The application of high-throughput technologies to fundamental crystallization research

Edward H. Snell, Joseph R. Luft, Michael Malkowski George DeTitta
and the staff of the Center for High-Throughput Structural Biology.

Hauptman Woodward Medical Research Institute, Buffalo, NY 14203, USA.

An introduction to the screening laboratory at the Hauptman-Woodward Medical Research Institute

Since February of 2000 the High Throughput Search (HTS) laboratory has been screening potential crystallization conditions for the general biomedical community and two Protein Structure Initiative large-scale structure production centers (NESG, Montelione, PI; SGPP/MSGPP, Hol, PI) and one PSI specialized PSI-2 center (CHTSB, DeTitta, PI).

The HTS lab screens samples against an incomplete factorial screen of two categories of crystallizing agents:

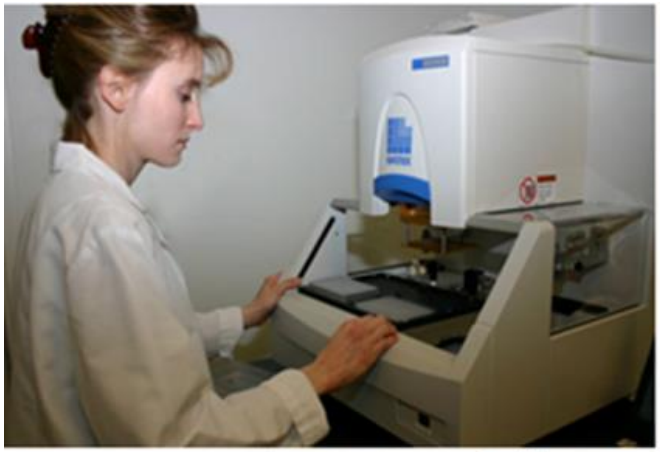
1. buffered (4<pH< 10), highly concentrated salts (35 salts total, sampling 18 different cations and 20 anions) – XXXX conditions.
2. PEG/salt/buffer solutions (eight buffers (4<pH< 10), six molecular weight PEGs at three concentrations, and 35 salts at fixed 200 mM concentration) – XXX conditions.

Added to this is a screen of some XXX conditions encompassing screens commercially available from Hampton Research.

The crystallization method used is micro-batch under oil with 200 nl of protein solution being added to 200 nl of precipitant cocktail in each well of a 1536 well plate.

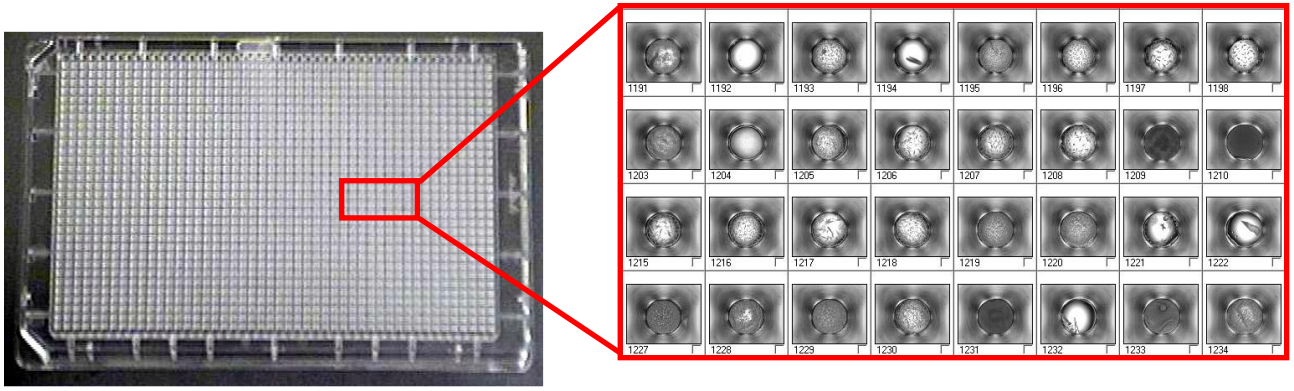
Wells are imaged before filling, immediately after filling then weekly for six weeks duration with images available immediately on a secure ftp server.

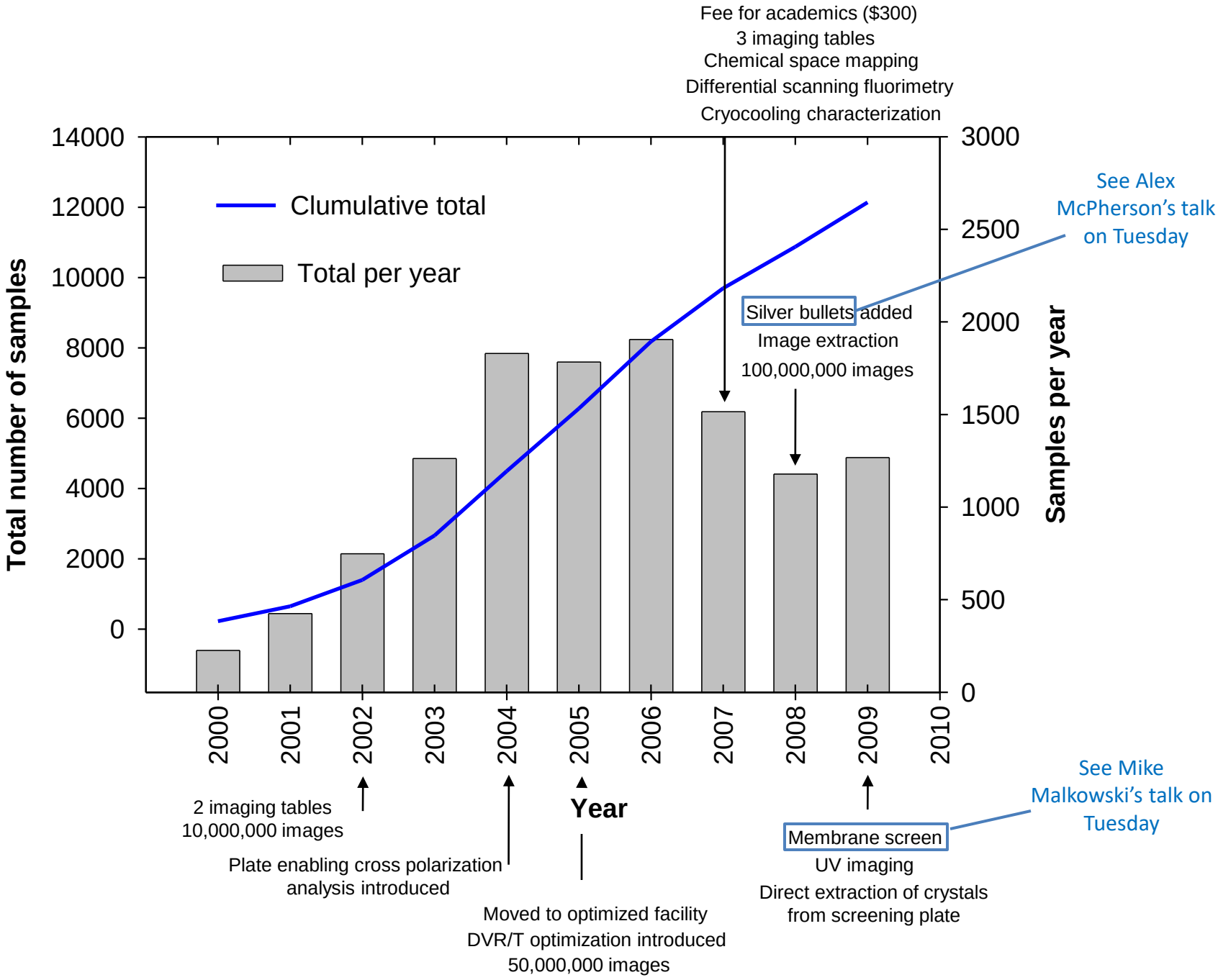
The HTSlab has investigated the crystallization properties of over 12,500 individual proteins archiving over 115,000,000 images of crystallization experiments.

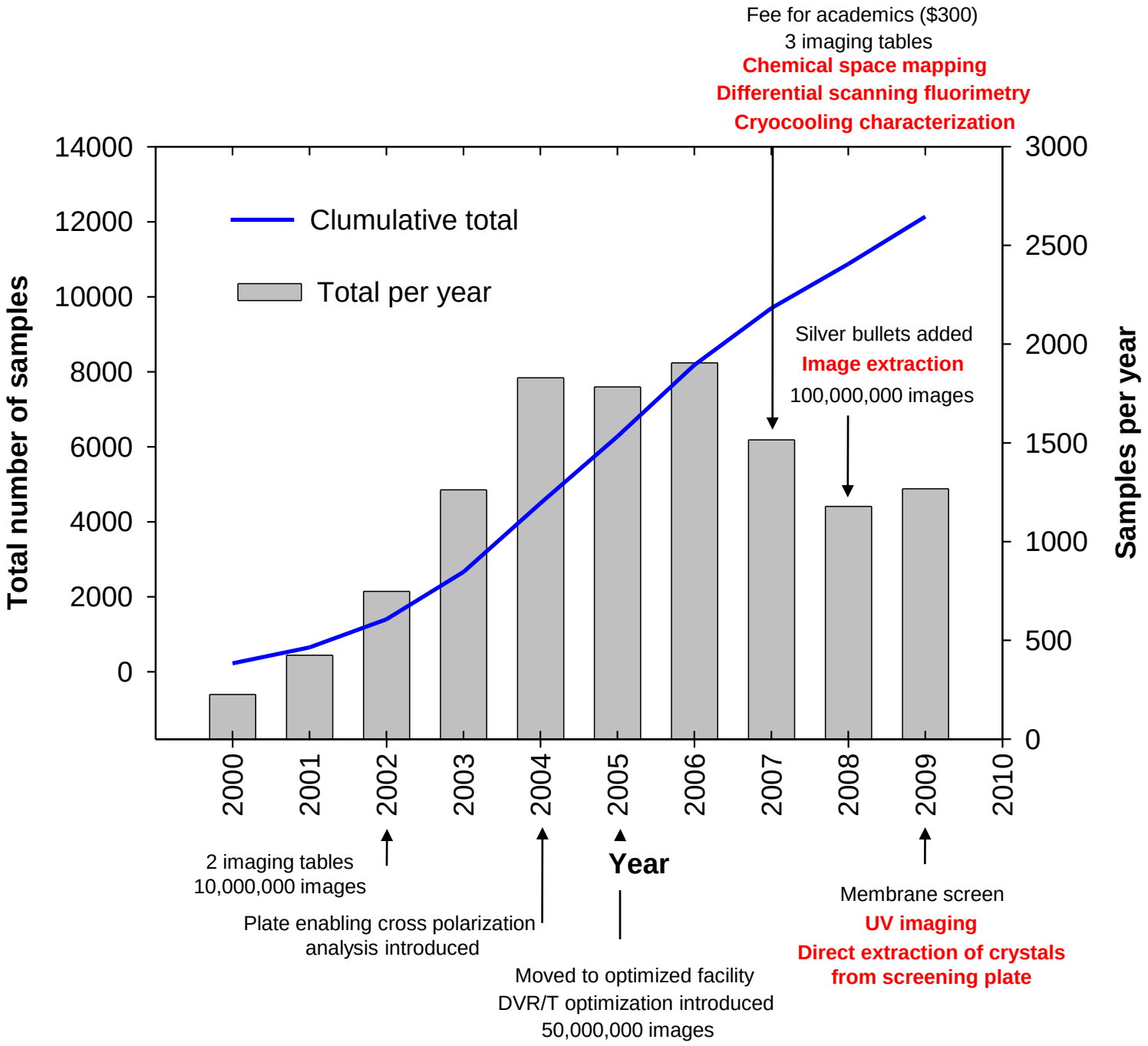


The staff, instrumentation and crystallization plate used

See the poster by Joseph Luft for full details of the laboratory.




Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Total number of samples
Samples per year
Year
14000
12000
10000
8000
6000
4000
2000
0
3000
2500
2000
1500
1000
500
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
See Alex McPherson's talk on Tuesday
Silver bullets added
Image extraction
100,000,000 images
2 imaging tables
10,000,000 images
Plate enabling cross polarization analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals from screening plate
See Mike Malkowski's talk on Tuesday


Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate

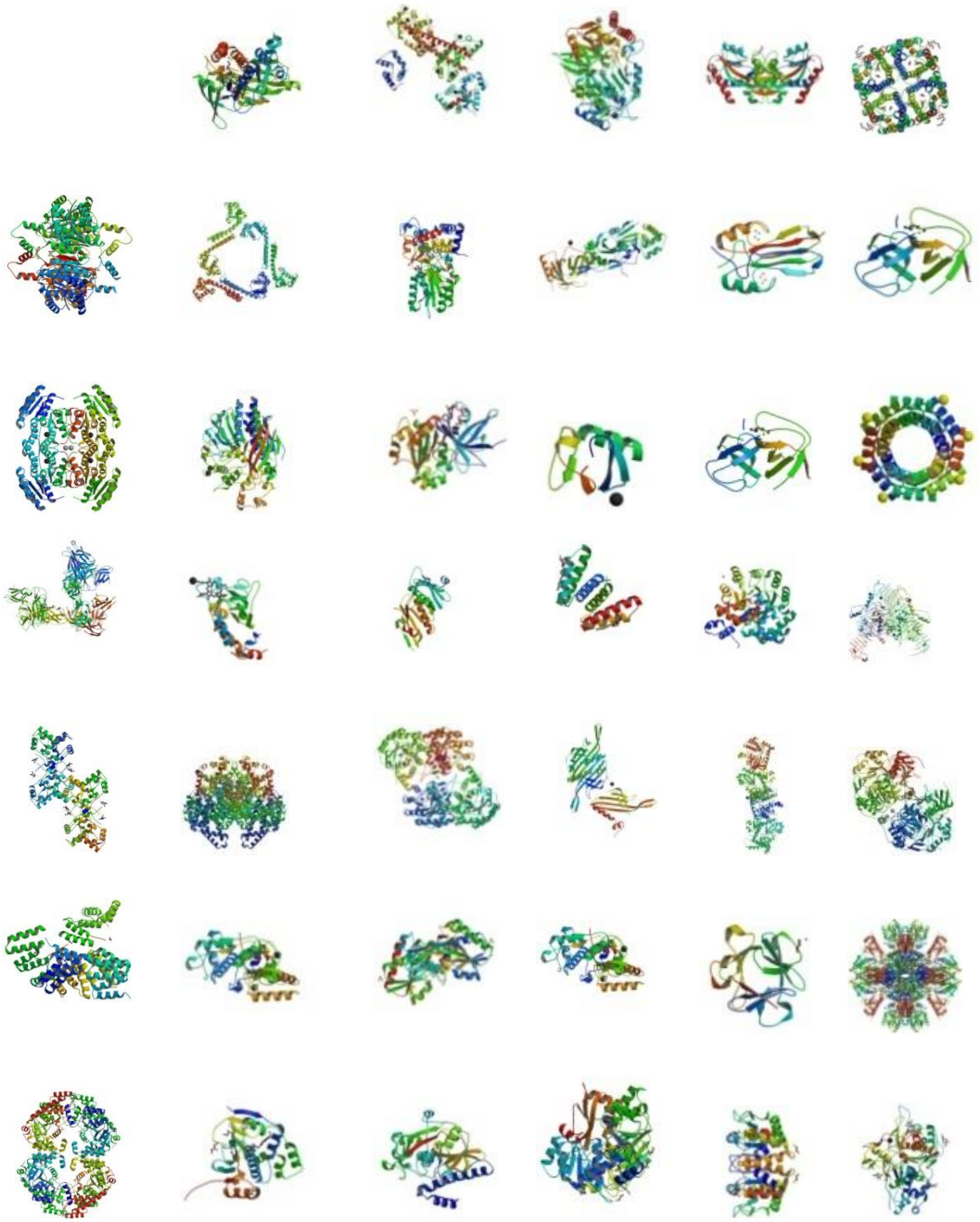
Born in Buffalo

Over 1,000 general biomedical laboratories world wide use the crystallization screening service with approximately 2,000 unique investigators.

Investigators are sent photographs of the results, analyze these images and perform their own optimization of any hits observed.

No information is released on targets. Progress is tracked by acknowledgements and citation searches. Currently no other metrics are used to measure success rates for the general biomedical community.

These images represent examples of structures from initial hits in the HTS laboratory.



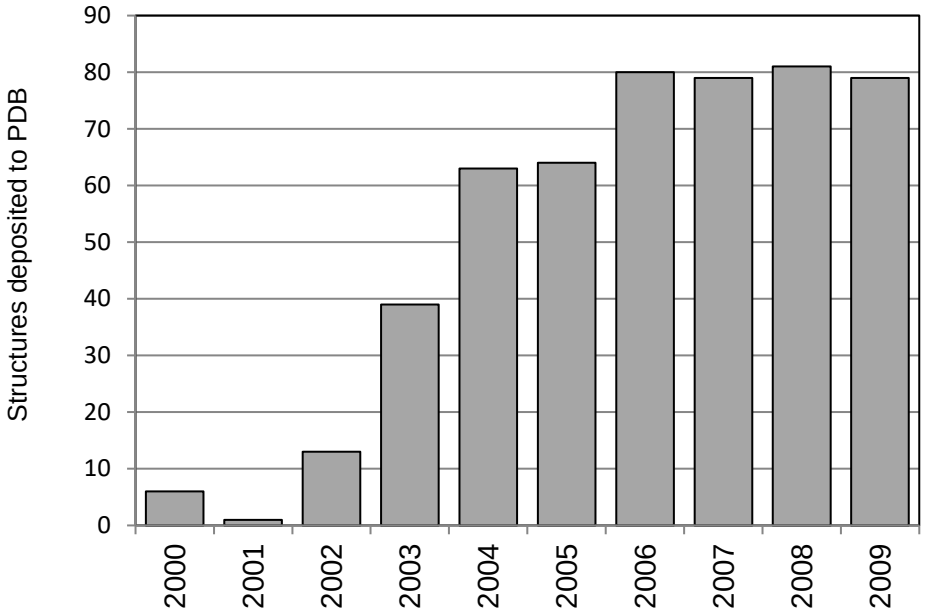
Where success is tracked.

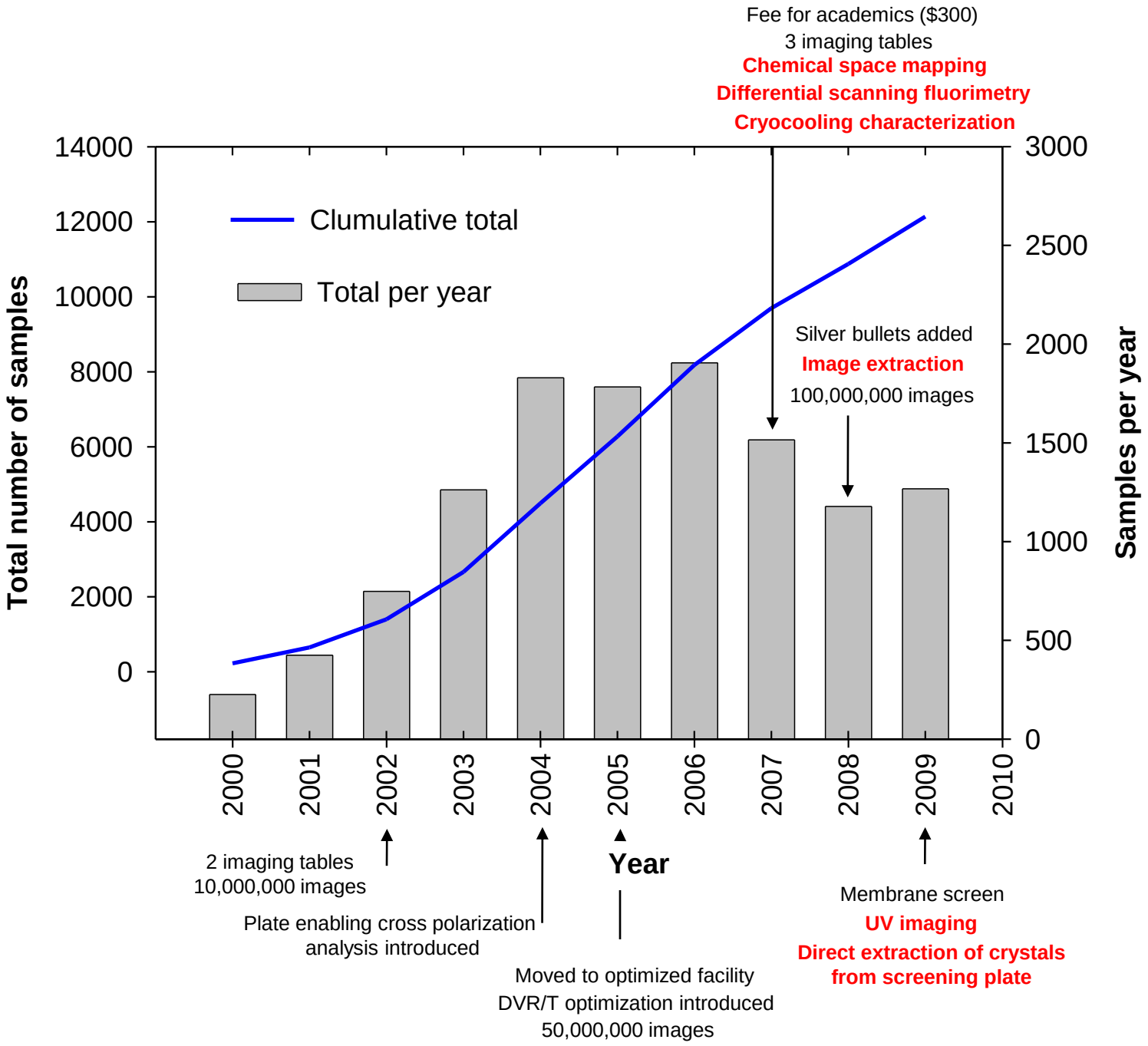
For our Protein Structure Initiative partners both success and failure is tracked. In the case of NESG our initial screening hits enable on average 80 structures per year to be deposited to the PDB.

The graph demonstrates the ramp up of operations with maximum success reached from 2006 onward.

Our success rate from protein in the door to a crystallization hit leading to a PDB deposition is **22%**.

The NESG samples represent a special case in that they are well characterized beforehand ……




Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate

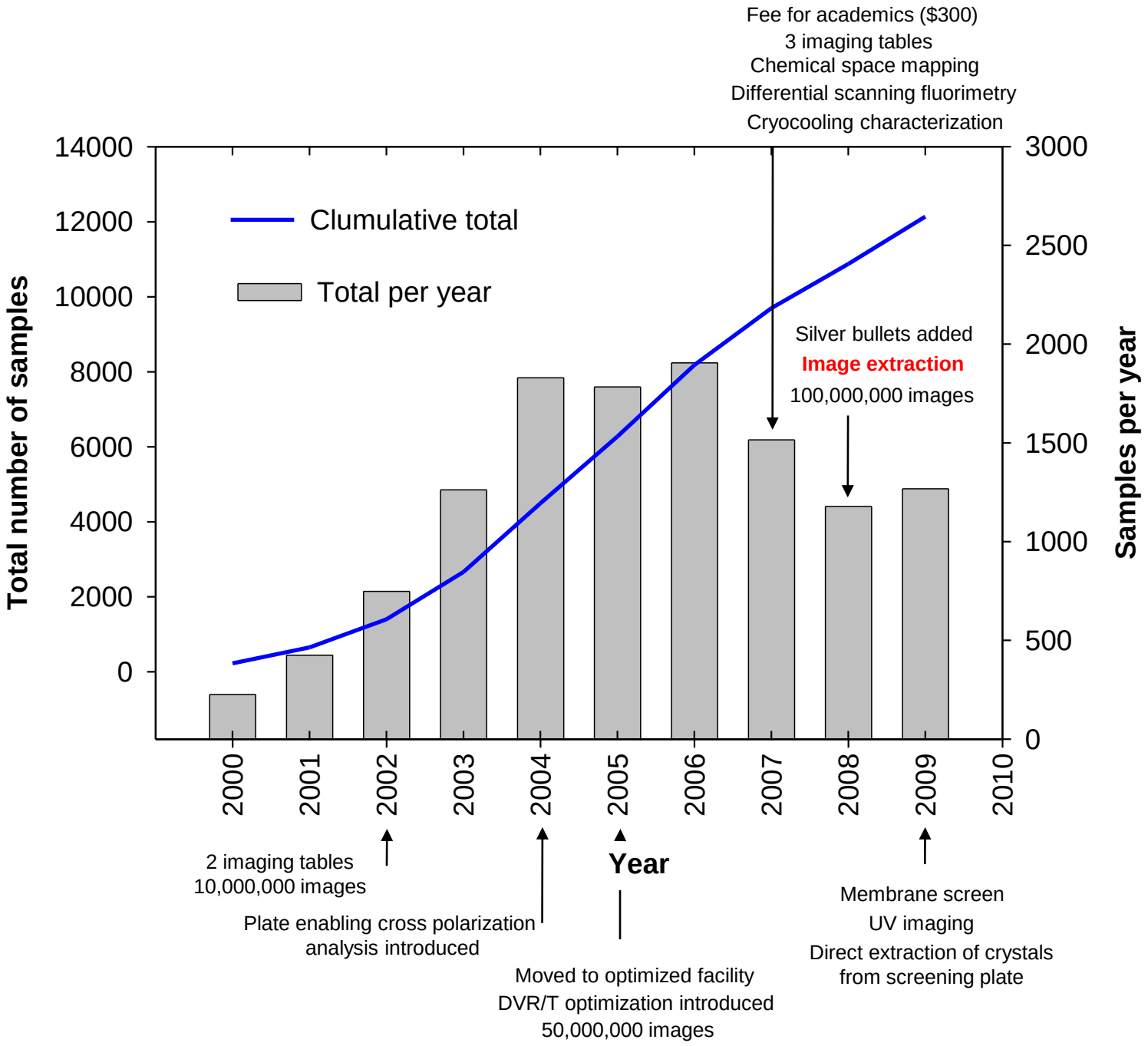
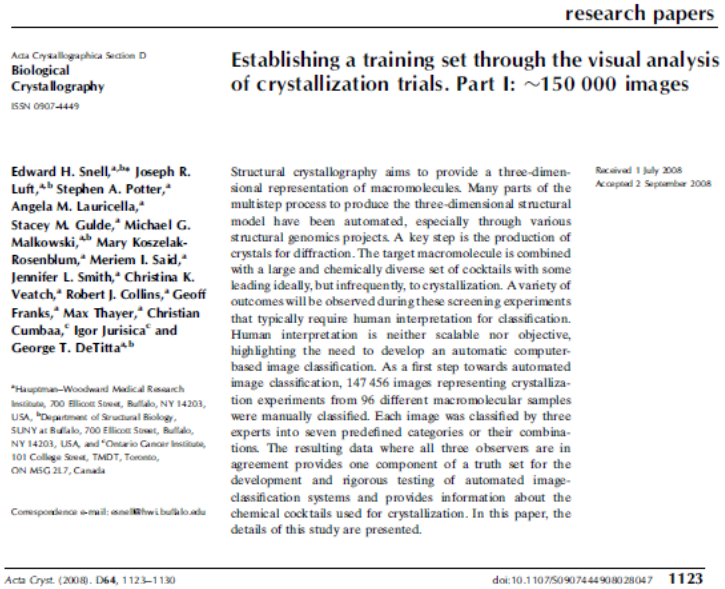

Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate

Image analysis (training set)

We have manually classified 147,456 images representing crystallization experiments from 96 different macromolecular samples. Each image has been classified by three people into seven predefined categories, or their combinations. These were clear, phase separation, precipitate, skin, crystals, junk and unsure representing a compromise between fidelity and practicality.

Crystal images were selected from 269 macromolecules that showed clear crystal hits (out of 823) provided by the NESG and SGPP structural genomics centers. These were classified by a single viewer and represent a data set of some 2.4 million images (2 week and 4 week reads) that have been manually viewed.

These data sets were combined, a test set extracted, and the remainder used to train an image classifier program.



research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Establishing a training set through the visual analysis of crystallization trials. Part I: ~150 000 images

Edward H. Snell,[a,b]* Joseph R. Luft,[a,b] Stephen A. Potter,[a] Angela M. Lauricella,[a] Stacey M. Gulde,[a] Michael G. Malkowski,[a,b] Mary Koszelak-Rosenblum,[a] Meriem I. Said,[a] Jennifer L. Smith,[a] Christina K. Veatch,[a] Robert J. Collins,[a] Geoff Franks,[a] Max Thayer,[a] Christian Cumbaa,[c] Igor Jurisica[c] and George T. DeTitta[a,b]

[a]Hauptman–Woodward Medical Research Institute, 700 Ellicott Street, Buffalo, NY 14203, USA, [b]Department of Structural Biology, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA, and [c]Ontario Cancer Institute, 101 College Street, TMDT, Toronto, ON M5G 2L7, Canada

Correspondence e-mail: esnell@hwi.buffalo.edu

Structural crystallography aims to provide a three-dimensional representation of macromolecules. Many parts of the multistep process to produce the three-dimensional structural model have been automated, especially through various structural genomics projects. A key step is the production of crystals for diffraction. The target macromolecule is combined with a large and chemically diverse set of cocktails with some leading ideally, but infrequently, to crystallization. A variety of outcomes will be observed during these screening experiments that typically require human interpretation for classification. Human interpretation is neither scalable nor objective, highlighting the need to develop an automatic computer-based image classification. As a first step towards automated image classification, 147 456 images representing crystallization experiments from 96 different macromolecular samples were manually classified. Each image was classified by three experts into seven predefined categories or their combinations. The resulting data where all three observers are in agreement provides one component of a truth set for the development and rigorous testing of automated image-classification systems and provides information about the chemical cocktails used for crystallization. In this paper, the details of this study are presented.

Received 1 July 2008
Accepted 2 September 2008

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Establishing a training set through the visual analysis of crystallization trials. Part II: crystal examples

Edward H. Snell,[a,b]* Angela M. Lauricella,[a] Stephen A. Potter,[a] Joseph R. Luft,[a,b] Stacey M. Gulde,[a] Robert J. Collins,[a] Geoff Franks,[a] Michael G. Malkowski,[a,b] Christian Cumbaa,[c] Igor Jurisica[c] and George T. DeTitta[a,b]

[a]Hauptman–Woodward Medical Research Institute, 700 Ellicott Street, Buffalo, NY 14203, USA, [b]Department of Structural Biology, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA, and [c]Ontario Cancer Institute, 101 College Street, TMDT, Toronto, ON M5G 2L7, Canada

In the automated image analysis of crystallization experiments, representative examples of outcomes can be obtained rapidly. However, while the outcomes appear to be diverse, the number of crystalline outcomes can be small. To complement a training set from the visual observation of 147 456 crystallization outcomes, a set of crystal images was produced from 106 and 163 macromolecules under study for the North East Structural Genomics Consortium (NESG) and Structural Genomics of Pathogenic Protozoa (SGPP) groups, respectively. These crystal images have been combined with the initial training set. A description of the crystal-enriched data set and a preliminary analysis of outcomes from the data are described.

Received 1 July 2008
Accepted 2 September 2008

1. Introduction

Image analysis (feature extraction and classifier)

Truth	Machine classification			Total	Recall
	Clear	Has crystal	Other		
Clear	2,841	20	26	2,887	0.984
Has crystal	99	1,507	273	1,879	0.802
Other	327	1,132	7,605	9,064	0.839
Total	3,267	2,659	7,904	13,830	
Precision	0.870	0.567	0.962		

Truth	Machine classification			Total	Recall
	Clear	Precip only	Other		
Clear	2,825	7	55	2,887	0.979
Precip only	22	2,571	304	2,897	0.887
Other	290	455	3,127	3,872	0.808
Total	3,137	3,033	3,486	9,656	
Precision	0.901	0.848	0.897		

J Struct Funct Genomics (2010) 11:61–69
DOI 10.1007/s10969-009-9076-9

Protein crystallization analysis on the World Community Grid

Christian A. Cumbaa · Igor Jurisica

Received: 1 September 2009 / Accepted: 30 December 2009 / Published online: 14 January 2010

Abstract We have developed an image-analysis and classification system for automatically scoring images from high-throughput protein crystallization trials. Image analysis for this system is performed by the *Help Conquer Cancer* (HCC) project on the World Community Grid. HCC calculates 12,375 distinct image features on microbatch-under-oil images from the Hauptman-Woodward Medical Research Institute's High-Throughput Screening Laboratory. Using HCC-computed image features and a massive training set of 165,351 hand-scored images, we have trained multiple Random Forest classifiers that accurately recognize multiple crystallization outcomes, including crystals, clear drops, precipitate, and others. The system successfully recognizes 80% of crystal-bearing images, 89% of precipitate images, and 98% of clear drops.

Keywords Image analysis · High-throughput protein crystallization

Abbreviations

HWI	Hauptman-woodward medical research institute
WCG	World community grid
HCC	Help conquer cancer
GLCM	Grey-level co-occurrence matrix
RF	Random forests

Introduction

Protein crystallization is a difficult step in the structural-crystallographic pipeline. Lacking specific theories that map a target protein's physico-chemical properties to a successful crystallization cocktail, the structural genomics community uses high-throughput protein crystallization screens to test targets against hundreds or thousands of cocktails. The Hauptman-Woodward Medical Research Institute's (HWI) High-Throughput Screening Laboratory uses the microbatch-under-oil technique to test 1,536 cocktails per protein on a single plate [9]. Robotic pipetting and imaging systems efficiently process dozens of protein samples (and thus tens of thousands of images) per day. The bottleneck in this process is in the scoring of each image—recognizing crystal growth or other outcomes in an image currently requires visual review by a human expert. To-date, HWI has generated over 100 million images, representing more than 15 million distinct protein/cocktail trials over 12,000 proteins.

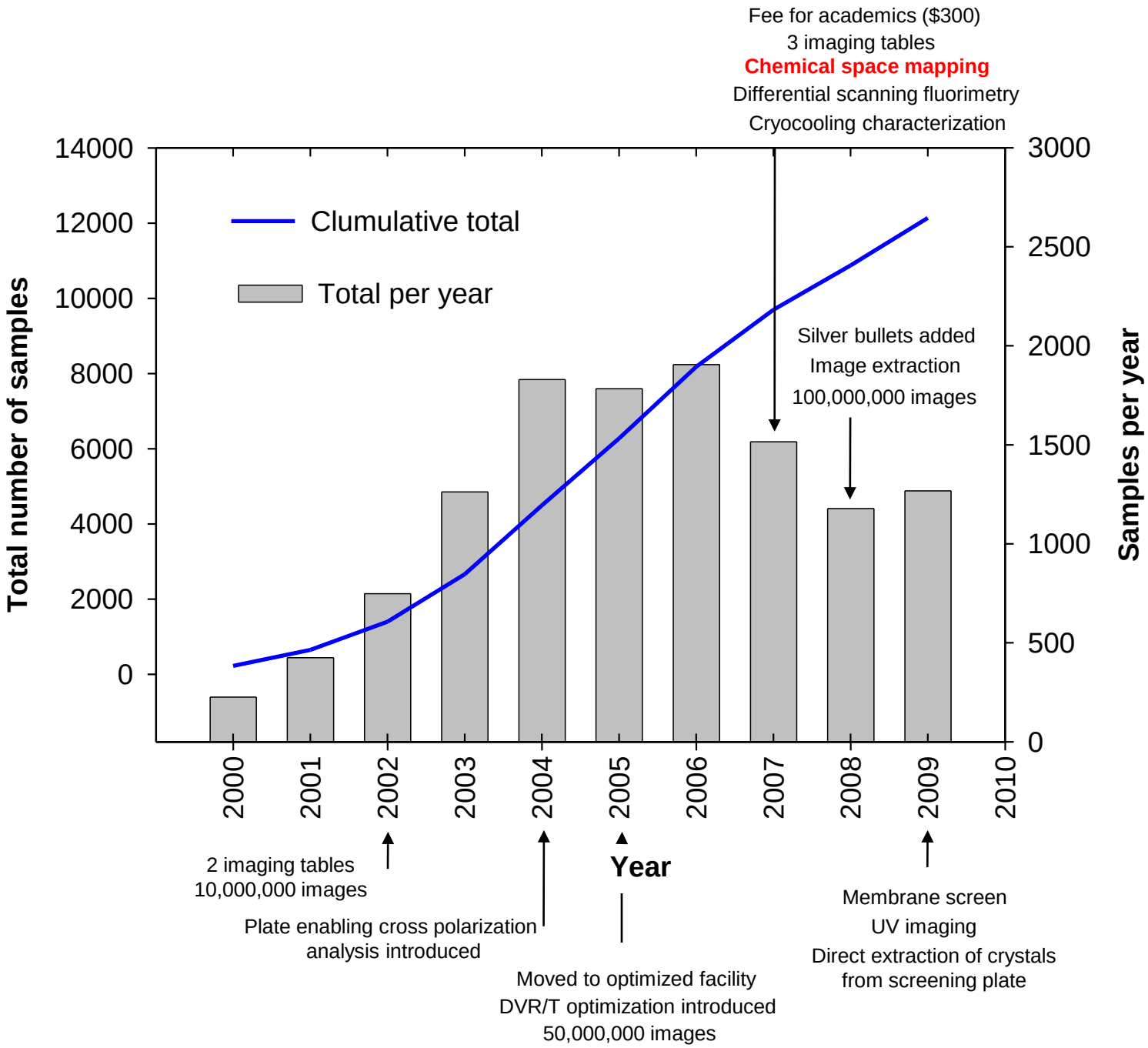
We describe here a method developed for automatically scoring protein-crystallization-trial images against multiple crystallization outcomes. Accurate, automated scoring of protein crystallization trials improves the protein crystallization process in several ways. The technology immediately improves throughput in existing screens by removing or

Electronic supplementary material The online version of this article (doi:10.1007/s10969-009-9076-9) contains supplementary material, which is available to authorized users.

C. A. Cumbaa · I. Jurisica (✉)
Division of Signaling Biology, Ontario Cancer Institute, University Health Network, Toronto Medical Discovery Tower, 9-305, 101 College Street, Toronto, ON M5G 1L7, Canada
e-mail: juris@ai.utoronto.ca

I. Jurisica
Department of Computer Science, University of Toronto, Toronto, ON, Canada

I. Jurisica
Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada


Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
14000
12000
10000
8000
6000
4000
2000
0
3000
2500
2000
1500
1000
500
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate

Chemical space mapping

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The application and use of chemical space mapping to interpret crystallization screening results

Edward H. Snell,[a,b]* Ray M. Nagel,[a] Ann Wojtaszcyk,[a] Hugh O'Neill,[c] Jennifer L. Wolfley[a] and Joseph R. Luft[a,b]

[a]Hauptman–Woodward Medical Research Institute, 700 Ellicott Street, Buffalo, NY 14203, USA, [b]Department of Structural Biology, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA, and [c]Center for Structural Molecular Biology, Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Correspondence e-mail: esnell@hwi.buffalo.edu

Received 17 June 2008
Accepted 7 October 2008

Macromolecular crystallization screening is an empirical process. It often begins by setting up experiments with a number of chemically diverse cocktails designed to sample chemical space known to promote crystallization. Where a potential crystal is seen a refined screen is set up, optimizing around that condition. By using an incomplete factorial sampling of chemical space to formulate the cocktails and presenting the results graphically, it is possible to readily identify trends relevant to crystallization, coarsely sample the phase diagram and help guide the optimization process. In this paper, chemical space mapping is applied to both single macromolecules and to a diverse set of macromolecules in order to illustrate how visual information is more readily understood and assimilated than the same information presented textually.

computer programs

Journal of
Applied Crystallography
ISSN 0021-8898

Received 17 June 2008
Accepted 9 September 2008

AutoSherlock: a program for effective crystallization data analysis

Raymond M. Nagel,[a] Joseph R. Luft[a,b] and Edward H. Snell[a,b]*

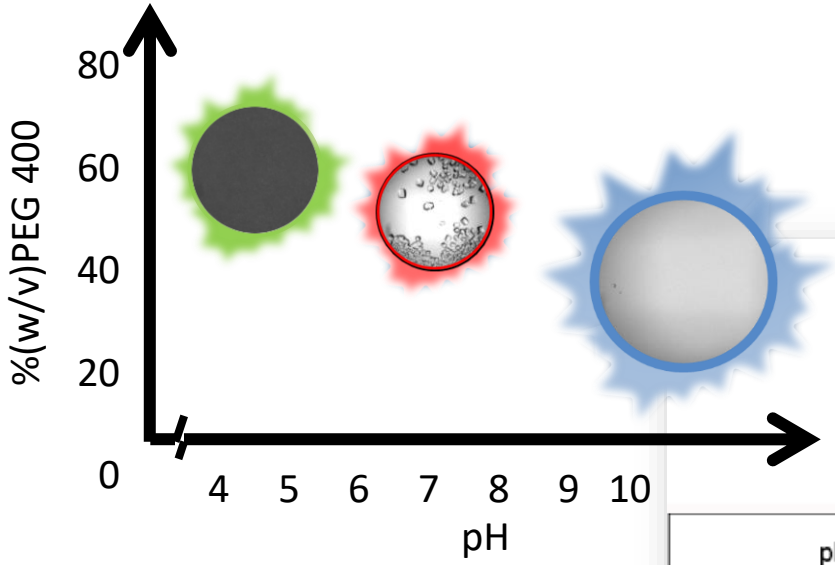
[a]Hauptman–Woodward Medical Research Institute, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA, and [b]Department of Structural Biology, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA. Correspondence e-mail: esnell@hwi.buffalo.edu

A program, *AutoSherlock*, has been developed to present crystallization screening results in terms of chemical space. This facilitates identification of lead conditions, rational interpretation of results and directions for the optimization of crystallization conditions.

Chemical space mapping

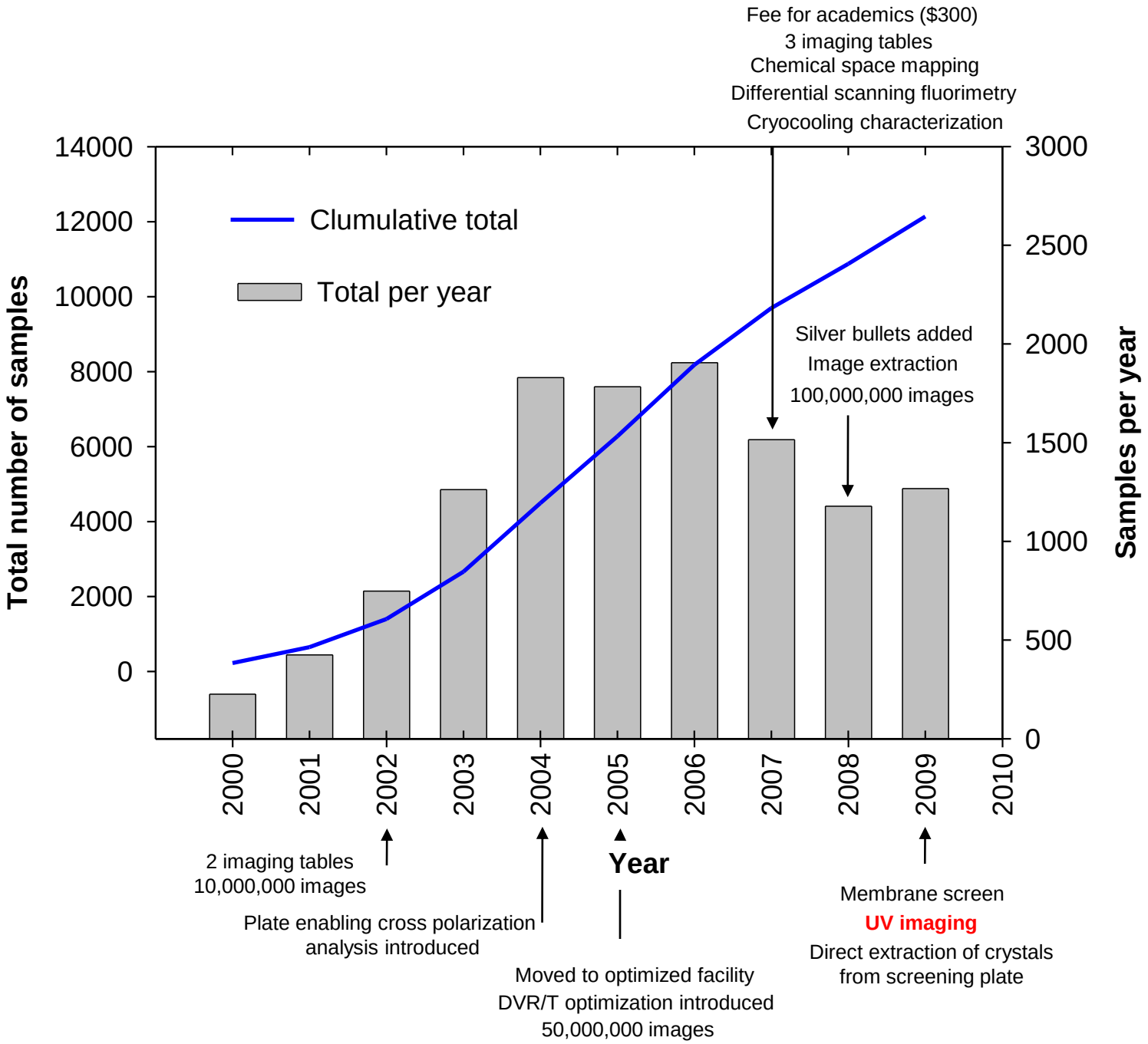
Chemical space mapping (analysis)

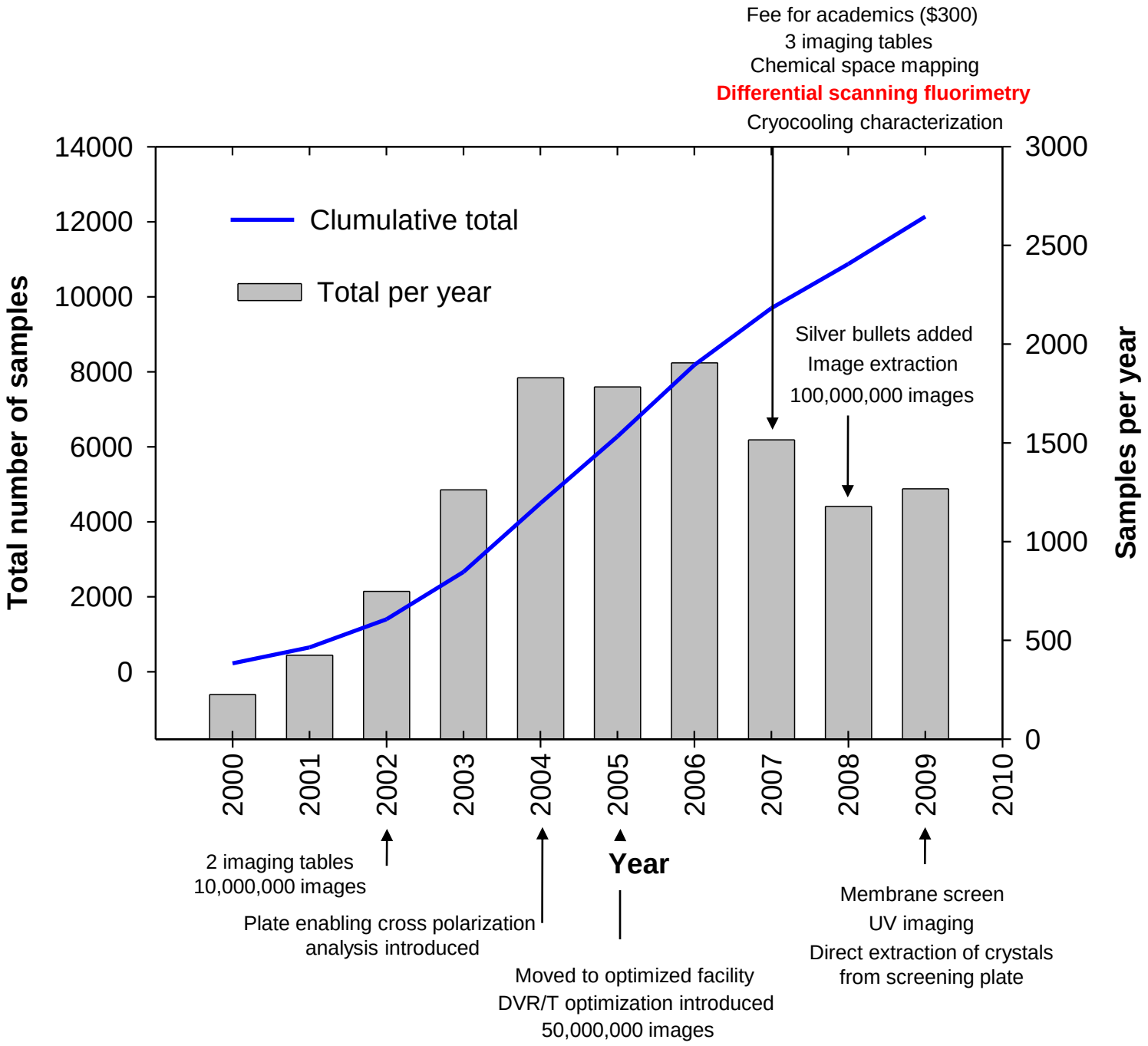
Chemical space mapping (analysis)



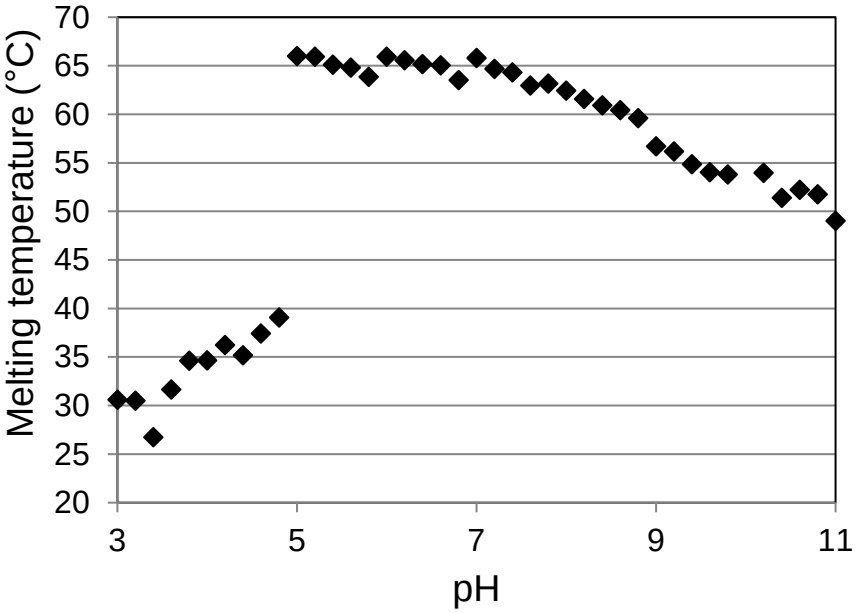
Un-sampled
Crystals
Clear
Precipitate

		PEG 400	Na Citrate	Na Acetate	MES	Bis-Tris Propane	HEPES	Tris	TAPS	CAPS
		Lithium								
pH			4.2	5	6	7	7.5	8	9	10
bromide	20%									●
	40%		●		●			●	●	
	80%							●		


Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
14000
12000
10000
8000
6000
4000
2000
0
3000
2500
2000
1500
1000
500
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate


Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
14000
12000
10000
8000
6000
4000
2000
0
3000
2500
2000
1500
1000
500
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate

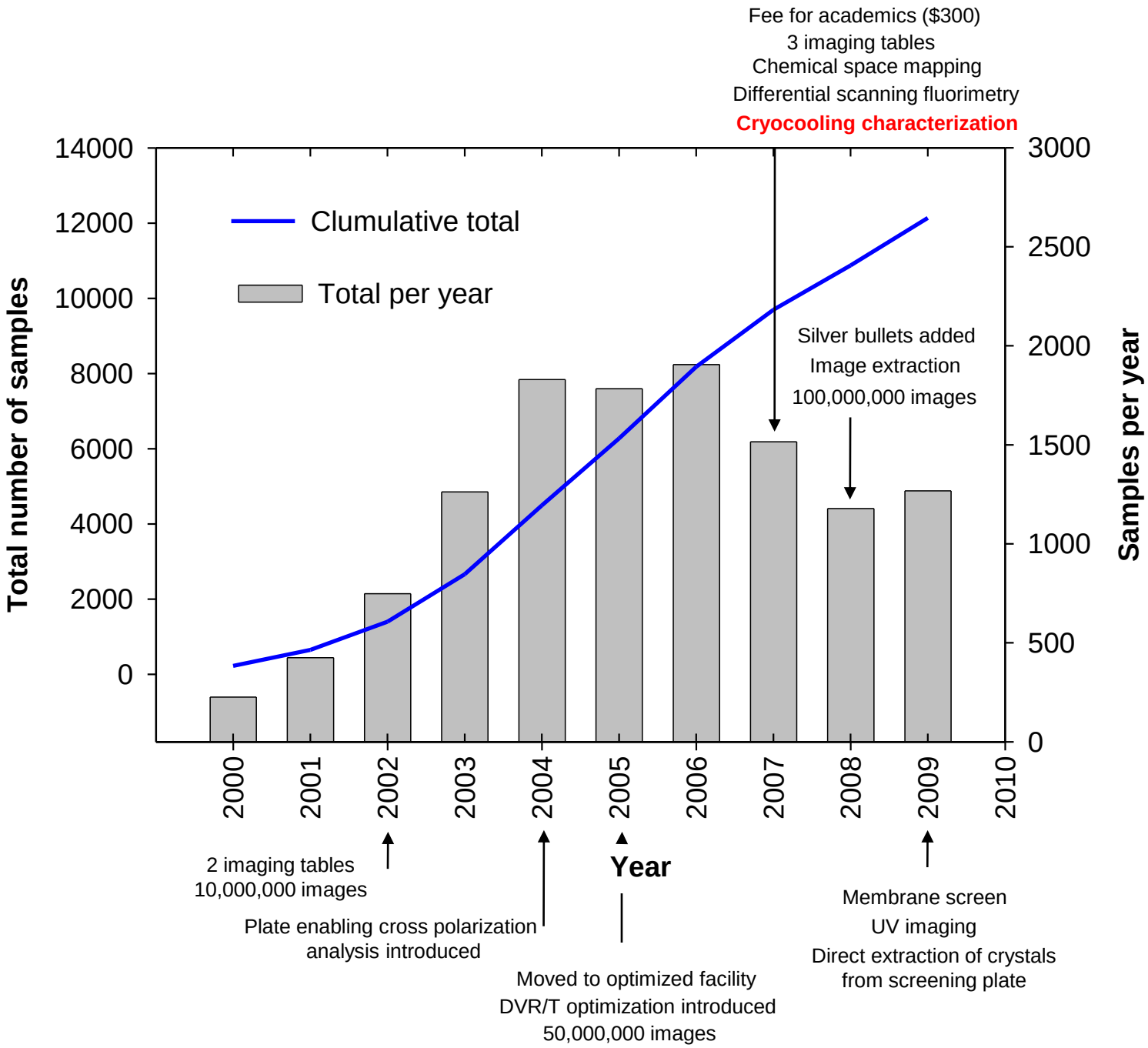
Should we look at multiple hits?

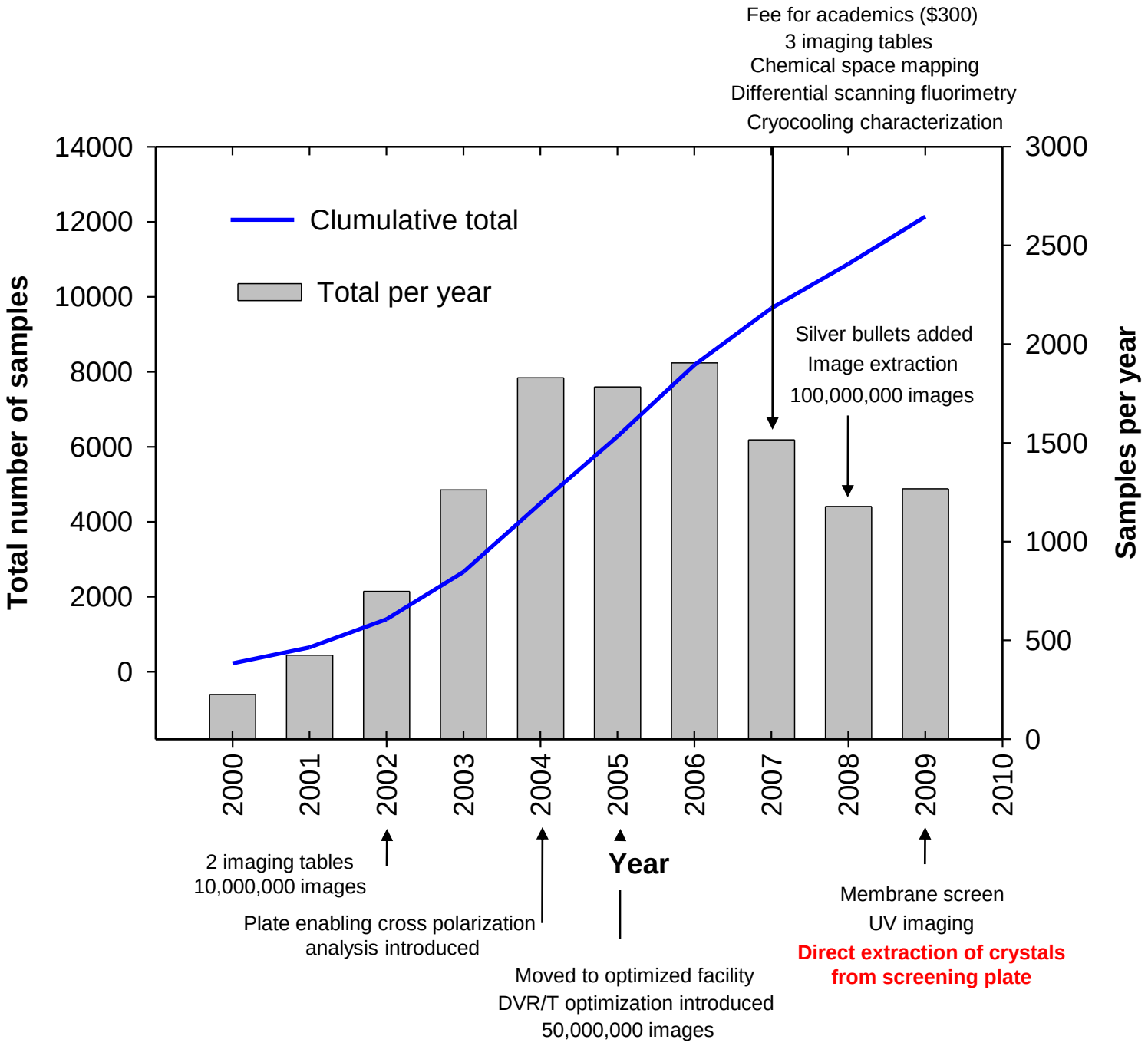


Structure	PO_4 buffer, pH 4.2 (%)	KAc buffer, pH 5.0 (%)	PO_4 buffer, pH 6.0 (%)	Cacod buffer, pH 7.0[a] (%)	PO_4 buffer, pH 9.0 (%)
α-Helix	32	38	37	38	30
β-Strand	21	20	22	20	25
Turns	15	14	11	18	15
Other	31	28	30	24	30

McCabe at al. Enzyme and Microbial Technology,36,70-74 (2005).

Differential scanning fluorimetry has identified a pH driven structural transition. This is in agreement with CD data. Our structural knowledge is of the low pH form.


Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
Year
14000
12000
10000
8000
6000
4000
2000
0
3000
2500
2000
1500
1000
500
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate


Fee for academics ($300)
3 imaging tables
Chemical space mapping
Differential scanning fluorimetry
Cryocooling characterization
Clumulative total
Total per year
Silver bullets added
Image extraction
100,000,000 images
Total number of samples
Samples per year
14000
12000
10000
8000
6000
4000
2000
0
3000
2500
2000
1500
1000
500
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year
2 imaging tables
10,000,000 images
Plate enabling cross polarization
analysis introduced
Moved to optimized facility
DVR/T optimization introduced
50,000,000 images
Membrane screen
UV imaging
Direct extraction of crystals
from screening plate

Does it diffract?

A unique data set

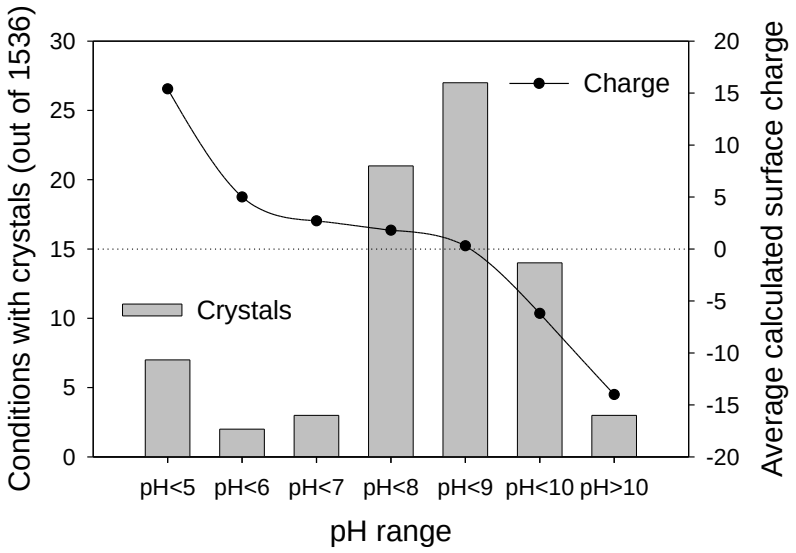
Turing high-throughput to high output

The current success rate is 22%, i.e. 1 out of every 5 samples coming through the laboratory door lead to a structure deposited in the PDB.

Optimist – structural information is obtained in a greater than average number of cases.

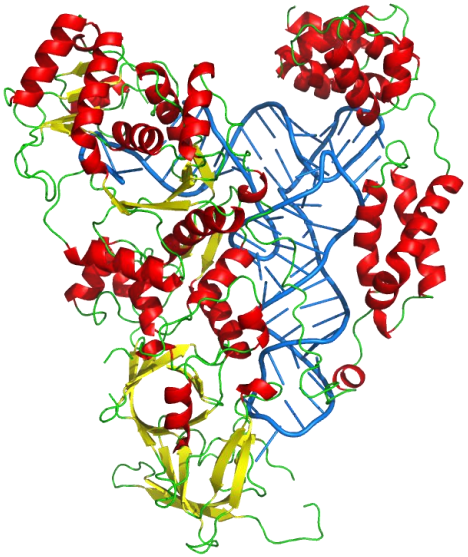
Pessimist – despite having soluble pure samples ~80% of the time we fail.

Predictive patterns



Case study

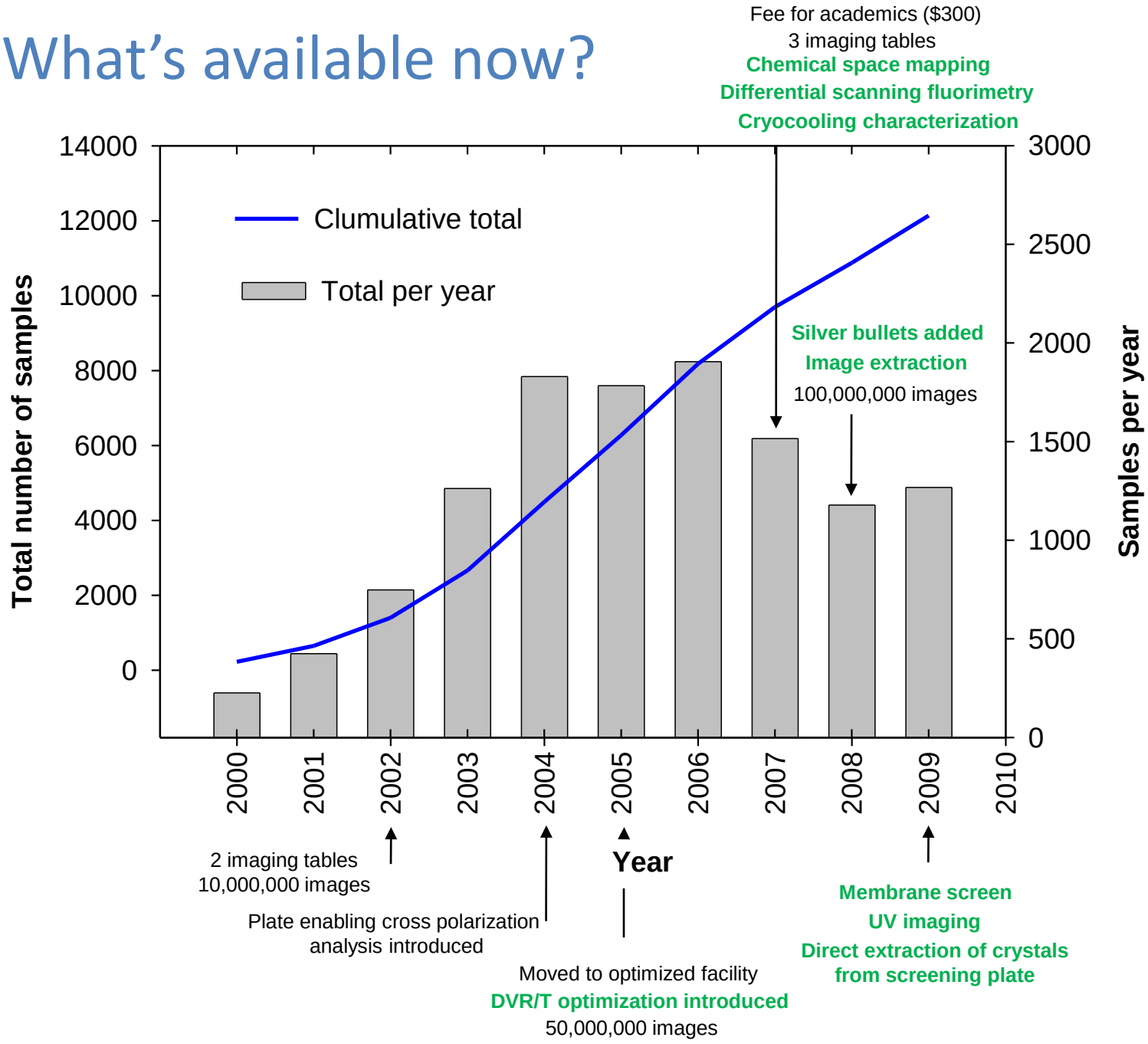
A eukaryotic tRNA transferase



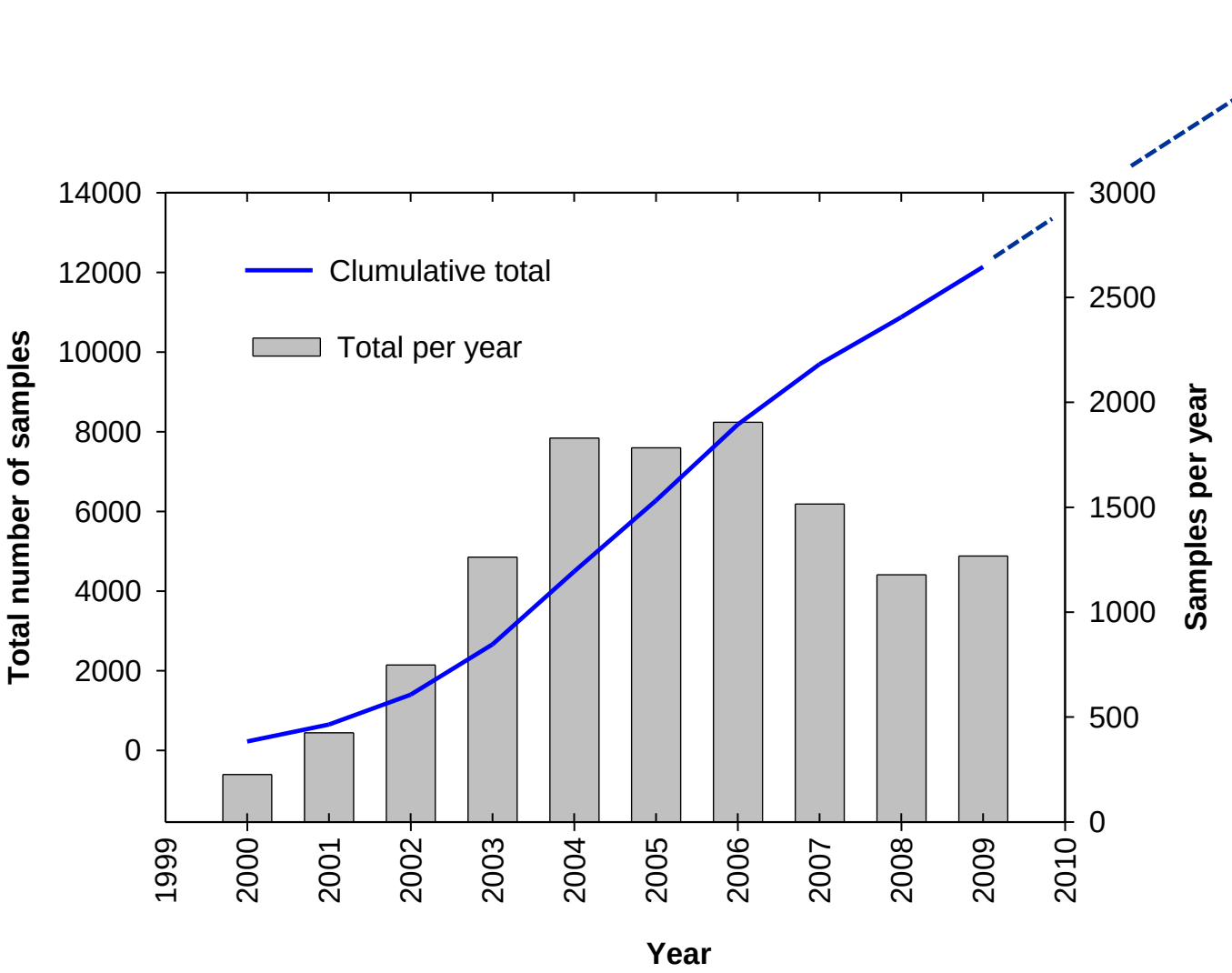
Crystallized in the standard screen but missing some 200 residues known to be in the crystal

Complete pipeline

What's available now?



What we're working on for the future?



SAXS characterization of all samples

Integration of all elements of the pipeline

Addition of color imaging

Protein specific screens

www.xtuition.org

Roles and Acknowledgements

University at Buffalo
The State University of New York

Molecular biology, protein production
Beth Grayhack, Eric Phizicky, Erin Quartley, Stephanie Corretore

Crystal growth, X-ray crystallography
Edward Snell, Joseph Luft, Jen Wolfley, Elizabeth Snell

Small Angle X-ray Scattering
Tom Grant, Edward Snell, Joseph Luft, Hiro Tsuruta

Computational modeling
Tom Grant, Edward Snell

Support and Funding

NIH and DoD support to Edward Snell, NIH PSI support to George DeTitta

With special thanks to the core team now analyzing these results:
Tom Grant, Joseph Luft, Eric Phizicky, and Beth Grayhack

Questions?

esnell@hwi.buffalo.edu