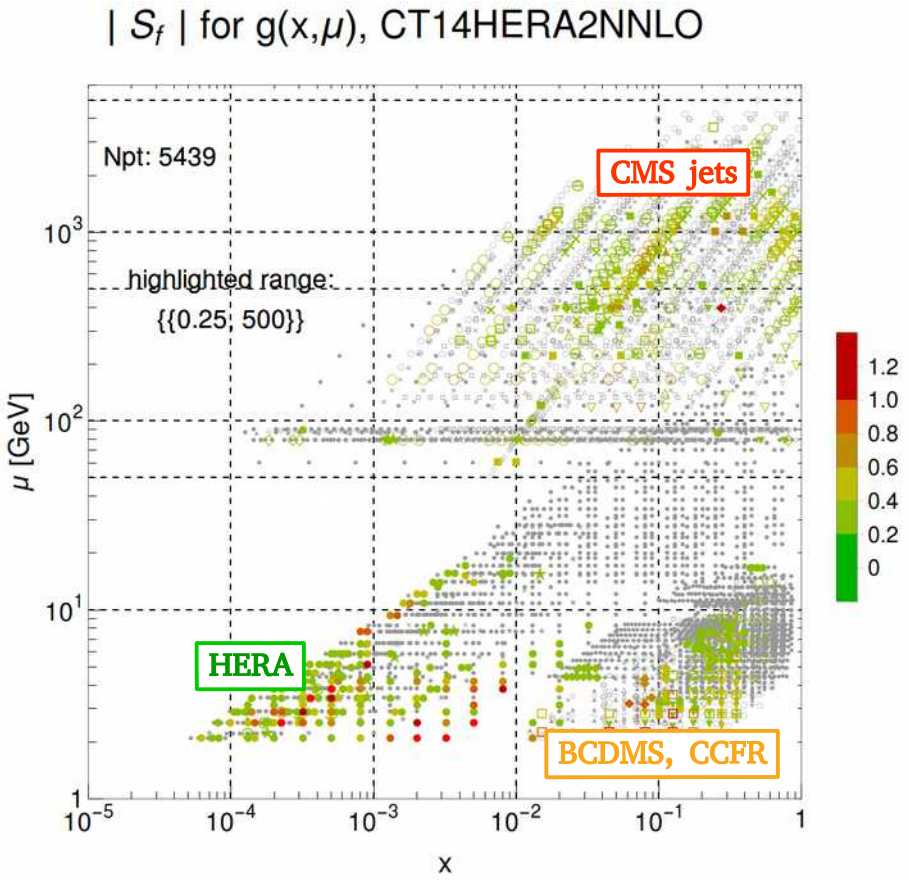
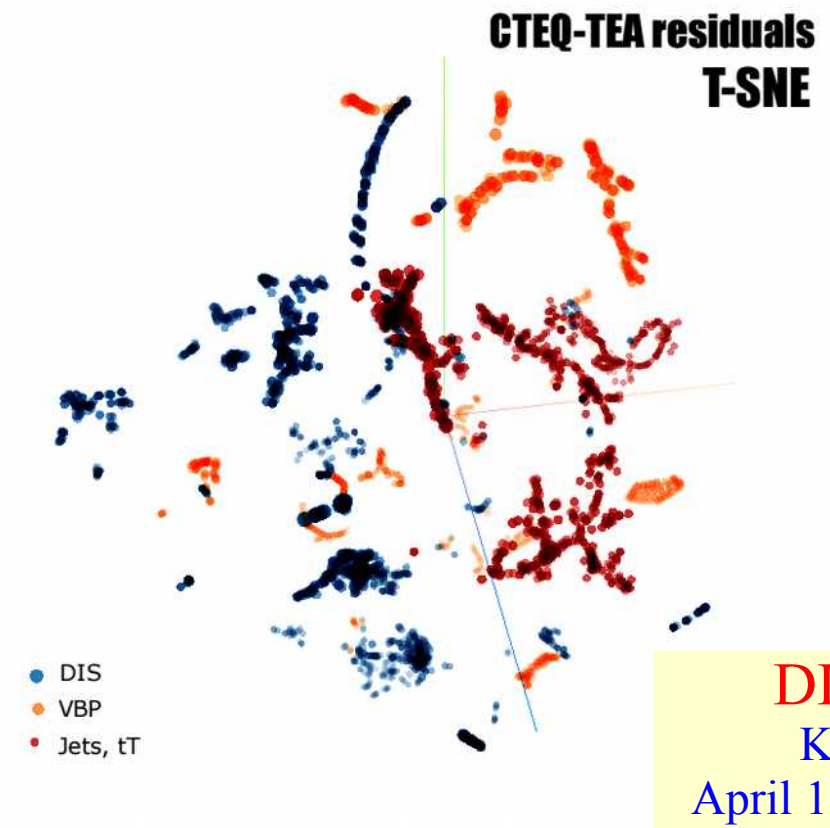
PDFSense:

Visualizing sensitivity of hadronic experiments to the nucleon structure

Fred Olness
SMU

with ...
Bo-Ting Wang, Tim Hobbs, S. Doyle, J. Gao, T.-J. Hou, & Pavel Nadolsky





DIS2018
Kobe, Japan
April 16-20, 2018

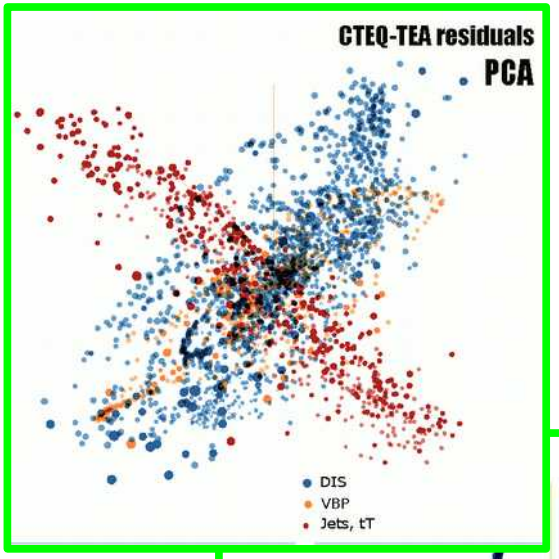
PDFSense:

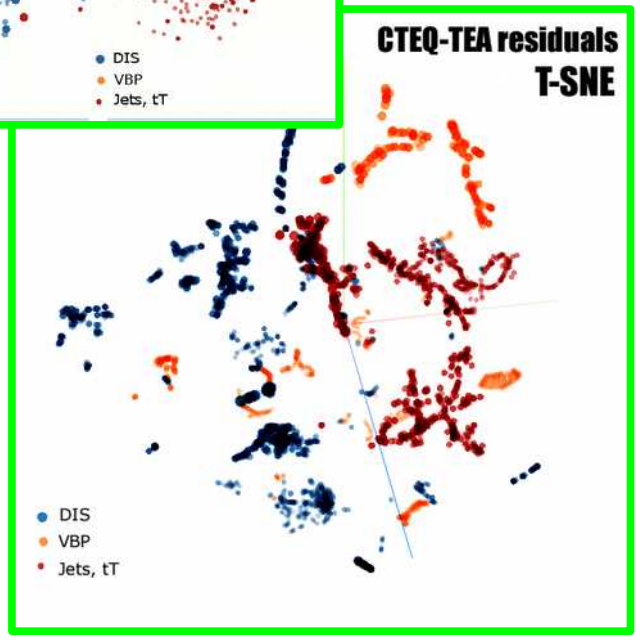
http://metapdf.hepforge.org/PDFSense/

Visualizing sensitivity of hadronic experiments to the nucleon structure

arXiv:1803.02777

Bo-Ting Wang, T.J. Hobbs, Sean Doyle,
Jun Gao, Tie-Jiun Hou,
Pavel Nadolsky, Fred Olness.





PDFSENSE project: sensitivities to PDFs by Expt measurements

On this webpage, we present the sensitivity of hadronic experiments to PDFs using **PDFsense [download here]**. The PDFsense enables users to compute the sensitivity and correlation of experimental data sets and CTEQ PDF sets.

Citation policy:

if you use results from this website, please cite

Visualizing the sensitivity of hadronic experiments to nucleon structure
Bo-Ting Wang, T. J. Hobbs, Sean Doyle, Jun Gao, Tie-Jiun Hou, Pavel Nadolsky, and Fredrick I. Olness
arXiv:1803.02777

Website development: **Bo-Ting Wang, Pavel Nadolsky**

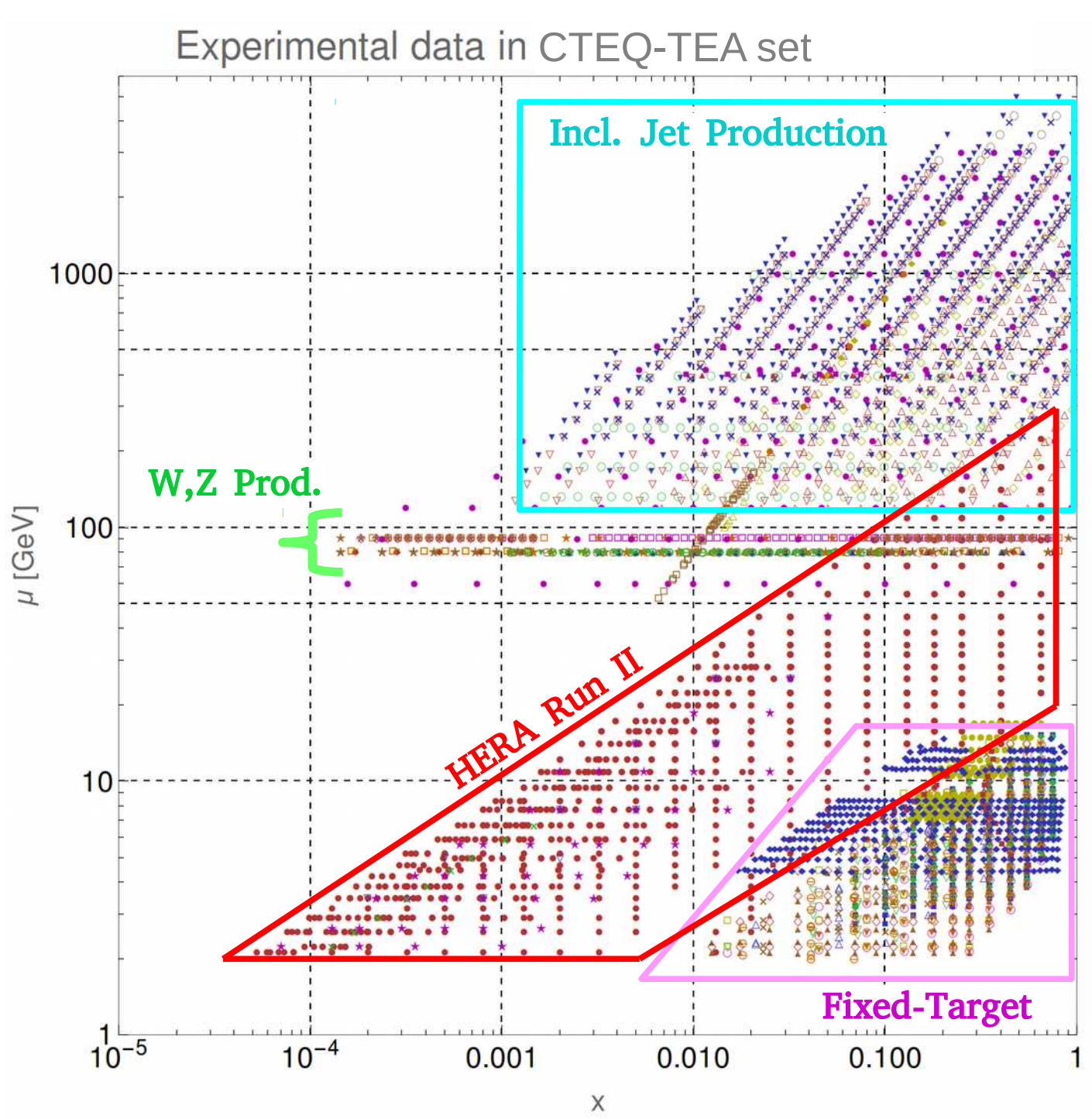
TSV Files

The .tsv file records the residuals of all replicas (normalized by the root-mean-square of th central values of residuals in each experiment as shown in the paper) for each point in each data set. Users can play the .tsv file with Excel or load .tsv file into **Embedding Projector** for PCA and T-SNE analysis. **[download]**

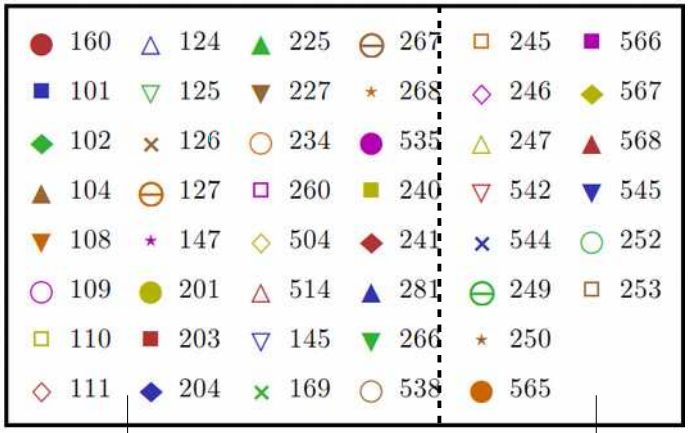
Figures

CT14HERA2 NNLO sensitivities

all experiments
Jet measurements at the LHC
pT distribution of Z measurements at the LHC
ttbar measurements at the LHC
lepton asymmetry in W measurements at the LHC



the problem : modern global PDF analyses must contend with LARGE data sets involving many physical processes and channels!



individual experiments contributing to CT14 analysis of HERA Run II ("CT14H2")

new, unfitted datasets from, e.g., LHC at 7, 8 TeV

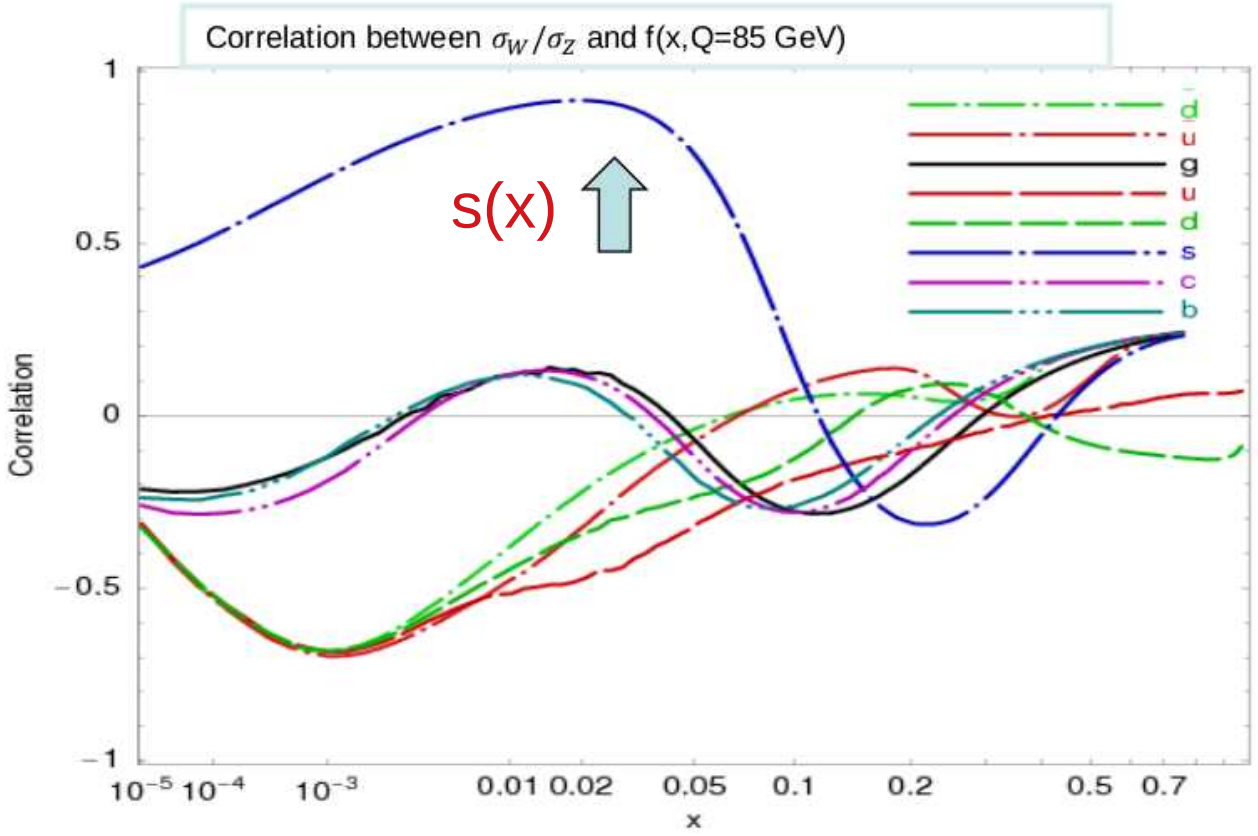
- Can we weigh the influence of datasets **WITHOUT** each time performing a full global analysis? **. . . we could then predict the impact of unfitted data and guide fits . . .**

How sensitive is an experiment to a PDF?

Can we know it **before** doing the global fit?

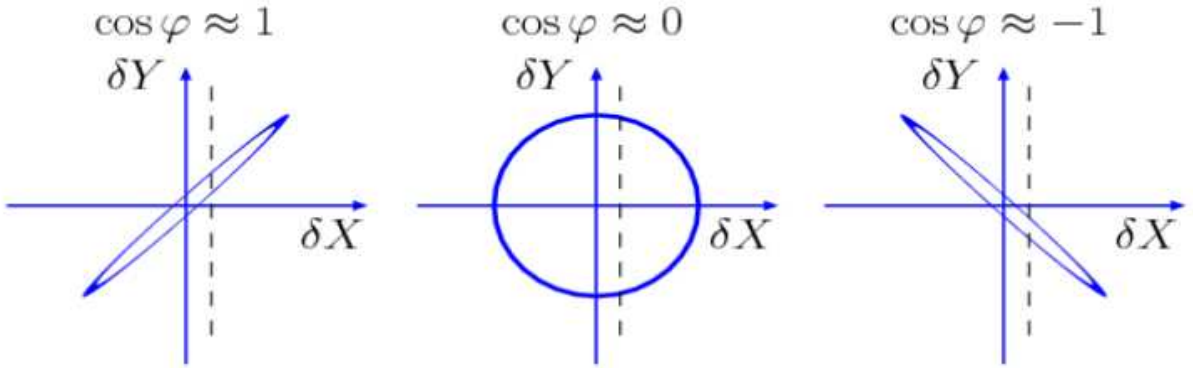
...we may turn to the Pearson correlations between PDFs and δ_i, but we first note

Correlations carry useful, but limited information



CTEQ6.6 [arXiv:0802.0007]:
$\cos\varphi > 0.7$ shows that the ratio σ_W/σ_Z at the LHC must be sensitive to the strange PDF $s(x, Q)$

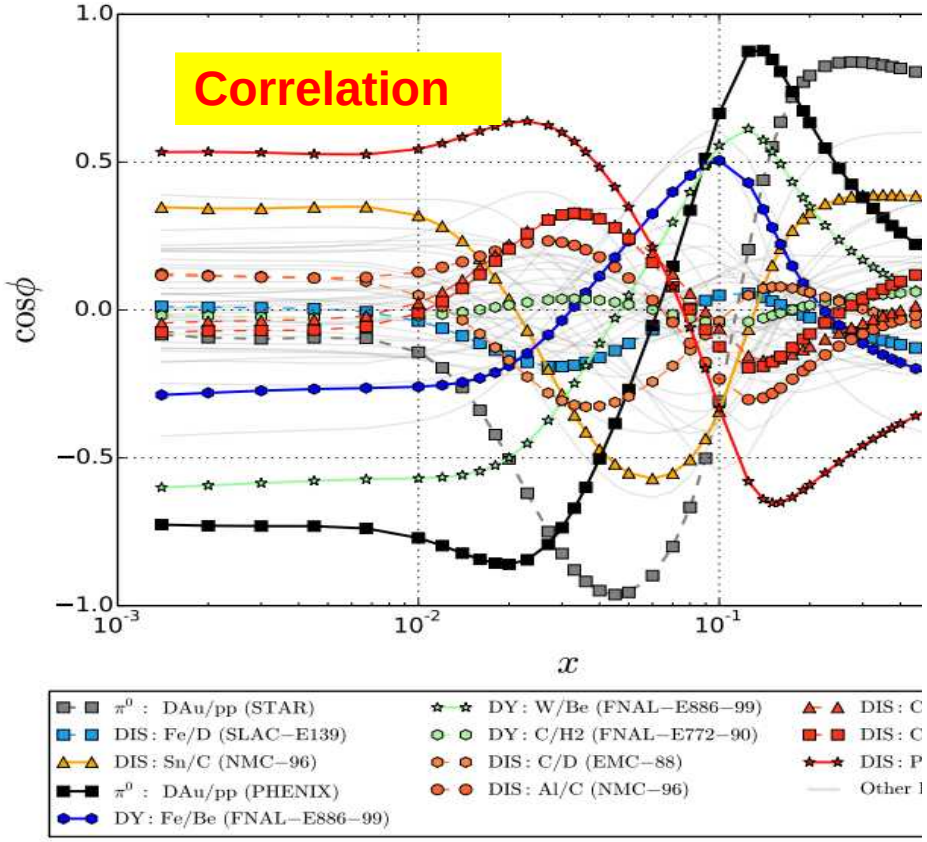
$\cos\varphi \approx \pm 1$ suggests that a measurement of X **may** impose tight constraints on Y

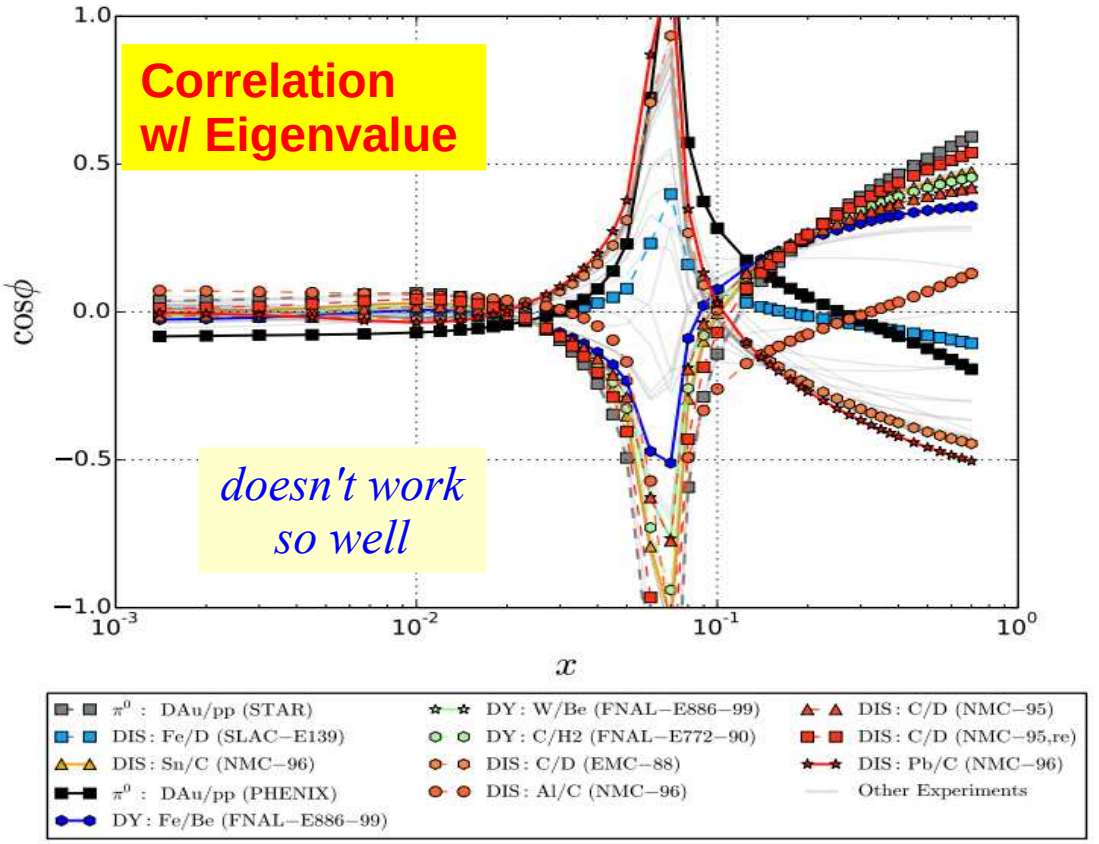


But, Corr[X,Y] between **theory** cross sections X and Y does not tell us about **experimental** uncertainties

Correlations carry useful but limited information

An idea: weight by Hessian eigenvalue λ_i





$$\cos\phi[X,Y] = \frac{\vec{\nabla}X \cdot \vec{\nabla}Y}{\Delta X\, \Delta Y}$$

$$\vec{\nabla}X_i = \left(X_i - \overline{X}\right)$$
$$\vec{\nabla}X_i = {}^1\!/_2\left(X_i^+ - X_i^-\right)$$
$$\Delta X^2 = \vec{\nabla}X \cdot \vec{\nabla}X$$

The goal is to **quantify the strength of the constraints** placed on a particular set of PDFs by both individual and aggregated measurements of physical processes

- for single-particle hadroproduction of gauge bosons at, e.g., LHC, factorization gives

$$\sigma(AB \to W/Z+X) = \sum_n \alpha_s^n(\mu_R^2) \sum_{a,b} \int dx_a dx_b \times f_{a/A}(x_a, \mu^2)\, \hat{\sigma}^{(n)}_{ab\to W/Z+X}(\hat{s}, \mu^2, \mu_R^2)\, f_{b/B}(x_b, \mu^2)$$

PDFs determined by fits to data; e.g., "CT14H2"

pQCD matrix elements – specified by theoretical formalism in a given fit

- **Idea**: study the **statistical <u>correlation</u>** between PDFs and the quality of the fit at to a measured data point(s); fit quality encoded in a (**T**heory) – (shifted **D**ata) ***residual*** :

$$r_i(\vec{a}) = \frac{1}{s_i}\left(T_i(\vec{a}) - D_{i,sh}(\vec{a})\right)$$

s_i : uncorrelated uncert. | $\vec{a}$: PDF parameters

a brief statistical aside

- the CTEQ-TEA global analysis relies on the **Hessian formalism** for its error treatment

$$\chi^2_E(\vec{a}) = \sum_{i=1}^{N_{pt}} r_i^2(\vec{a}) + \sum_{\alpha=1}^{N_\lambda} \overline{\lambda}^2_\alpha(\vec{a})$$

→ nuisance parameters to handle correlated errors

$$r_i(\vec{a}) = \frac{1}{s_i}\left(T_i(\vec{a}) - D_{i,sh}(\vec{a})\right)$$

these result in systematic shifts to data central values:

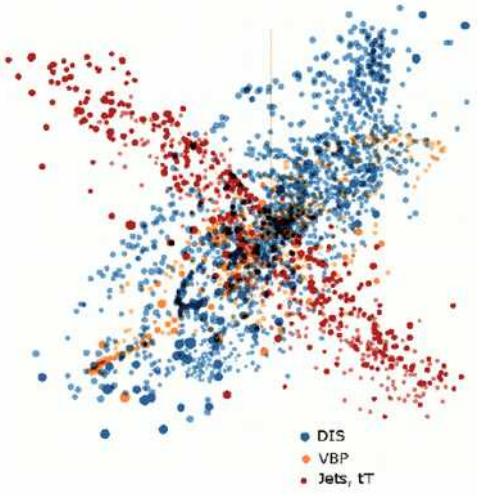
$$D_i \to D_{i,sh}(\vec{a}) = D_i - \sum_{\alpha=1}^{N_\lambda} \beta_{i\alpha}\overline{\lambda}_\alpha(\vec{a})$$

- a 56-dimensional parametric basis $\vec{a}$ is obtained by diagonalizing the Hessian matrix H determined from χ^2

→ use this basis to compute 56-component "normalized" residuals :

$$\delta^{\pm}_{i,l} \equiv \left(r_i(\vec{a}^{\pm}_l) - r_i(\vec{a}_0)\right) / \langle r_0 \rangle_E$$

where $\langle r_0 \rangle_E \equiv \sqrt{\frac{1}{N_d}\sum_{i=1}^{N_d} r_i^2(\vec{a}_0)}$

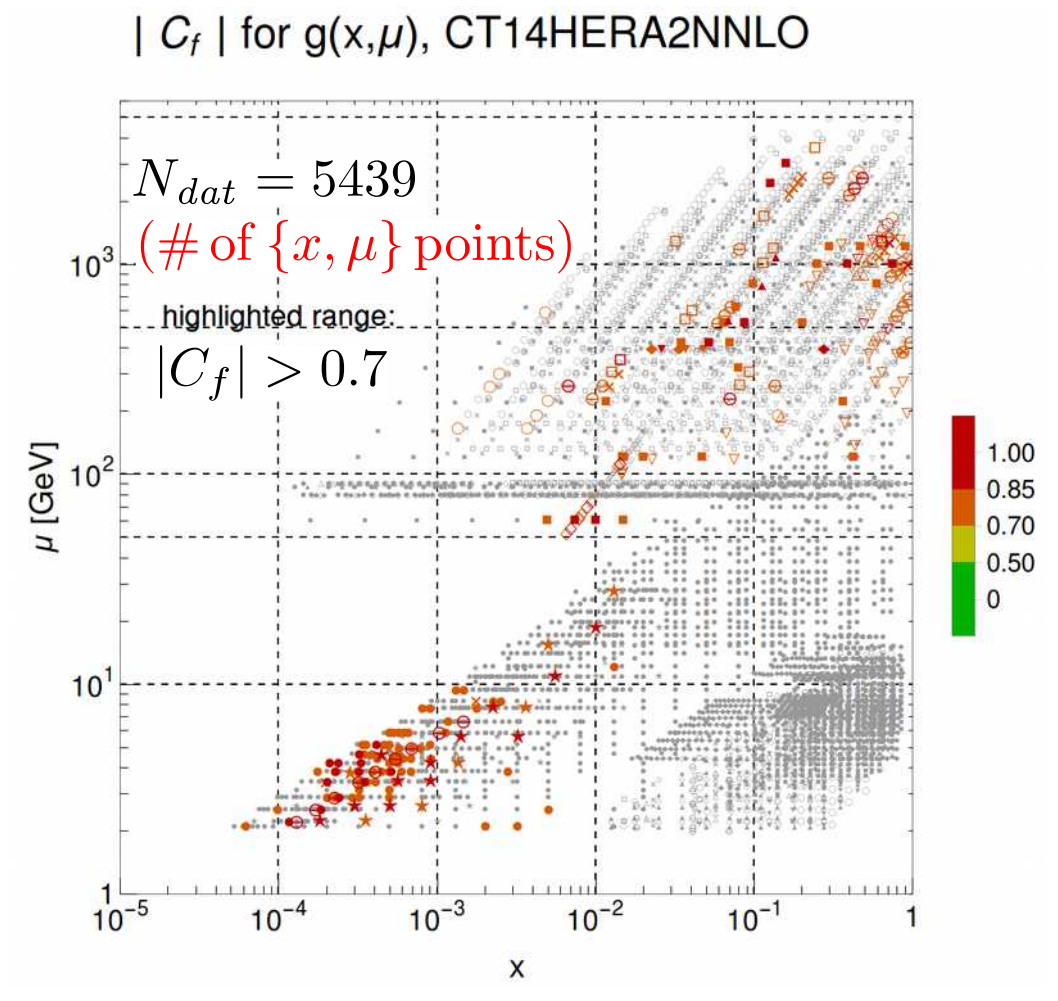


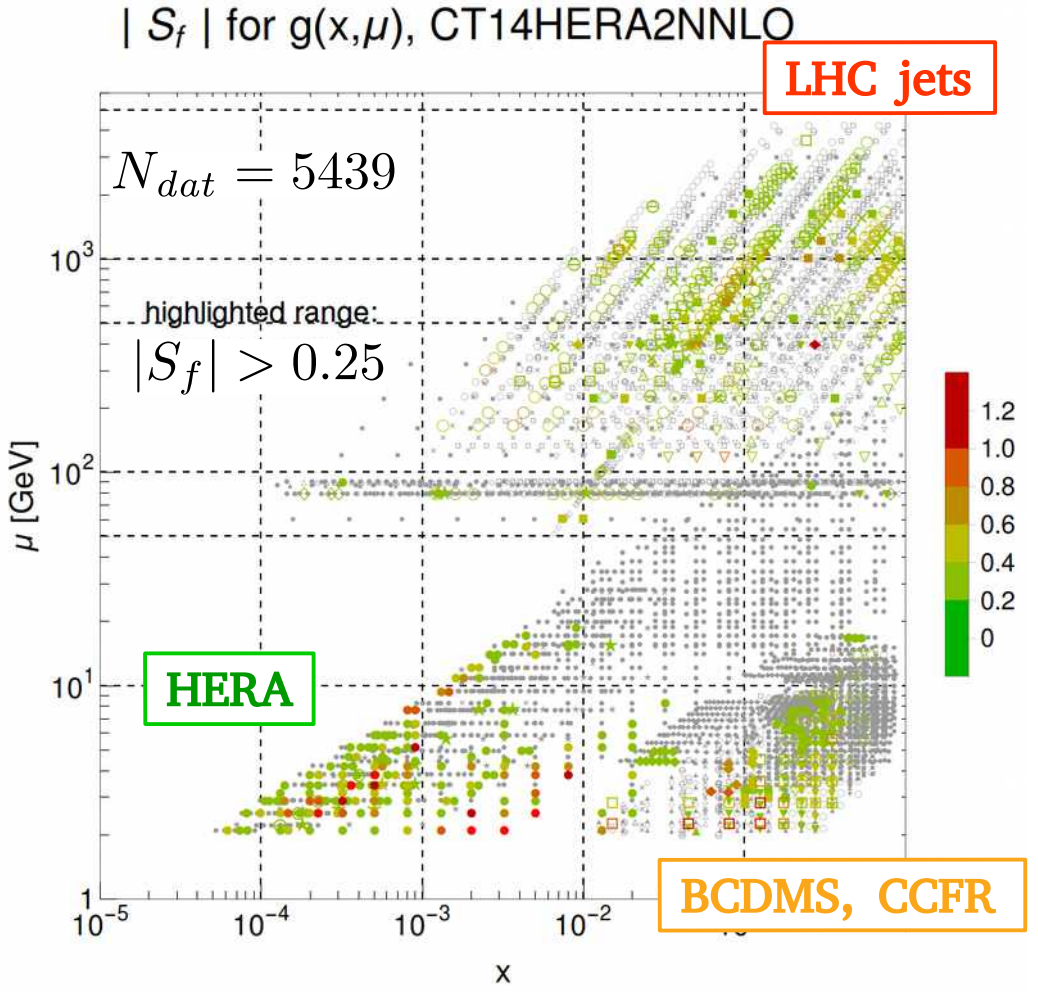
the sensitivity reveals a richer landscape than the correlation!

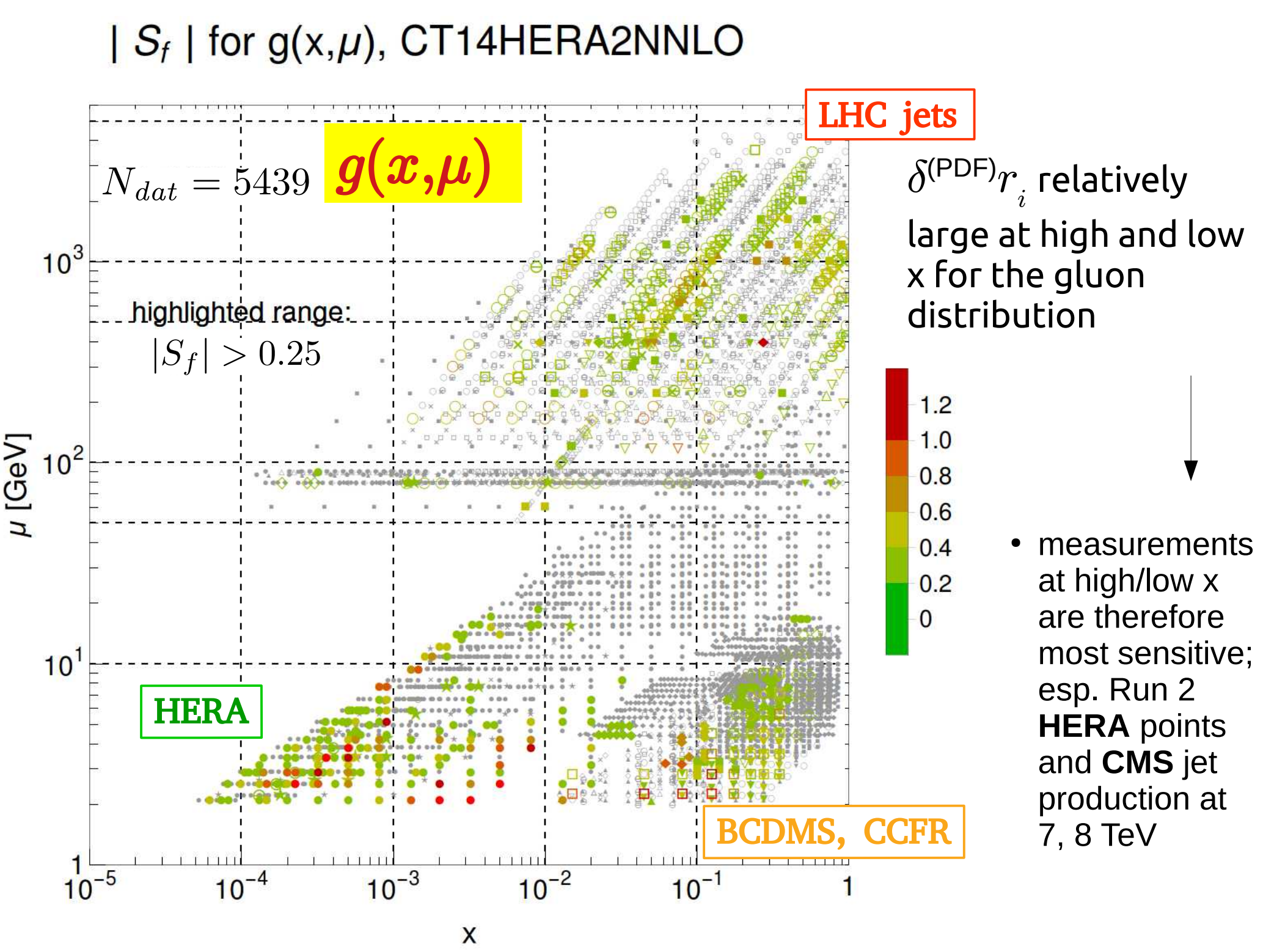
$$C_f(x_i, \mu_i) \equiv \mathrm{Corr}[f(x_i, \mu_i), r_i]$$

$$S_f(x_i, \mu_i) \equiv \frac{\delta^{(\mathrm{PDF})} r_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2}} C_f(x_i, \mu_i)$$

- broad outlays of the experimental parameter space are shown to be sensitive, but are **missed by the correlation**

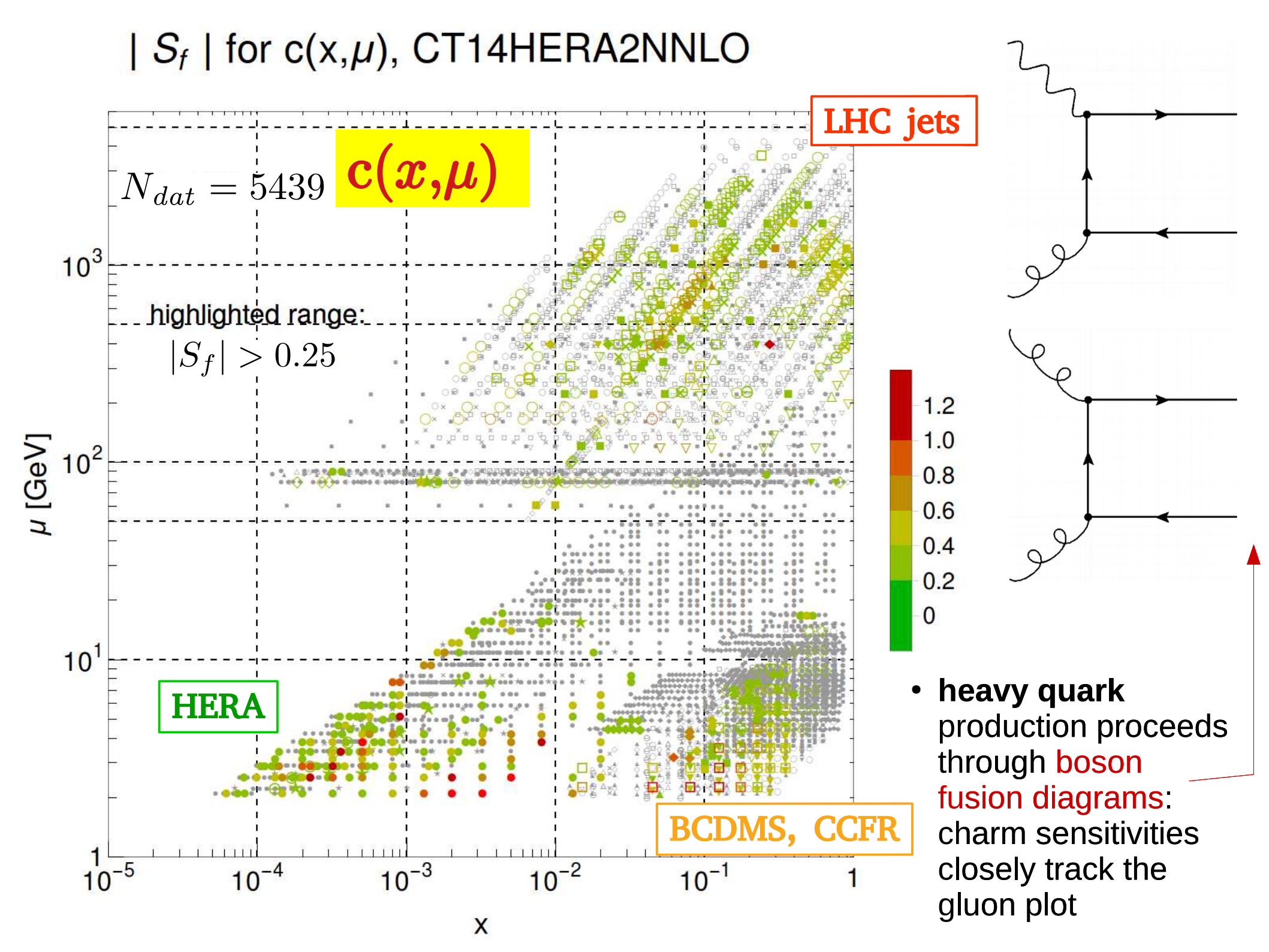




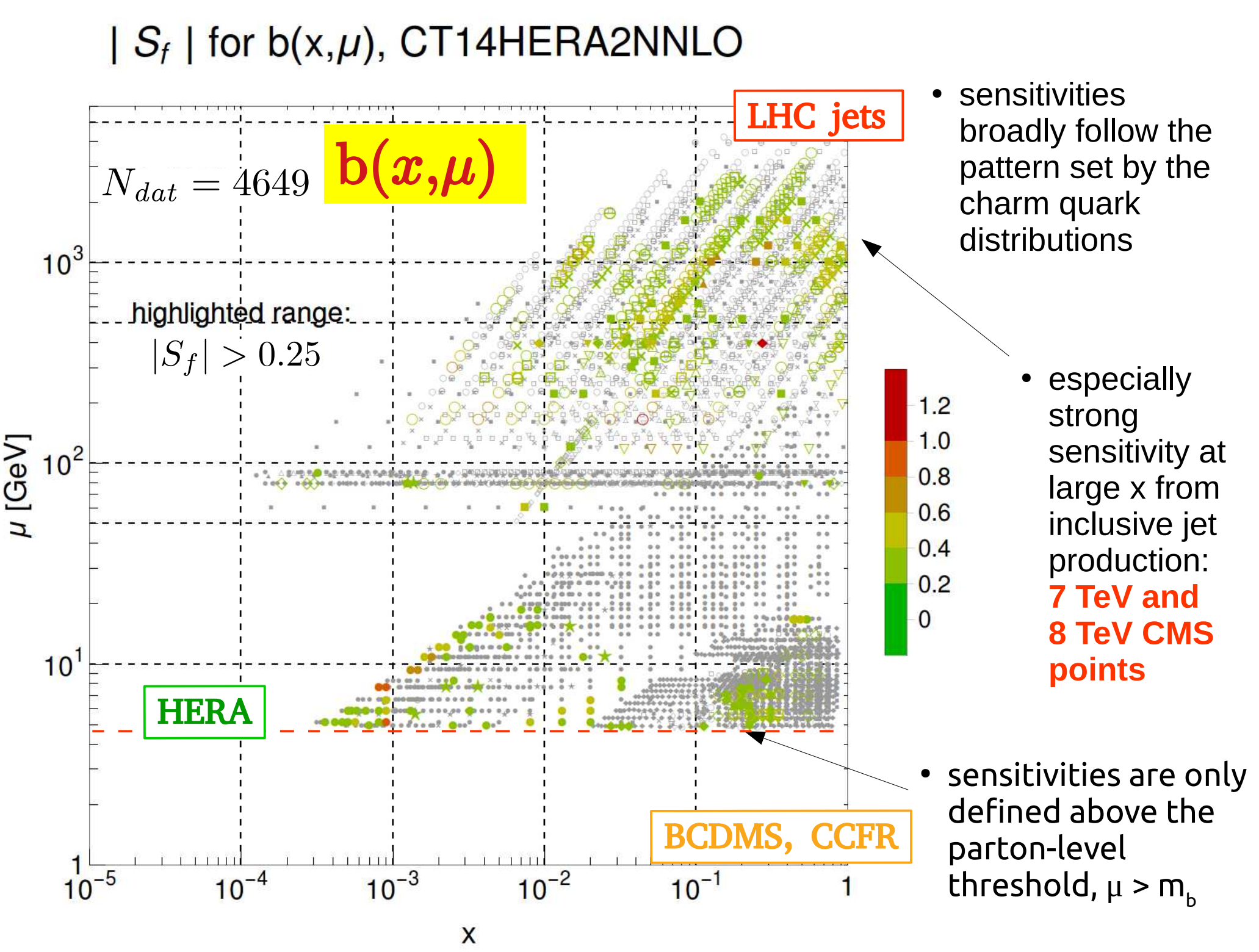


$\delta^{(\mathrm{PDF})} r_i$ relatively large at high and low x for the gluon distribution

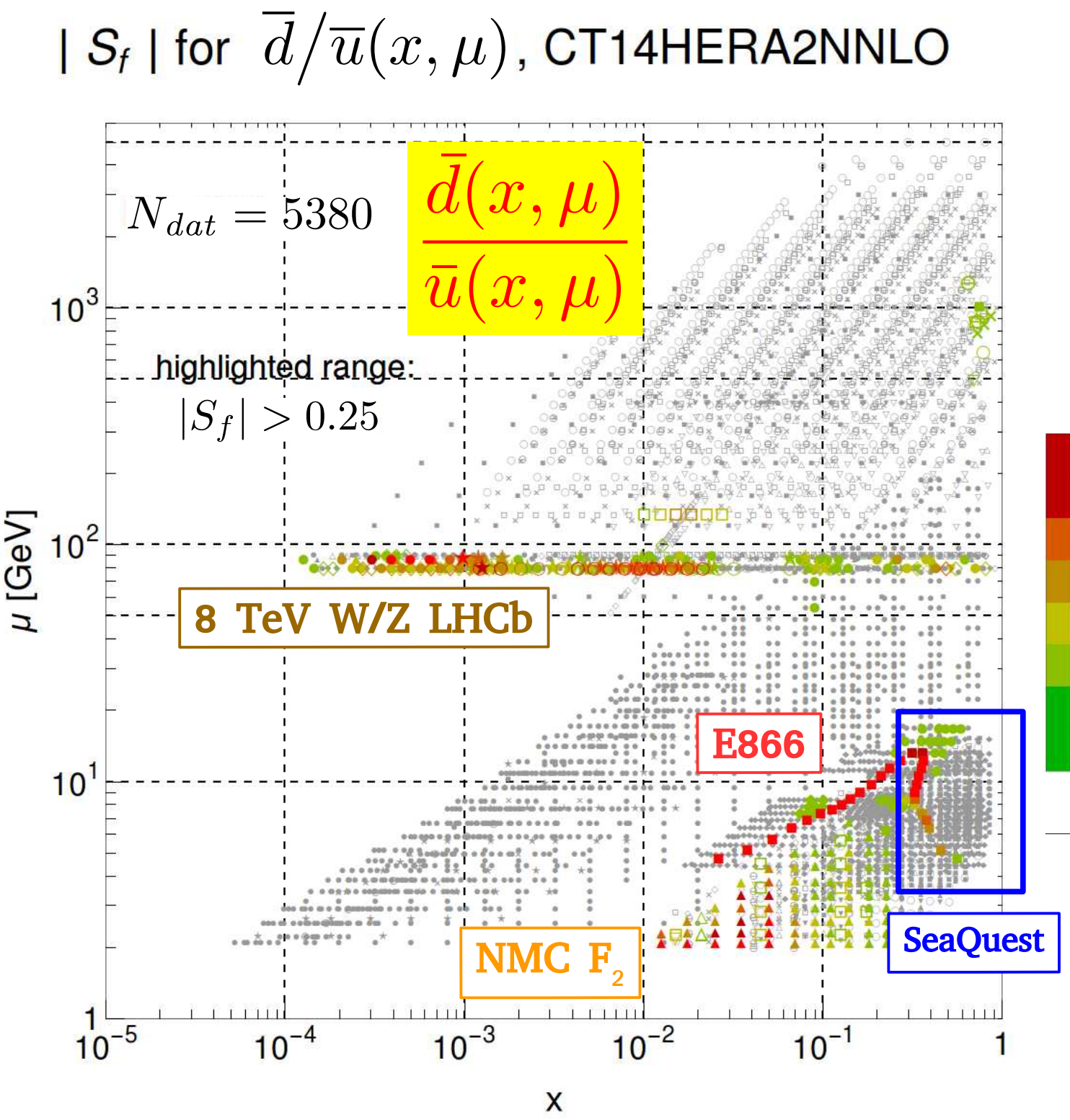
- measurements at high/low x are therefore most sensitive; esp. Run 2 **HERA** points and **CMS** jet production at 7, 8 TeV



- **heavy quark** production proceeds through boson fusion diagrams: charm sensitivities closely track the gluon plot



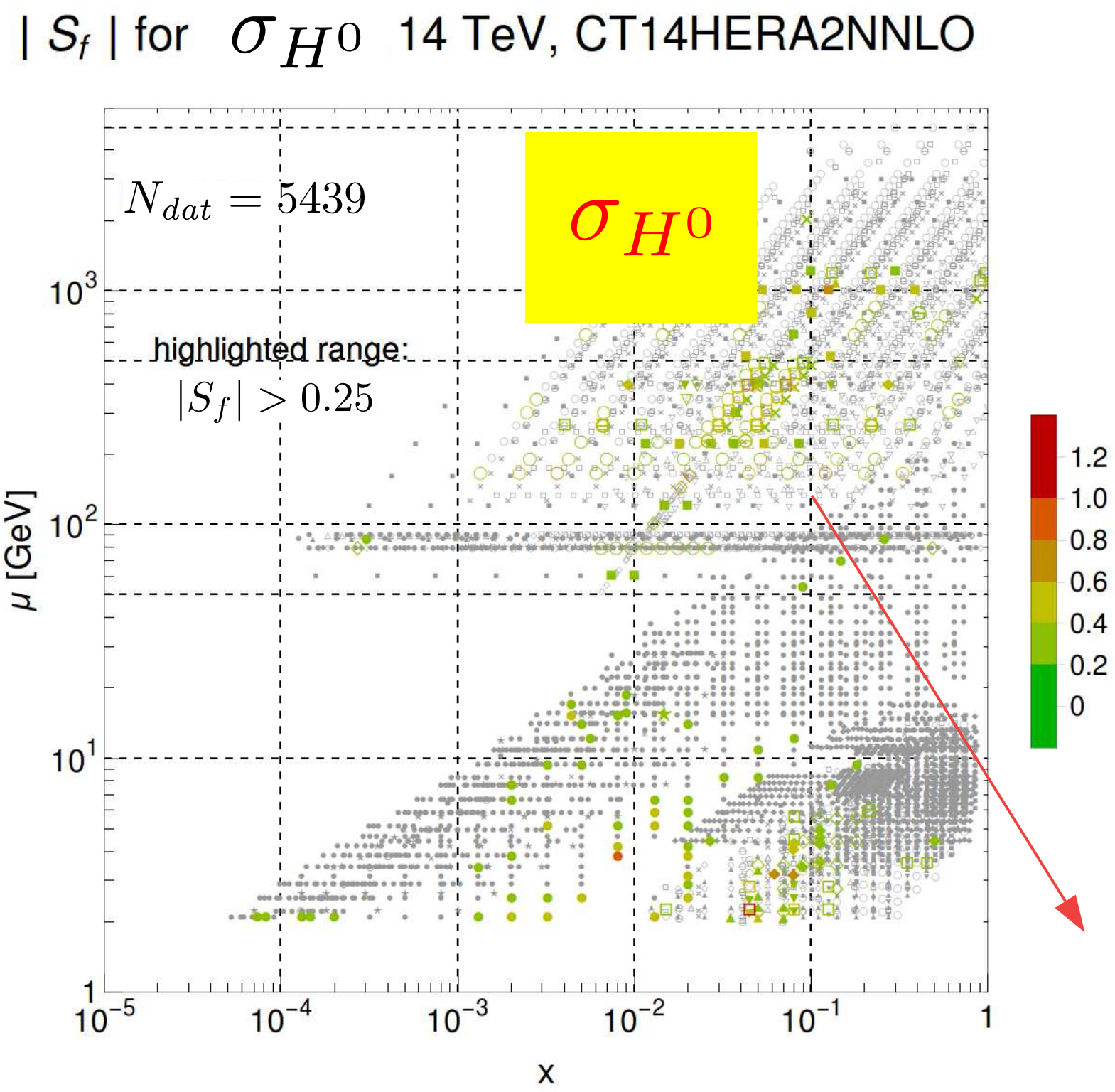
- sensitivities broadly follow the pattern set by the charm quark distributions
- especially strong sensitivity at large x from inclusive jet production: **7 TeV and 8 TeV CMS points**
- sensitivities are only defined above the parton-level threshold, $\mu > m_b$



- PDF ratio is sensitive to flavor symmetry breaking in the light quark sea
- the large E866 sensitivity *degrades* at larger x

this is a prime motivation for higher x DY measurements at **E906 (SeaQuest)**

- some contribution at high x from CMS inclusive jet production



- several processes (high p_T Z prod., top prod.) have been suggested as providing leverage on the Higgs cross section

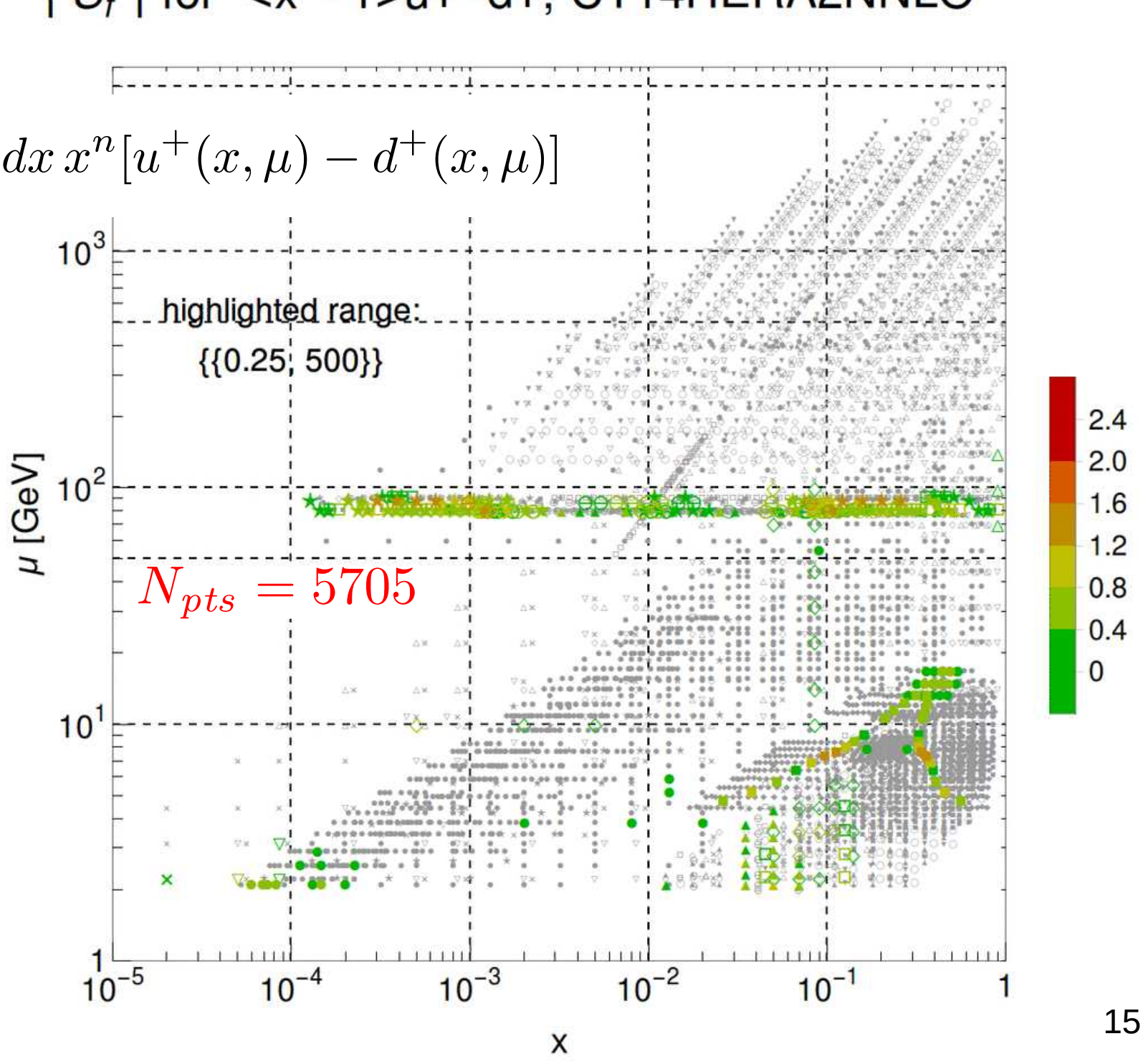
- in fact, we find **inclusive jet production** to have the broadest overall sensitivity!

… moments from the lattice in future

$$M^n_{u^+-d^+} = \int_0^1 dx\, x^n [u^+(x,\mu) - d^+(x,\mu)]$$

- 1st several PDF moments can be computed on the **lattice**; knowledge of sensitivities suggests where lattice calculations might constrain PDFs

- higher moments shift sensitivities *rightward*



Ranking Tables

… to assess the impact of separate experiments

No.	Exp. ID	N_d	Rankings													
			$\sum_f \lvert S_f^E \rvert$	$\langle \sum_f \lvert S_f^E \rvert \rangle$	$\lvert S_{\bar{d}}^E \rvert$	$\langle \lvert S_{\bar{d}}^E \rvert \rangle$	$\lvert S_{\bar{u}}^E \rvert$	$\langle \lvert S_{\bar{u}}^E \rvert \rangle$	$\lvert S_g^E \rvert$	$\langle \lvert S_g^E \rvert \rangle$	$\lvert S_u^E \rvert$	$\langle \lvert S_u^E \rvert \rangle$	$\lvert S_d^E \rvert$	$\langle \lvert S_d^E \rvert \rangle$	$\lvert S_s^E \rvert$	$\langle \lvert S_s^E \rvert \rangle$
1	160	1120.	620.	0.0922	B		**A**	3	**A**	3	**A**	3	B		C	
2	111	86	218.	0.423	C	**1**	C	**1**		3	B	**1**	C	2		
3	101	337	184.	0.0909			C		C		B	3	C			
4	104	123	169.	0.229	C	2					C	2	B	2		
5	102	250	141.	0.0938	C				C	3	C	3	C	3		
6	109	96	115.	0.199	C	2	C	2		3	C	2	C	3		
7	201	119	113.	0.158	C	2	C	2				3				
19	268	41	39.6	0.161		3		3				3		3		3
20	**249**	33	39.2	0.198		2		3				3		2		3
21	514	110	36.8	0.0557						3						
22	125	33	36.7	0.185		3		3				3		3		2
23	**252**	48	34.5	0.12		3		3				3				3
24	203	15	33.3	0.37		**1**		**1**				3		2		
25	535	90	30.2	0.056						3						
26	**245**	33	30.2	0.152		3		3		3		3		3		

Experiments are listed in the descending order of the summed sensitivities to $\bar{d}, \bar{u}, g, u, d, s$

For each flavor, A and 1 indicate the strongest total sensitivity and strongest sensitivity per point

C and 3 indicate marginal sensitivities; low sensitivities are not shown

illustration: a high energy EIC, "**LHeC**"

... special thanks to Tim Hobbs on this study

- an electron-proton (or electron-ion) collider to achieve **high luminosities** $\gtrsim 1000$ times that of HERA

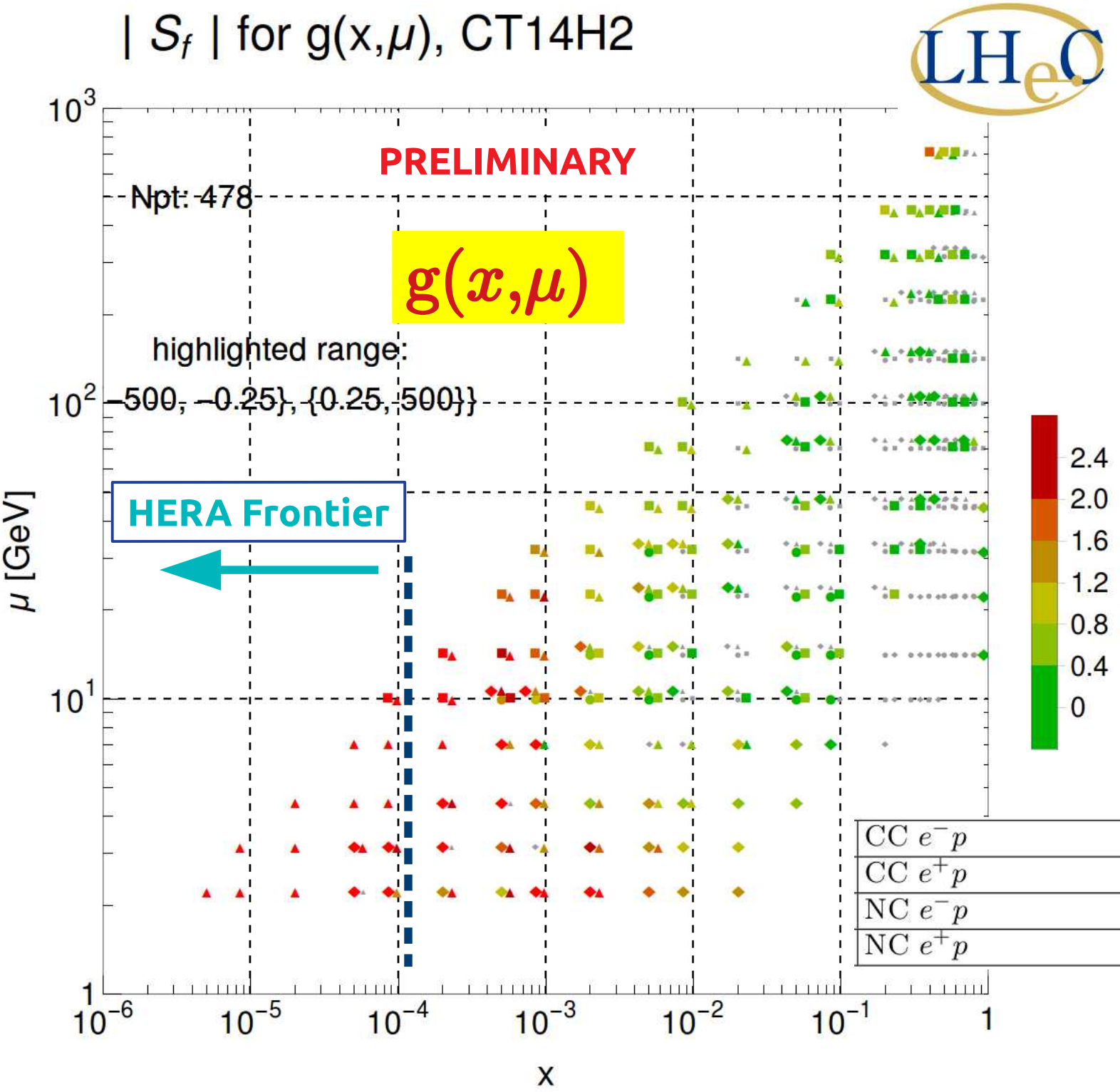
  → access a wide range of x, including $x \simeq 10^{-6}$

  → explore the dynamics of gluon saturation; greatly improve PDF precision; perform SM tests; and many other physics goals

- can perform a sensitivity analysis of Monte Carlo generated ep reduced NC/CC cross sections (Klein & Radescu, LHeC-Note-2013-002 PHY)

$$60\,\mathrm{GeV}\ e^{\pm}\ \mathrm{on}\ 1\ \mathrm{or}\ 7\,\mathrm{TeV}\ p$$

- to minimize the impact of large χ^2 of unfitted data (especially at low x), we study the sensitivities for **fluctuated data** – i.e., pseudodata randomly fluctuated about the CT14 prediction according to putative LHeC uncorrelated errors – based on **10 fb^{-1}** of data from a hypothetical **year of data-taking**



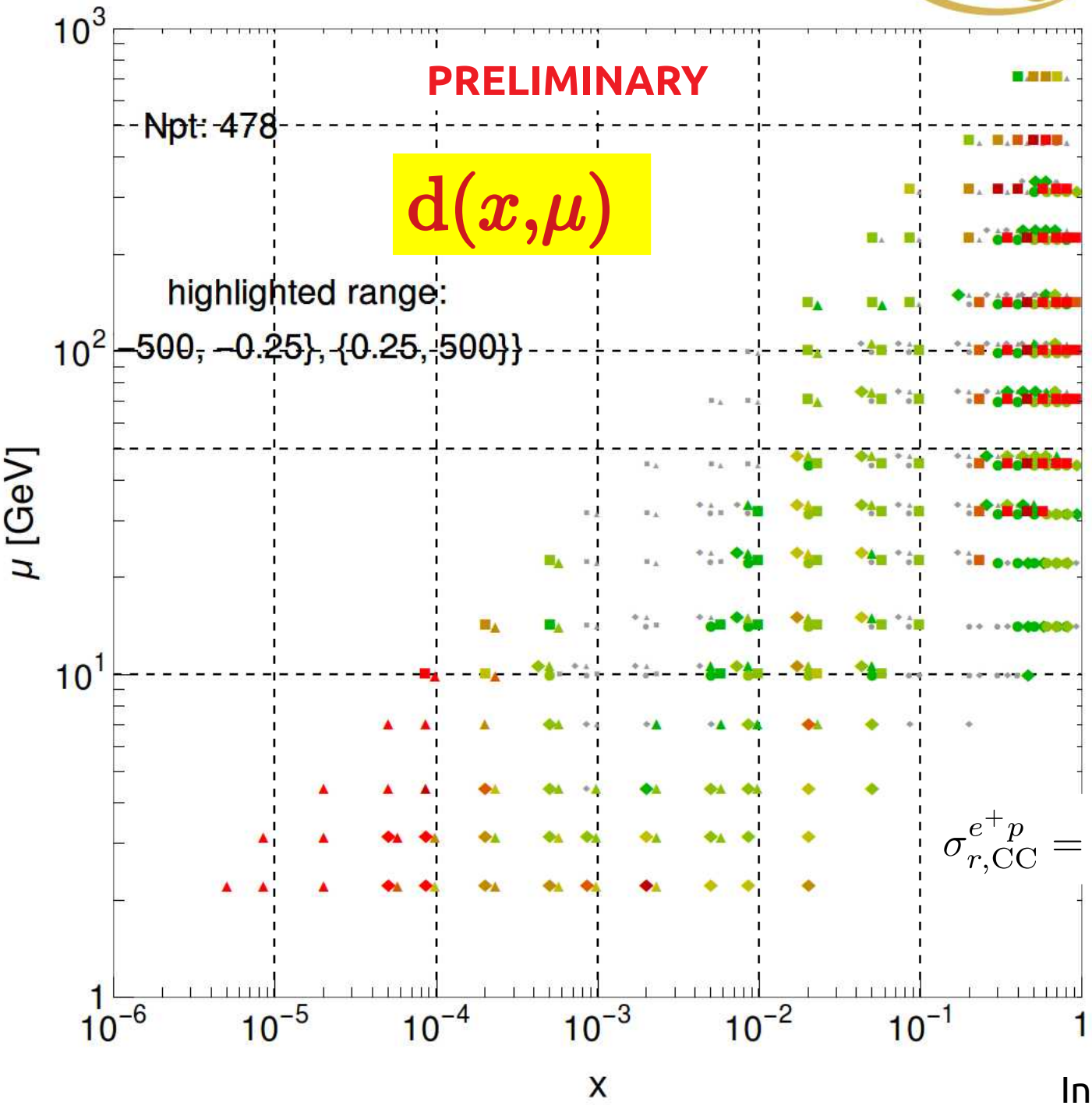
- extremely strong sensitivities along the frontier of the HERA Run II data, $x \lesssim 10^{-4}$

however, there are stringent constraints in general for $x \lesssim 10^{-2}$

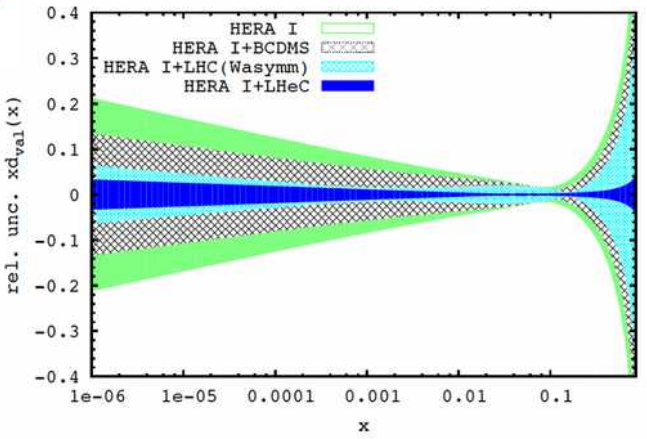
...and high x, μ

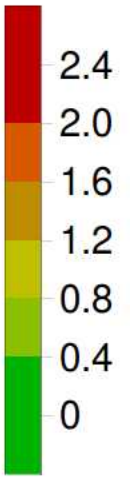
	N_d	
CC e^-p	93	circle
CC e^+p	109	square
NC e^-p	128	diamond
NC e^+p	148	triangle

$|S_f|$ for d(x,μ), CT14H2



EIC Whitepaper:1206.2913





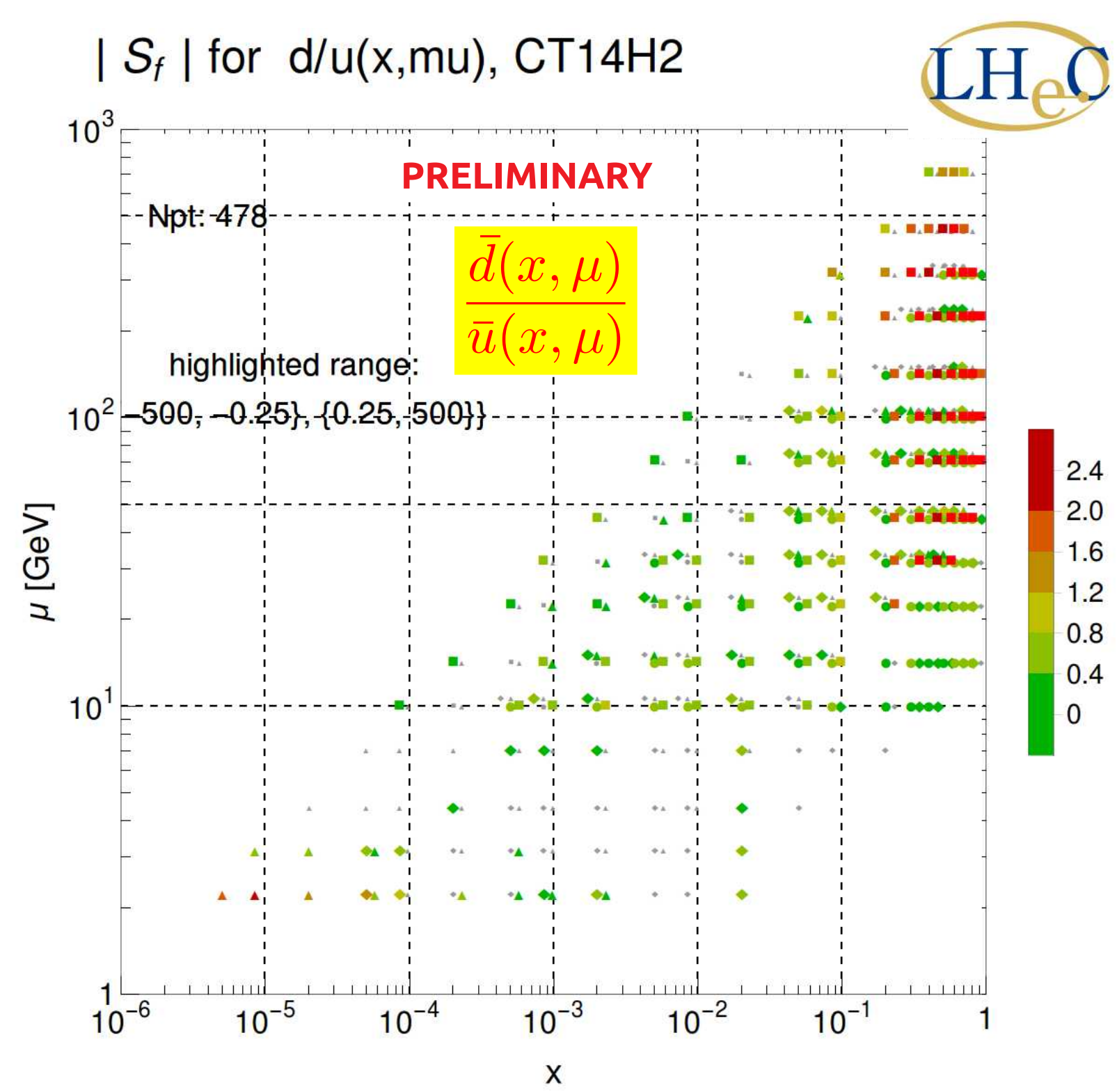
- The **strongest sensitivities** are realized for those regions with the **weakest current PDF constraints**; here, very high and very low x

for $x \to 1$

$$\sigma^{e^+p}_{r,\mathrm{CC}} = \frac{Y_+}{2}W_2^+ \mp \frac{Y_-}{2}xW_3^+ - \frac{y^2}{2}W_L^+$$

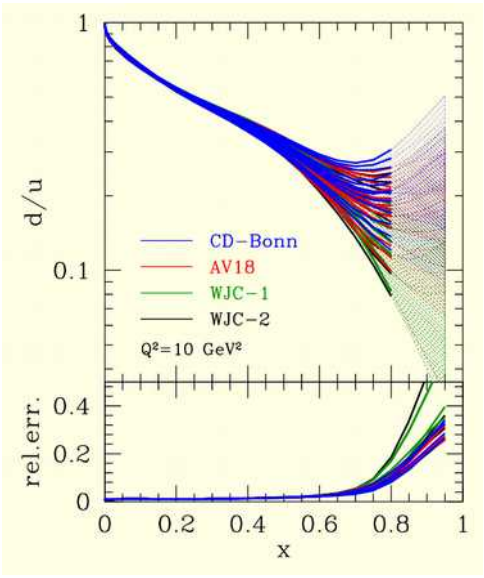
$$\simeq [1-y]^2\, x(d+s)$$

In the **LO quark-parton model**



- LHeC's high luminosity may give it a reach to high enough x to help resolve the stubborn d/u question

…without a nuclear target…

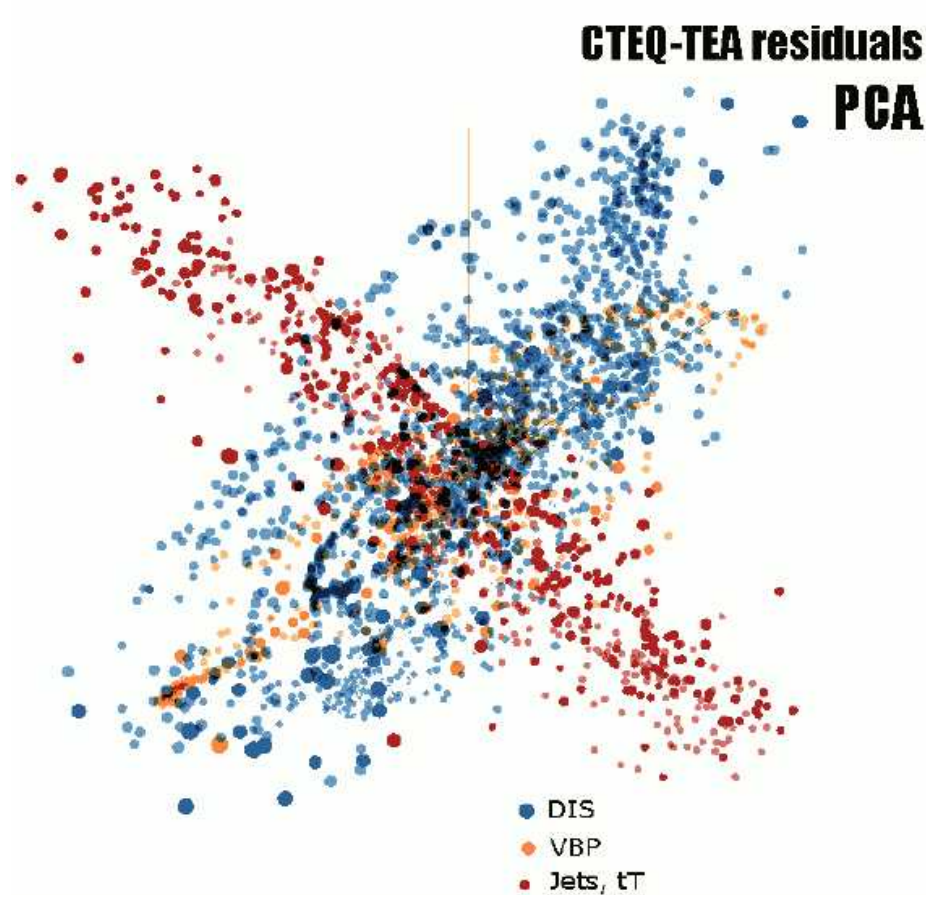


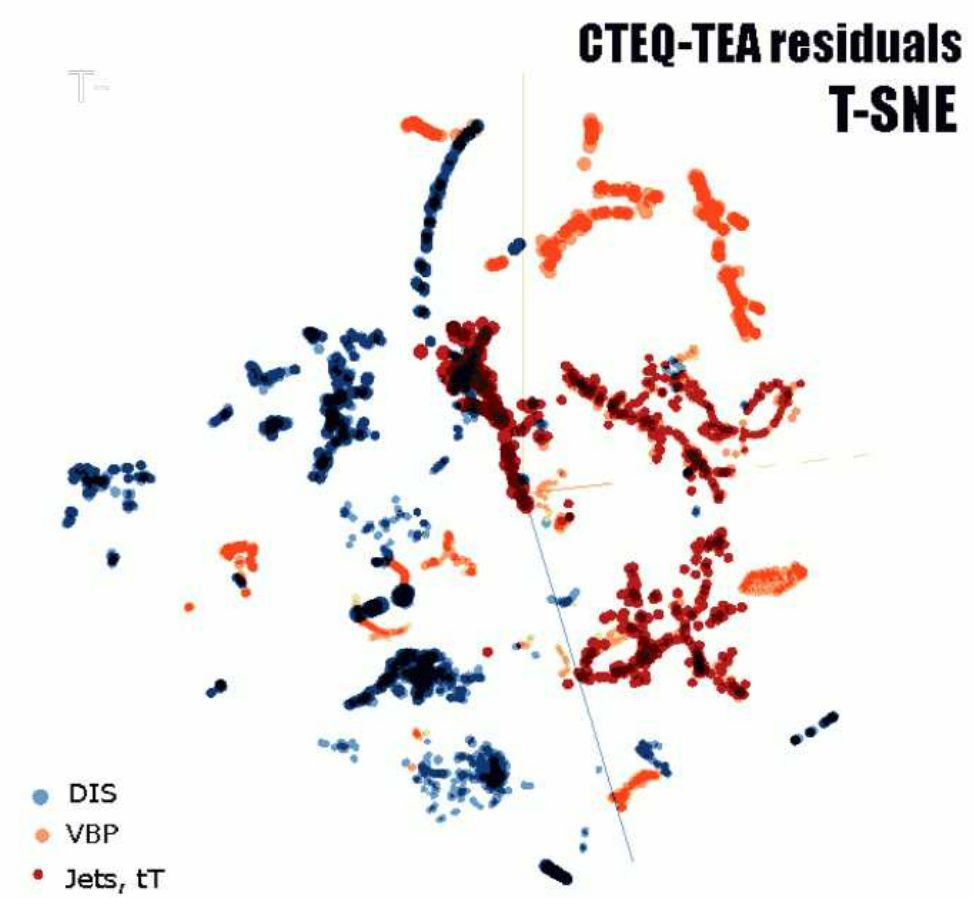
… one final topic

TensorFlow Embedding Projector

http://projector.tensorflow.org

Reads 2 .tsv files with vectors and metadata (descriptions of data points)





Principal Component Analysis (PCA) visualizes the 56-dim. manifold by reducing it to 10 dimensions (à la META PDFs)

t-distributed stochastic neighbor embedding (**t-SNE**) sorts vectors according to their similarity

$$r_i(\vec{a}) = \frac{1}{s_i}\left(T_i(\vec{a}) - D_{i,sh}(\vec{a})\right),$$

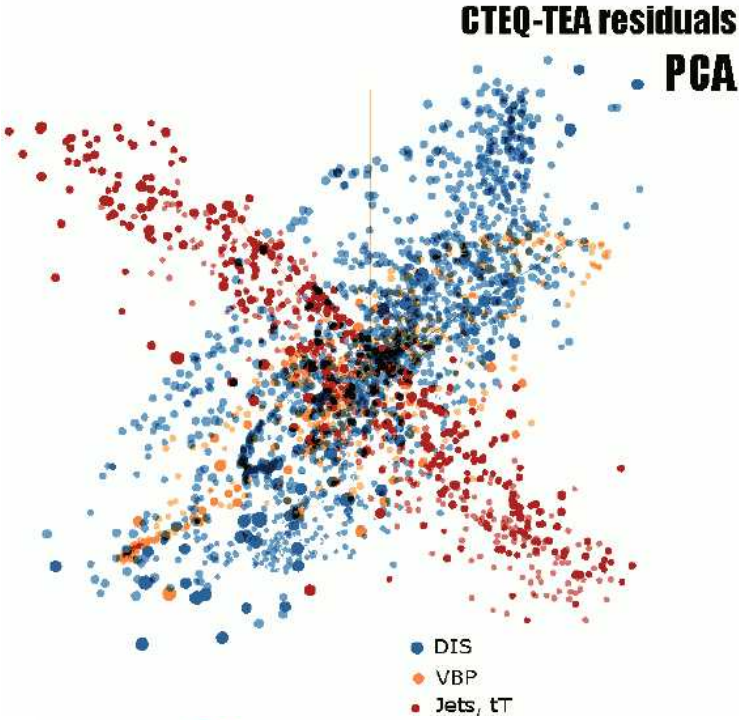
Manifolds of data residuals

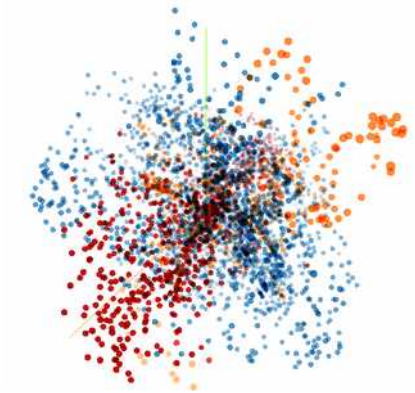
The 2N-dimensional distribution of δ_i is easy to analyze with data-mining tools…

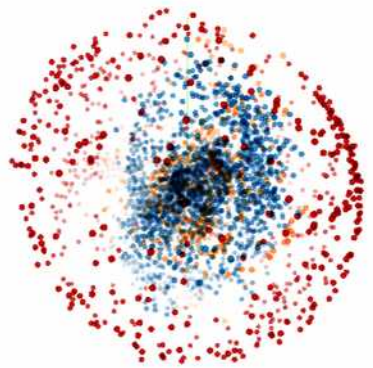
…to sort the fitted data points according to their PDF dependence (expressed by lengths and directions of δ_i);

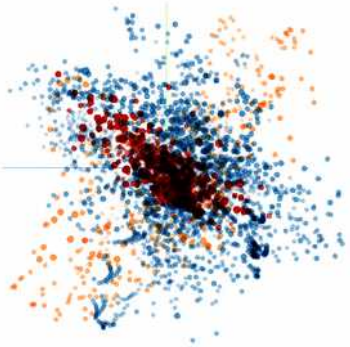
…to identify high-value data points (having long δ_i that point away from the rest of vectors).

Some projections separate DIS, DY, jet and tt-bar data residuals according to their PDF dependence.









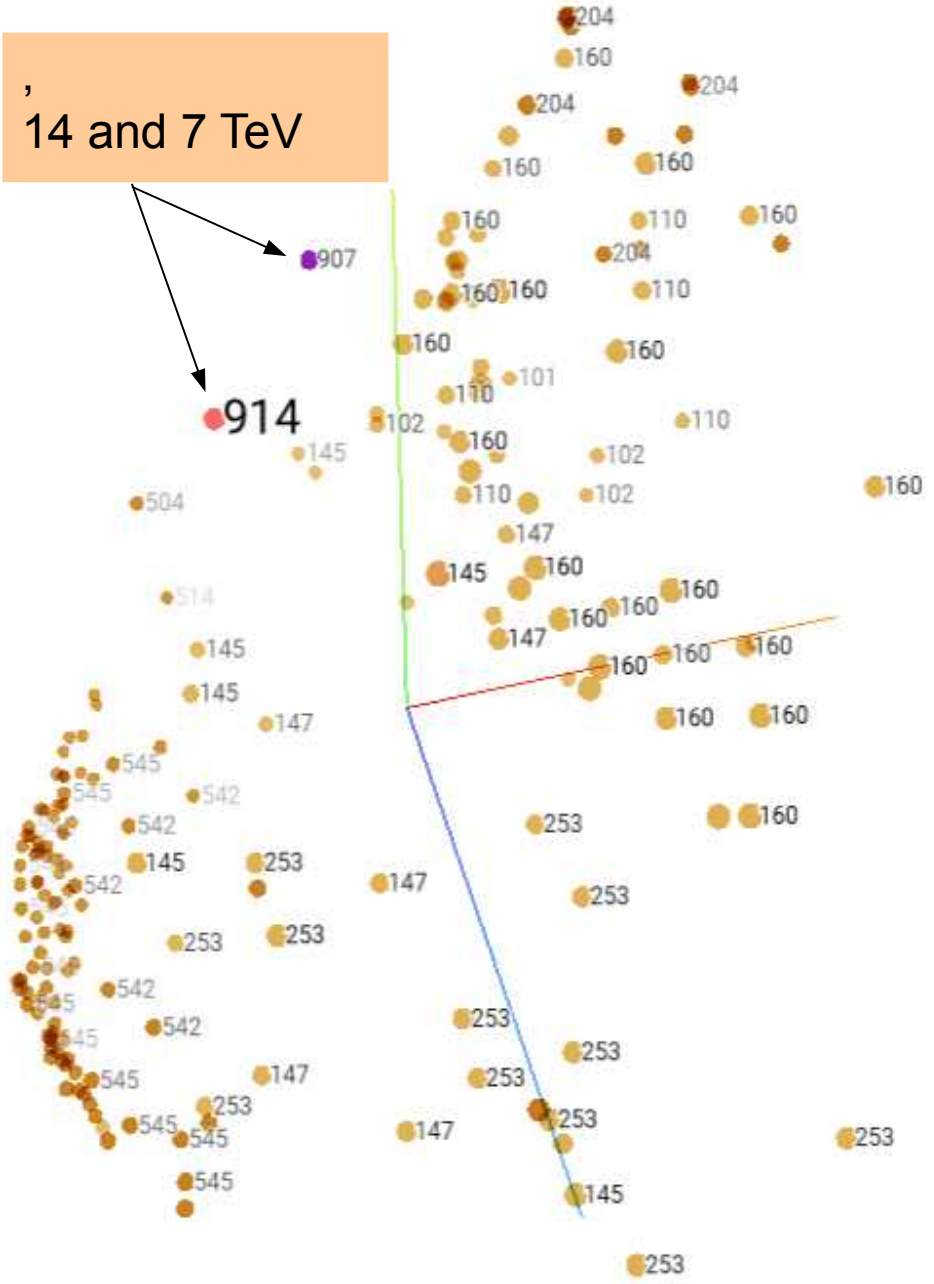
A PDF-dependent quantity *f* such as the Higgs cross section at 7 or 14 TeV (ID=907, 914), defines a direction δ_f in the (2)N-dim space.

The 3-dim projection on the right shows 300 vectors δ_f of the CT14HERA2 global set whose directions are closest to $\delta_f[\sigma\ (H)]$.

These vectors are given by the experiments:

160=HERA I+II; 101, 102=BCDMS;
110=CCFR F2p; 147, 145=HERA I+II ; 204=E866 ;
253= 8 TeV; 542, 545=CMS jets 7, 8 TeV; 504, 514=Tevatron jets

The net constraint of the i-th point on including systematic errors, is quantified by the projection of δ_f on $\delta_f[\sigma\ (H)]$ called the sensitivity $S_{f,i}$.



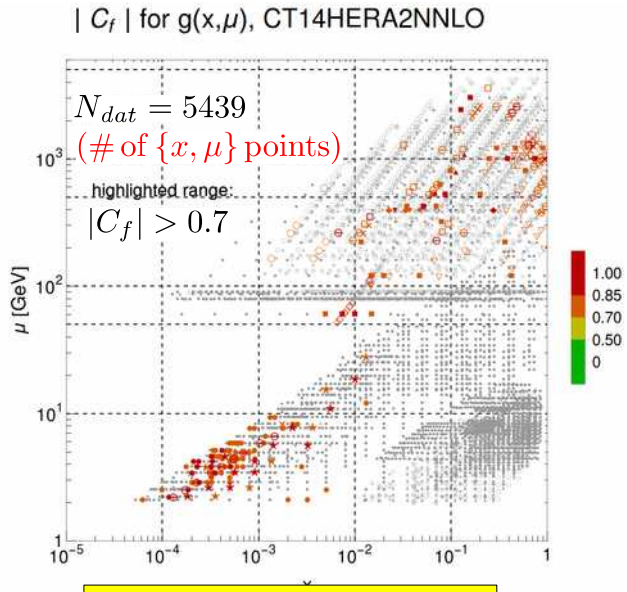
Sensitivity of expt E = sum of $S_{f,i}$ over data points in E

Thanks & Conclusions

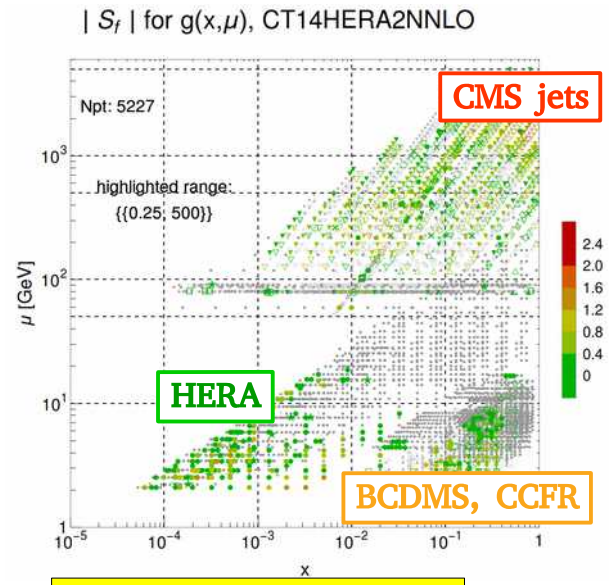
Bo-Ting Wang, Tim Hobbs, S. Doyle,
J. Gao, T.-J. Hou, & Pavel Nadolsky

C T E Q

New Theoretical Tools



Correlation

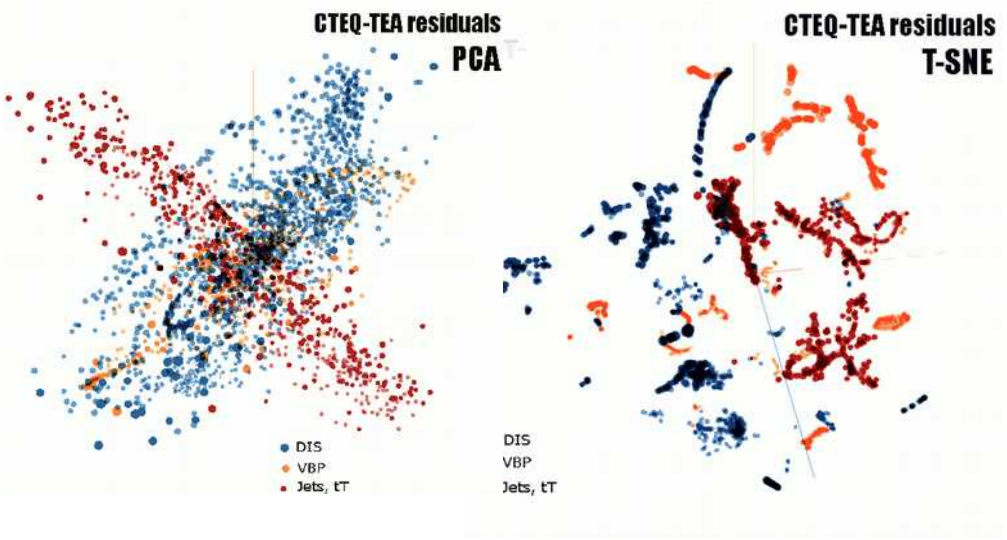


Sensitivity

Ranking Tables

No.	Exp. ID	N_d	$\sum_f \lvert S_f^E \rvert$	$\langle \sum_f \lvert S$
1	160	1120.	620.	0.092
2	111	86	218.	0.423
3	101	337	184.	0.090
4	104	123	169.	0.229
5	102	250	141.	0.093
6	109	96	115.	0.199
7	201	110	113	0.158

Ranking Tables



Innovative Approaches

Give it a try ...

http://metapdf.hepforge.org/PDFSense/

conclusions and future directions

- we have developed a very general framework that can be **extended to other datasets and PDF parametrizations**

  → explore the impact of future high energy datasets – e.g., LHC information

  → sensitivities of pseudodata sets (LHeC, EIC); for these, at least require Monte Carlo statistics and systematic uncertainties in bins of definite x and Q^2

  → extend to other processes?? e.g., **semi-inclusive** production of light mesons (π, K) – guide efforts to map meson structure

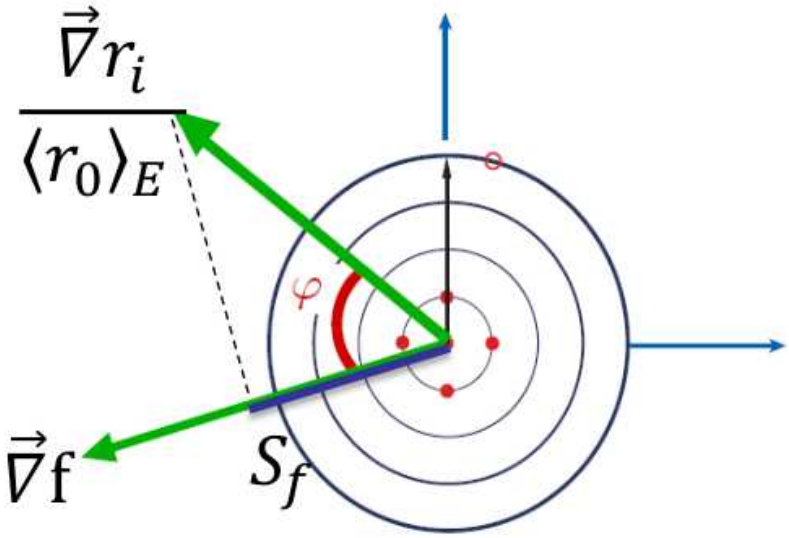
    … or nuclear process? e.g., sensitivity to different scenarios of the deuteron wave function…

- we can compute sensitivities for many derived quantities: e.g., physical cross sections; PDF combinations; lattice-calculable moments, …

  → are there **other physically interesting quantities** to which an future facility might be sensitive?

Correlation C_f and sensitivity S_f

The relation of data point i on the PDF dependence of f can be estimated by:



- $C_f \equiv \mathrm{Corr}[\rho_i(\vec{a})), f(\vec{a})] = cos\varphi$

$\vec{\rho}_i \equiv \vec{\nabla} r_i / \langle r_0 \rangle_E$ -- gradient of r_i normalized to the r.m.s. average residual in expt E;

$$\left(\vec{\nabla} r_i\right)_k = \left(r_i(\vec{a}_k^+) - r_i(\vec{a}_k^-)\right)/2$$

C_f is **independent** of the experimental and PDF uncertainties. In the figures, take $|C_f| \gtrsim 0.7$ to indicate a large correlation.

- $S_f \equiv |\vec{\rho}_i| cos\varphi = C_f \frac{\Delta r_i}{\langle r_0 \rangle_E}$ -- projection of $\vec{\rho}_i(\vec{a})$ on $\vec{\nabla} f$

S_f is proportional to $cos\varphi$ and the ratio of the PDF uncertainty to the experimental uncertainty. We can sum $|S_f|$.

In the figures, take $|S_f| > 0.25$ to be significant.

Vectors of data residuals

For every data point i ,
construct a vector of residuals $r_i(a_k^{\pm})$
for 2N Hessian eigenvectors. k=1, ..., N,
with N=28 for CT14 NNLO.

For example, **define**

$$\vec{\delta}_i = \{\delta^+_{i,1}, \delta^-_{i,1}, ...\delta^+_{i,N}, \delta^-_{i,N}, \} \qquad \{N = 28\}$$

$$\delta^{\pm}_{i,k} \equiv \left(r_i(\vec{a}^{\pm}_k) - r_i(\vec{a}_0)\right) / \langle r_0 \rangle_E$$

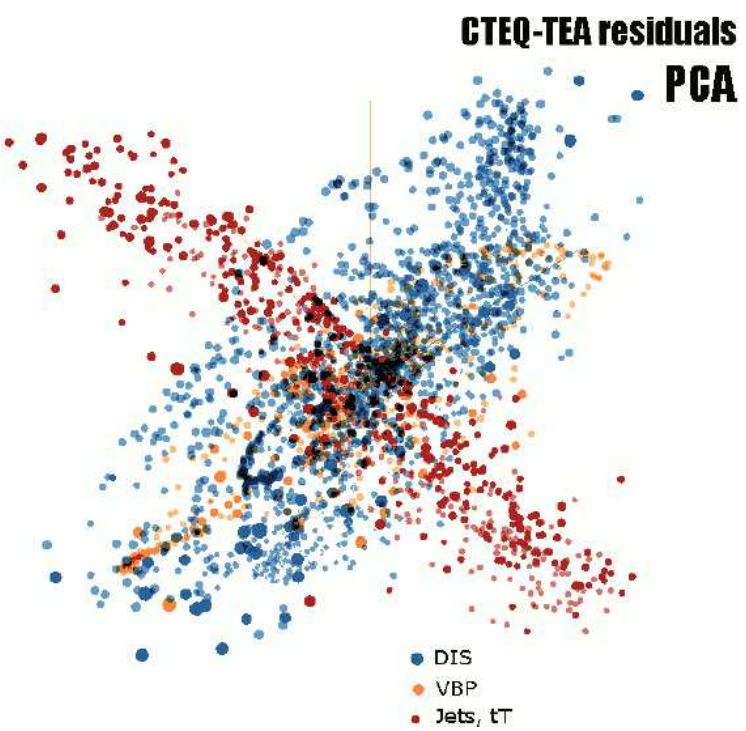
-- a 56-dim vector normalized to $\langle r_0 \rangle_E$, the root-mean-squared residual for the experiment *E* for the central fit a_0

$$\langle r_0 \rangle_E \equiv \sqrt{\frac{1}{N_{pt}} \sum_{i=1}^{N_{pt}} r_i^2(\vec{a}_0)} \sim \sqrt{\frac{\chi^2_E(\vec{a}_0)}{N_{pt}}}$$

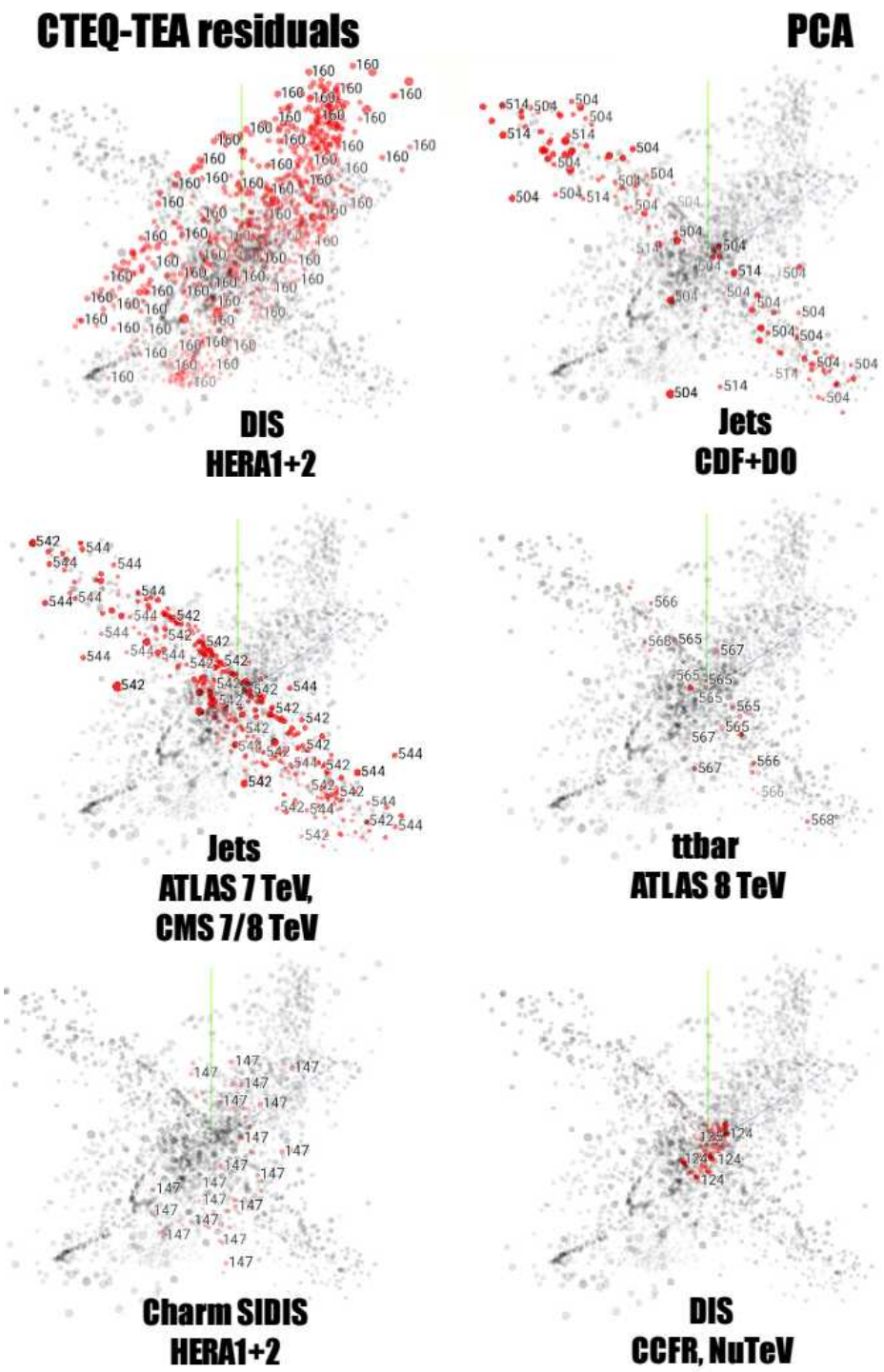
with

$$\langle r_0 \rangle_E \sim 1$$

in a good fit to *E*



The TensorFlow Embedding Projector (http://projector.tensorflow.org) represents CT14HERA2 vectors by their 10 principal components indicated by scatter points. A sample 3-dim. projection of the 56-dim. manifold is shown above. A symmetric 28-dim. representation can be alternatively used.


CTEQ-TEA residuals
PCA
DIS
HERA1+2
Jets
CDF+D0
Jets
ATLAS 7 TeV,
CMS 7/8 TeV
ttbar
ATLAS 8 TeV
Charm SIDIS
HERA1+2
DIS
CCFR, NuTeV