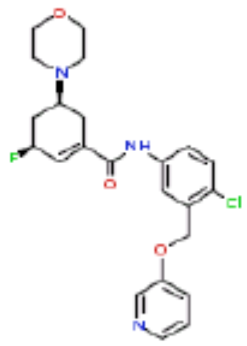

PRC

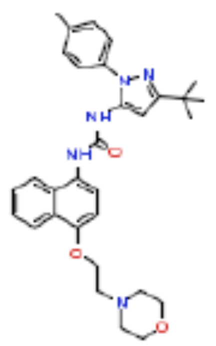

C19

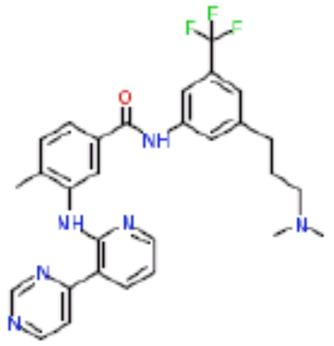

KIN


608

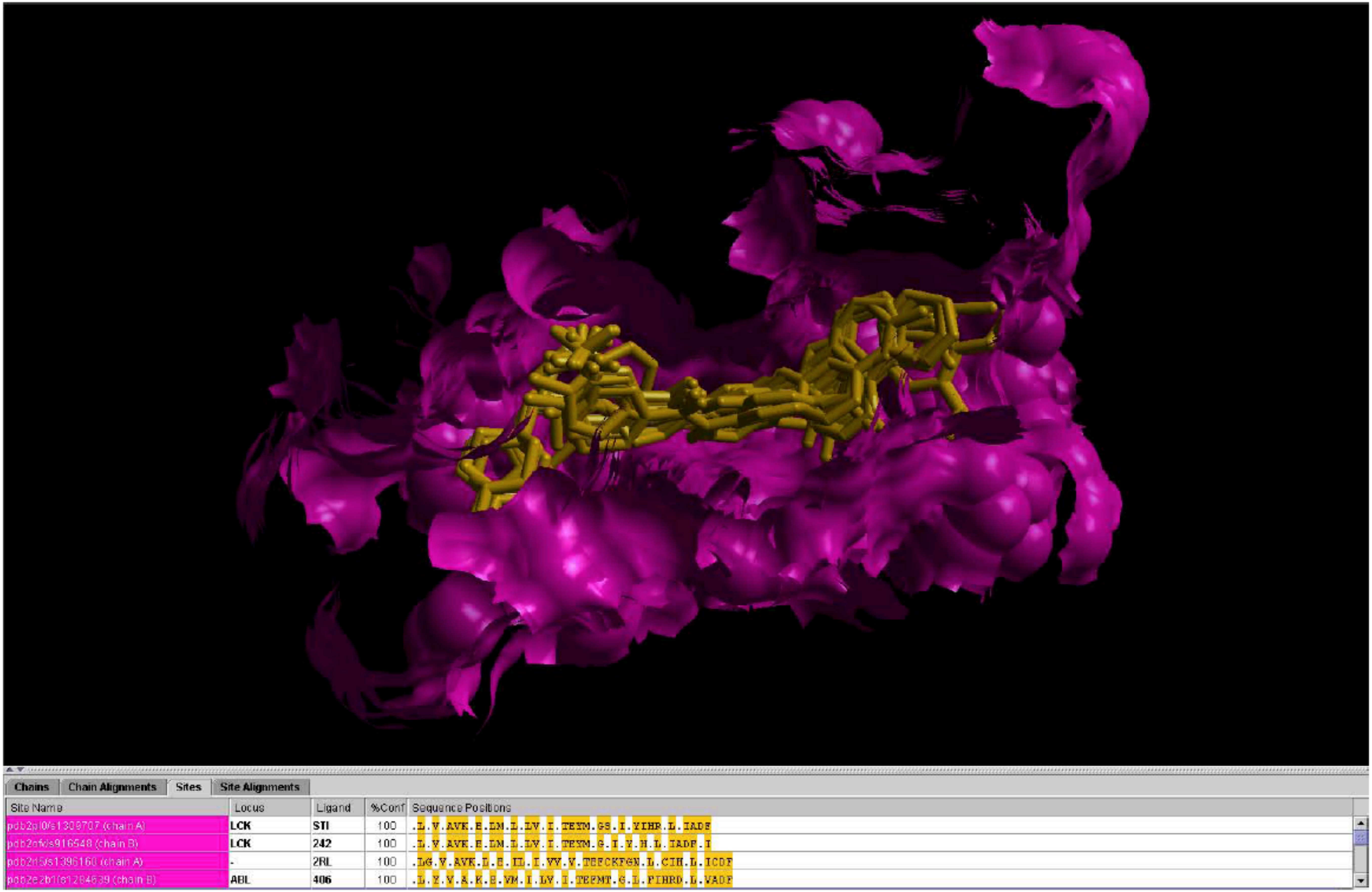

GIG


L11


B96

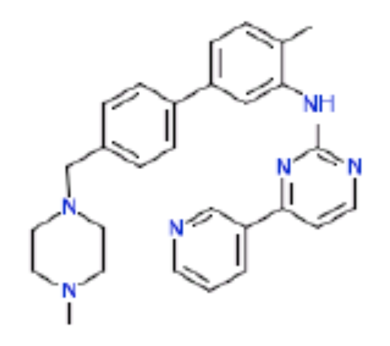
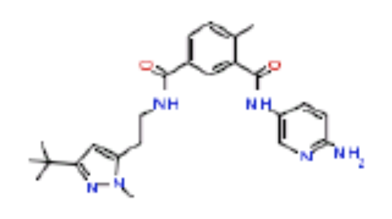
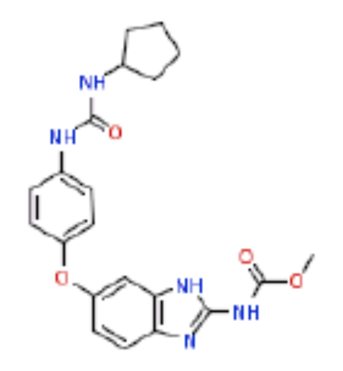
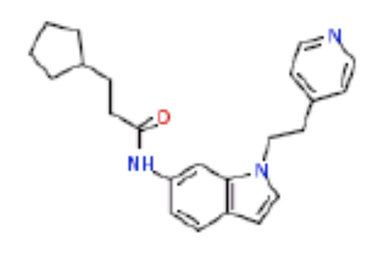
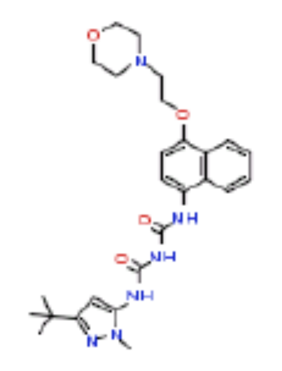
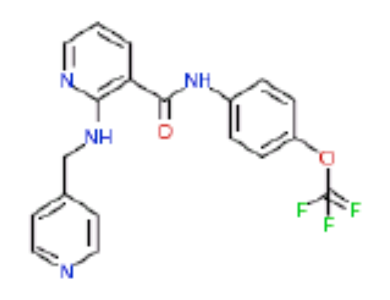
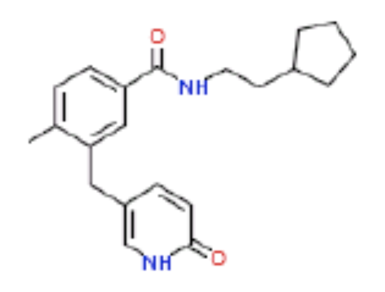
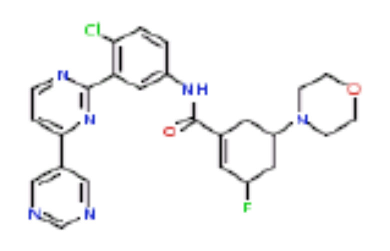
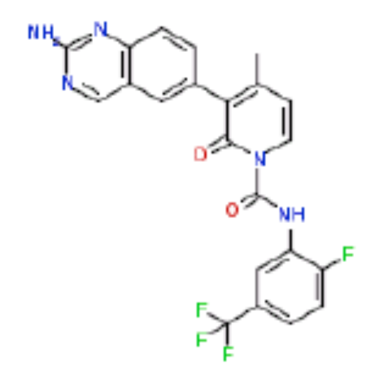
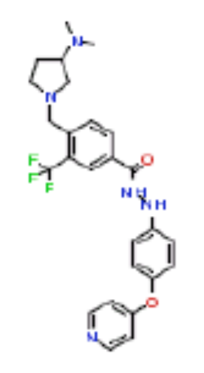
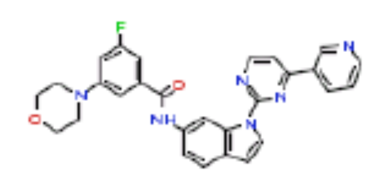
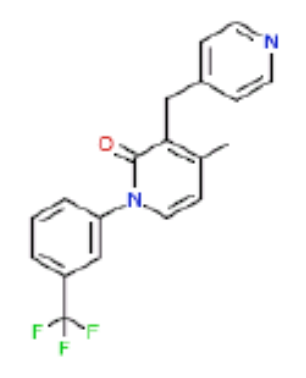
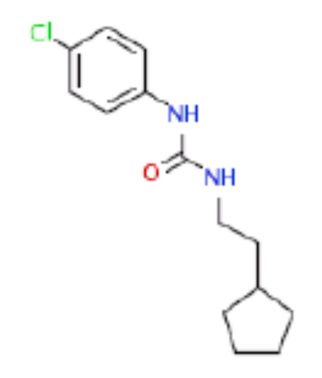
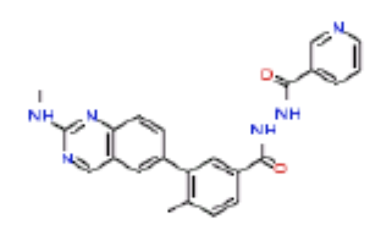
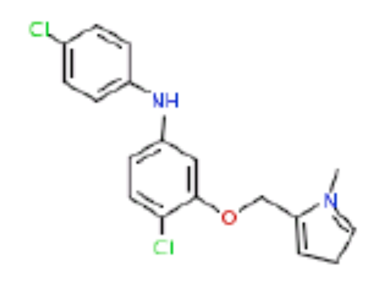
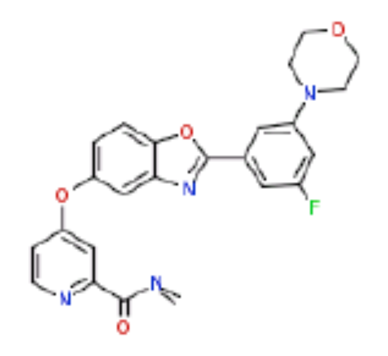
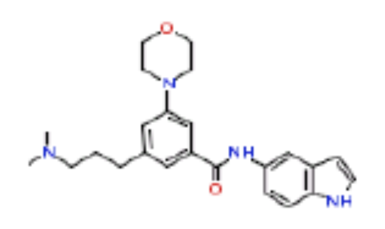
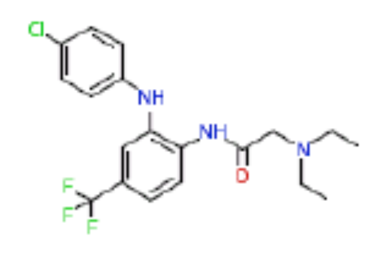
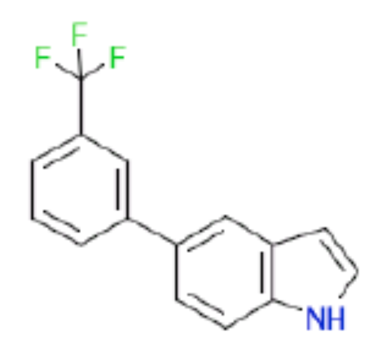
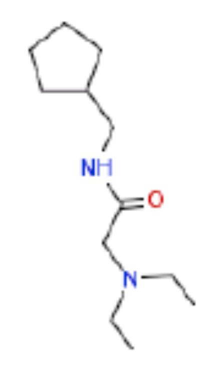

RAJ

Step 3: LigandCross: Shuffle Ligand Features from Aligned Sites



Site Name	Locus	Ligand	%Conf	Sequence Positions
pdb2pl0/s1309707 (chain A)	**LCK**	**STI**	100	.L.V.AVK.E.LM.L.LV.I.TEYM.GS.I.YIHR.L.IADF
pdb2ofv/s916548 (chain B)	**LCK**	**242**	100	.L.V.AVK.E.LM.L.LV.I.TEYM.G.I.Y.H.L.IADF.I
pdb2il5/s1396160 (chain A)	-	2RL	100	.LG.V.AVK.L.E.IL.I.VV.V.TEFCKFGN.L.CIH.L.ICDF
pdb2e2b1/s1284639 (chain B)	ABL	**406**	100	.L.Y.V.A.K.E.VM.I.LV.I.TEFMT.G.L.FIHRD.L.VADF

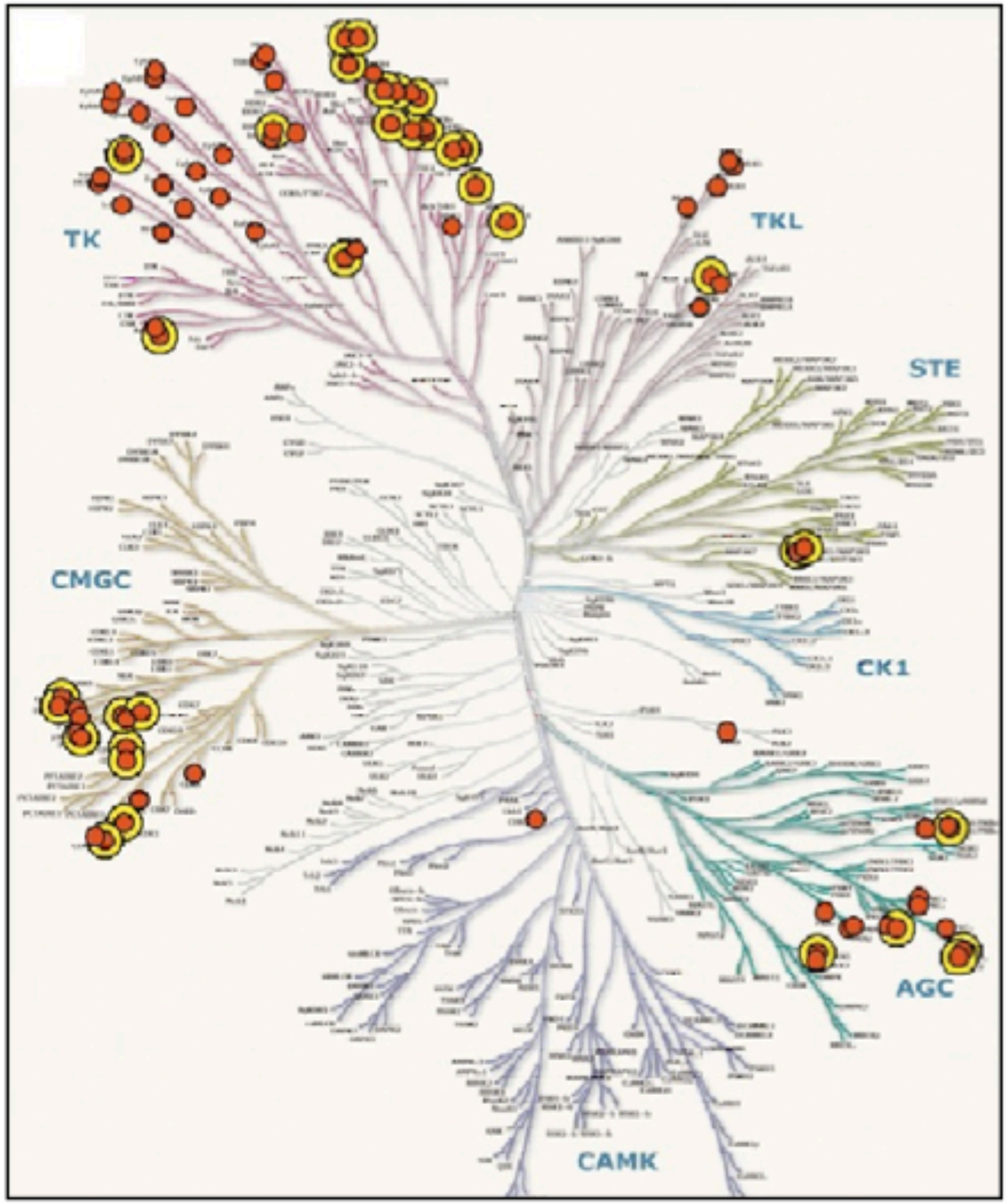
Example LigandCross Results

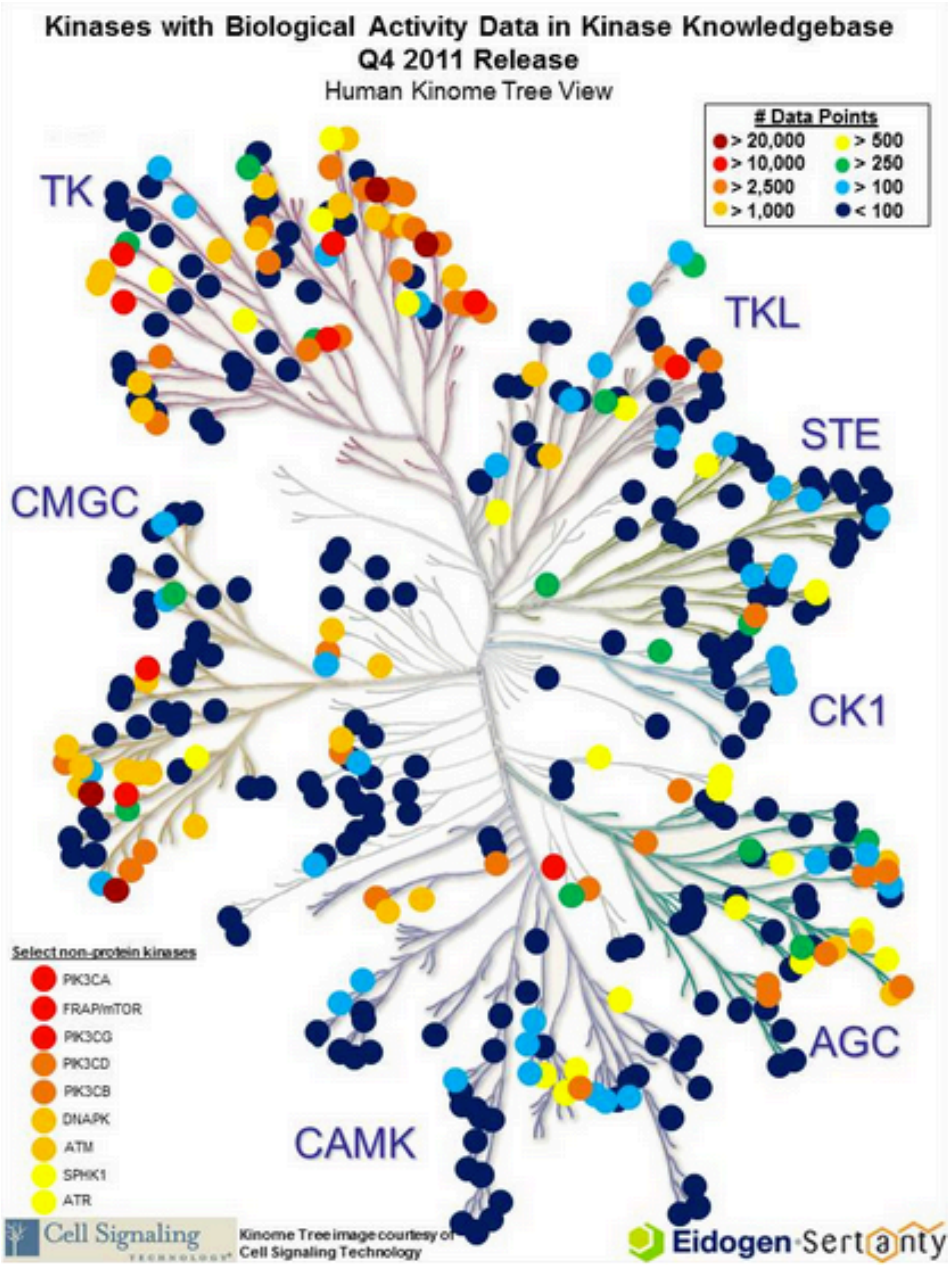
 STI_PRC_2 0.667	 C92_BMU_5 0.635	 C92_GIG_3 0.633	 C92_WBT_1 0.625	 B96_BMU_2 0.623
 608_276_3 0.608	 C92_GIN_7 0.608	 406_L11_6 0.577	 GIG_C52_1 0.574	 406_KIN_2 0.545
 NIL_WBT_6 0.538	 608_C52_2 0.529	 C92_BMU_1 0.520	 1N8_PRC_3 0.491	 857_BMU_4 0.480
 857_WBT_2 0.472	 RAJ_LI3_1 0.462	 1N8_BMU_2 0.449	 LI3_C52_2 0.385	 C92_1N8_1 0.375

Kinase Knowledgebase (KKB)

Kinase Targets of Clinical Interest

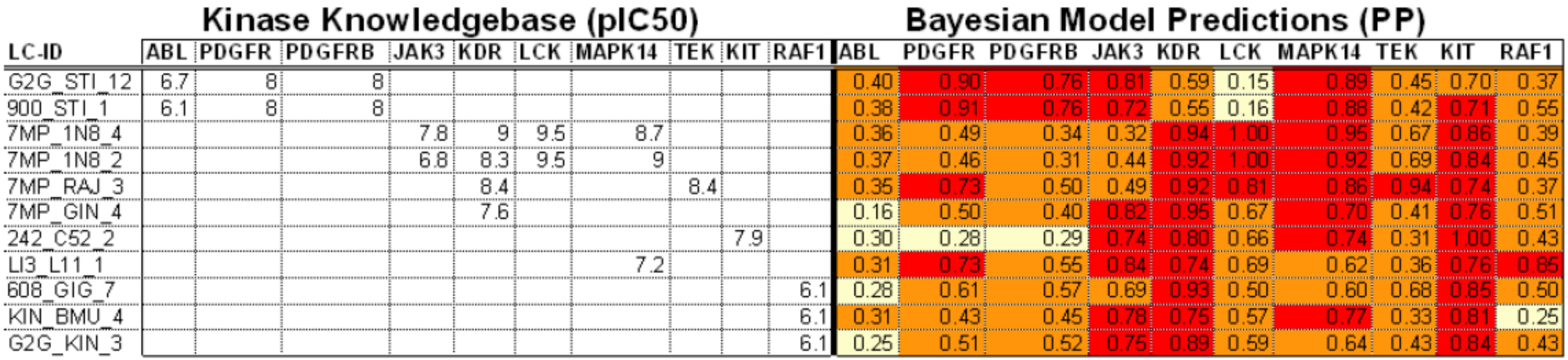
from Vieth *et al.* *Drug Disc. Today* **10**, 839 (2005).



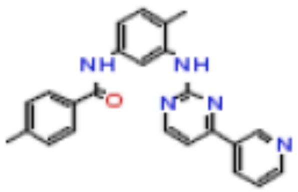


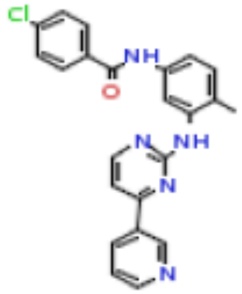
> 583K SAR data points curated from
> 7700 articles and patents

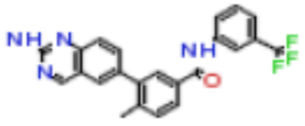
LigandCross Validation

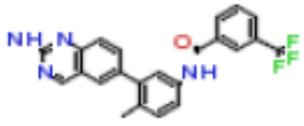


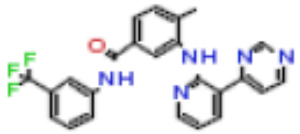
	Kinase Knowledgebase (pIC50)										Bayesian Model Predictions (PP)									
LC-ID	ABL	PDGFR	PDGFRB	JAK3	KDR	LCK	MAPK14	TEK	KIT	RAF1	ABL	PDGFR	PDGFRB	JAK3	KDR	LCK	MAPK14	TEK	KIT	RAF1
G2G_STI_12	6.7	8	8								0.40	0.90	0.76	0.81	0.59	0.15	0.89	0.45	0.70	0.37
900_STI_1	6.1	8	8								0.38	0.91	0.76	0.72	0.55	0.16	0.88	0.42	0.71	0.55
7MP_1N8_4				7.8	9	9.5	8.7				0.36	0.49	0.34	0.32	0.94	1.00	0.95	0.67	0.86	0.39
7MP_1N8_2				6.8	8.3	9.5	9				0.37	0.46	0.31	0.44	0.92	1.00	0.92	0.69	0.84	0.45
7MP_RAJ_3					8.4			8.4			0.35	0.73	0.50	0.49	0.92	0.81	0.86	0.94	0.74	0.37
7MP_GIN_4					7.6						0.16	0.50	0.40	0.82	0.95	0.67	0.70	0.41	0.76	0.51
242_C52_2									7.9		0.30	0.28	0.29	0.74	0.80	0.66	0.74	0.31	1.00	0.43
LI3_L11_1							7.2				0.31	0.73	0.55	0.84	0.74	0.69	0.62	0.36	0.76	0.85
608_GIG_7										6.1	0.28	0.61	0.57	0.69	0.93	0.50	0.60	0.68	0.85	0.50
KIN_BMU_4										6.1	0.31	0.43	0.45	0.78	0.75	0.57	0.77	0.33	0.81	0.25
G2G_KIN_3										6.1	0.25	0.51	0.52	0.75	0.89	0.59	0.64	0.43	0.84	0.43

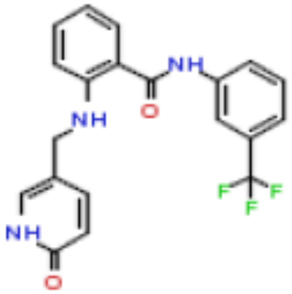

1: G2G_STI_12

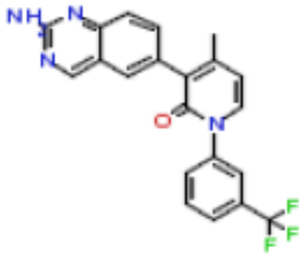

2: 900_STI_1

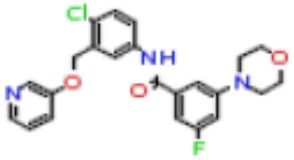

3: 7MP_1N8_4

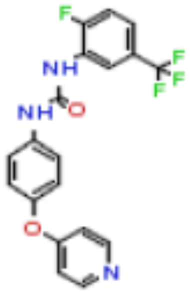

4: 7MP_1N8_2

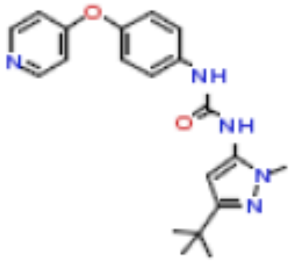

5: 7MP_RAJ_3

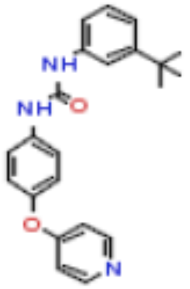

6: 7MP_GIN_4


7: 242_C52_2


8: LI3_L11_1

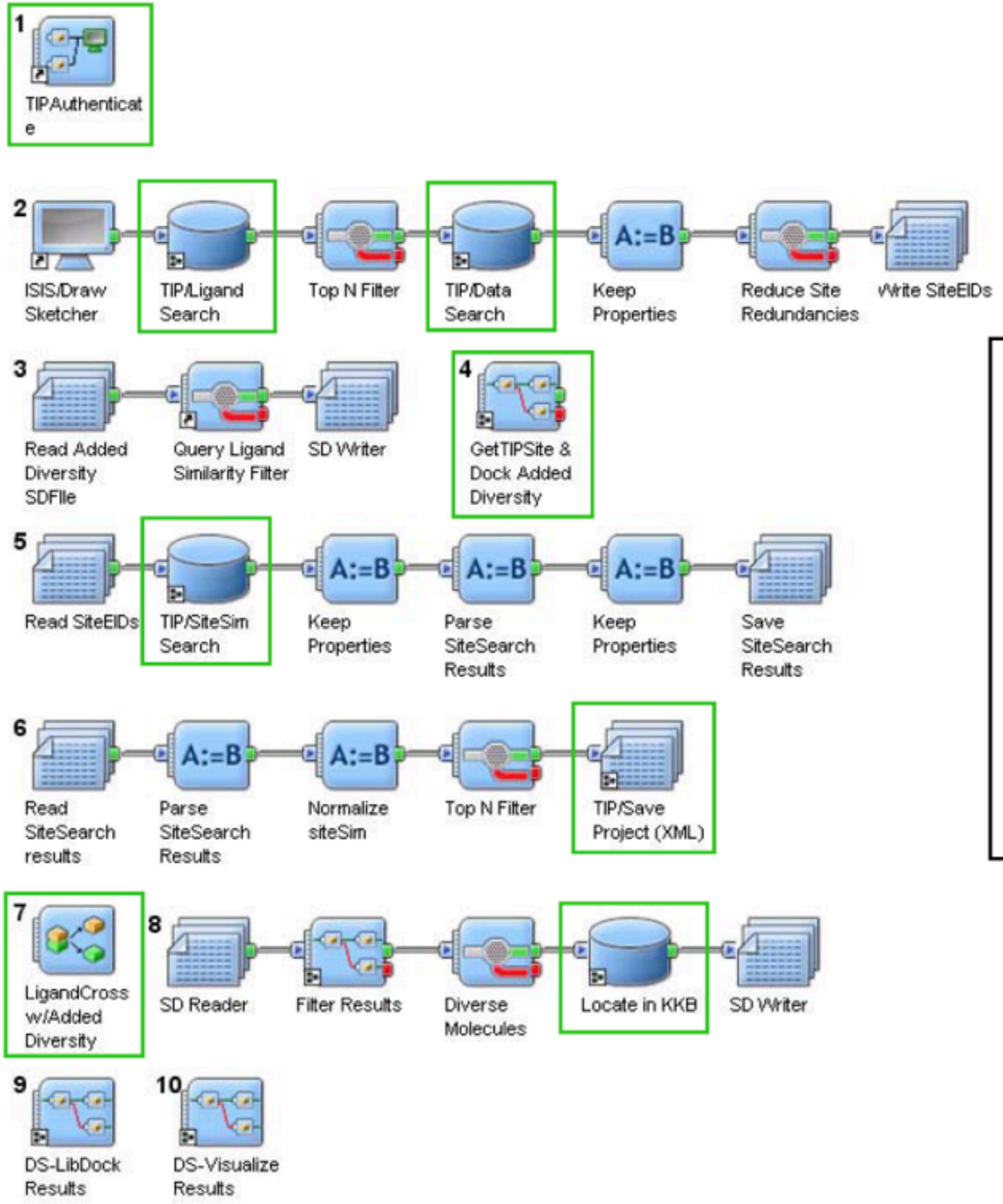

9: 608_GIG_7


10: KIN_BMU_4


11: G2G_KIN_3

Eidogen-Sertanty

Enhance LigandCross with Added Diversity



> Issue TIP/LigandSearch

> Identify/Dock “AddedDiversity”

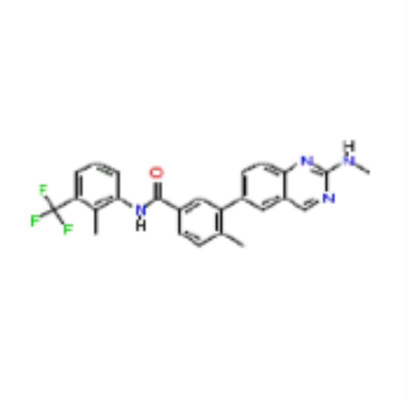
> Issue TIP/SiteSimSearch

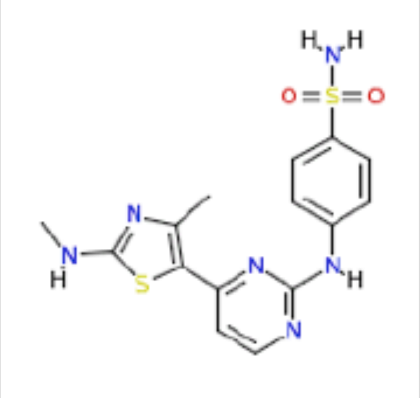
> LigandCross w/AddedDiversity

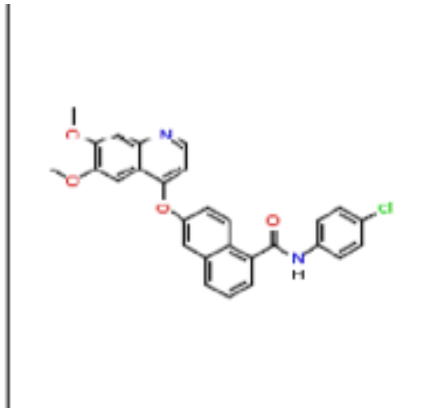
> Filter and locate results in KKB

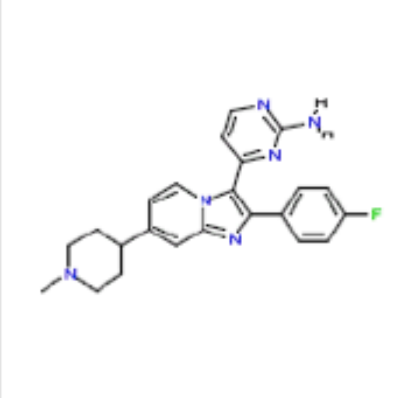
> Dock and visualize results

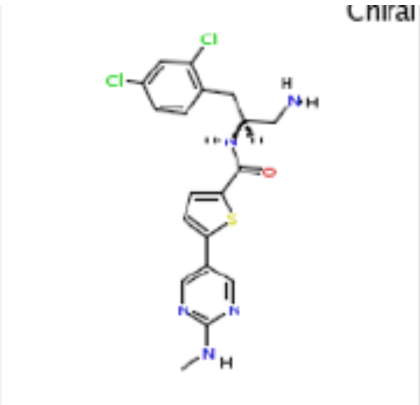
Example Potent Kinase Inhibitors (From KKB)


4336533 LCK pval: 11.00

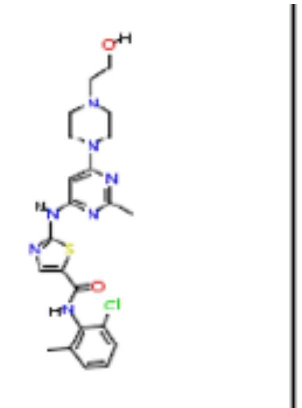

4302493 CDK9 pval: 10.54

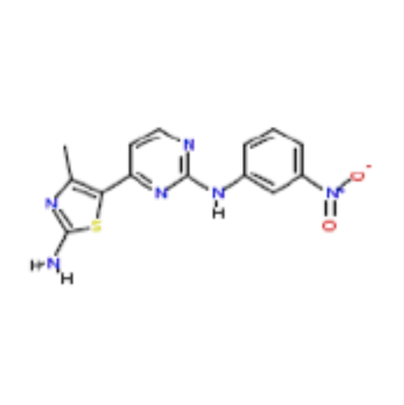

4332561 KDR pval: 10.52

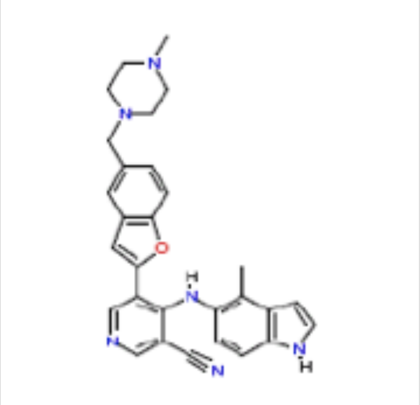

4318145 PKG pval: 10.40

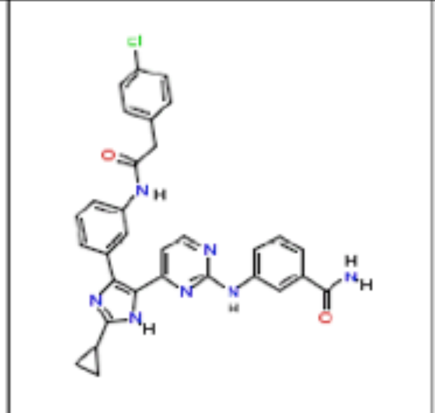


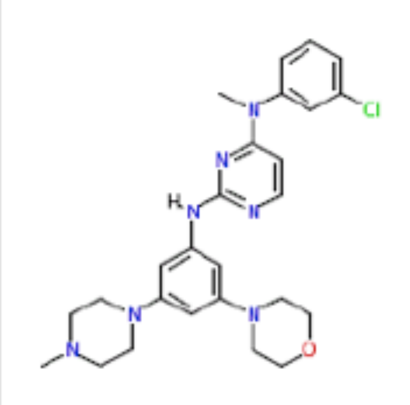
4336686 PKA pval: 10.00

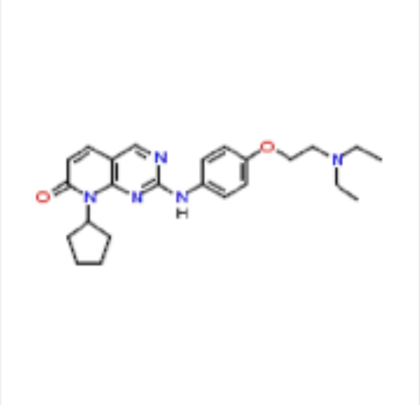

4272835 ABL1 pval: 10.00

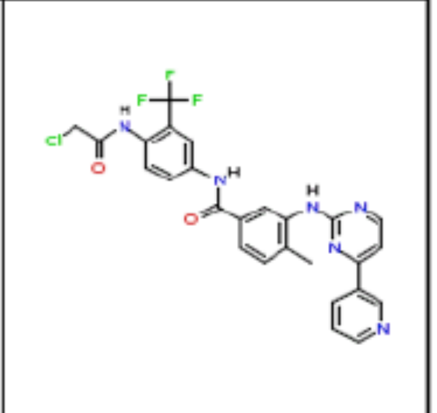

894611 CDK2 pval: 9.70

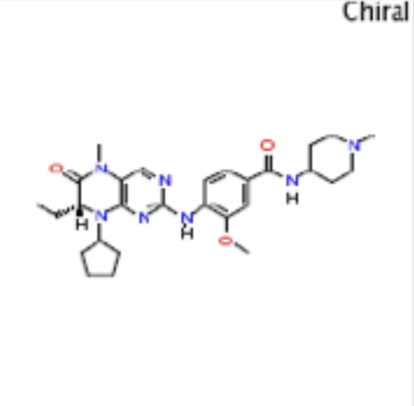

4358565 PRKCG pval: 9.70


4363734 RAF1 pval: 9.30

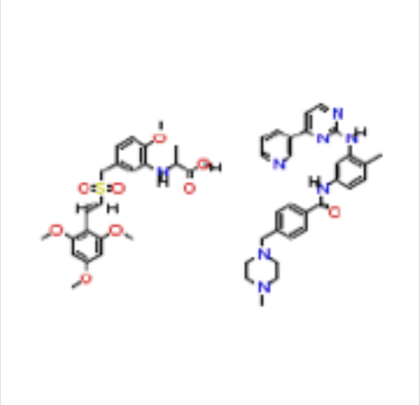

4369892 EPHB4 pval: 9.24

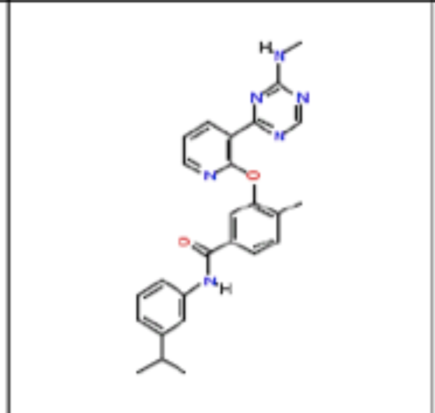

809 CDK4 pval: 9.15

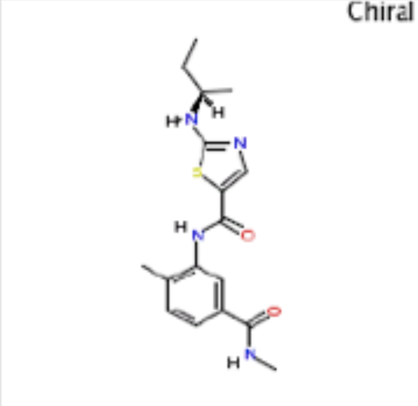

4374385 PDGFRA pval: 9.14



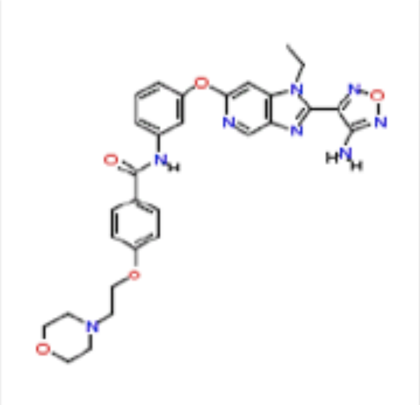
4366691 PLK1 pval: 9.10

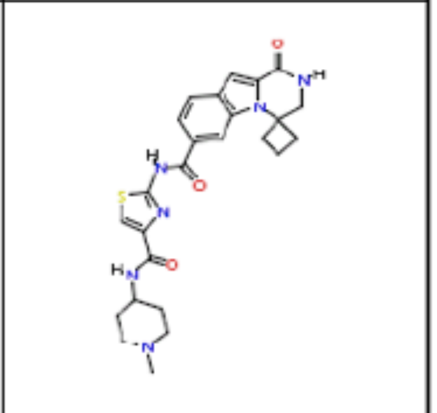

4301886 BCR_ABL pval: 9.08

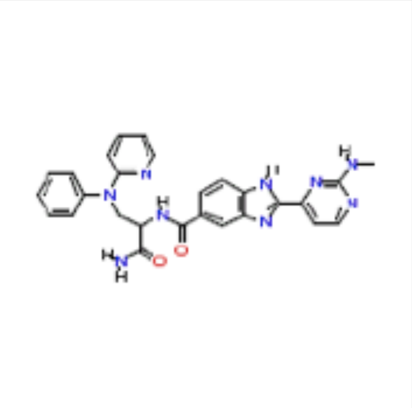

4307551 TEK pval: 9.00

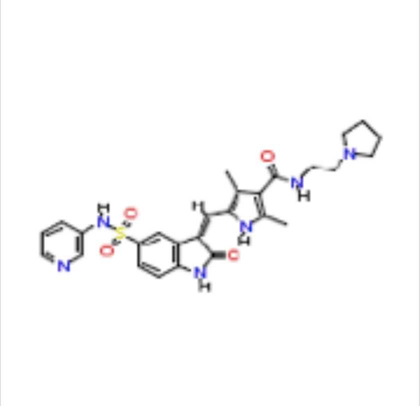


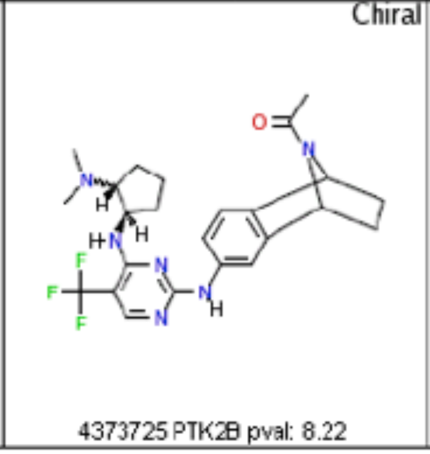
4363016 MAPK11 pval: 8.82


4343448 ROCK1 pval: 8.74

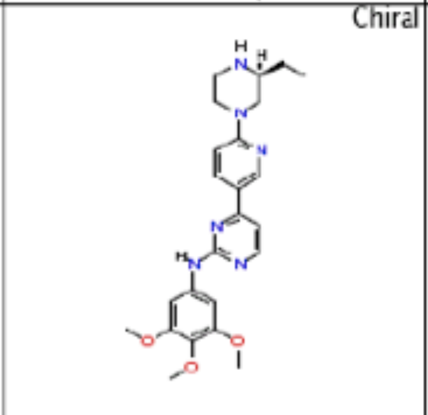

4363247 MAPKAPK2 pval: 8.70


4291996 IKB pval: 8.70

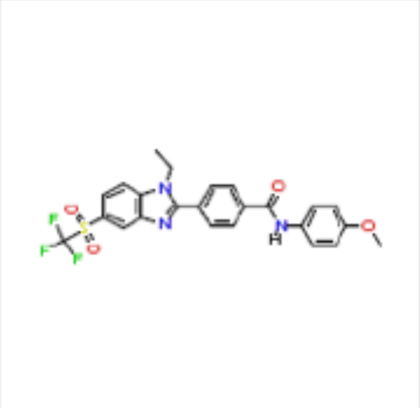

4208857 FAK2 pval: 8.22

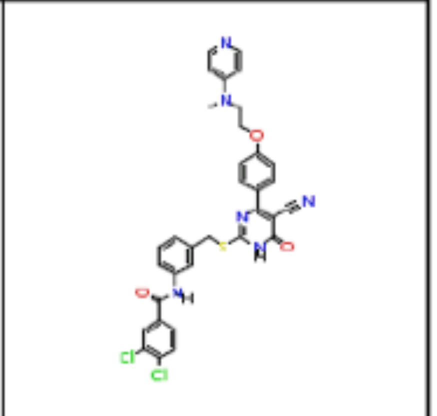


4373725 PTK2B pval: 8.22

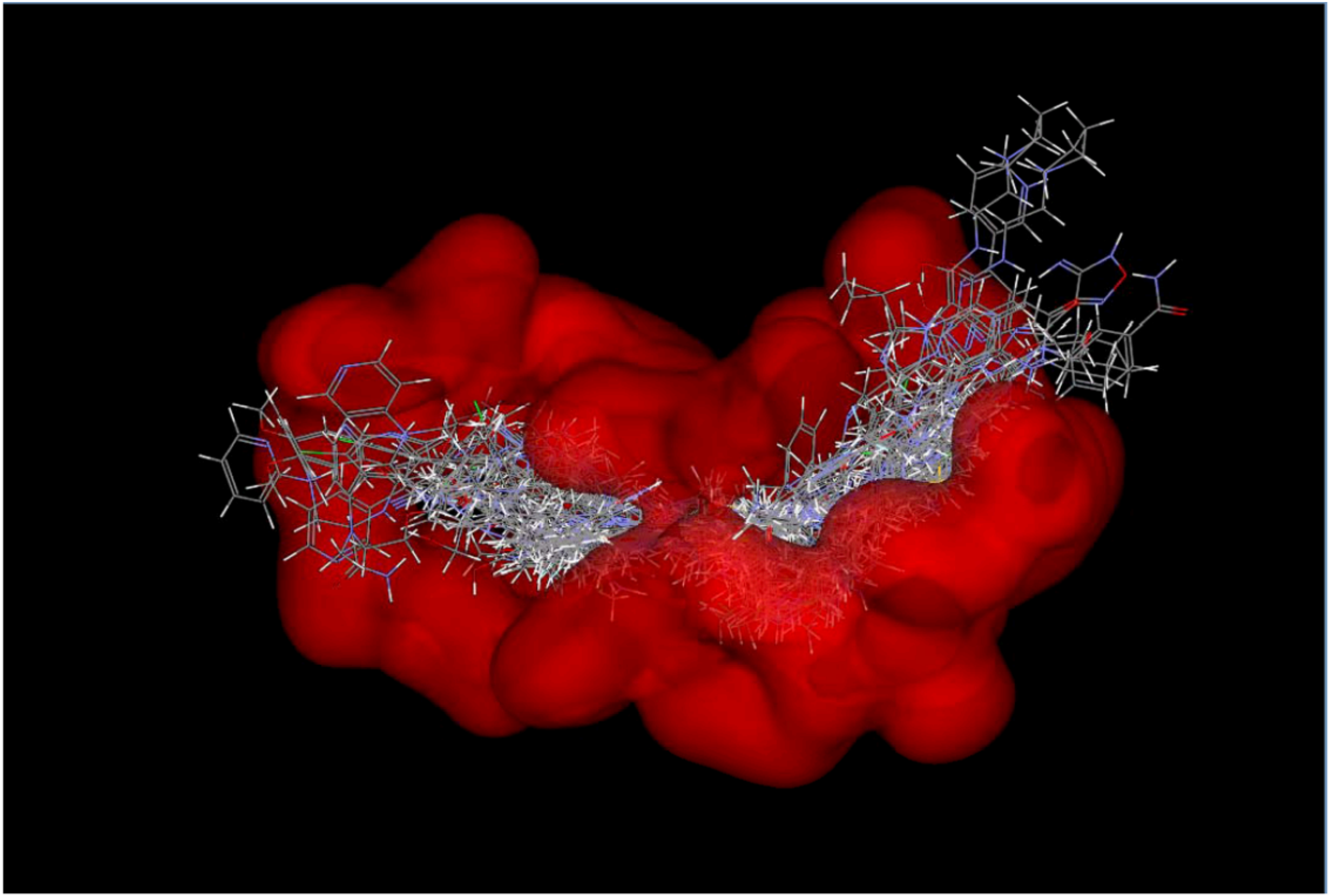


1788 ZAP70 pval: 8.10

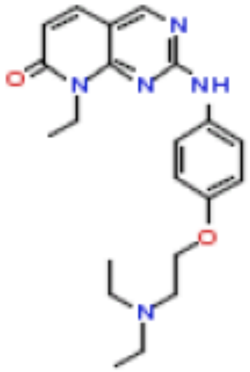

2425813 PTPN9 pval: 5.96

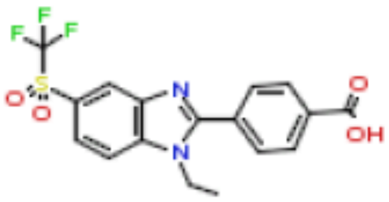

4303129 MAP3K2 pval: 4.70

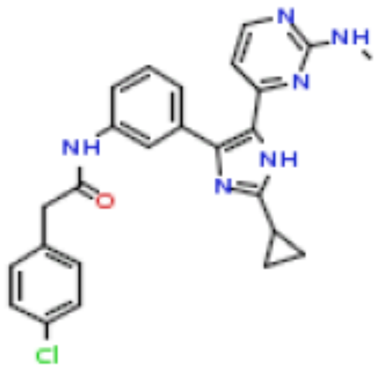
Potent Kinase Inhibitors Docked (s1309707)

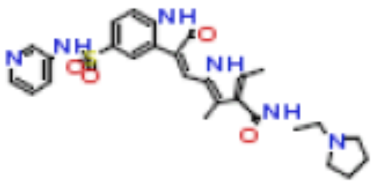


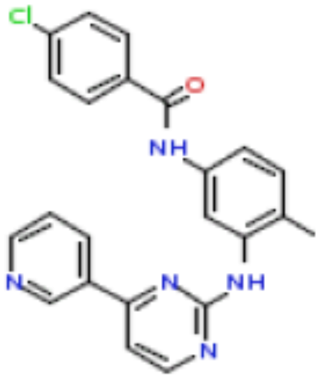
Added Diversity LigandCross Validation

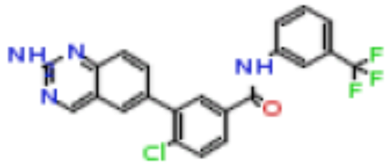

4343448_809_27

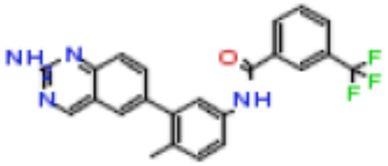

4272835_2425813_23

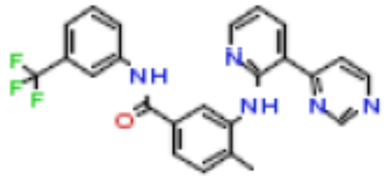

4363734_4291996_2


4208857_4208857_1


900_STI_1


242_A96_5


242_MUH_1


242_MUH_2


406_STI_1

4343448_809_27:
CDK4: 6.80 CDK2: 5.63 CDK2: 6.12 CDC2: 5.58 CSK: 5.99 CDK5: 6.81 CDK4: 6.80 CDK2: 5.63 CDK2: 6.12 CDC2: 5.58 CDK4: 6.80

4272835_2425813_23:
PTPN1: 4.24 PTPRA: 4.21

4363734_4291996_2:
RAF1: 9.00 MAPK1: 5.29 BRAF: 8.05 BRAF: 8.52

4208857_4208857_1:
FAK2: 8.22 KDR: 5.86 PDGFRB: 4.90 EGFR: 4.17 ERBB2: 5.23

900_STI_1:
PDGFR: 8.00 PDGFR: 8.00 ABL: 6.10 PDGFRB: 8.00 PDGFR: 8.00 ABL: 6.10

242_A96_5:
LCK: 9.40

242_MUH_1:
LCK: 9.40 TEK: 7.68 KDR: 8.22 MAPK14: 9.00 JAK3: 6.81

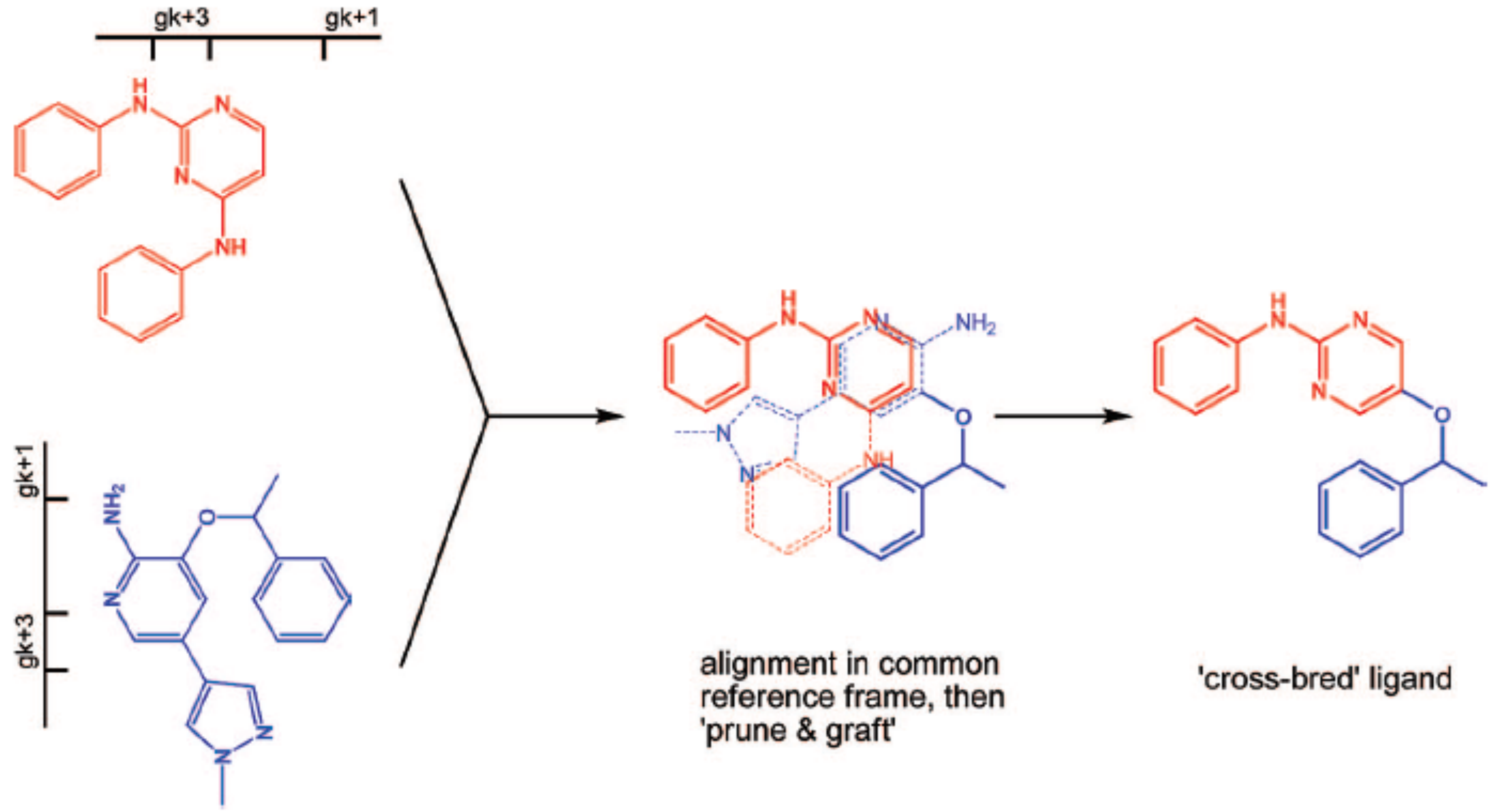
242_MUH_2:
KDR: 8.40 TEK: 8.40 TEK: 8.40 KDR: 8.40 TEK: 8.40 KDR: 8.40

406_STI_1:
BCR_ABL: 8.40 BCR_ABL: 5.30 LYN: 8.06 ABL1: 8.07 ABL1: 8.40

Generation of Cross-Bred Kinase Ligands

Arup K. Ghose *et al. J. Med. Chem*. 2008, 51, 5149-5171

1. Align all relevant PDB kinases on a reference kinase structure.
2. Optional step: ‘Refine’ the binding poses of the ligands for the target kinase using any docking pose refinement program (e.g. Schrodinger: Glide/Refine).
3. Apply Ligand-Cross to generate first generation ligands.



alignment in common reference frame, then 'prune & graft'

'cross-bred' ligand

Before alignment

A Few Useful Comments on Using LigandCross

- A few steps of steepest descent followed by conjugate gradient of the ligand in the protein binding site, often fixes distorted geometry problems.

- Use both Type-I and Type-II ligands.

- Use aligned DFG-in kinase structure/model during minimization and binding mode evaluation for generating Type-I ligands.

- Use aligned DFG-out kinase structure/model during minimization and binding mode evaluation for generating Type-II ligands.

- The method was successfully used to design novel Type-I and Type-II ligands for ALK, JAK2, FAK and TAM/RTK kinases.

- For further information contact Arup K. Ghose (akghose@msn.com)

Conclusions

- The structurally resolved and modelable proteome is a very rich source for new matter ideas

- Receptor-site similarity can rapidly identify “feed-stock” functionality for favorable ligand decoration

- LigandCross can be an effective approach to generate novel, bioactive matter using co-complexes, known inhibitors, and/or fragment-based information.

Eidogen-Sertanty’s iPhone and iPad Apps

iKinase

iKinasePro

Mobile Reagents

Reaction101

Yield101

SPRESImobile

iProtein

iKinasePro

Mobile Reagents

Reaction101

Yield101

SPRESImobile

** Reaction101,Yield101, and SPRESImobile - built collaboratively with Molecular Materials Informatics, Inc.*

Eidogen Sertanty

MobileApps: Worldwide Marketing Vehicles!

iKinase

iKinasePro

iProtein

Mobile Reagents

Reaction101

Yield101

SPRESImobile

~ 25,000 People Use Eidogen Mobile Apps