# Supporting Information for ”Augmentation of FTIR Spectral Datasets Using Wasserstein Generative Adversarial Networks For Cancer Liquid Biopsies”

Rose G. McHardy,[a,b] Georgios Antoniou,[b] Justin J. A. Conn,[b] Matthew J. Baker,[b,c] and David S. Palmer,[*a,b]

[a] Department of Pure and Applied Chemistry, Thomas Graham Building, 295 Cathedral Street, University of Strathclyde, Glasgow, G1 1XL, UK

[b] Dxcover Ltd, Royal College Building, 204 George Street, Glasgow, G1 1XW, UK

[c] School of Medicine, Faculty of Clinical and Biomedical Sciences, University of Central Lancashire, Preston, PR1 2HE, UK

Table S1: Distribution of full pancreatic cancer dataset

| Full PC | | C | NC | Total |
|---|---|---|---|---|
| Age, years | Mean | 65 | 57 | 59 |
| | Min-max | 39-87 | 20-90 | 20-90 |
| Sex, n(%) | Female | 69 (42) | 260 (57) | 329 (53) |
| | Male | 97 (58) | 199 (43) | 296 (47) |
| Cancer stage, n(%) | I | 6 (4) | - | 6 (1) |
| | II | 64 (39) | - | 64 (10) |
| | III | 69 (42) | - | 69 (11) |
| | IV | 19 (11) | - | 19 (3) |
| | Unknown | 8 (5) | - | 8 (1) |

Table S2: Distribution of 525 patient pancreatic cancer dataset

| 525 PC | | C | NC | Total |
|---|---|---|---|---|
| Age, years | Mean | 65 | 57 | 59 |
| | Min-max | 39-87 | 20-90 | 20-90 |
| Sex, n(%) | Female | 44 (38) | 230 (56) | 274 (52) |
| | Male | 72 (62) | 179 (44) | 251 (48) |
| Cancer stage, n(%) | I | 4 (3) | - | 4 (1) |
| | II | 44 (38) | - | 44 (8) |
| | III | 47 (41) | - | 47 (9) |
| | IV | 13 (11) | - | 13 (2) |
| | Unknown | 8 (7) | - | 8 (2) |

Table S3: Distribution of full colorectal cancer dataset

| Full PC | | C | NC | Total |
|---|---|---|---|---|
| Age, years | Mean | 66 | 57 | 60 |
| | Min-max | 29-85 | 20-90 | 20-90 |
| Sex, n(%) | Female | 64 (32) | 260 (57) | 324 (49) |
| | Male | 136 (68) | 199 (43) | 335 (51) |
| Cancer stage, n(%) | I | 36 (18) | - | 36 (5) |
| | II | 70 (35) | - | 70 (11) |
| | III | 67 (34) | - | 67 (10) |
| | IV | 27 (14) | - | 27 (4) |

Table S4: Distribution of 559 patient colorectal cancer dataset

| 559 Colorectal | | C | NC | Total |
|---|---|---|---|---|
| Age, years | Mean | 66 | 57 | 59 |
| | Min-max | 29-85 | 20-90 | 20-90 |
| Sex, n(%) | Female | 48 (32) | 234 (57) | 282 (50) |
| | Male | 103 (68) | 174 (43) | 277 (50) |
| Cancer stage, n(%) | I | 27 (18) | - | 27 (5) |
| | II | 52 (34) | - | 52 (9) |
| | III | 51 (34) | - | 51 (9) |
| | IV | 21 (14) | - | 21 (4) |

Table S5: Distribution of 100 patient colorectal dataset

| 100 Colorectal | | C | NC | Total |
|---|---|---|---|---|
| Age, years | Mean | 67 | 58 | 62 |
| | Min-max | 37-84 | 20-80 | 20-84 |
| Sex, n(%) | Female | 16 (32) | 25 (50) | 41 (41) |
| | Male | 34 (68) | 25 (50) | 59 (59) |
| Cancer stage, n(%) | I | 9 (18) | - | 9 (9) |
| | II | 18 (36) | - | 18 (18) |
| | III | 17 (34) | - | 17 (17) |
| | IV | 6 (12) | - | 6 (6) |

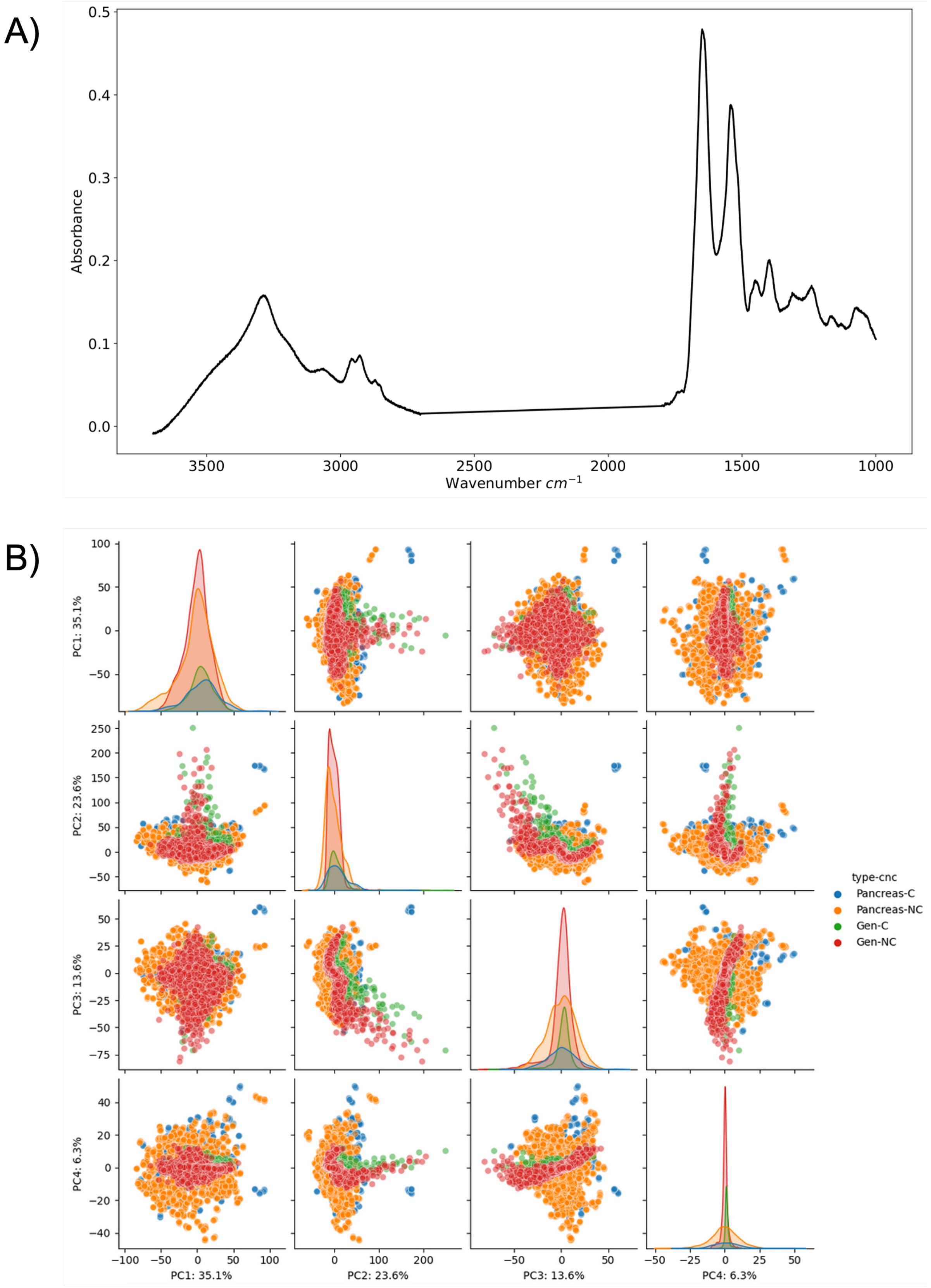

Figure S1: Output from WGAN with 3 generator layers and 1 critic layers with 512 units in the dense layer: A) example of generated spectra from WGAN, B) generated spectra projected onto real spectra PCA space.

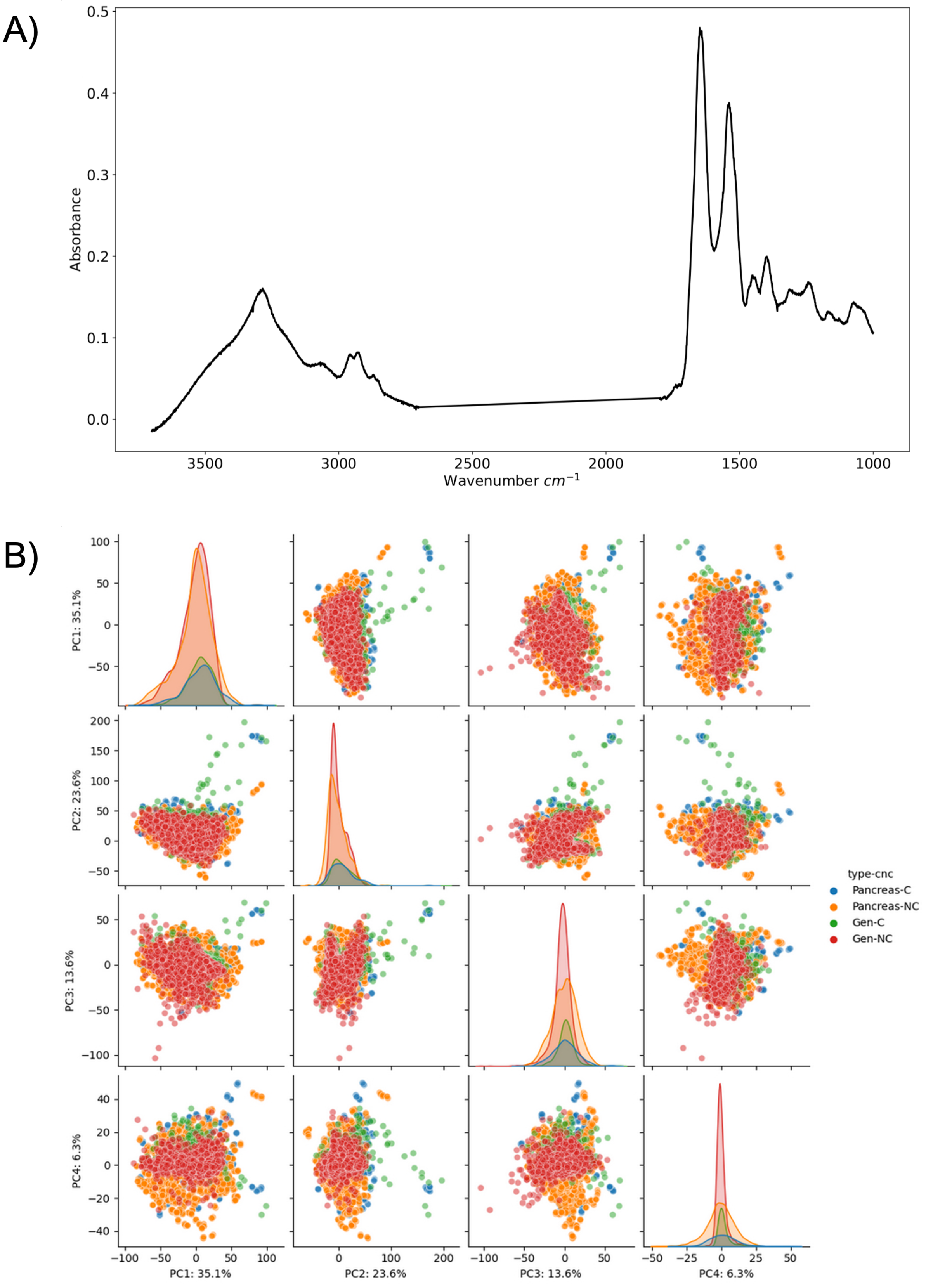

Figure S2: Output from WGAN with 3 generator layers and 2 critic layers with 1024 and 512 units in the dense layers: A) example of generated spectra from WGAN, B) generated spectra projected onto real spectra PCA space.

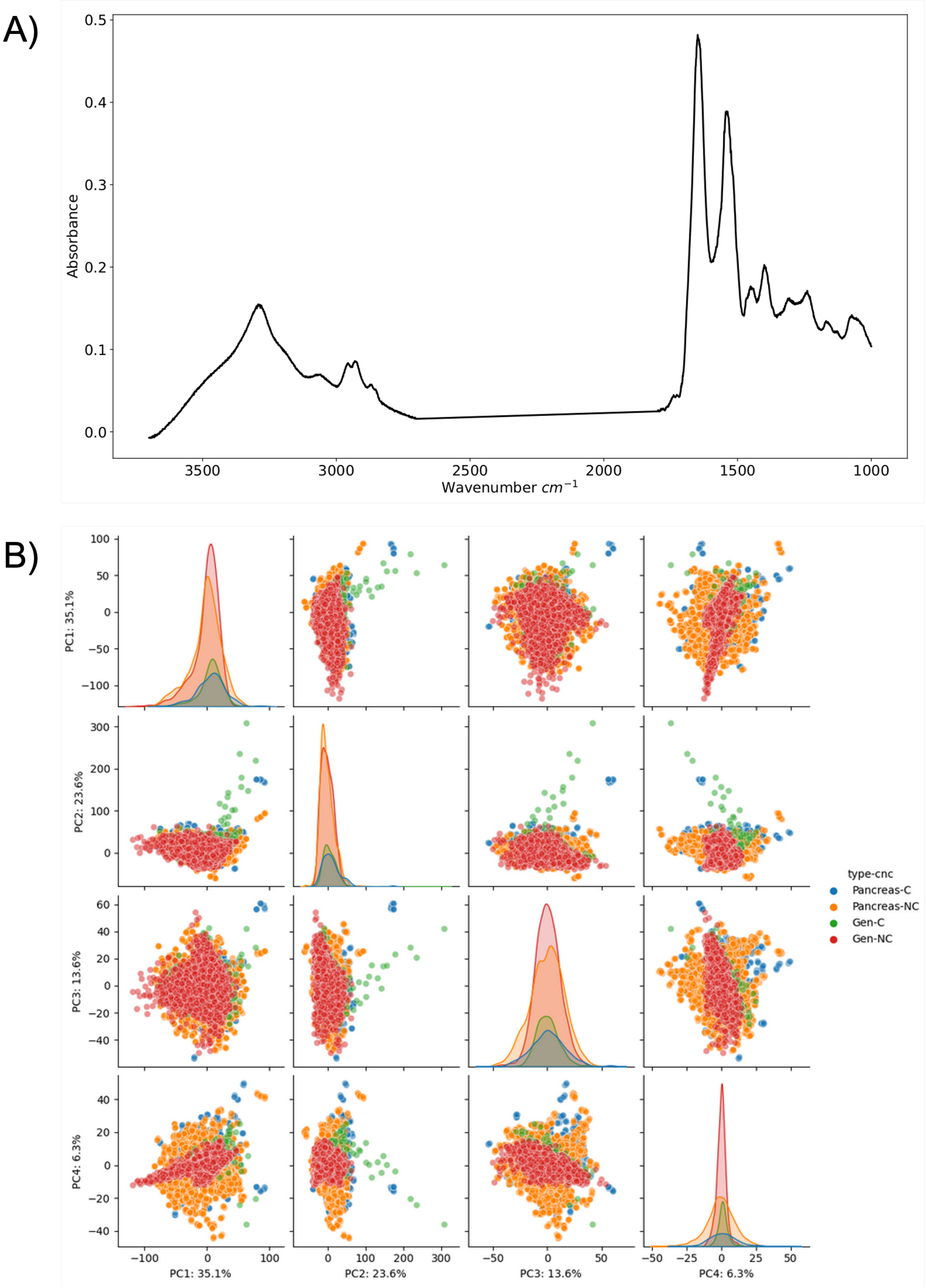

Figure S3: Output from WGAN with 3 generator layers and 2 critic layers with 512 and 256 units in the dense layers: A) example of generated spectra from WGAN, B) generated spectra projected onto real spectra PCA space.

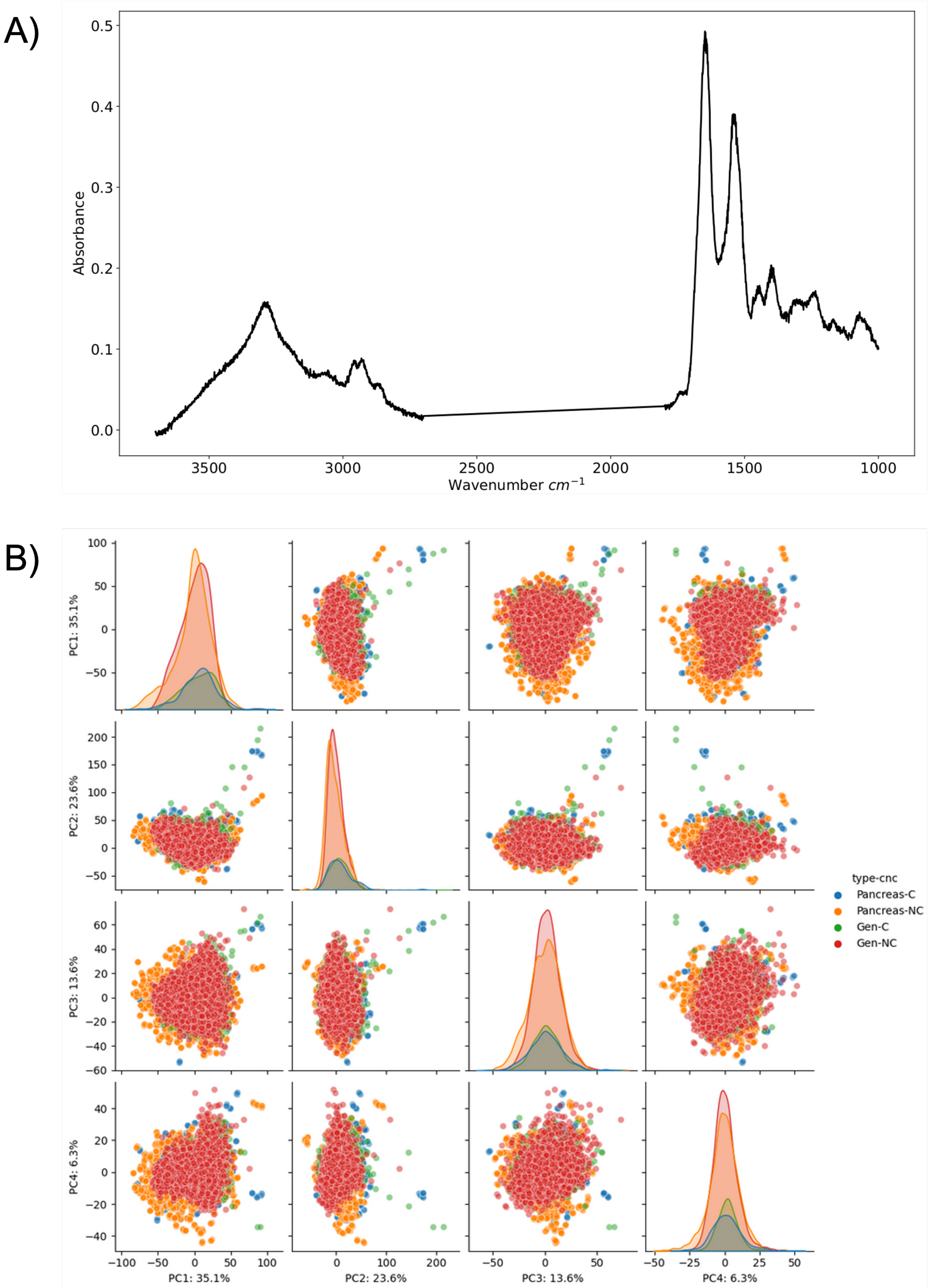

Figure S4: Output from WGAN with 3 generator layers and 3 critic layers with 2048, 1024, and 512 units in the dense layers: A) example of generated spectra from WGAN, B) generated spectra projected onto real spectra PCA space.

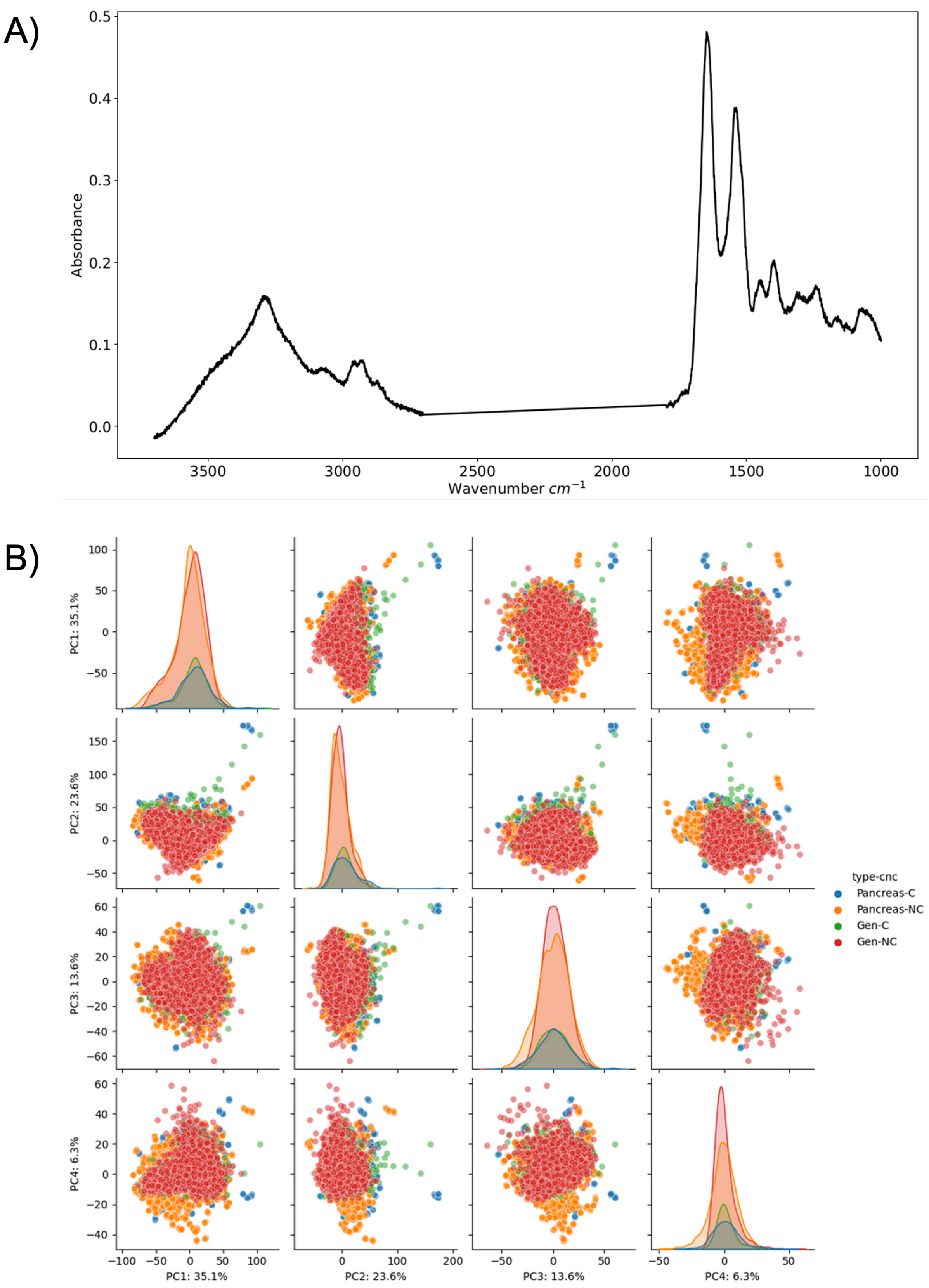

Figure S5: Output from WGAN with 3 generator layers and 3 critic layers with 1024, 512, and 256 units in the dense layers: A) example of generated spectra from WGAN, B) generated spectra projected onto real spectra PCA space.

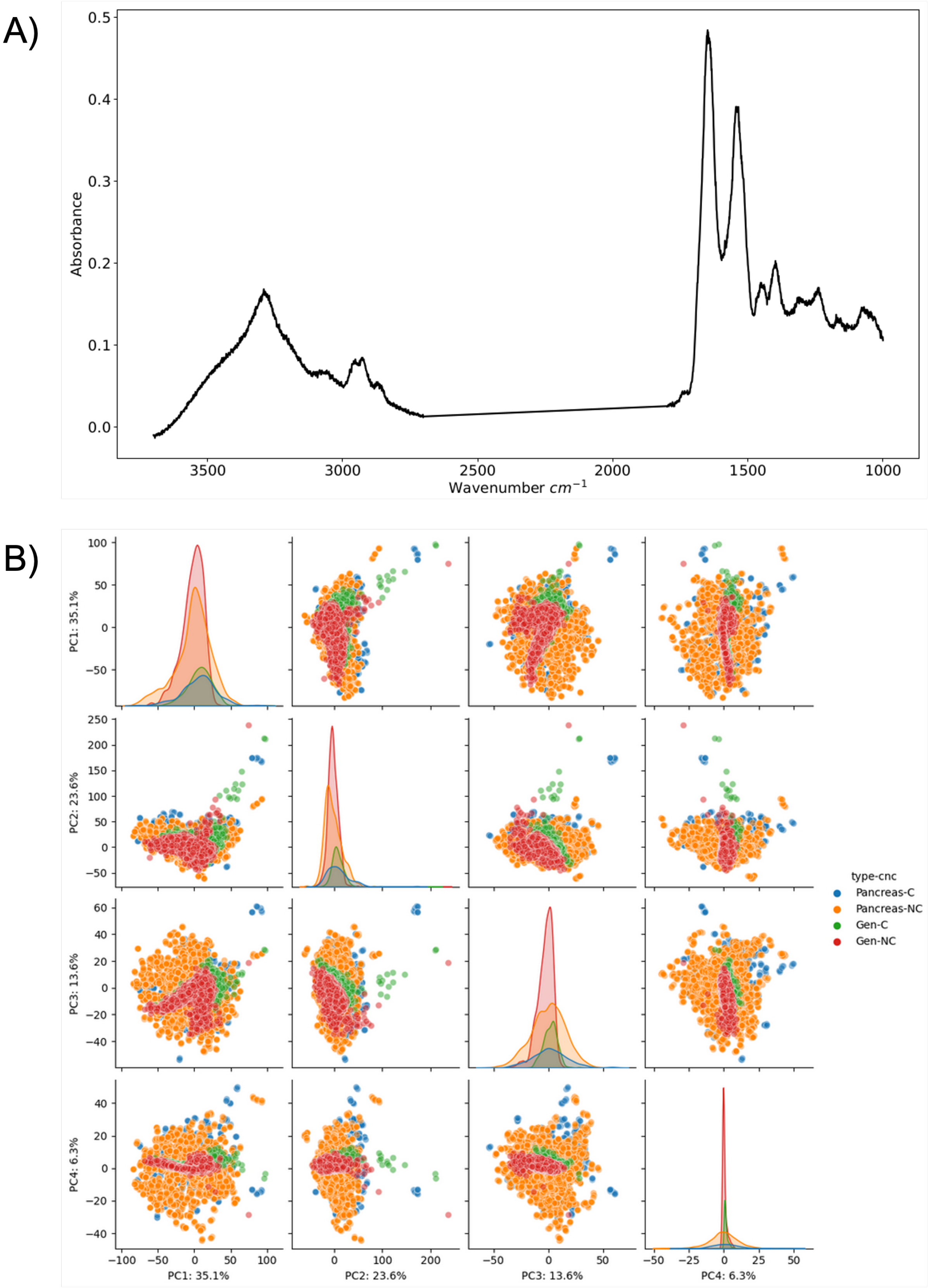

Figure S6: Output from WGAN with 3 generator layers and 3 critic layers with 512, 256, and 128 units in the dense layers: A) example of generated spectra from WGAN, B) generated spectra projected onto real spectra PCA space.

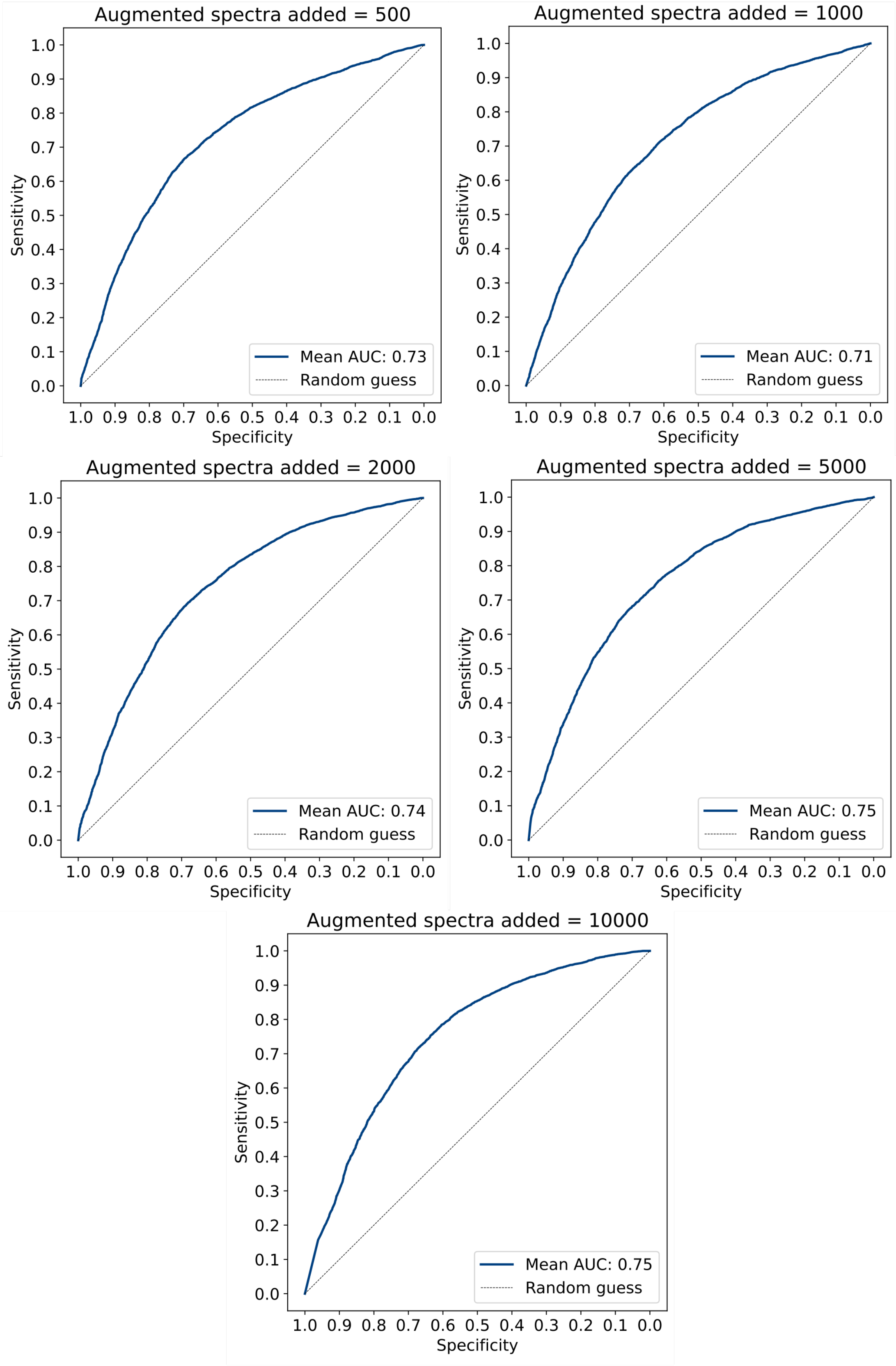

Figure S7: ROC curves for CNN models trained on augmented spectra generated by adding noise to spectra with the validation set containing 100-patient dataset only.

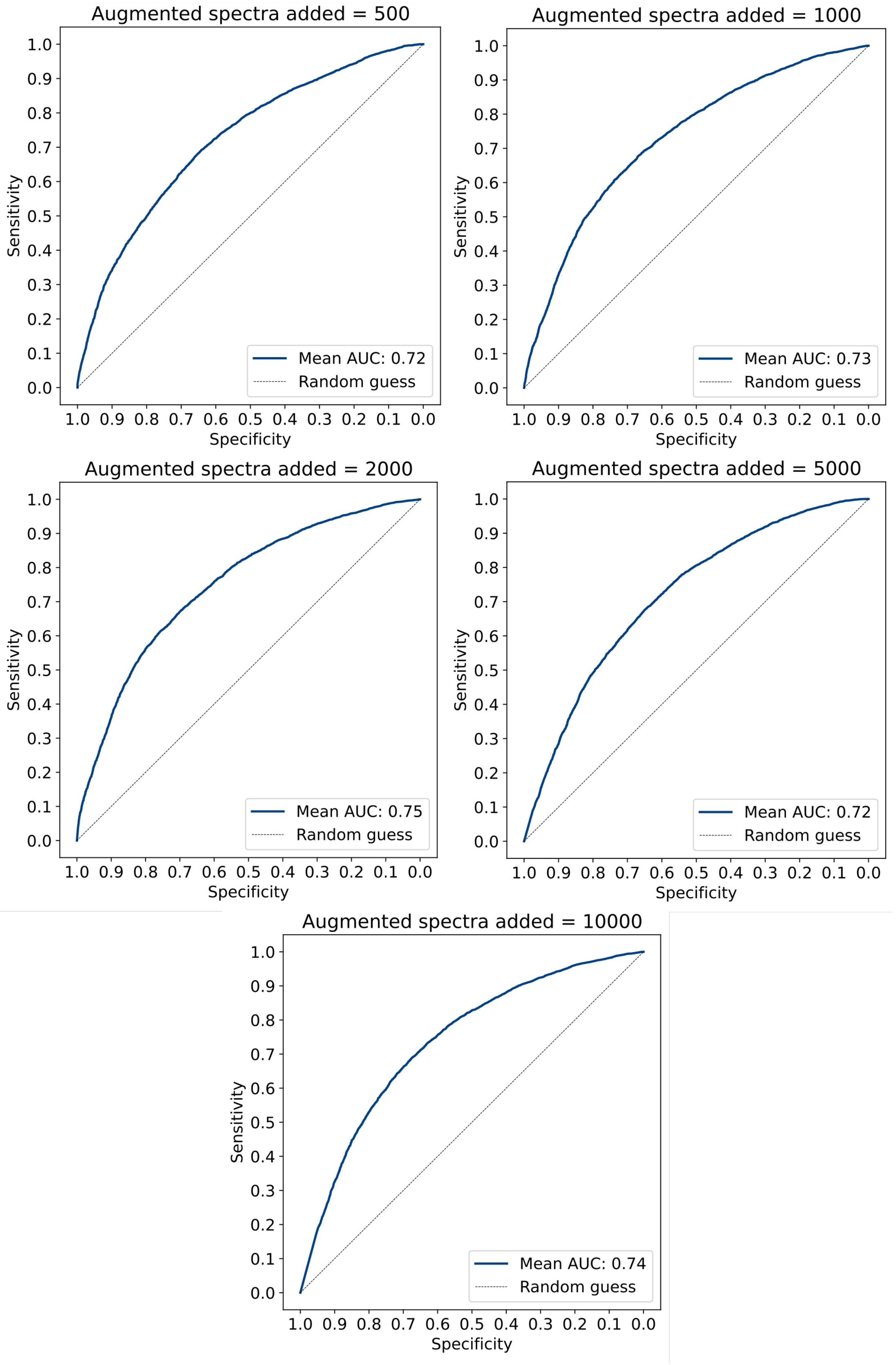

Figure S8: ROC curves for CNN models trained on augmented spectra generated by averaging spectra with the validation set containing 100-patient dataset only.

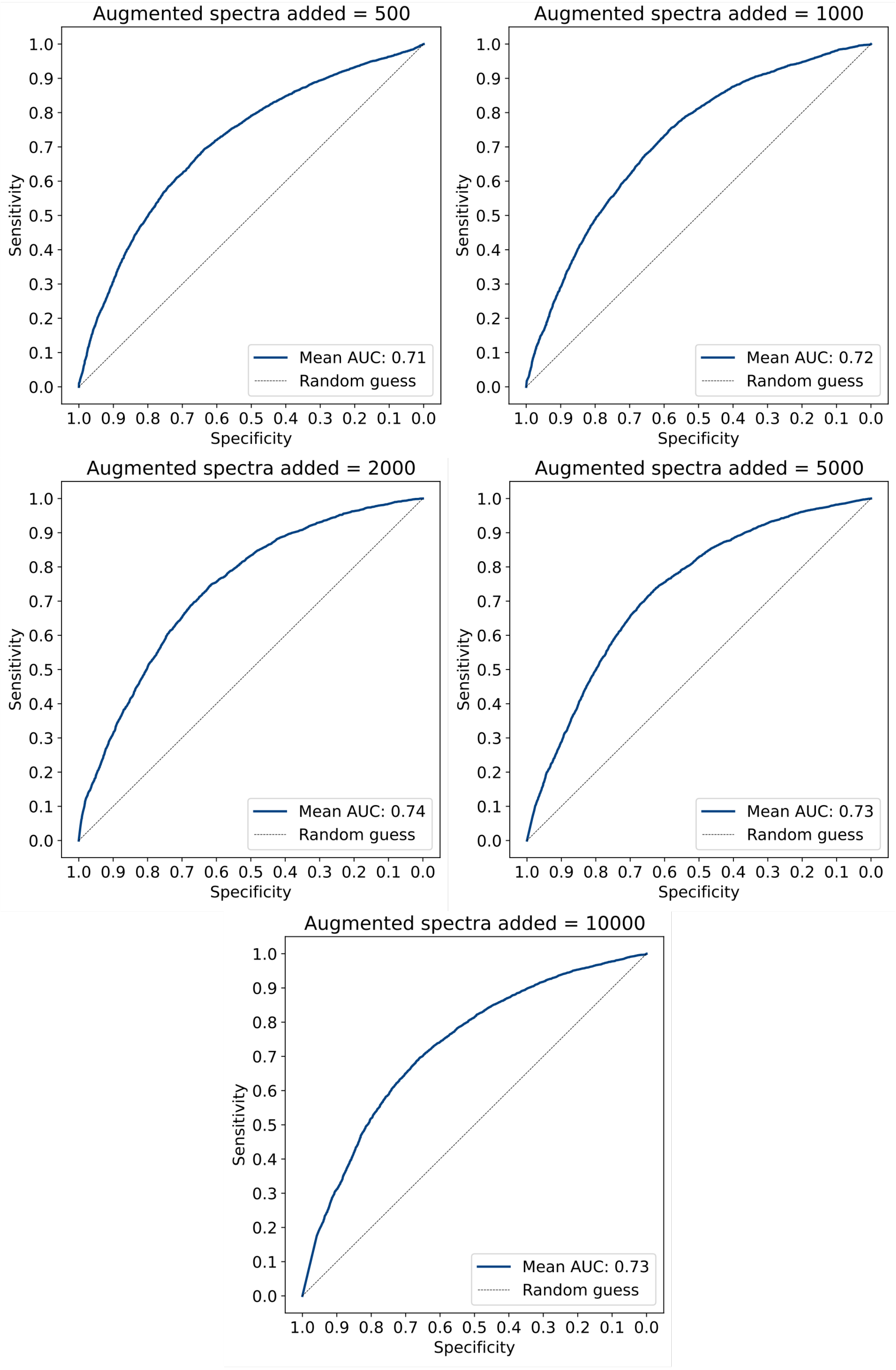

Figure S9: ROC curves for CNN models trained on augmented spectra generated by splicing spectra with the validation set containing 100-patient dataset only.

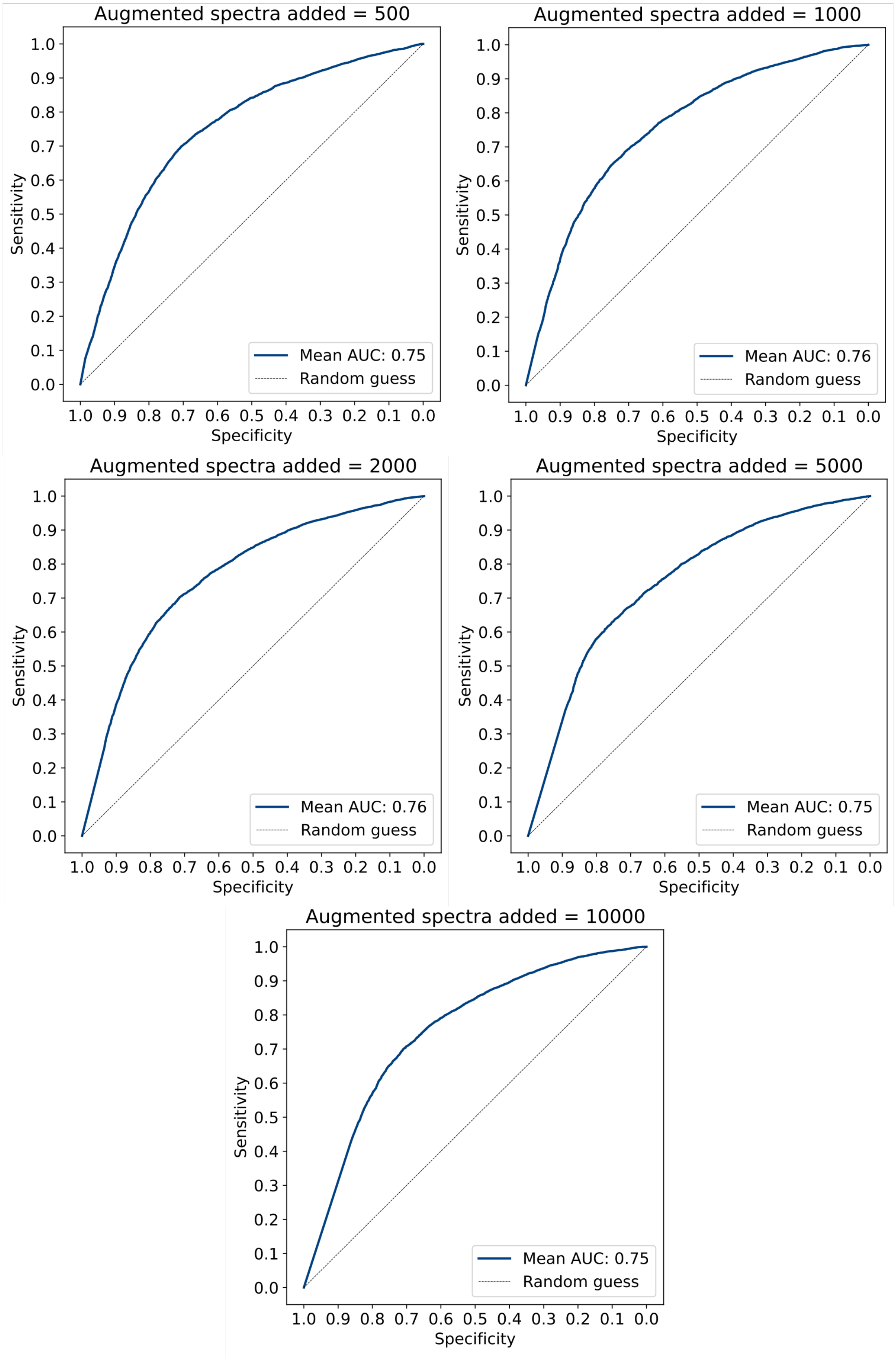

Figure S10: ROC curves for CNN models trained on WGAN augmented spectra with the validation set containing 100-patient dataset only.

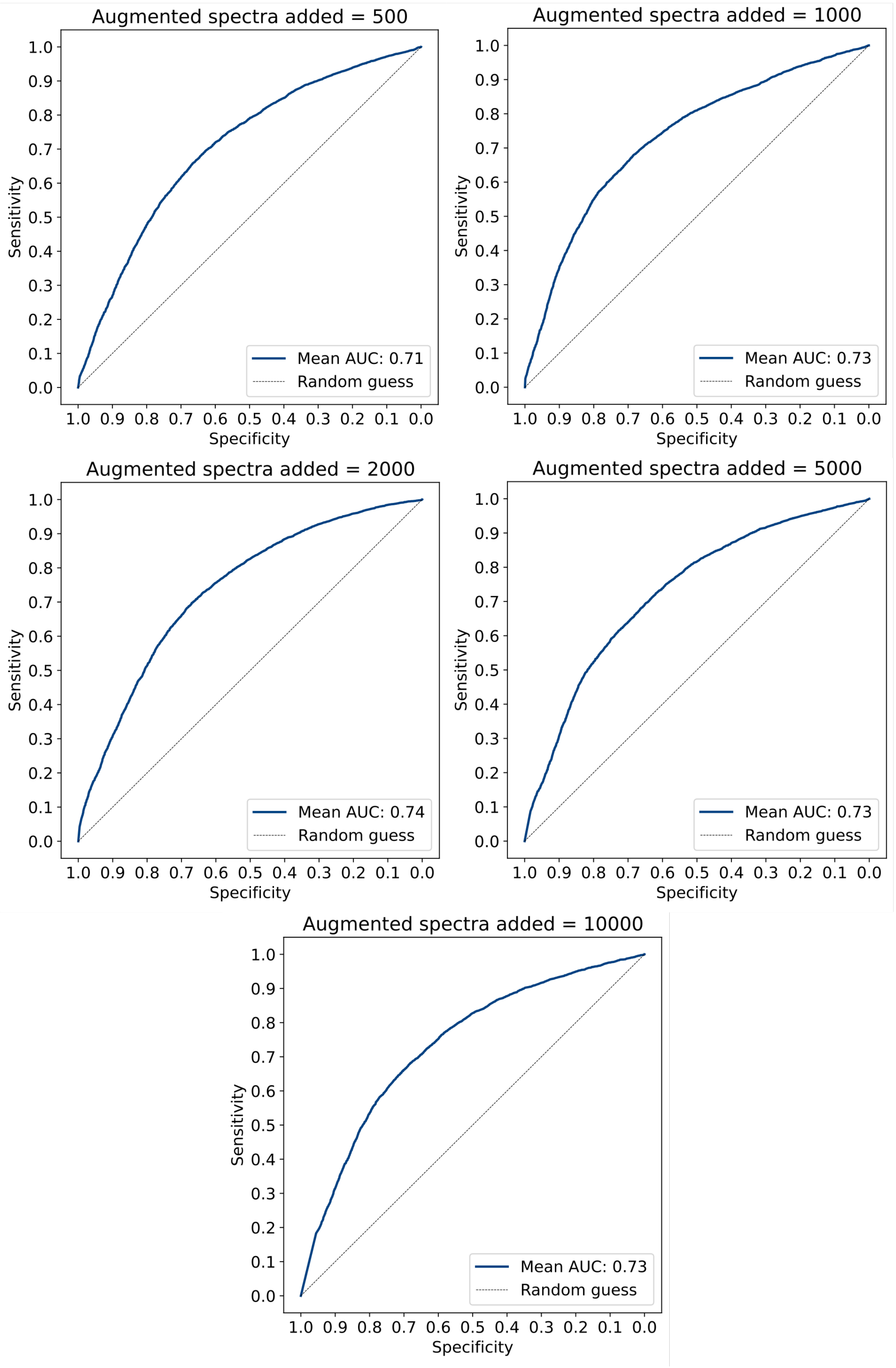

Figure S11: ROC curves for CNN models trained on augmented spectra generated by adding noise to spectra with the validation set containing full dataset.

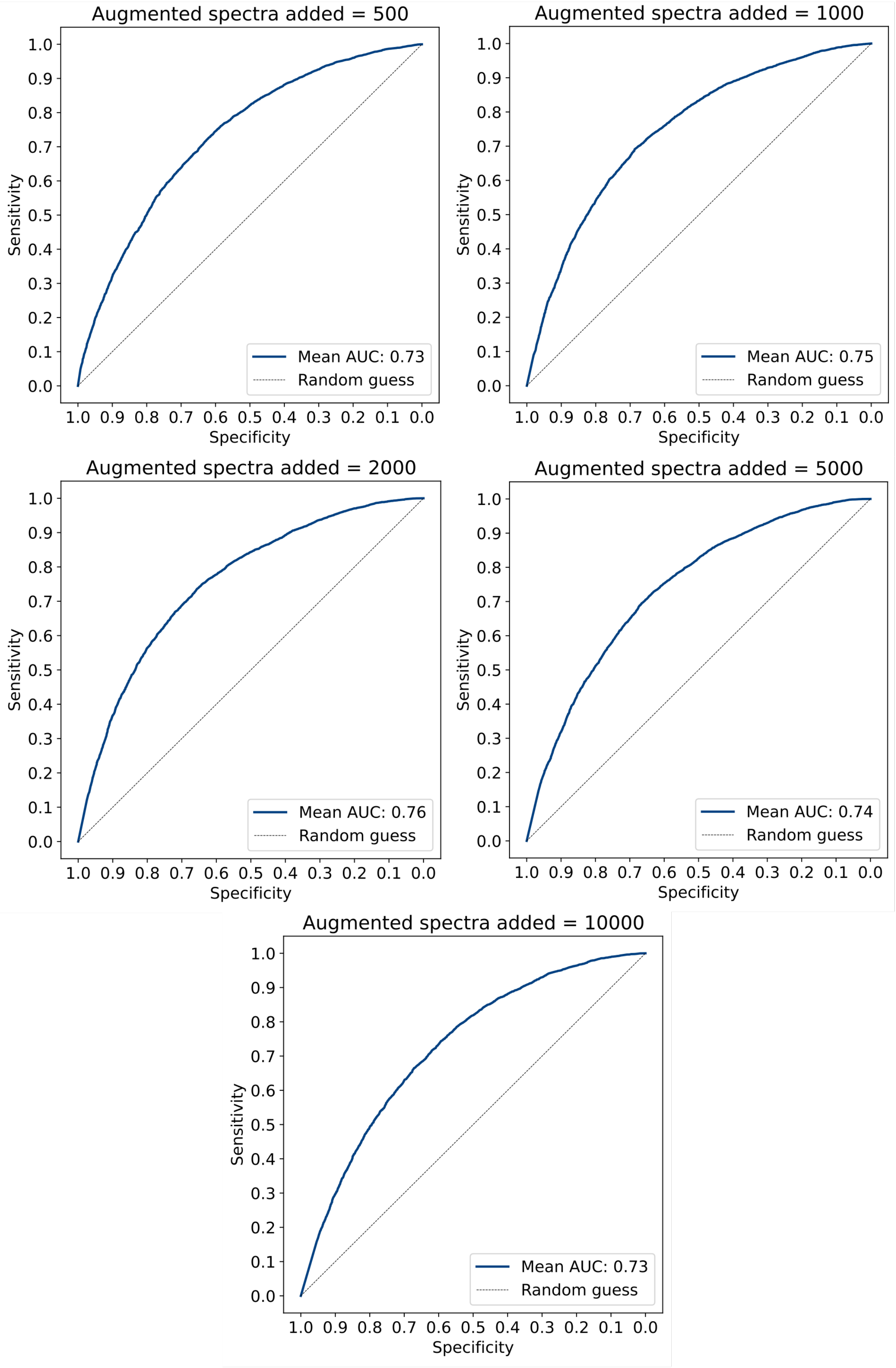

Figure S12: ROC curves for CNN models trained on augmented spectra generated by averaging spectra with the validation set containing full dataset.

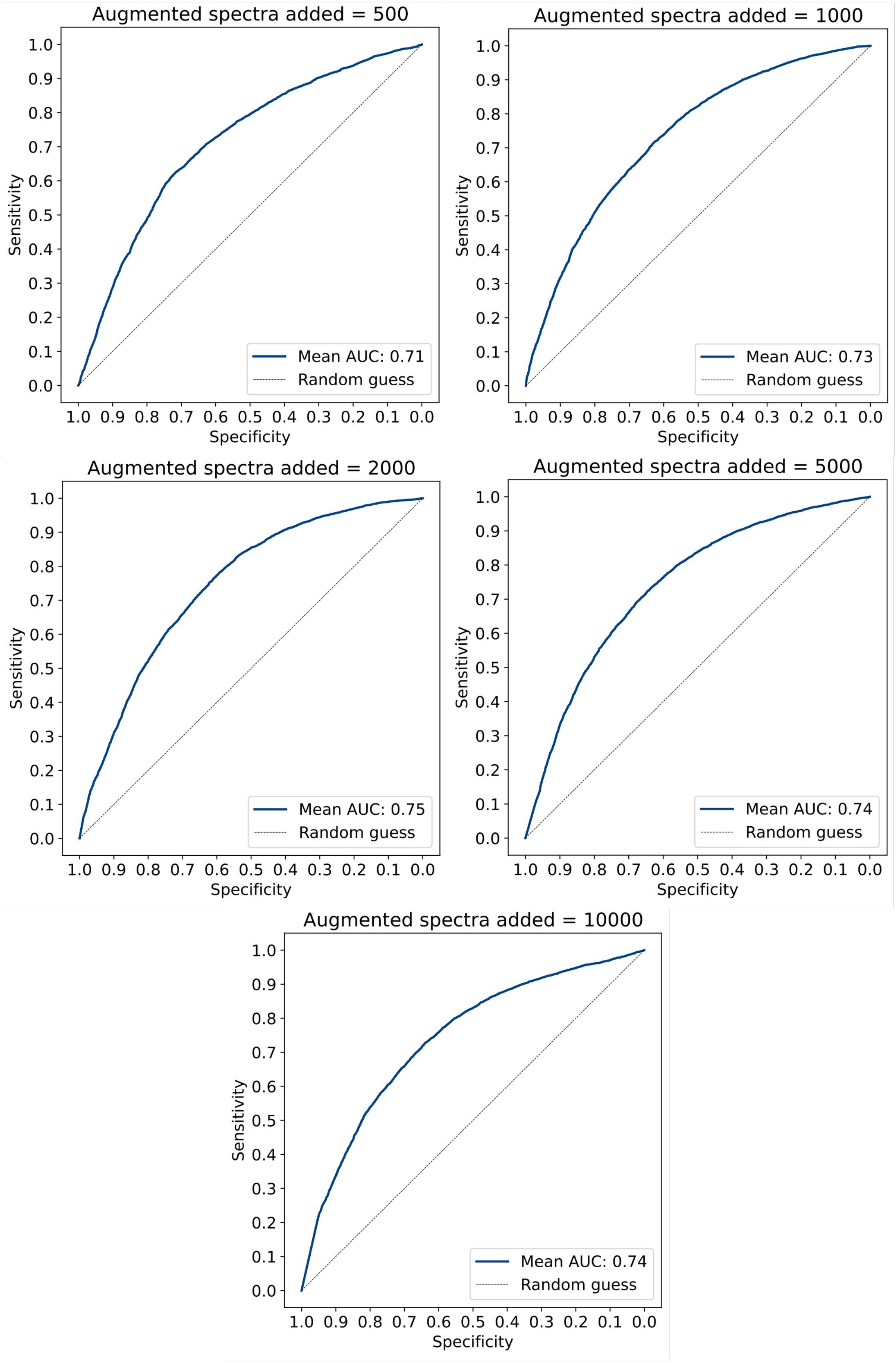

Figure S13: ROC curves for CNN models trained on augmented spectra generated by splicing spectra with the validation set containing full dataset.

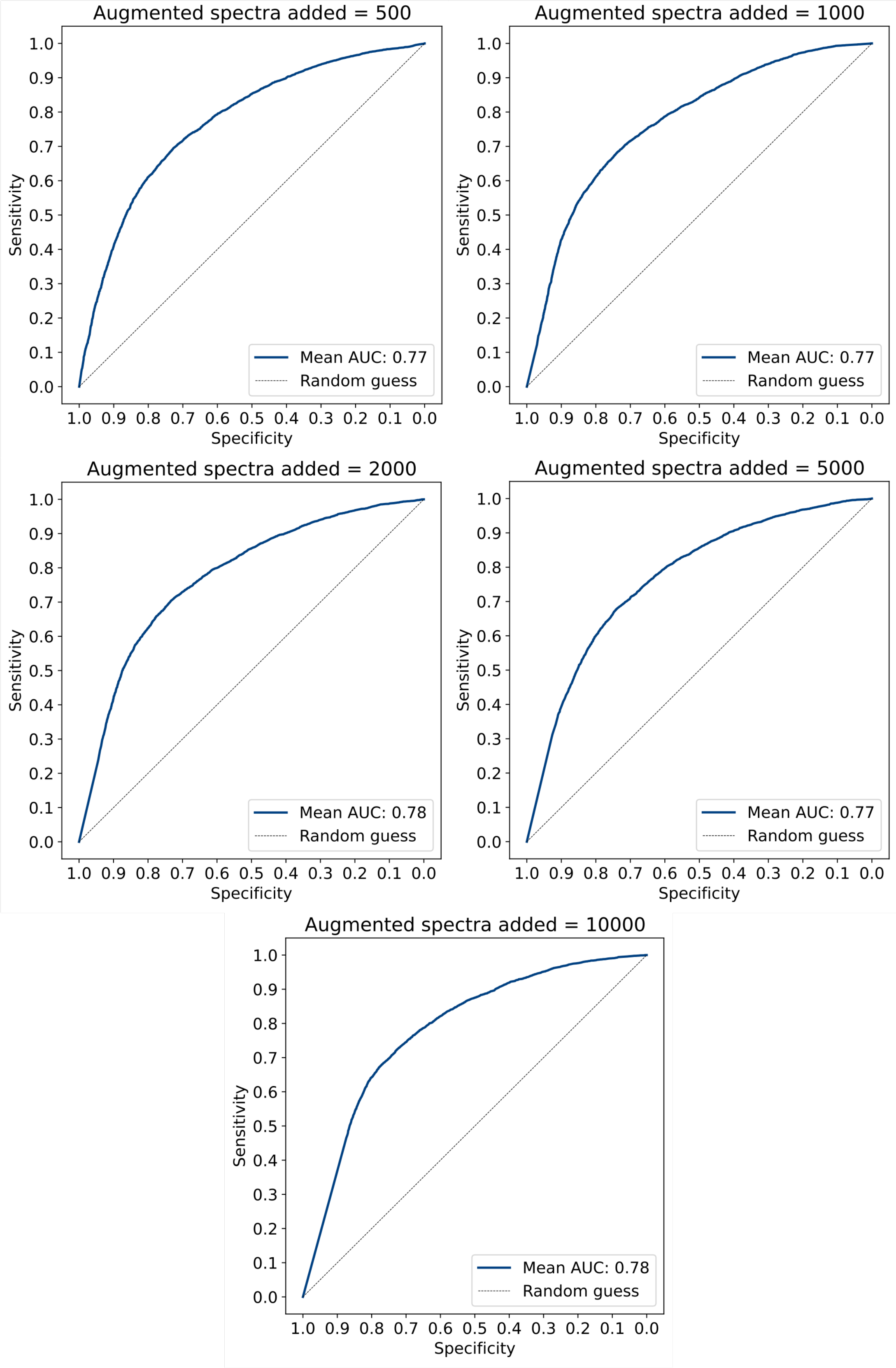

Figure S14: ROC curves for CNN models trained on WGAN augmented spectra with the validation set containing full dataset.