Genomic prediction with missing pedigrees in single-step GBLUP for production traits in US Holstein

Y. Masuda[1], S. Tsuruta[1], E Nicolazzi[2], A. Legarra[3], D. A. L. Lourenco[1], and I. Misztal[1]

1 University of Georgia, USA
2 Council on Dairy Cattle Breeding (CDCB), USA
3 INRA, France

EAAP Annual Meeting, Aug 29, 2019, Ghent, Belgium

Single-step genomic BLUP (ssGBLUP)

- Combination of genomic and pedigree relationships
- Required: compatibility in scale among matrices

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

- Issue: missing pedigrees
  - Missing elements in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$
  - Compensation by unknown parent groups (UPG) or metafounders (MF)
  - MF: a generalization of UPG
- Several models for UPG/MF in $\mathbf{H}^{-1}$

UPG/MF in ssGBLUP

- Classical UPG
  - $\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$\mathbf{A}^{*}$: $\mathbf{A}^{-1}$ with UPG
(Quaas, 1988)

- Full pedigree UPG (Bradford et al., 2019)
  - $\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & \mathbf{A}_{22}^{-1}\mathbf{Q}_2 \\ 0 & \mathbf{Q}_2'\mathbf{A}_{22}^{-1} & -\mathbf{Q}_2'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \end{bmatrix}$

$\mathbf{Q}_2$: UPG-composite
for genotypes

- Metafounders (Legarra et al., 2015)
  - $\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}_{05}^{-1} - \mathbf{A}_{22}^{\Gamma-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$\mathbf{A}^{\Gamma-1}$: $\mathbf{A}^{-1}$ with MF
$\mathbf{A}_{22}^{\Gamma-1}$: $\mathbf{A}_{22}^{-1}$ with MF
$\mathbf{G}_{05}^{-1}$: with 0.5 allele frequency

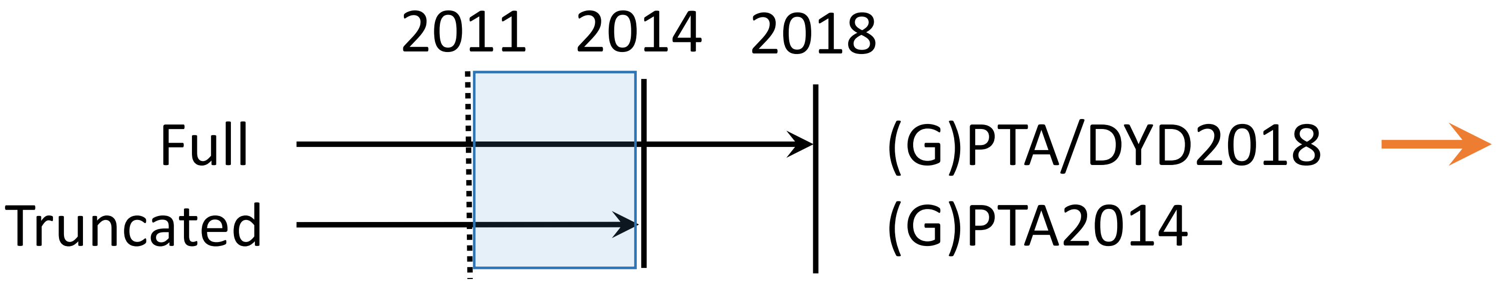
Objectives

To validate the genetic trends and the predictability of young-bull predictions for protein yield in US Holstein with 2.3M genotypes using UPG/MF

Full data in Holstein

	Description	Number of records/animals
Phenotype	**Protein yield** (305-d basis) for US Holstein cows recorded between Jan. 1990 and Dec. 2018	61,229,782
Pedigree	Animals born in Dec. 2018 or earlier (3-gen. back from phenotyped cows) **16 UPGs: sex by 4-yr group**	35,857,897
Genotype	Animals born in Dec. 2018 or earlier (79,294 markers) **15%** with missing sire and/or dam	2,334,951

Cross-validation for bulls



2011 2014 2018

Full → (G)PTA/DYD2018

Truncated → (G)PTA2014

Validation bulls (N=2,315)

- Genotyped bulls with no tested daughters in 2014
- With at least 50 tested daughters in 2018

$$DYD/GPTA2018 = b_1 \times GPTA2014 + b_0$$

- R^2 : validation reliability
- Slope (b_1): Inflation of prediction

Benchmarks

1. DYD classical BLUP
2. DYD ssGBLUP / UPG
3. DYD ssGBLUP / MF

   VanRaden and Wiggans (1991)

4. PTA classical BLUP
5. GPTA ssGBLUP / UPG
6. GPTA ssGBLUP / MF

Model

- Same model as the official evaluation (repeatability model)
- APY for genomic relationships
  - 15,000 core animals (randomly chosen)
- UPG: sex by year-group
  - Full data (16): -1986, -1990, 1994, -1998, -2002, -2006, -2010, and 2011-
  - Truncated (14): -1986, -1990, 1994, -1998, -2002, -2006, and 2007-
- $\mathbf{\Gamma}$: calculated with GLS (Garcia-Baccino et al., 2017)
- Genetic base: all phenotyped cows born in 2005

Solving MME in this study

	OpenMP-based solver	MPI-based solver
Parallelism	OpenMP	MPI and OpenMP
CPU-cores used	6	8
Iteration on data	Data and pedigree files	APY G-inverse files
Genotypes (core animals)	2.3 M (15K)	2.3 M (15K)
Max. memory usage	> 267 GB	> 17 GB
Wall-clock time per PCG round	35 s	39 s
WC time for 600 rounds	5.8 h	6.5 h

The software development is still going on especially for efficiency.

R^2 and b1: DYD/GPTA2018 on GPTA2014

	R2		b1	
Benchmark	**pedUPG**	**MF**	**pedUPG**	**MF**
DYD2018 BLUP	0.67	**±0.00**	0.79	**-0.01**

For validation bulls with at least 50 daughters (N=2315)

R^2 and b1: DYD/GPTA2018 on GPTA2014

	R2		b1	
Benchmark	**pedUPG**	**MF**	**pedUPG**	**MF**
DYD2018 BLUP	0.67	±0.00	0.79	-0.01
DYD2018 pedUPG	0.77	**+0.01**	0.85	**-0.01**
DYD2018 MF	0.78	**+0.01**	0.86	**-0.01**

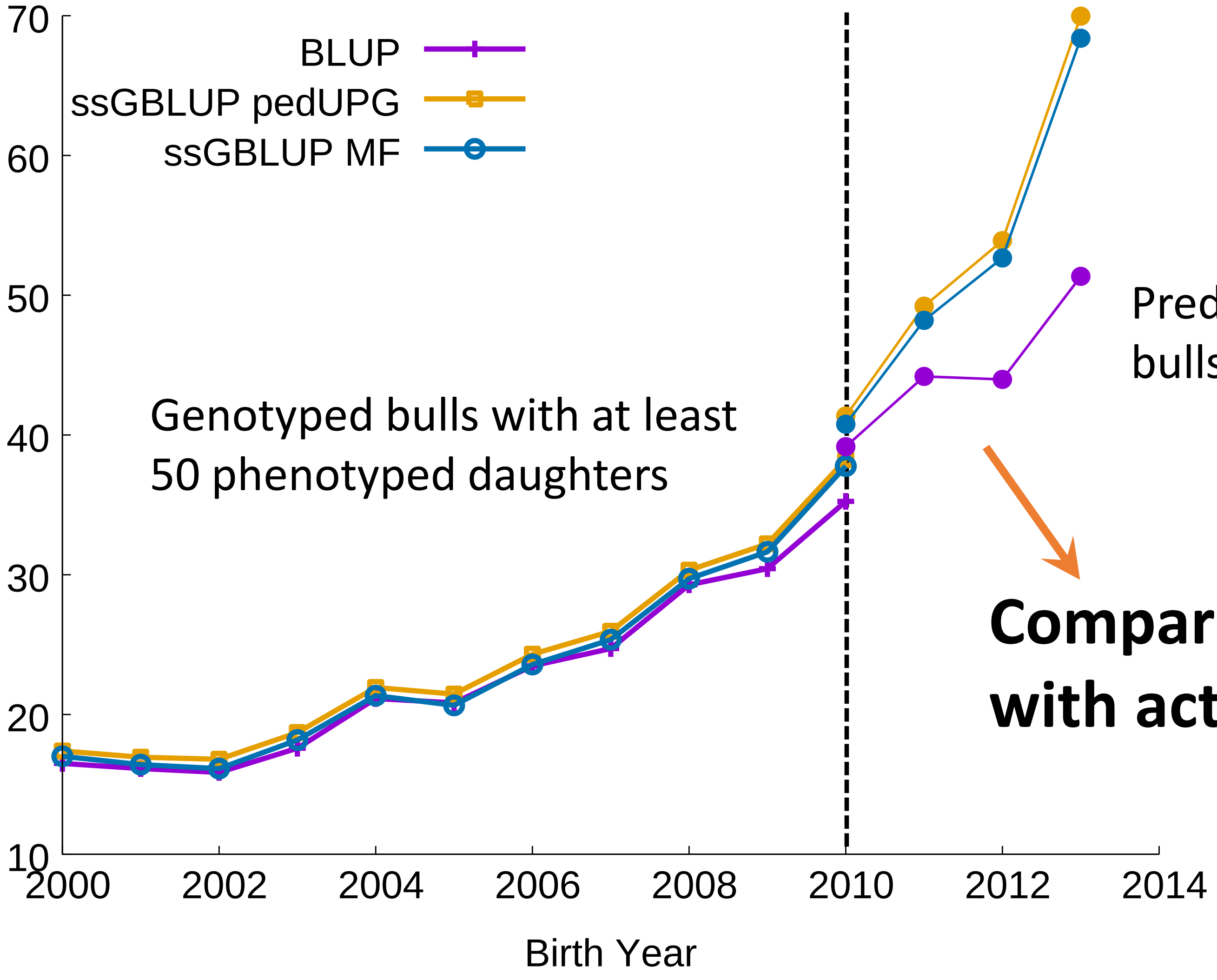
For validation bulls with at least 50 daughters (N=2315)

R^2 and b1: DYD/GPTA2018 on GPTA2014

	R2		**b1**	
Benchmark	**pedUPG**	**MF**	**pedUPG**	**MF**
DYD2018 BLUP	0.67	±0.00	0.79	-0.01
DYD2018 pedUPG	0.77	+0.01	0.85	-0.01
DYD2018 MF	0.78	+0.01	0.86	-0.01
PTA2018 BLUP	0.68	**±0.00**	0.83	**±0.00**
GPTA2018 pedUPG	0.82	**+0.01**	0.90	**-0.01**
GPTA2018 MF	0.83	**+0.01**	0.91	**-0.01**

For validation bulls with at least 50 daughters (N=2315)

Genetic trend for genotyped bulls in 2014

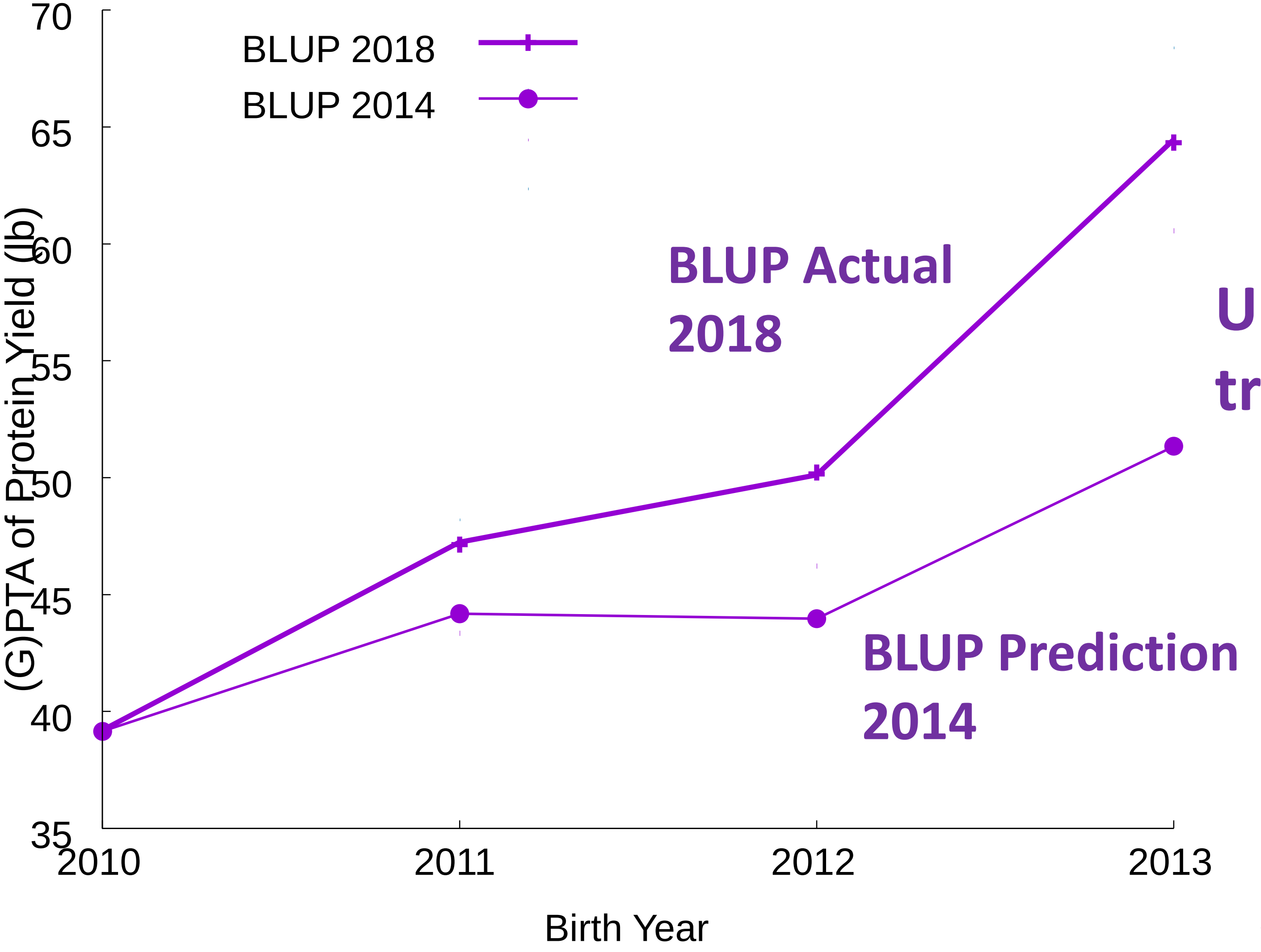


Genotyped bulls with at least 50 phenotyped daughters

Prediction for validation bulls (N=2,315)

Comparison of predictions with actual trends

Predicted and actual trends: BLUP



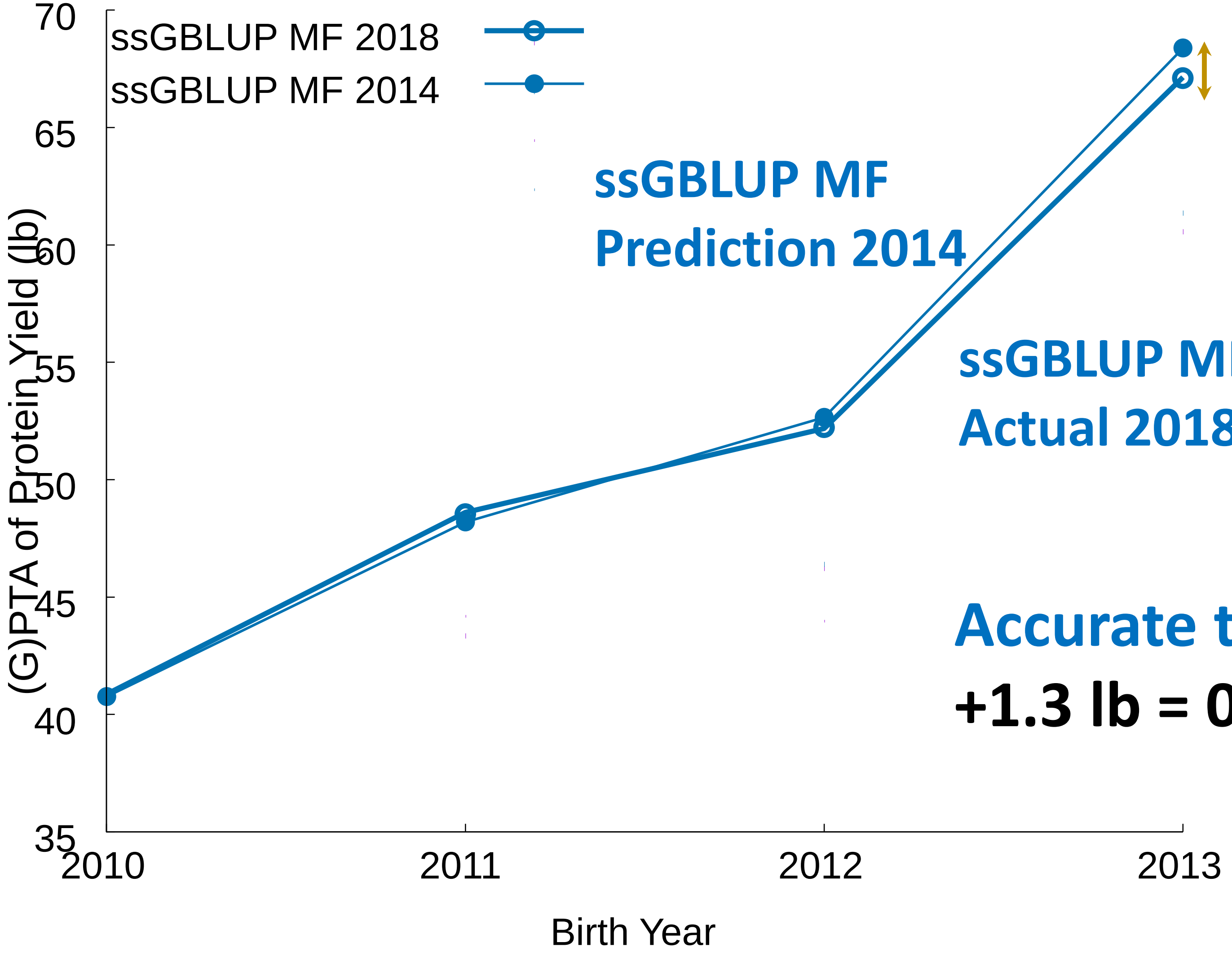
BLUP Actual 2018

BLUP Prediction 2014

Underestimated trend-prediction

* Trends adjusted to 2010 genetic base

Predicted and actual trends: ssGBLUP



ssGBLUP MF Prediction 2014

ssGBLUP MF Actual 2018

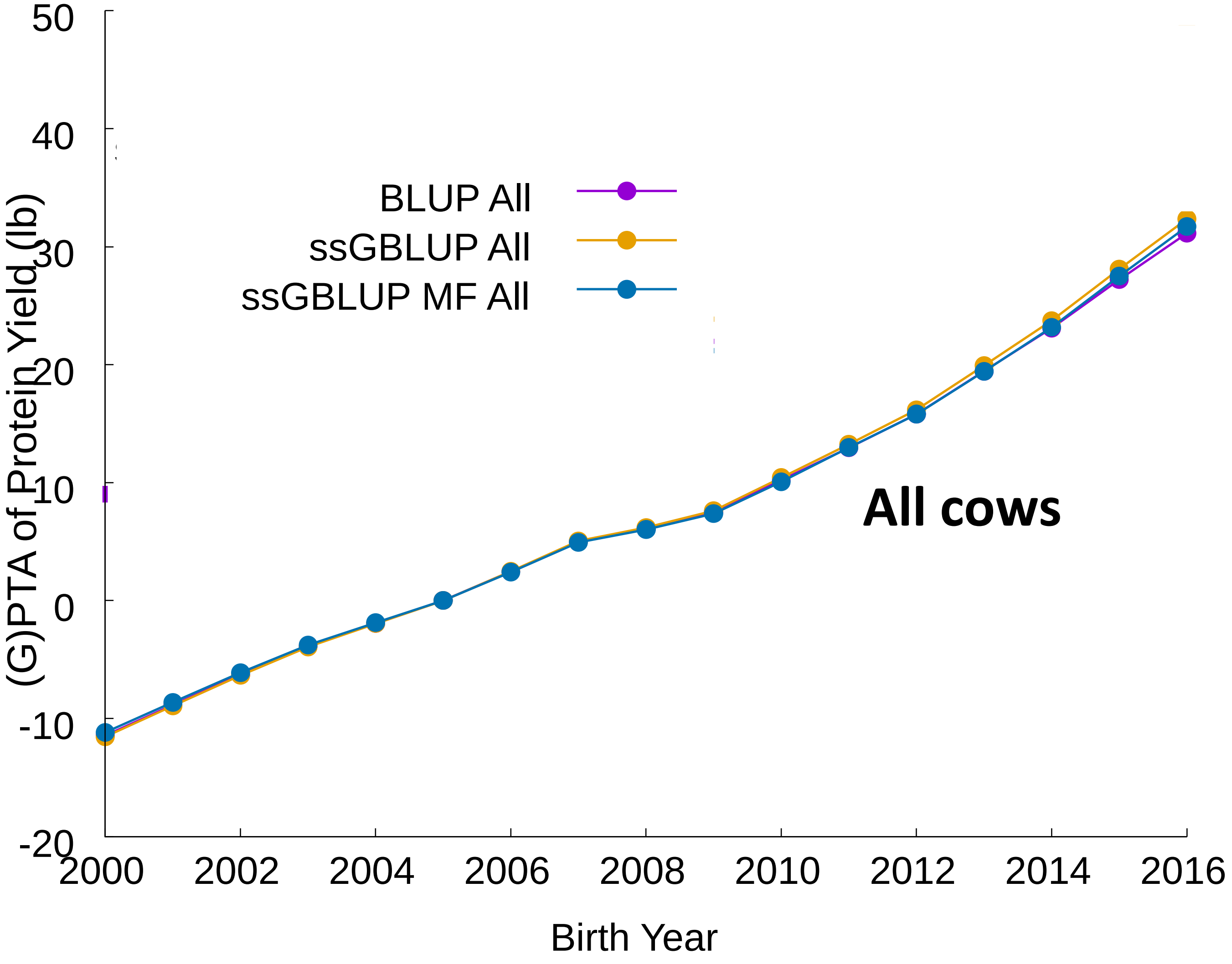
1.3 lb (0.6kg)

Accurate trend prediction

+1.3 lb = 0.07 GSD

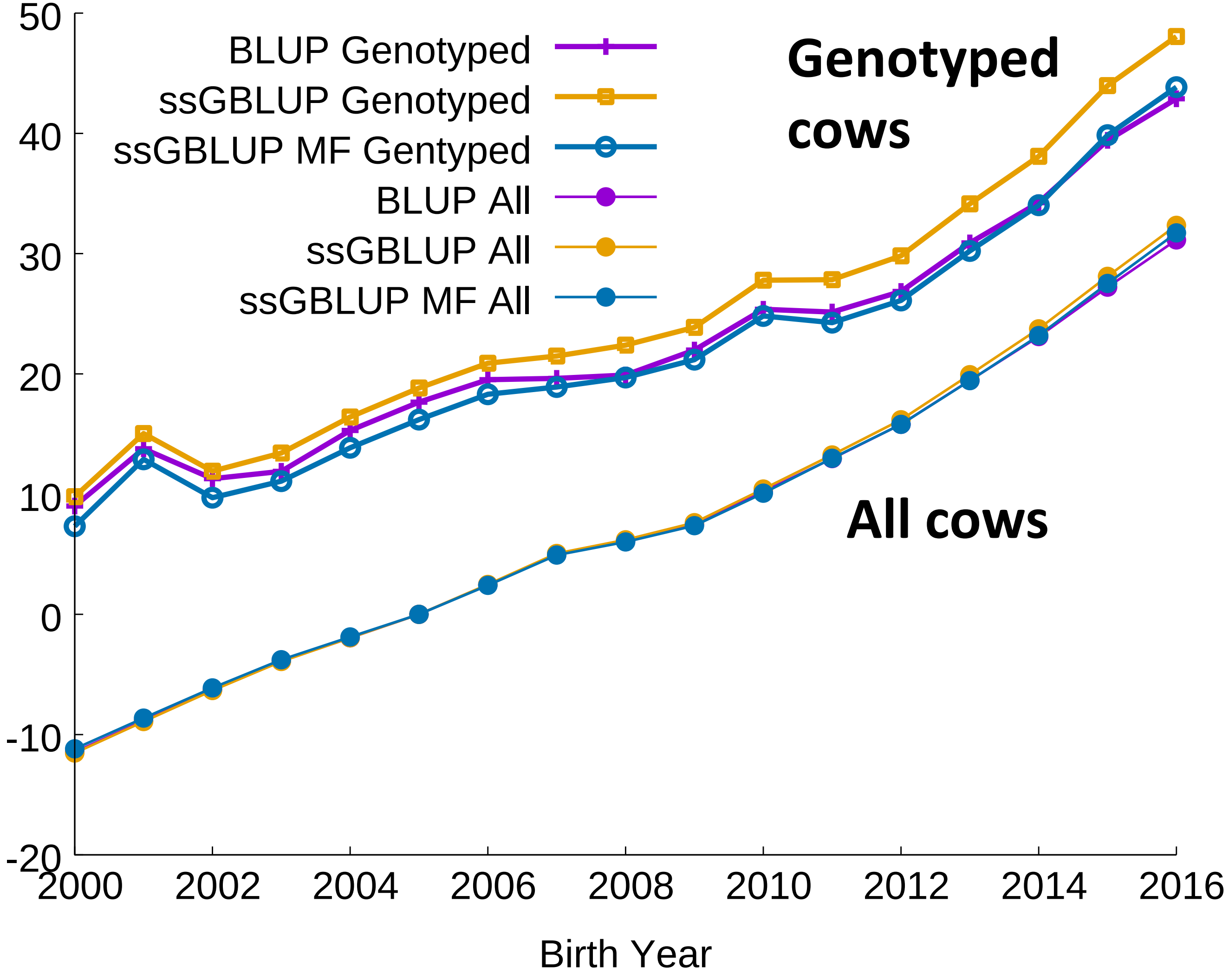
* Trends adjusted to 2010 genetic base

All/genotyped cows with record(s) in 2018



All cows

All/genotyped cows with record(s) in 2018



pedUPG: over-predicted?

- Incorrect scaling of G?

MF: under-predicted?

- Biased Γ ?

Summary

- The UPG/MF model is reasonable in generic trends, predictability, and inflation for young bulls.
- The traditional DYD is biased down by pre-selection.
- Single-step GBLUP with >2M genotypes is computationally feasible.
- Additional research in progress:
    - Reasonable scaling (“tuning”) on G
    - Refinement on benchmarks in validation
    - Development of more efficient and optimized tools
    - Various models (e.g., 18-trait model)

Acknowledgement

- Council of Dairy Cattle Breeding (CDCB) for phenotype, genotype, and pedigree data.
- Holstein Association USA for financial support.
- Paul VanRaden (AGIL, USDA) for discussion on UPG in genetic evaluation
- Ignacio Aguilar (INIA) for discussion on various topics in ssGBLUP