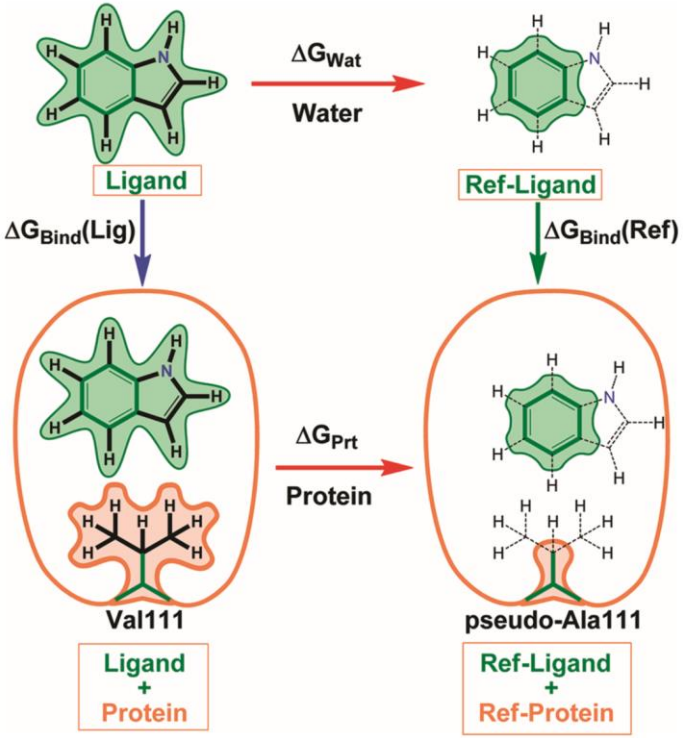
Binding Energy Calculations for Drug Discovery with NVIDIA GPUs

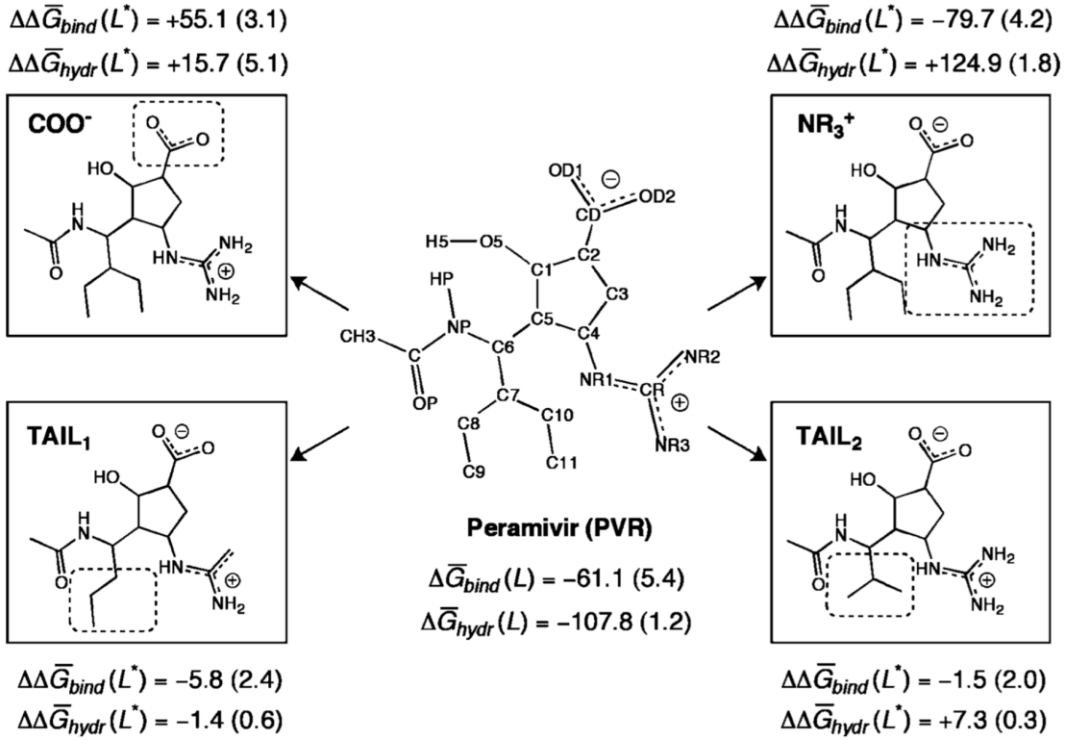
David S. Cerutti and Taisung Lee
Rutgers, the State University of New Jersey

Free Energy Governs Biochemistry, too!

- Fundamental principles of chemistry drive biochemical reactions and recognition.
- Protein-ligand and protein-protein interactions occur amidst a plethora of small molecules and other chemicals.
- The systems are more intricate and harder to quantify than bench chemistry involving only a few compounds in a neat solvent.

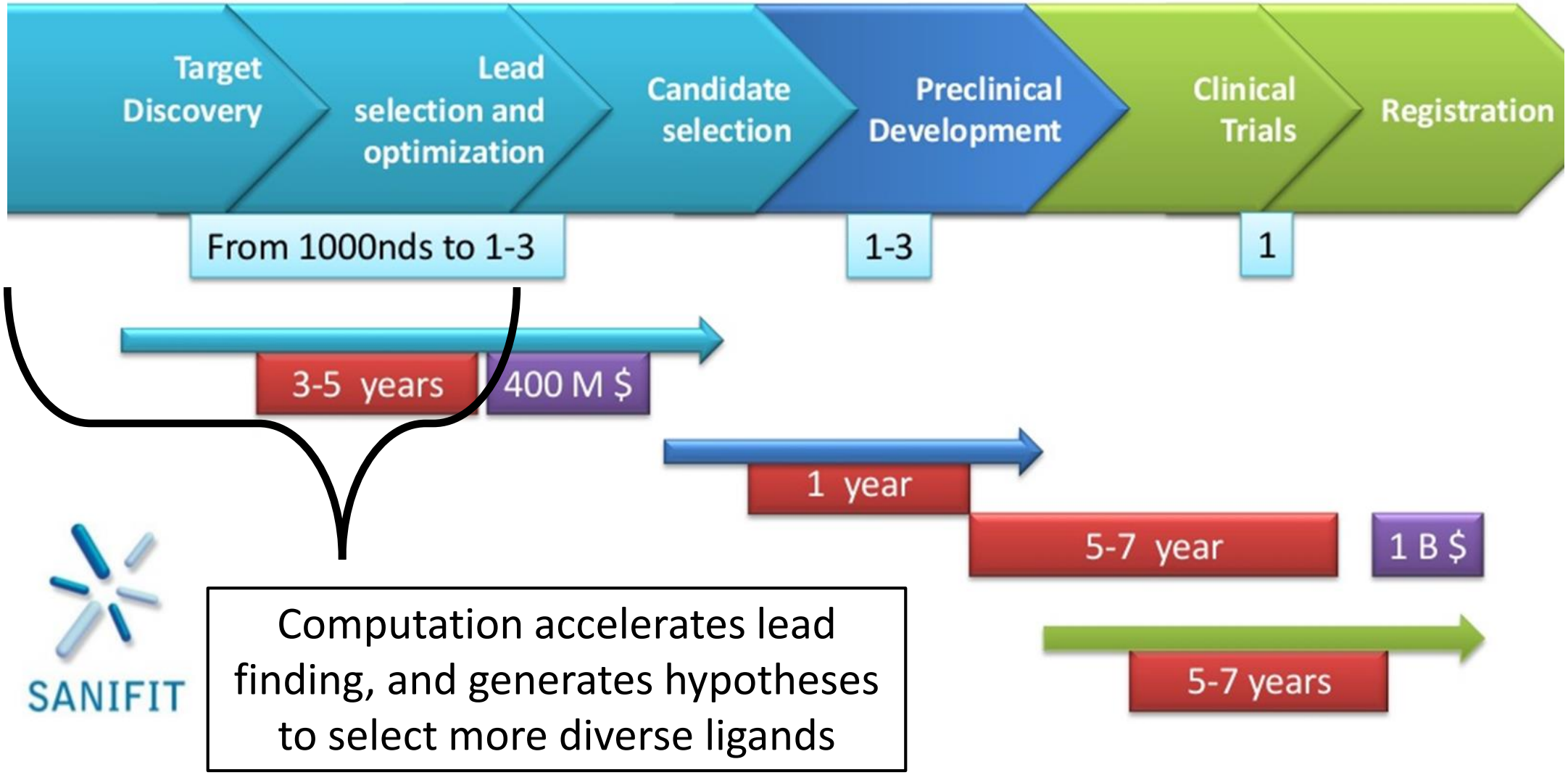


Source: Khavrutskii, I. and Wallqvist, A. (2011) J. Chem. Theory Comput. 7:3001-3011.

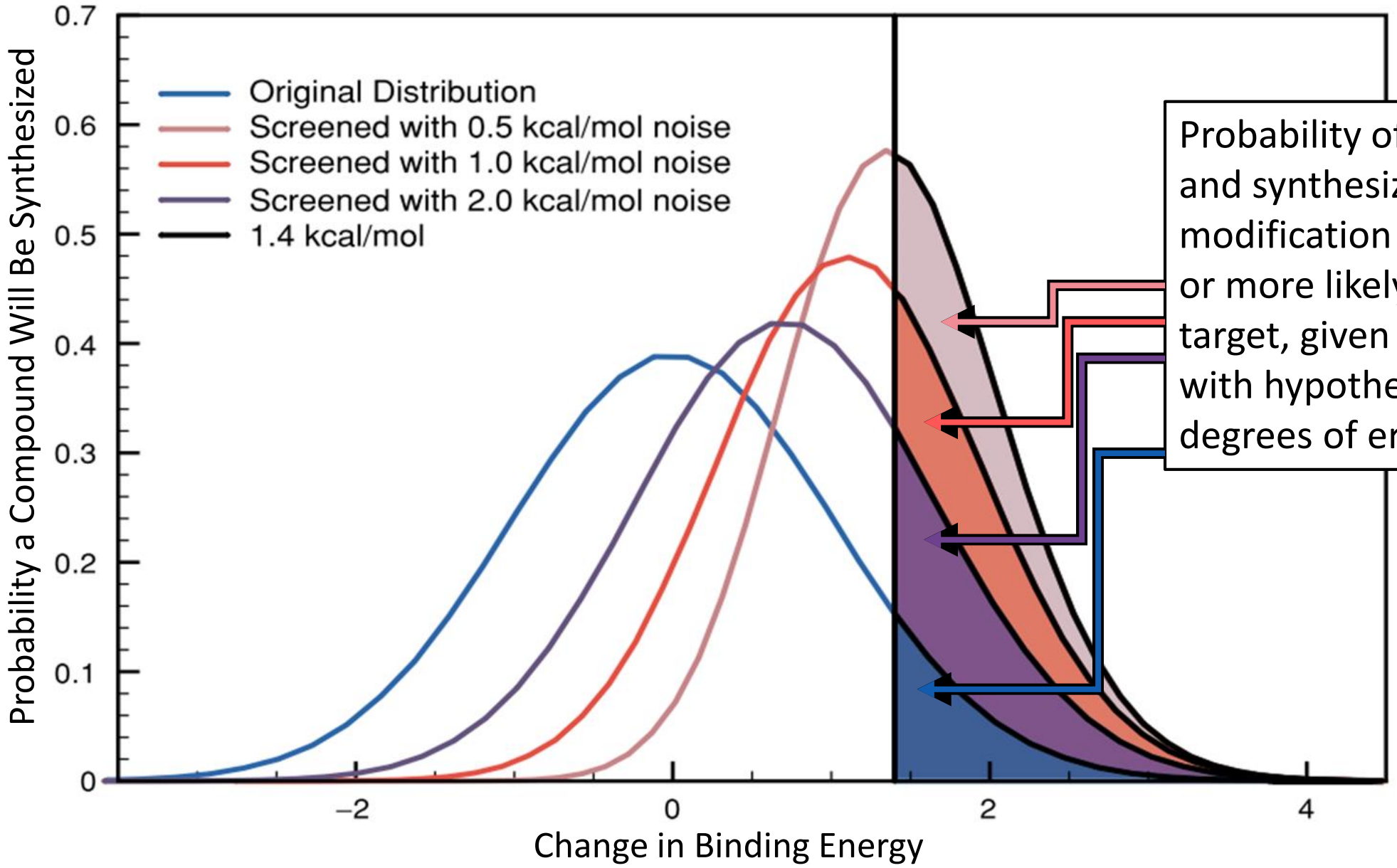


Source: Lawrenz M., Baron R., and McCammon J.A. (2009) J. Chem. Theory Comput. 5:1106-1116.

Most Drugs Take More than a Decade to Develop



Accuracy Increases the Impact of Computation



Probability of proposing and synthesizing a modification that is 10x or more likely to bind the target, given a predictor with hypothetical degrees of error

The Promise, and Reality, of Molecular Simulations

- Molecular simulations offer a powerful alternative to chemical assays when investigating biomolecular interactions.
  - CPU hardware is expensive
  - Human expertise is scarce
  - Power consumption increases as the square of the statistical precision factor
  - Power consumption increases as the square of the chip cycle speed

CPU: 2.5-4GHz, huge cache to feed arithmetic logic units through a handful of threads

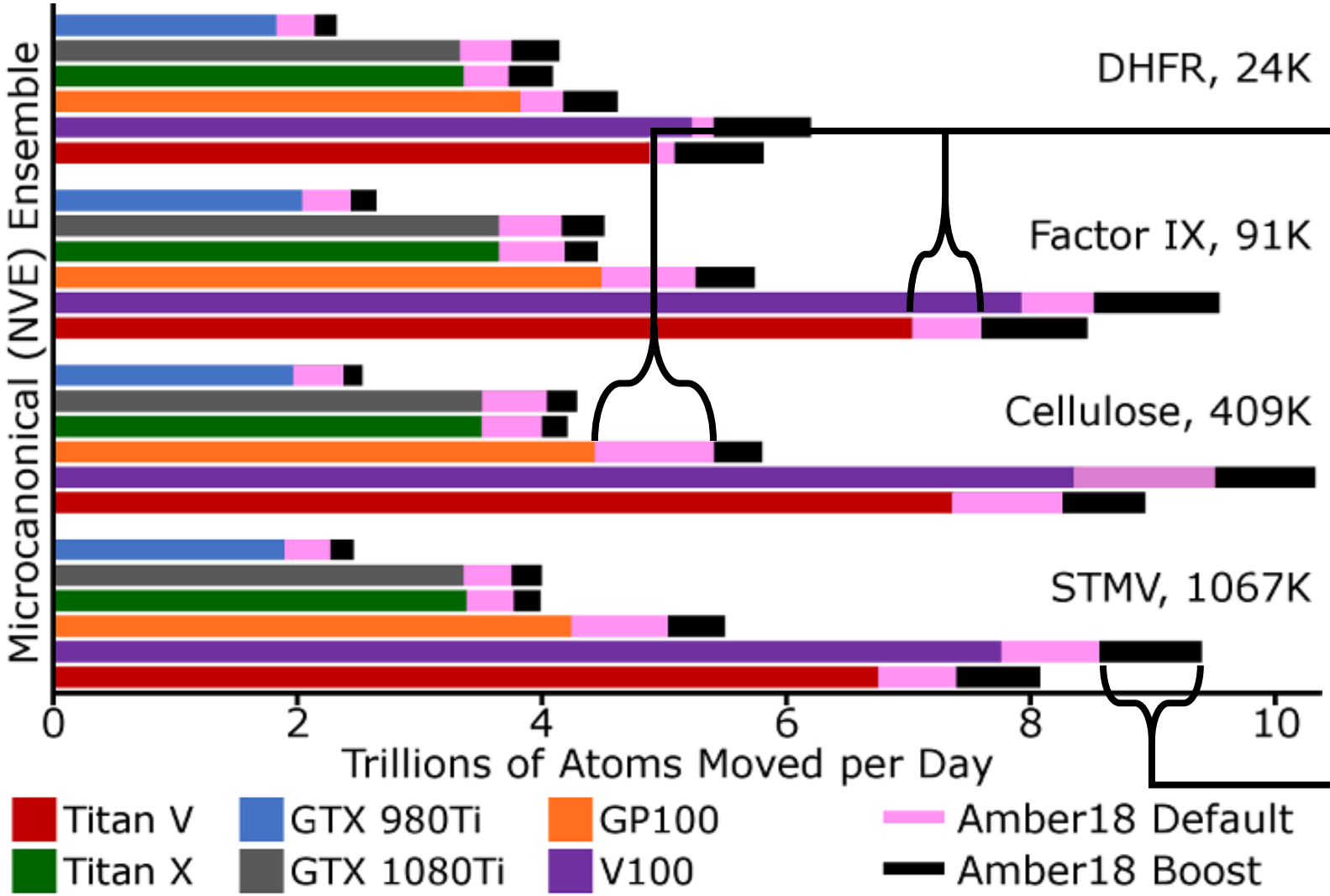
GPU: 1.2-1.6 GHz, scarce cache to free up silicon for thousands of ALUs operated by tens of thousands of threads

The Amber pmemd.cuda Engine

- Scott Legrand's enduring contributions to the Amber community have grown with NVIDIA's device capability.
- Vectorization of algorithms on a GPU is simple to understand and easier to implement than massively parallel MPI
  - Much faster calculations
  - Lower power consumption
  - Most simulations make efficient use of modern GPUs, and NVIDIA's OS-level enhancements maximize throughput for a broad range of system sizes
- The recent Amber18 release includes enhancement of the underlying engine and tremendous expansion of its applicability to computing free energies
  - Thermodynamic integration
  - Replica Exchange and Nudged Elastic Band (enhanced sampling methods)
  - Versatile programming model for a dynamic and growing developer base

Baseline Improvements in Kernel Design

- Mathematical identities and numerical approximations have been rigorously tested to reduce the arithmetic cost and data transfer requirements of the basic engine.



Amber18 improves over Amber16 performance on a range of architectures and systems

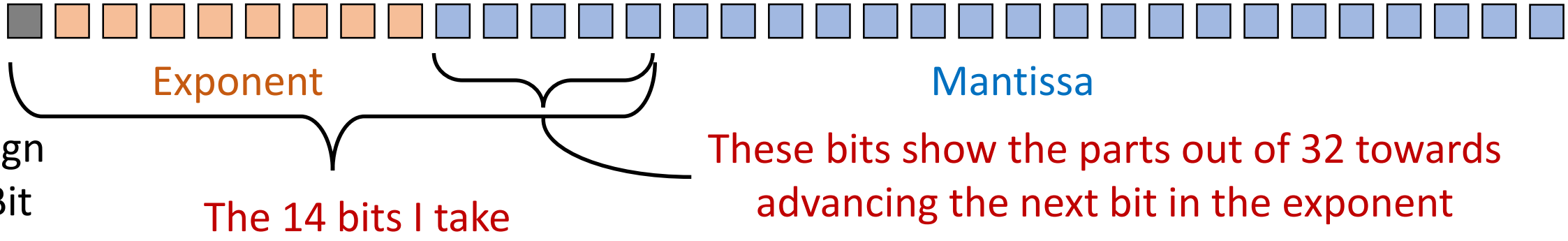
New code features are more permissive, providing alternatives to costly, exact methods. Any run settings contain approximations: the goal is an even distribution of precision.

Spline Approximation to Short-Ranged Functions

- The non-bonded calculations in PME calculations were aided by a spline correction:

$$\frac{kq_iq_j}{r_{ij}} = k\left[\left(\frac{1-\mathrm{erf}(\alpha r_{ij})}{r_{ij}}\right) + \left(\frac{\mathrm{erf}(\alpha r_{ij})}{r_{ij}}\right)\right]$$

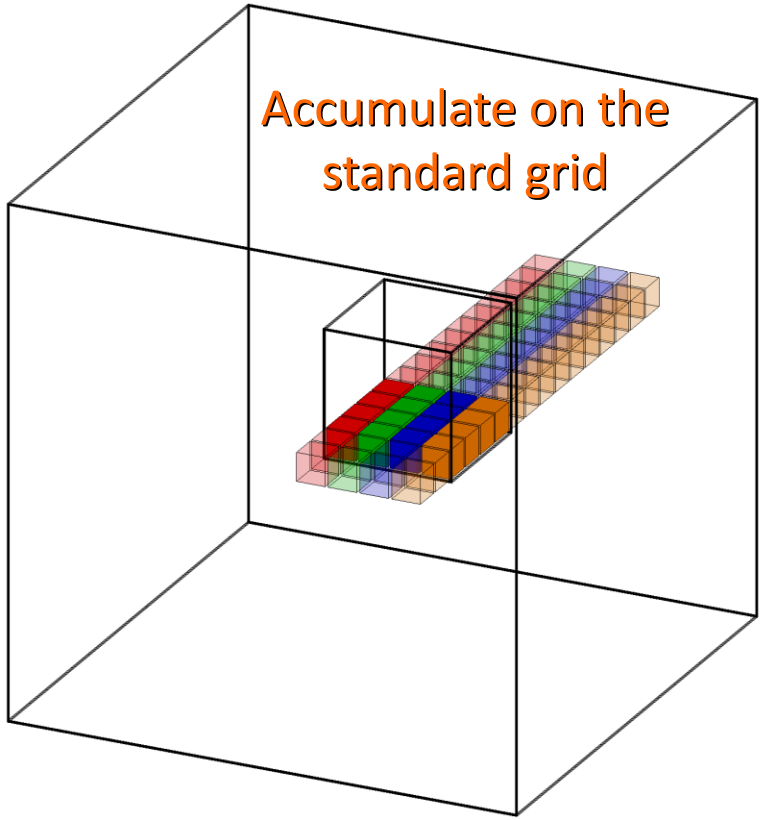
- To differentiate the (1-erf) term on a GPU, we don't have 200kB of cache to devote.



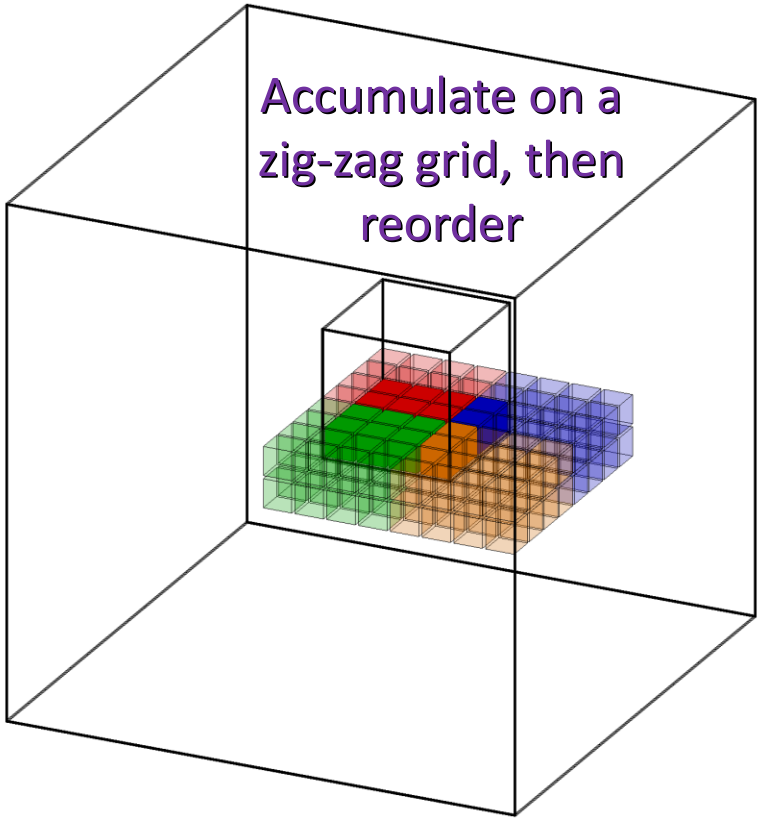

```
float r2, float invr2 = 1.0/r2, float invr4 = invr2*invr2;
int index = __float_as_int(r2) >> 18;
float4 coef = func_table[index];
float result = r2*coef.x + invr2*coef.y + invr4*coef.z +
               coef.w;
```

Improved Coherence in Particle-Mesh Interpolation

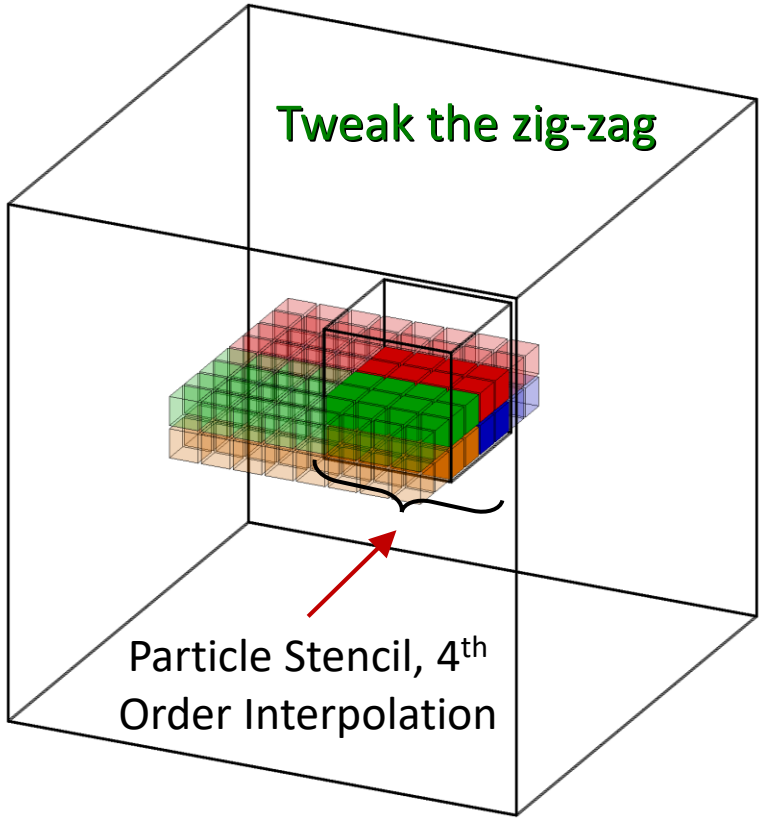
- Early on, the FFT was a bottleneck, but cufft has really come along.
- The problem is getting particles onto the mesh.



Naïve method: each atom writes to 16 sectors



Revised method: atoms write to 2-12 sectors



Current method: atoms write to 4-8 sectors

Changes to the SPFP Precision Model

- The biggest limitation on SPFP precision is the conversion of fractional coordinates calculated in fp64 to fp32. Atom positions must be represented to within 1 part in 16,777,216 of the box size for various non-bonded computations.

Lose 3 bits' precision in pixelation

Changes to the SPFP Precision Model

- In Amber16, the charge mesh is calculated in fp32, accumulated as int64, then converted back to fp32 prior to performing the FFT.
- Accumulating as int32 is a negligible loss of precision here, for half the bandwidth.

Lose 6 bits' precision in coloration

For the Wee Ones: CUDA Multi-Process Service

- The MPS feature has been available since Kepler (2012)
- Designed to aid MPI programs running parallel host (CPU) threads that each launch their own kernels on the GPU
- Also enables multiple programs to launch kernels on the same device with better utilization
- Very modest (1-2%) degradation of single program performance on a V100:

System (NVE, 4fs time step, 9Å cutoff)	Serial Performance (ns / day)	MPS, Multiple Jobs Single Job Rate (ns/day) / Total Throughput (% of Serial)			
		1	2	4	8
TrpCage (304 atom GB)	2719	2634 / 97%	2420 / 178%	1899 / 279%	1281 / 377%
Myoglobin (2492 atom GB)	1812	1804 / 100%	1188 / 131%	683 / 151%	344 / 152%
DHFR (24k atom PME)	1061	1045 / 99%	711 / 135%	408 / 154%	222 / 168%
STMV (1067k atom PME)	34.4	34.2 / 99%	17.4 / 101%	8.8 / 102%	~~Memory~~

For the Wee Ones: CUDA Multi-Process Service

- MPS also benefits the Turing architecture, specifically RTX-2080 Ti:

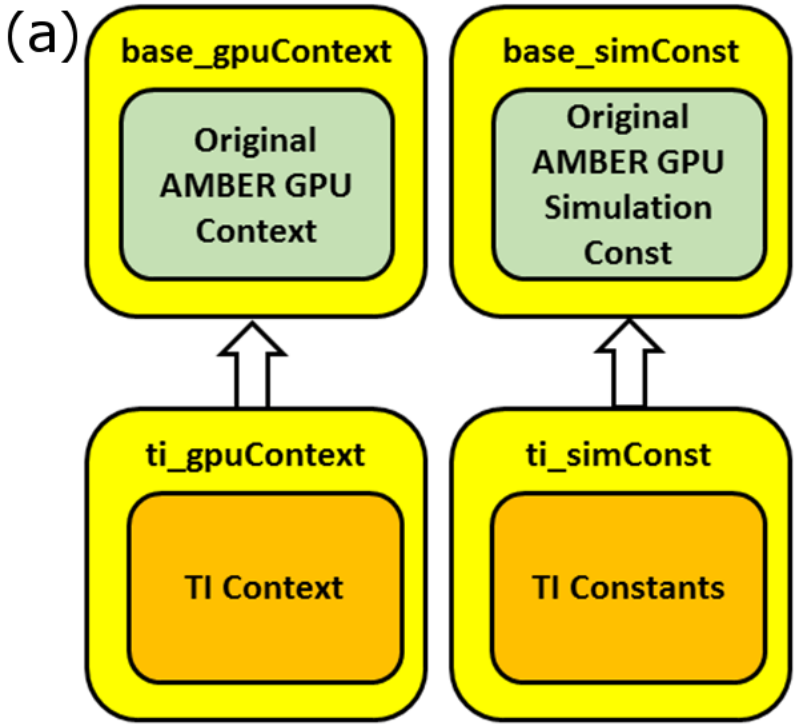
System (NVE, 4fs time step, 9Å cutoff)	Serial Performance (ns / day)	MPS, Multiple Jobs Single Job Rate (ns/day) / Total Throughput (% of Serial)			
		1	2	4	8
TrpCage (304 atom GB)	2317	2272 / 98%	1979 / 171%	1689 / 291%	1209 / 418%
Myoglobin (2492 atom GB)	1100	1104 / 100%	692 / 124%	392 / 143%	202 / 147%
DHFR (24k atom PME)	882	880 / 100%	574 / 130%	325 / 147%	167 / 151%
STMV (1067k atom PME)	25.0	24.0 / 96%	12.0 / 96%	~~Memory~~	~~Memory~~

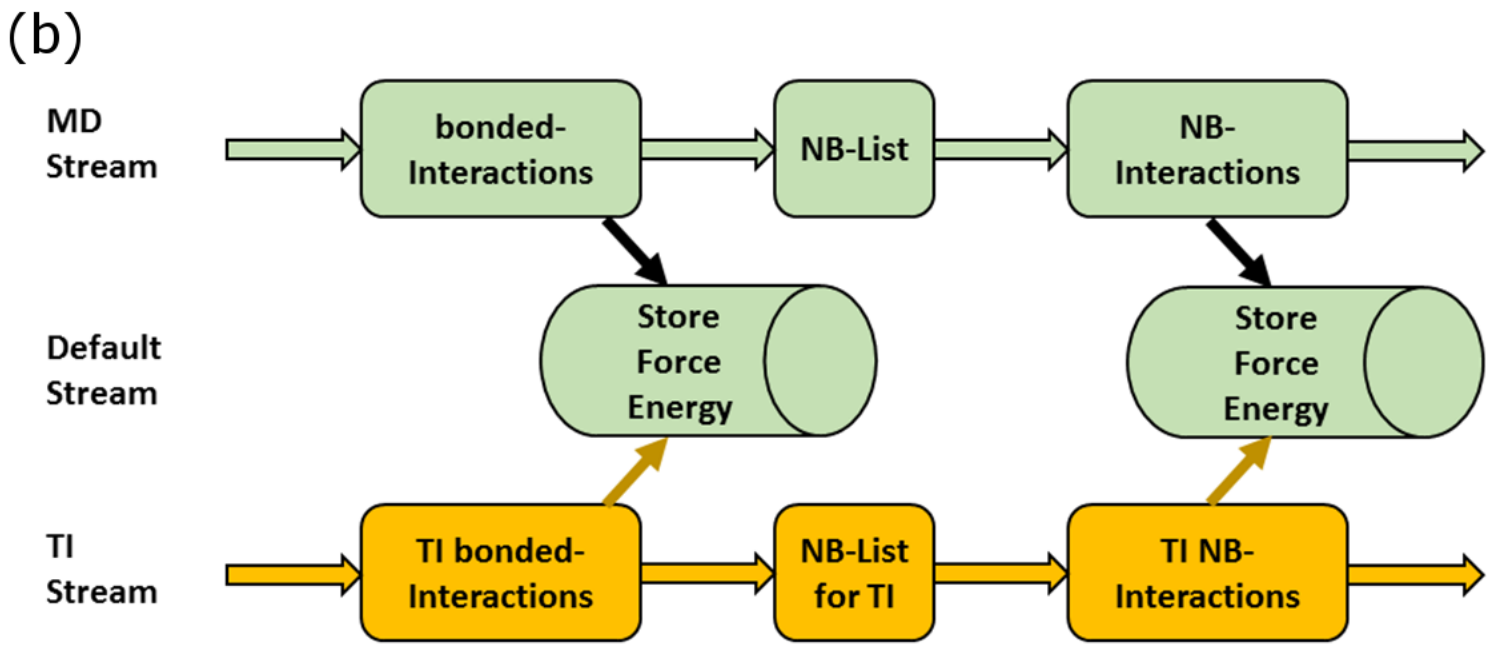
- Changing the block size for the GB non-bonded kernels to improve granularity also helps with MPS throughput on small GB systems.

TrpCage (304 atom GB)	2317	2323 / 100%	2124 / 183%	1808 / 312%	1401 / 484%
Myoglobin (2492 atom GB)	1100	1118 / 101%	730 / 133%	394 / 143%	204 / 148%

Free Energy Calculations with pmemd.cuda GTI

- Design space between moving parts. GTI accesses and benefits from separate development efforts within the pmemd engine but does not interfere with them.
  - C++ class inheritance in (a)
  - Separate CUDA streams in (b)





- Keep the TI module entensible, simplify maintenance and optimization of the engine

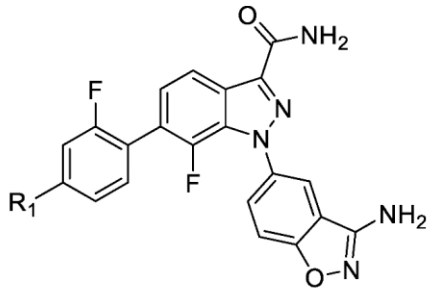
How the Problem Looks to a Pharmaceutical Chemist

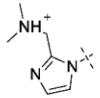
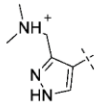
- A ladder of putative ligands, perhaps bearing a common pharmacophore or target site, needs analysis to predict effective characteristics of the drug

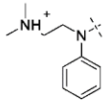
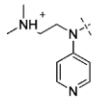
Factor Xa with L51 ligands

Calculation Details:

- 11 λ-windows (5-ns)
- 1fs step
- Constant volume, T=300K
- TIP4P-Ew water
- 46,000 atoms



Ligand	R_1	K_i[b]	ΔG[b]
L51a		15.9	-10.71
L51b		1.4	-12.15
L51bt[a]		1.4	-12.15
L51c		8.1	-11.11
L51d		2255.0	-7.75
L51e		147.0	-9.38

Ligand	R_1	K_i[b]	ΔG[c]
L51f		3.0	-11.70
L51g		4.4	-11.47
L51h[c]		4.1	-11.51
L51i		690.0	-8.46
L51k		4.9	-11.41

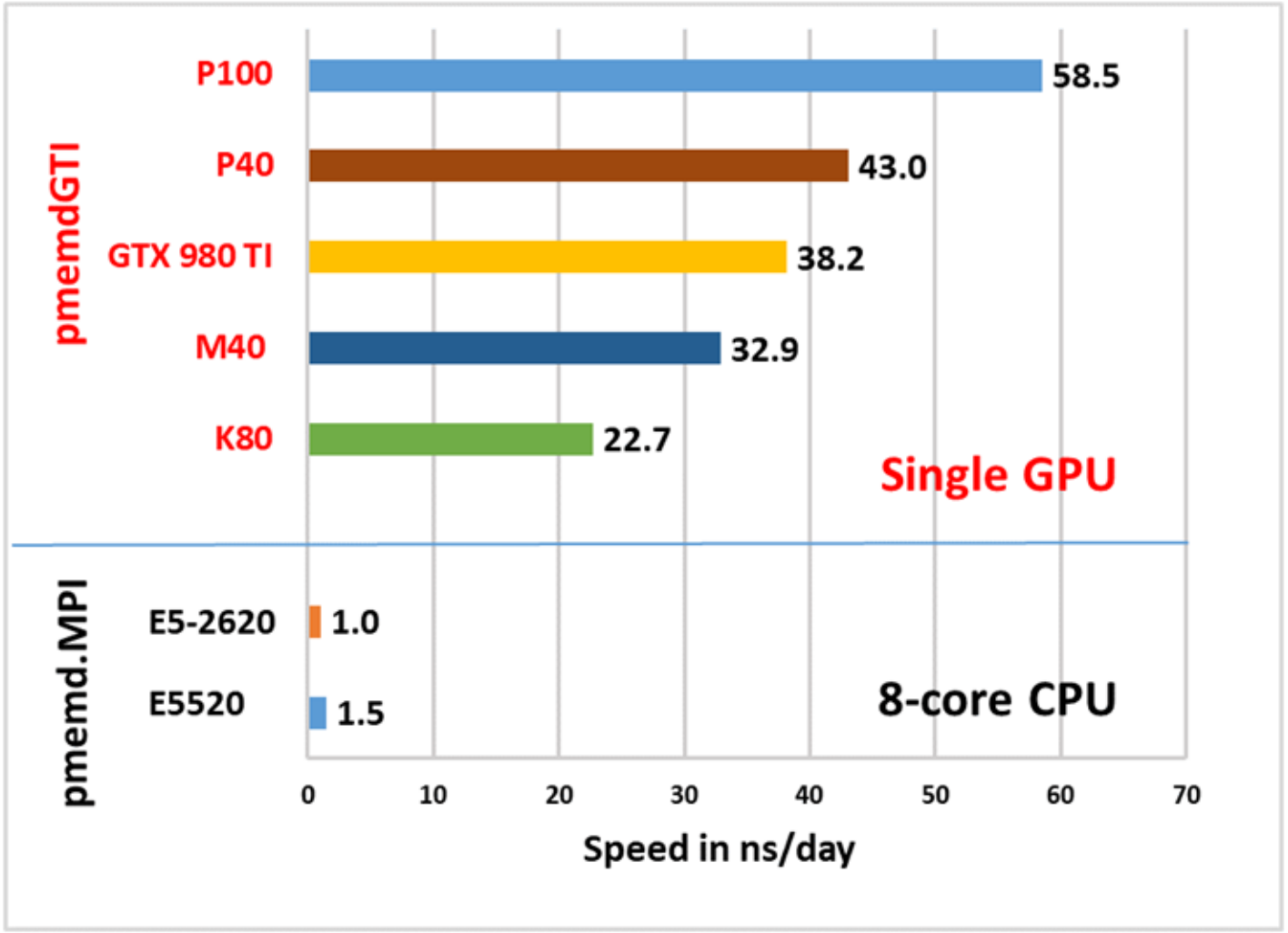
Advantages of GPU Computing

- GPUs offer hundreds of times the performance of a single CPU core, tens of times the performance of a typical multicore compute node, at a fraction of the cost.

Factor Xa with L51 ligands

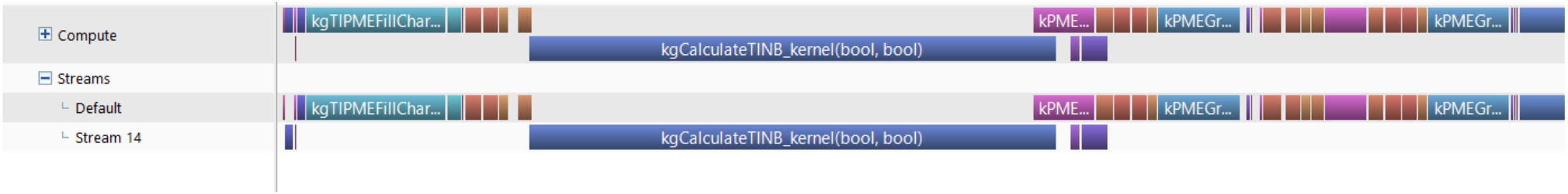
Calculation Details:

- 11 λ-windows (5-ns)
- 1fs step
- Constant volume, T=300K
- TIP4P-Ew water
- 46,000 atoms



Advantages of CUDA Streams

- A drawback of the independent modules is the need to keep separate lists of standard particle pairs and TI particle pairs.
- Separate streams mitigate the utilization problems posed by small TI thread counts



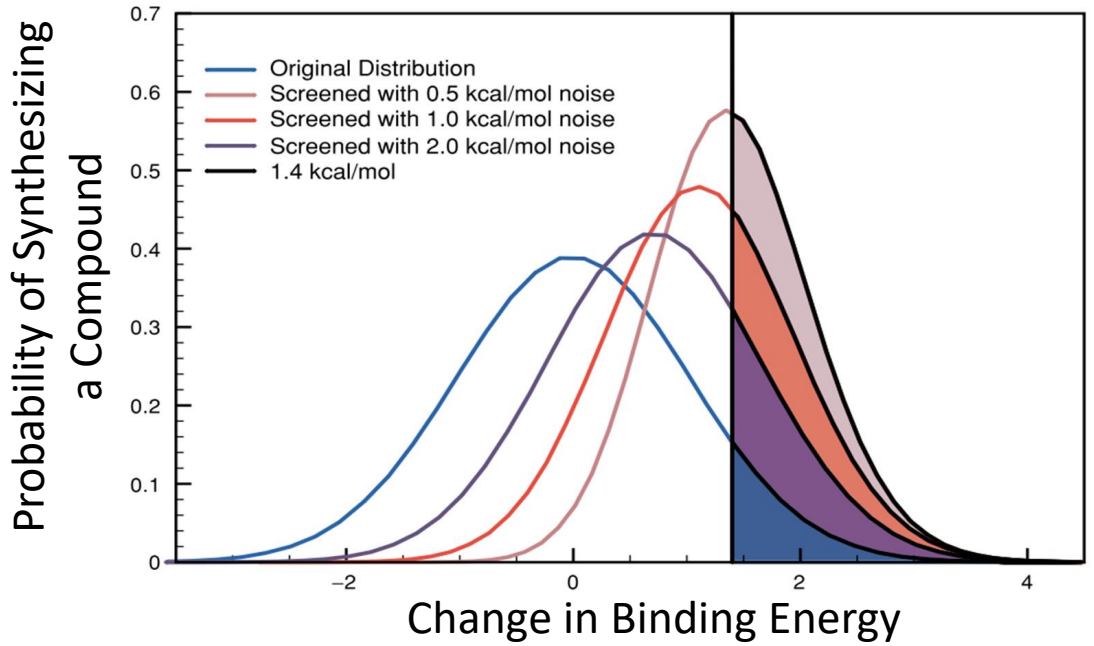
- Challenges: GTI may be independent from the pmemd engine, but results must still be synchronized, which costs time of its own. Determine the optimal sync points.

Performance TI Case: Factor X-a Ligand Mutation, L51a → L51b (Quadro GP100)		
	Amber 16	Amber 18
Standard MD	101.62	113.42
GTI	67.16	70.31
GTI / Multi-Stream:		76.17

The Other Side of the Coin: the Chemical Model

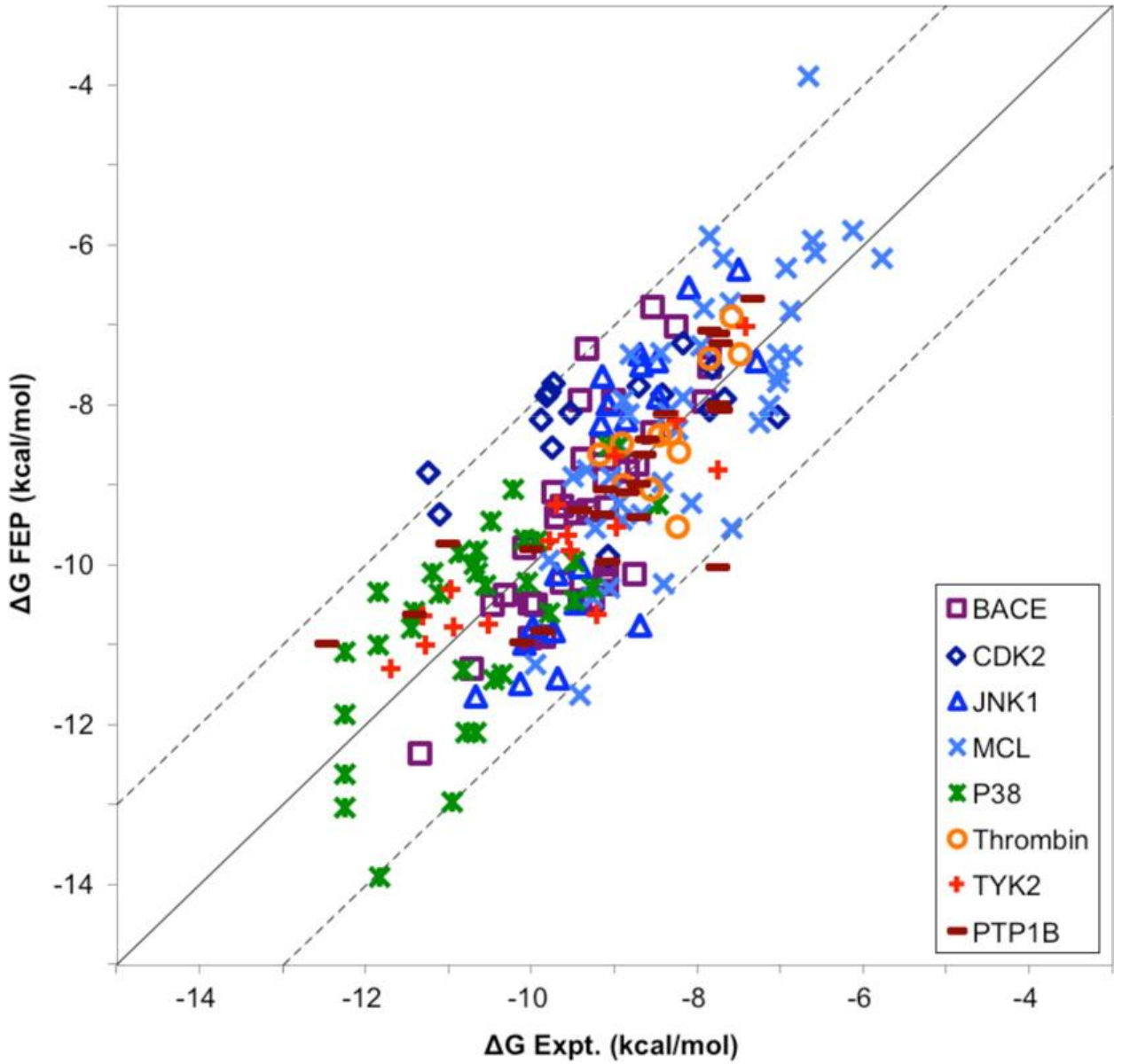
- The GPU TI calculations mentioned in prior slides deliver results to within 0.1 kcal/mol precision. That is well below the accuracy of the chemical model.
  - GPUs enable rapid discrimination between different chemical models for a given problem.
  - Tuning the chemical model is a separate field of study.
  - CUDA can implement chemical models with different features for little more effort than modifying the underlying CPU code.

Precision is necessary but not sufficient. Accuracy unhindered by imprecision will move this distribution to the right, enriching drug lead compounds.



Alternative Technology: Schrödinger's FEP+

- Correlation between FEP-predicted binding free energies and experimental data for all eight systems studied. FEP-predicted binding free energies for most of the ligands are within 1.0 kcal/mol of their experimental values, and only nine of 199 studied ligands deviate from their experimental free energies by more than 2 kcal/mol.

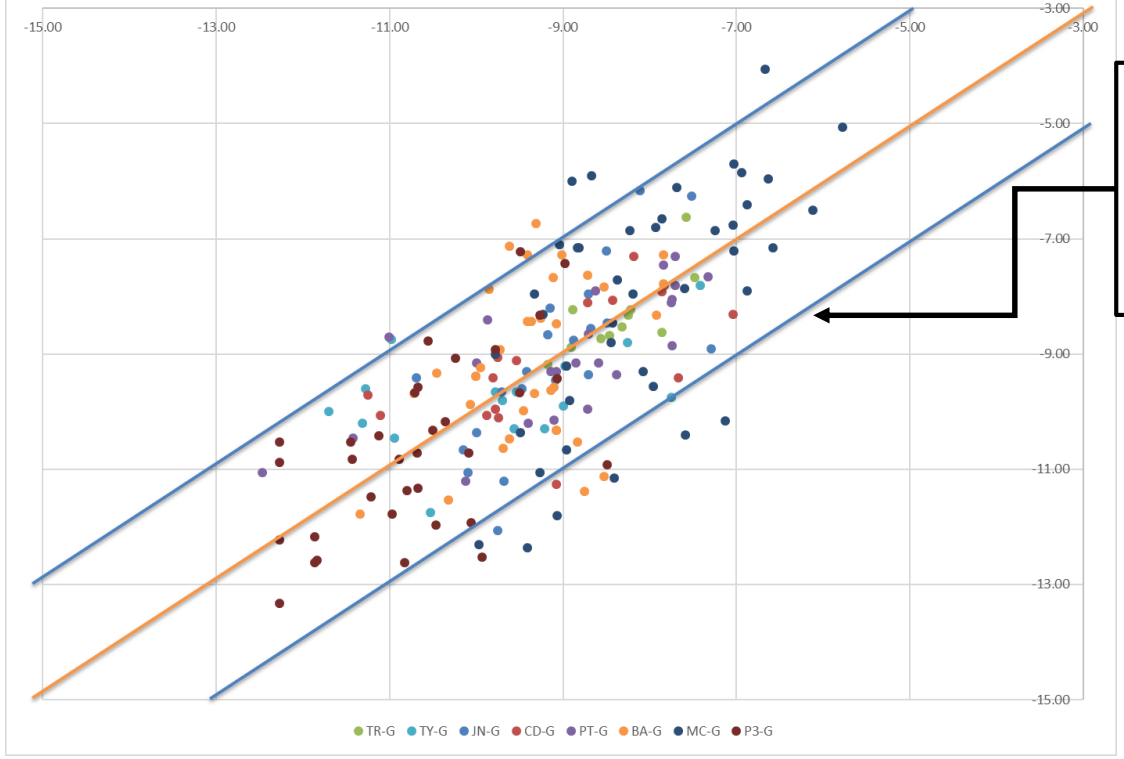


Published in: Lingle Wang; Yujie Wu; Yuqing Deng; Byungchan Kim; Levi Pierce; Goran Krilov; Dmitry Lupyan; Shaughnessy Robinson; Markus K. Dahlgren; Jeremy Greenwood; Donna L. Romero; Craig Masse; Jennifer L. Knight; Thomas Steinbrecher; Thijs Beuming; Wolfgang Damm; Ed Harder; Woody Sherman; Mark Brewer; Ron Wester; Mark Murcko; Leah Frye; Ramy Farid; Teng Lin; David L. Mobley; William L. Jorgensen; Bruce J. Berne; Richard A. Friesner; Robert Abel; *J. Am. Chem. Soc.* **2015,** 137, 2695-2703.
DOI: 10.1021/ja512751q

Software Availability Increases Success Chance

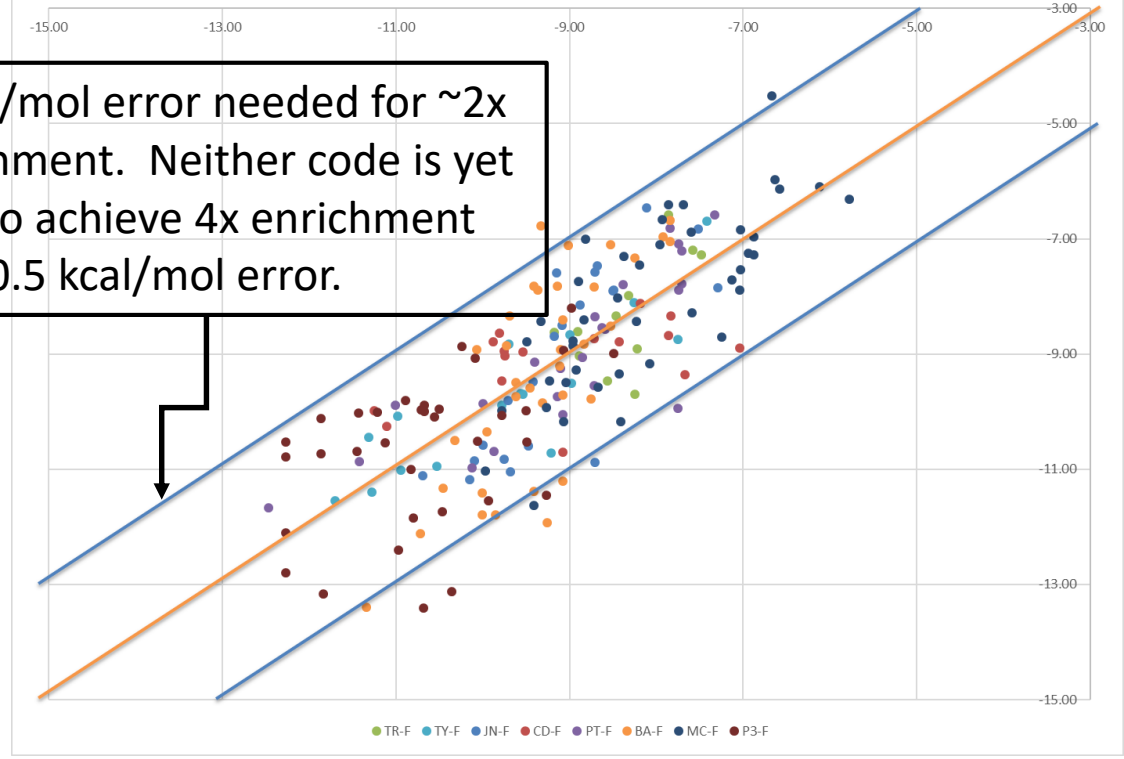
- Amber's pmemd.cuda GTI is an affordable, academic licensed alternative to FEP+. It lacks the FEP Mapper capabilities that are a boon to the performance of Schrödinger's code, and the underlying force field may not be as optimized for drug problems.

AMBER18



2 kcal/mol error needed for ~2x enrichment. Neither code is yet able to achieve 4x enrichment with 0.5 kcal/mol error.

FEP+



Conclusions: A Governing Equation for Our Problem

- CUDA is a powerful and accessible code base for computational chemists
- Optimized FFT libraries, CUDA streams, and L1 cache design by NVIDIA engineers have enabled great leaps in our productivity
- [Taisung : please add your own conclusions here.]

Accuracy (chemical model) and precision (a function of code performance)

X

Affordability and throughput (reduced technical staff labor)

=

Enrichment in the set of lead compounds for synthetic chemists to test

References for pmemd.cuda GTI

GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features.
Lee TS, Cerutti DS, Mermelstein D, Lin C, LeGrand S, Giese TJ, Roitberg A, Case DA, Walker RC, York DM.
J Chem Inf Model. 2018 Oct 22;58(10):2043-2050.

Toward Fast and Accurate Binding Affinity Prediction with pmemdGTI: An Efficient Implementation of GPU-Accelerated Thermodynamic Integration.
Lee TS, Hu Y, Sherborne B, Guo Z, York DM.
J Chem Theory Comput. 2017 Jul 11;13(7):3077-3084

Acknowledgement

- Rutgers, the State University of New Jersey
  - Professors David A. Case and Darrin York
- NVIDIA Corporation
  - Mark Berger, Jon Lefman
  - Peng Wang, Ke Li
  - Norbert Juffa (ret.)
  - Scott Legrand
- Ross Walker, GlaxoSmithKline
- Funding from NIH grants GM122086, (… Taisung state grant number here …)