Collaborative Phenotype Inference from Comorbid Substance Use Disorders and Genotypes

BIBM 2017@Kansas City

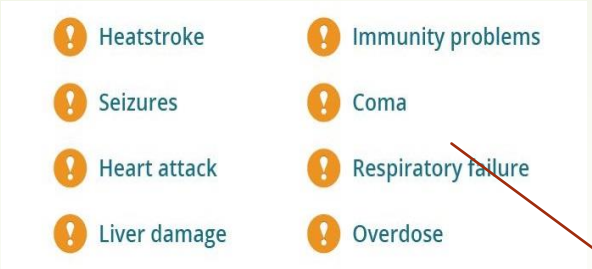
Presenter: Jin Liu

University of Connecticut

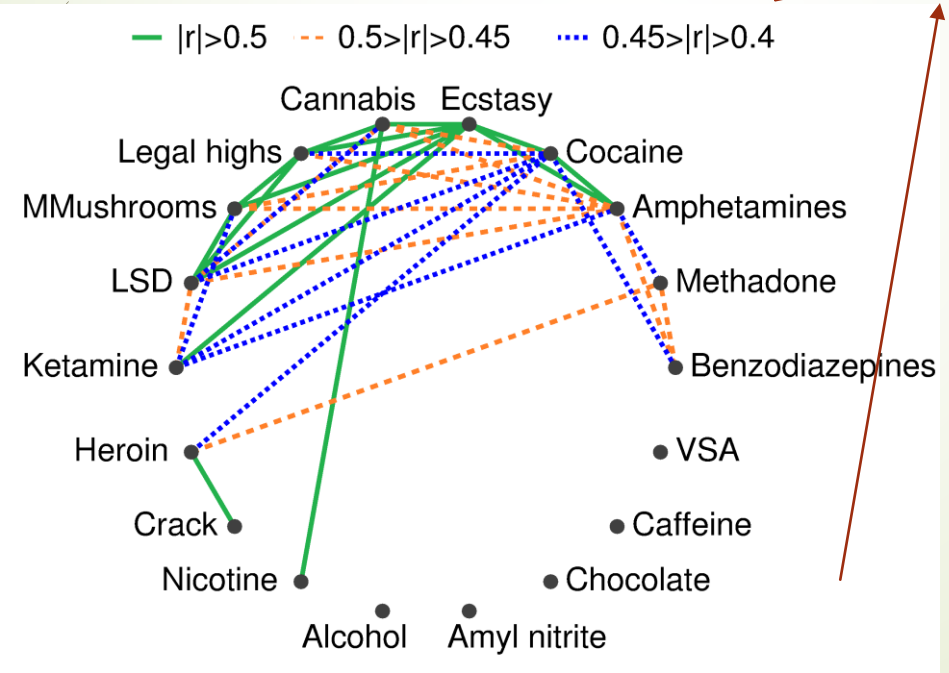
Joint work with Jiangwen Sun, Xinyu Wang, Henry Kranzler, Joel Gelernter and Jinbo Bi

Comorbid substance use disorders(CSUD)

- very dangerous



- validated by strong correlations

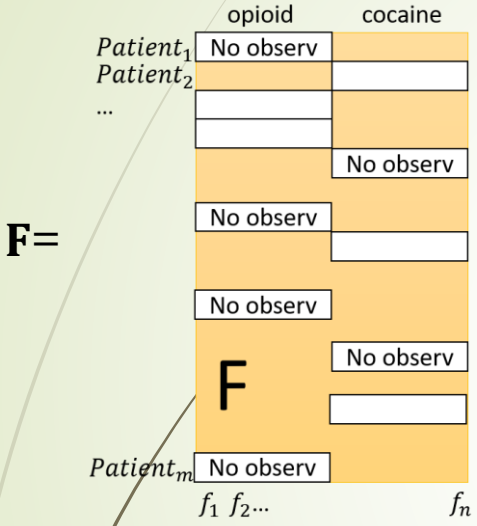


USE BOTH
GENOTYPES
RELATED PHENOTYPES
TO INFER
DIAGNOSTIC
CRITERIA FOR
UNREPORTED SUD

- Dependence on different substances are correlated both phenomenologically and biologically.

Inferring SUD diagnostic criteria

- Our phenotypic imputation problem



$\mathbf{F}=$

additional useful information:
associated genetic variants; known similarities between comorbid disorders.
Often referred to as side or auxiliary information in matrix completion

- Classical low-rank matrix completion problem

$\mathbf{F}=$

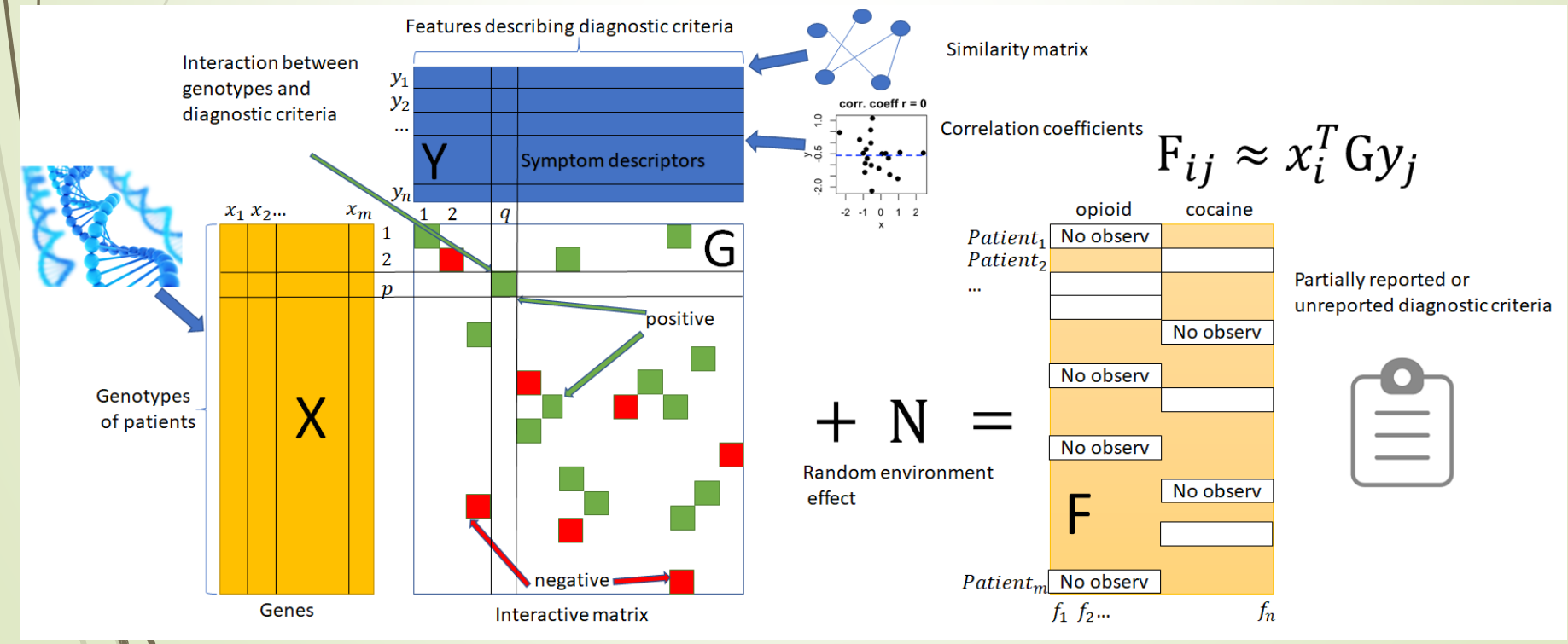
Optimization Problem:

$$\min_{\mathbf{E}} \|\mathbf{E}\|_* \quad \text{subject to} \quad R_\Omega(\mathbf{E}) = R_\Omega(\mathbf{F})$$

where Ω is the set of indices of **observed entries** in $\mathbf{F}$ and $\|\cdot\|_*$ computes the **nuclear norm (low-rank regularizer)**.

The proposed method

- Our collaborative inference method of CSUD diagnostic criteria using side information.
- **Side information of patients: Genotypes**
- **Side information of diagnostic criteria: Sampled Corr Coef matrix; Similarity matrix**



The proposed method

$$\min_{\mathbf{G}} \frac{1}{2}\left\|\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{Y}-\mathbf{E}\right\|_F^2 + \lambda_G g(\mathbf{G}) + \lambda_E \|\mathbf{E}\|_*,$$
$$\text{s.t.} \quad R_\Omega(\mathbf{E}) = R_\Omega(\mathbf{F})$$

- The proposed method uses a low-rank matrix E to directly approximate matrix F and then estimates E from matrix X and Y.

Proposed method

$$\min_{\mathbf{G}} \frac{1}{2} \left\| \mathbf{X}^{\mathrm{T}} \mathbf{G} \mathbf{Y} - \mathbf{E} \right\|_F^2 + \lambda_G g(\mathbf{G}) + \lambda_E \|\mathbf{E}\|_*,$$
$$\text{s.t.} \quad R_\Omega(\mathbf{E}) = R_\Omega(\mathbf{F})$$

- The proposed method uses a low-rank matrix E to directly approximate matrix F and then estimates E from matrix X and Y.

- The proposed model can identify crucial interactions between specific genotypes and diagnostic criteria by enforcing the **sparsity** in **G**. $(g(\mathbf{G}) = \|\mathbf{G}\|_1)$

Adaptive LADMM Algorithm

- We propose a new stochastic Linearized Alternative Direction Method of Multipliers (**StoLADMM**) algorithm
- by substituting $\mathbf{C} = \mathbf{E} - \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{Y}$.
- The augmented Lagrangian function is given by

$$\begin{aligned}
&\mathcal{L}(\mathbf{E}, \mathbf{G}, \mathbf{C}, \mathbf{M}_1, \mathbf{M}_2, \beta) \\
&= \frac{1}{2}\|\mathbf{C}\|_F^2 + \lambda_G\|\mathbf{G}\|_1 + \lambda_E\|\mathbf{E}\|_* + \langle \mathbf{M}_1, R_\Omega(\mathbf{E} - \mathbf{F})\rangle \\
&+ \langle \mathbf{M}_2, \mathbf{E} - \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{Y} - \mathbf{C}\rangle + \frac{\beta}{2}\|R_\Omega(\mathbf{E} - \mathbf{F})\|_F^2 \\
&+ \frac{\beta}{2}\left\|\mathbf{E} - \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{Y} - \mathbf{C}\right\|_F^2
\end{aligned}$$

- Solve each variable alternatively.

Our efficient stochastic algorithm

- Effectiveness

1. convergence in expectation

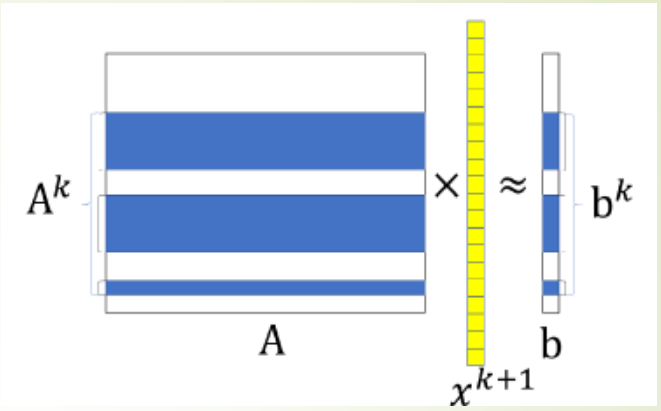
2. global optimal solution for our convex optimization problem

- Efficiency

1. Save memory costs

2. Can utilize parallel computing to speed up the algorithm

3. Without sacrificing performance notably.



Algorithm 1 The StoLADMM algorithm to solve $\mathbf{C}^k$, $\mathbf{G}^k$, $\mathbf{E}^k$, $k = 1, ..., K$

Input: $\mathbf{X}$, $\mathbf{Y}$ and $R_\Omega(\mathbf{F})$ with parameters λ_G, λ_E, τ_A, τ_B, ρ and β_{max}.

Output: $\mathbf{C}, \mathbf{G}, \mathbf{E}$;

Initialize $\mathbf{E}^0, \mathbf{G}^0, \mathbf{M}_1^0, \mathbf{M}_2^0$. Compute $\mathbf{A} = \mathbf{Y}^T \otimes \mathbf{X}^T$. $k = 0$,

repeat;

$\mathbf{C}^{k+1} = \frac{\beta}{\beta+1}(\mathbf{E}^k - \mathbf{X}^T\mathbf{G}^k\mathbf{Y} + \mathbf{M}_2^k/\beta)$;

$\mathbf{G}^{k+1} = reshape(\max(|\mathbf{g}^k - f_1^k/\tau_A| - \frac{\lambda_G}{\tau_A\beta}, 0) \odot sgn(\mathbf{g}^k - f_1^k/\tau_A))$ where f_1^k can be computed by (5);

$\mathbf{E}^{k+1} = SVT(\mathbf{E}^k - (f_2^k + f_3^k)/(2\tau_B), \lambda_E/2(\beta\tau_B))$ where f_2^k and f_3^k can be computed by (6);

$\mathbf{M}_1^{k+1} = \mathbf{M}_1^k + \beta(R_\Omega(\mathbf{E}^{k+1} - \mathbf{F}))$.

$\mathbf{M}_2^{k+1} = \mathbf{M}_2^k + \beta(\mathbf{E}^{k+1} - \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} - \mathbf{C}^{k+1})$.

$k = k + 1$ until convergence;

Return $\mathbf{C}, \mathbf{G}, \mathbf{E}$;

Experimental results

- Compared methods:
  - MAXIDE

    M. Xu, R. Jin, and Z. hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. Advances in Neural Information Processing Systems 26, pages 2301–2309, 2013

  - IMC

    N. Natarajan and I. S. Dhillon. Inductive matrix completion for predicting gene–disease associations. Bioinformatics, 30(12):i60–i68, 2014

  - DirtyIMC

    K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon. Matrix completion with noisy side information. Advances in Neural Information Processing Systems 28, pages 3429–3437, 2015.

- The relative mean squared error (**RMSE**) is used as the performance measurement.

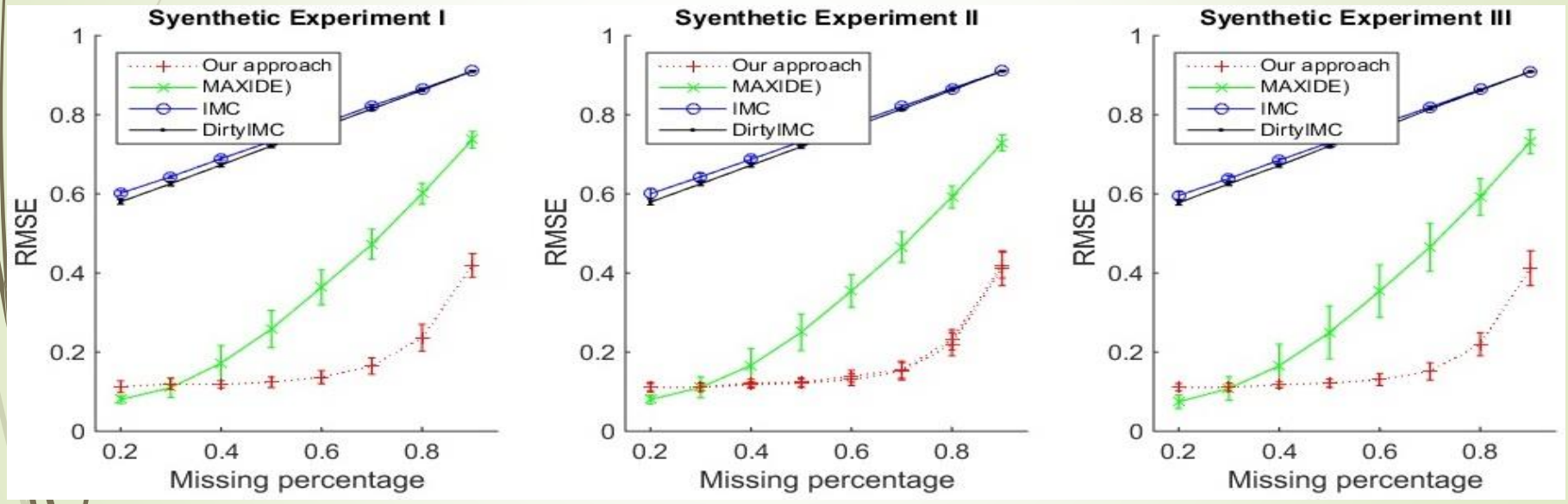
$$\mathrm{RMSE} = \frac{\left\| R_{\bar{\Omega}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{Y} - \mathbf{F}\right)\right\|_F^2}{\left\| R_{\bar{\Omega}}(\mathbf{F})\right\|_F^2}$$

Experimental results

- Synthetic Datasets:
  - $\mathbf{X}$ and $\mathbf{Y}$ were generated from **Gaussian**, **Poisson** and **Gamma** distributions.
  - $\mathbf{G}$ contains **20%** of non-zero components.
  - $\mathbf{F} = \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{Y} + \mathbf{N}$ where $\mathbf{N}$ represents the noise.
  - Then, the values of $\mathbf{F}$ were dropped by $[\mathbf{20\%} - \mathbf{80\%}]$ to test the recovery rate of the methods.

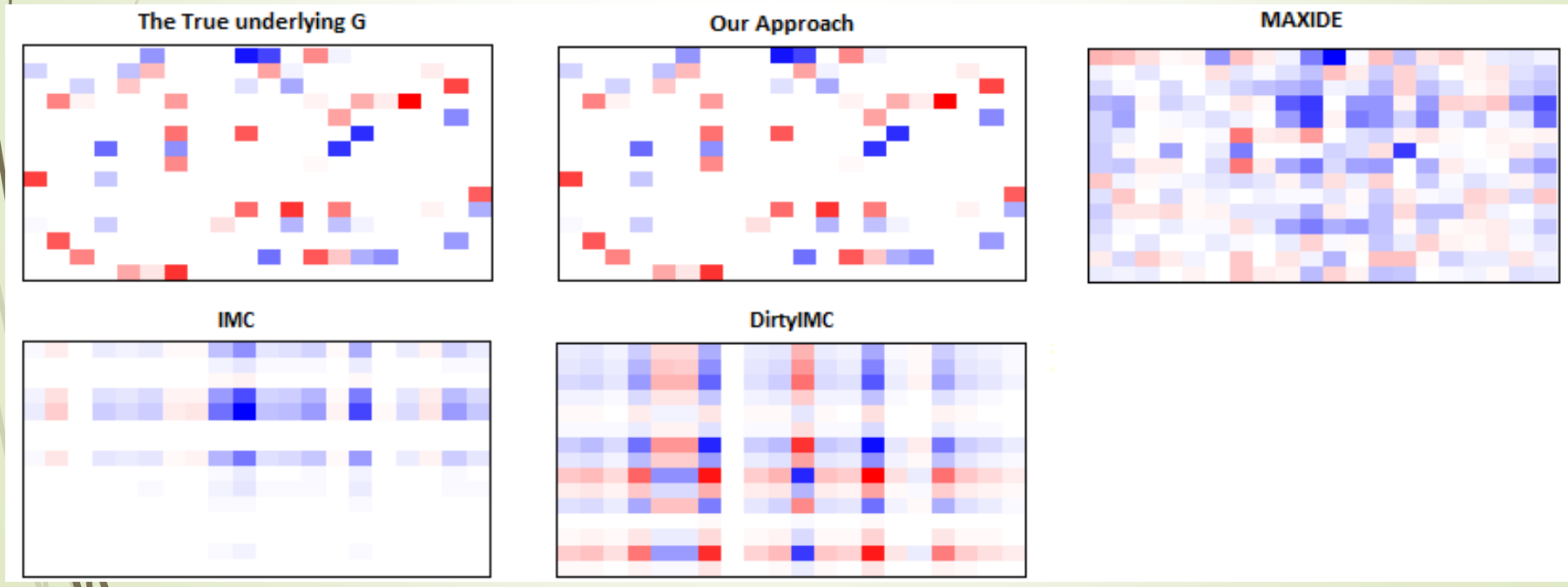
Experimental results

- Synthetic Datasets:
  - RMSE for all compared methods



Experimental results

- Synthetic Datasets:
  - RMSE for all compared methods
  - Recovery of true underlying $\mathbf{G}$.



Experimental results

- Synthetic Datasets
- Cormorbid Substance Use Data:
  - A total of **7,189 subjects** were aggregated from family and case-control based genetic studies of cocaine use disorder (CUD) and opioid use disorder (OUD).
  - The **383 genetic variants** identified in our GWAS were used as side feature matrix X with the size 7189 by 383.
  - The correlations between **22 CUD and OUD symptoms** formed a correlation matrix which was used as side features matrix Y with the size 22 by 22.
  - We randomly removed the phenotypes of q% CSUD patients associated with either opioid or cocaine use (not both). Then our partially observed F is the matrix with the size **7189 by 22**, which needs inference.

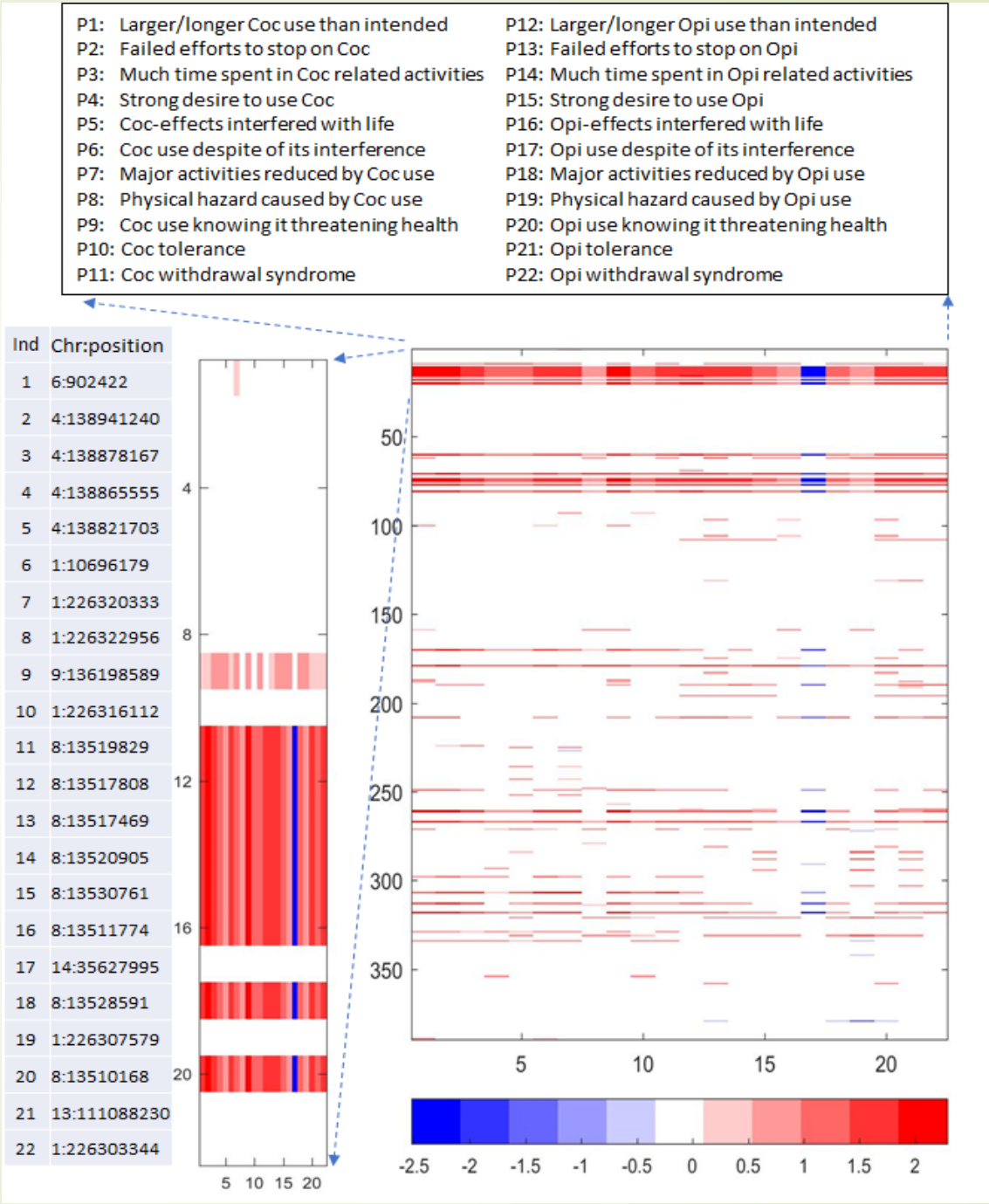
Experimental results

- Synthetic Datasets
- Cormorbid Substance Use Data:

q		StoLADMM	LADMM	DirtyIMC	IMC	MAXIDE	BM
20%	RMSE	0.236	0.231	0.297	**0.230**	0.235	0.567
	time(s)	30.938	664.515	45.366	**21.053**	4732.718	NaN
40%	RMSE	**0.226**	0.234	0.298	0.235	0.236	0.582
	time(s)	29.953	982.212	21.063	**20.803**	3772.202	NaN
60%	RMSE	**0.228**	0.236	0.301	0.237	0.235	0.581
	time(s)	28.719	815.841	**20.269**	36.737	4718.916	NaN
80%	RMSE	**0.236**	0.237	0.303	0.239	0.241	0.585
	time(s)	30.547	877.886	**23.906**	32.872	4011.692	NaN
100%	RMSE	**0.223**	0.239	0.303	0.246	0.242	0.574
	time(s)	30.172	489.770	**22.922**	24.653	3695.292	NaN

TABLE II: The inference results on the Opioid-Cocaine data.

Interaction Matrix



Conclusion

- We adopted a matrix completion approach to infer SUD criteria using both correlation among criteria of different conditions and genotypes as side information.
- By imposing sparse prior on the model parameters, the method can find a sparse interactive matrix that connects specific genotypes to diagnostic criteria.
- We introduced an efficient stochastic LADMM algorithm to solve the optimization problem in this method.
- The empirical evaluation shows that our method can significantly enhance the running efficiency with minimal adverse effects on the imputation accuracy.

Any Questions?

Thank you.